รายการอ้างอิง

ภาษาไทย

กัลยา วานิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 1. กรุงเทพมหานคร: ธรรมสาร, 2548.

ทัศนาพร จงเกตุกรณ์. การประมาณค่าพารามิเตอร์ของตัวแบบถดถอยโลจิสติคทวินาม. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2546.

บุษยมาส นันตสุคนธ์. การประมาณค่าฟังก์ชันอัตราภาวะภัยสำหรับข้อมูลสมบูรณ์. วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาประกันภัย คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2545.

ภาษาอังกฤษ

Allison, Paul D. Logistic Regression Using SAS: Theory and Application. 3rd printing. Cary, NC: SAS Institute Inc., 2001.

Davis, Jesse, and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. Proceeding of the 23rd International Conference on Machine Learning. 2006.

Gonen, Mithat. Receiver Operating Char..acteristic (ROC) Curves. Statistics and Data Analysis 31, 1 (2002): 210-31.

Hanley, J.A. and McNeil, B.J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, 1 (April 1982): 29-36.

Hosmer, David W., and Stanley Lemeshow. Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons, Inc., 2000.

Kiefer, Nicholas M. Economic Duration Data and Hazard Function. Journal of Economic Literature 26, 2 (June 1988): 646-679.

Shumway, Tyler. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. Journal of Business 74, 1 (January 2001): 101-124.

Tirapat, Sunti, and Seksan Kiatsupaibul. Credit Value at Risk VIA Credit Scoring Models. Simulation Society Research Workshop 2007: 1-4.

Zuiguoli, Shiyu Zhou, Suresh Choubey, and Crispian Sievenpiper. Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures. IIE Transactions 39, (2007): 303–315.

Zweig, Mark H., and Gregory Campbell. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry 39, 4 (1993): 561-577.

ภาคผนวก

ตัวอย่างการประยุกต์ใช้การวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตรกับข้อมูลจริง

ในส่วนนี้จะเป็นตัวอย่างการนำวิธีการวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตรมาใช้กับข้อมูลที่เกิดขึ้นจริงที่ได้มีการเก็บรวบรวมไว้แล้ว (secondary data) พร้อมทั้งการเปรียบเทียบประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบประมาณที่ได้จากวิธีการทั้งสอง ดังจะแสดงต่อไปนี้

ข้อกำหนดของข้อมูลจริง

1. เป็นข้อมูลตัวอย่างที่สุ่มจากข้อมูลการเช่าซื้อสินค้าชนิดหนึ่งของลูกค้าประเภทบุคคลที่ทำสัญญาเช่าซื้อในระหว่างเดือนพฤษภาคม พ.ศ. 2543 ถึง เดือนเมษายน พ.ศ. 2549 ซึ่งลักษณะของการเช่าซื้อ คือ การที่ลูกค้ารับสินค้าไปใช้โดยการจ่ายเงินบางส่วนจำนวนหนึ่ง ซึ่งเรียกว่า "เงินดาวน์" และส่วนที่เหลือจะจ่ายในภายหลังโดยจะแบ่งจ่ายเป็นงวดๆ ในแต่ละเดือนพร้อมดอกเบี้ยด้วยจำนวนเงินที่เท่าๆ กัน จนครบตามจำนวนและเวลาที่กำหนด แต่ในบางครั้งอาจเกิดปัญหาลูกค้าผิดนัดชำระหนี้ตามเวลาที่กำหนดหรืออาจไม่ชำระหนี้เลย จึงทำให้เกิดเป็นหนี้เสียหรือหนี้ NPL ขึ้น ซึ่งหากเกิดกรณีนี้เป็นจำนวนมากจะก่อให้เกิดความเสียหายต่อองค์กรที่เป็นเจ้าของเงินกู้ได้อย่างมากทีเดียว

2. เป็นข้อมูลทุติยภูมิ (secondary data) และไม่ได้ถูกเก็บตามแบบการวิเคราะห์แบบพลวัตร

3. จำนวนลูกค้าตัวอย่างทั้งหมดเท่ากับ 10,000 ราย โดยทำการสุ่มตัวอย่างจำนวน 2 ชุด ที่เป็นอิสระต่อกัน ดังนี้

- ข้อมูลชุดที่ 1 : ใช้เป็นข้อมูลตัวอย่างสำหรับการประมาณค่าพารามิเตอร์ในตัวแบบ (Training set) สุ่มจากลูกค้าที่เริ่มทำสัญญาเช่าซื้อในช่วงระหว่างเดือนพฤษภาคม พ.ศ. 2543 ถึง เดือนเมษายน พ.ศ. 2546 จำนวน 5,000 ราย
- ข้อมูลชุดที่ 2 : ใช้เป็นข้อมูลตัวอย่างสำหรับการทดสอบการพยากรณ์ (Validation set) สุ่มจากลูกค้าที่เริ่มทำสัญญาเช่าซื้อในช่วงระหว่างเดือนพฤษภาคม พ.ศ. 2546 ถึง เดือนเมษายน พ.ศ. 2549 จำนวน 5,000 ราย

4. กำหนดให้เริ่มต้นเก็บข้อมูลของแต่ละหน่วยตัวอย่างที่เวลาเริ่มเข้ามาของหน่วยตัวอย่างนั้นและกำหนดให้เป็น $t = 0$ และเก็บข้อมูลเป็นระยะเวลา 1 ปี นับจากวันที่เริ่มทำสัญญาเช่าซื้อของลูกค้าแต่ละราย โดยจะแบ่งออกเป็น 4 ช่วงเวลาๆ ละ 3 เดือน

5. จำนวนตัวแปรอิสระที่ใช้ในแต่ละตัวแบบเท่ากับ 3 ตัวแปร และเป็นตัวแปรเชิงปริมาณทั้งหมด ได้แก่

- PAY เป็นจำนวนเงินที่ต้องจ่ายต่อเดือน (Payment)
- LTV เป็นสัดส่วนเงินที่กู้ต่อมูลค่าของสินค้า (Loan / Value)
- RTERM เป็นจำนวนงวด (เดือน) ที่เหลืออยู่ (Remaining Term)

เนื่องจากถ้าจำนวนเงินที่ต้องจ่ายต่อเดือนมากจะทำให้ยอดเงินหมดในระยะเวลาสั้นลง และถ้าจำนวนงวดที่เหลือน้อยอาจจะทำให้มีช่วงเวลาที่จะเกิดความเสี่ยงของการเกิดหนี้ NPL น้อยลงได้ จึงเป็นเหตุผลที่ทำให้ผู้วิจัยเลือกใช้ตัวแปรเหล่านี้เป็นตัวแปรอิสระในตัวแบบที่จะศึกษา

6. ข้อมูลของตัวแปรอิสระ PAY และ LTV เป็นค่าคงที่ตลอดช่วงการศึกษา ส่วนข้อมูลของตัวแปร RTERM จะเปลี่ยนแปลงไปตามเวลาโดยมีค่าลดลงเมื่อเวลาผ่านไป

7. ตัวแปรตาม ได้แก่ ตัวแปร STATUS เป็นตัวแปรที่แสดงถึงการเกิดหนี้ NPL ของลูกค้าโดยกำหนดให้

6.1 นิยาม

STATUS = 1 หมายถึง ลูกค้าเกิดหนี้ NPL ได้แก่ ลูกค้าที่มีการค้างชำระหนี้ติดต่อกันตั้งแต่ 3 งวดหรือ 3 เดือนขึ้นไป

STATUS = 0 หมายถึง ลูกค้าไม่เกิดหนี้ NPL ได้แก่ ลูกค้าที่ไม่มีการค้างชำระหนี้หรือค้างชำระหนี้ติดต่อกันน้อยกว่า 3 เดือน

6.2 ค่าสังเกตของ STATUS ในแต่ละช่วงเวลาจะพิจารณาที่เวลาสิ้นสุดของช่วงเวลานั้น และหากลูกค้ารายใดมี STATUS = 1 (เกิดหนี้ NPL) หรือมี RTERM = 0 (หมดสัญญาเช่าซื้อ) ในช่วงเวลา t เราจะหยุดการสังเกตลูกค้ารายนั้นในช่วงเวลาถัดไป $t+1,\ldots,T$

8. ทำการวิเคราะห์ผลข้อมูลจริงด้วยโปรแกรม SAS ซึ่งเป็นโปรแกรมคำนวณทางสถิติ

ขั้นตอนการวิเคราะห์

1. จากข้อมูลตัวอย่างที่สุ่มมาจำนวน 2 ชุด จัดข้อมูลให้เหมาะสมสำหรับการวิเคราะห์ความถดถอยโลจิสติคในแต่ละแบบ

2. วิเคราะห์ข้อมูลเบื้องต้นด้วยสถิติพรรณนา (Descriptive Statistics)

3. วิเคราะห์ความถดถอยโลจิสติคด้วยข้อมูลที่ใช้สำหรับการประมาณค่าพารามิเตอร์ของแต่ละแบบ จากการวิเคราะห์จะได้ 2 ตัวแบบประมาณ คือ ตัวแบบสถิตย์และตัวแบบพลวัตร พร้อมทั้งค่าสถิติที่เกี่ยวข้อง

4. ทดสอบการพยากรณ์ของตัวแบบทั้งสองด้วยข้อมูลที่ใช้สำหรับทดสอบการพยากรณ์ของแต่ละแบบ ทำการคำนวณค่าความน่าจะเป็นของเหตุการณ์ที่สนใจจากการพยากรณ์ของแต่ละหน่วยตัวอย่าง $(\hat{p}_i)$ คำนวณค่า Sensitivity และ 1-Specificity ของแต่ละตัวแบบ

5. พล็อตโค้ง ROC พร้อมทั้งคำนวณพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสอง

6. เปรียบเทียบประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบทั้งสองจากพื้นที่ใต้โค้ง ROC โดยใช้สถิติทดสอบในสมการ (a)

ขั้นตอนของกระบวนการวิเคราะห์แสดงในแผนผังที่ 2

การเปรียบเทียบความแตกต่างของพื้นที่ใต้โค้ง ROC ระหว่าง 2 ตัวแบบสำหรับข้อมูลจริง

เนื่องจากการวิเคราะห์ข้อมูลจริงเราจะเปรียบเทียบความแตกต่างของพื้นที่ใต้โค้ง ROC ระหว่าง 2 ตัวแบบเพียง 1 คู่ หรือ 1 การทดลองเท่านั้น ซึ่งเราจะไม่สามารถใช้สถิติทดสอบ t-test ได้ Hanley, J.A. and McNeil, B.J.(1982) ได้แนะนำการคำนวณหาขนาดตัวอย่างที่เหมาะสมที่จะทำให้เกิดความแตกต่างอย่างมีนัยสำคัญทางสถิติของพื้นที่ใต้โค้ง ROC ระหว่าง 2 ตัวแบบใช้ในการเปรียบเทียบแทน โดยมีหลักเกณฑ์ว่า "สำหรับขนาดความแตกต่างของพื้นที่ใต้โค้ง ROC ที่เกิดขึ้นระหว่าง 2 ตัวแบบจะแตกต่างกันอย่างมีนัยสำคัญทางสถิติก็ต่อเมื่อขนาดตัวอย่างในแต่ละกลุ่มทั้ง 2 กลุ่มของตัวแปรตาม คือกลุ่มที่เกิดเหตุการณ์ (n_{y1}) และกลุ่มที่ไม่เกิดเหตุการณ์ (n_{y0}) จะต้องมีขนาดมากกว่าหรือเท่ากับ n ที่ได้จากการคำนวณ สมการคำนวณเป็นดังนี้

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

โดยที่ θ_1 คือ พื้นที่ใต้โค้ง ROC ของตัวแบบที่ 1 (ตัวแบบสถิตย์)

θ_2 คือ พื้นที่ใต้โค้ง ROC ของตัวแบบที่ 2 (ตัวแบบพลวัตร)

$\delta = \theta_2 - \theta_1$ คือ ความแตกต่างของพื้นที่ใต้โค้ง ROC ระหว่าง 2 ตัวแบบ

$Z_\alpha = 1.645$ สำหรับระดับนัยสำคัญ 0.05 ของการทดสอบทางเดียว

$Z_\beta = 0.84, 1.28, 1.645$ สำหรับอำนาจการทดสอบที่ 80% 90% และ 95% ตามลำดับ

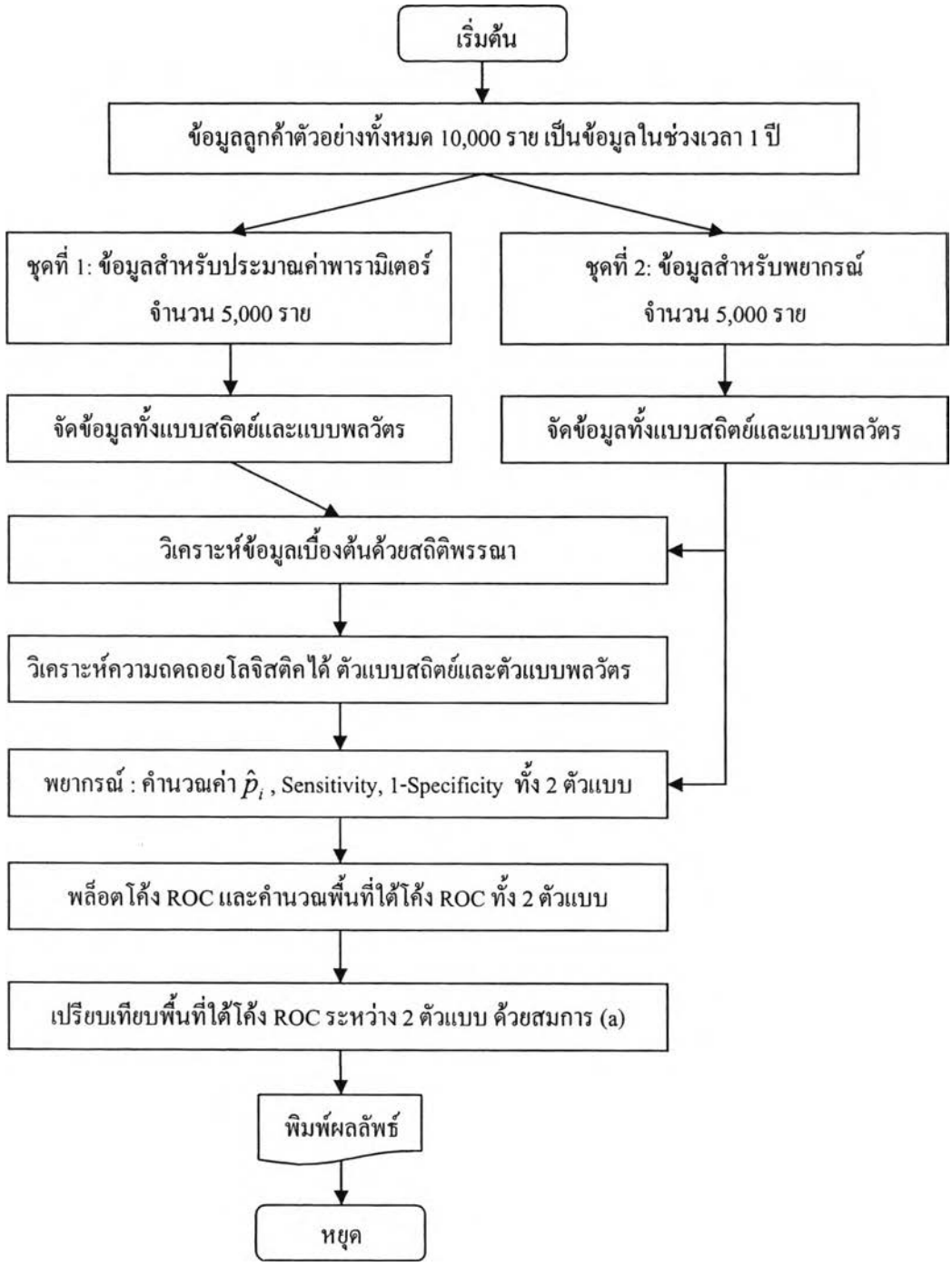
$V_1 = Q_1 + Q_2 - 2\theta_1^2$ เมื่อ Q_1 และ Q_1 ได้จากการแทนค่า θ_1 ในสมการ (b)

$V_2 = Q_1 + Q_2 - 2\theta_2^2$ เมื่อ Q_1 และ Q_1 ได้จากการแทนค่า θ_2 ในสมการ (b)

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

หมายเหตุ : สำหรับขนาดตัวอย่างที่มากกว่าจะต้องการความแตกต่างของพื้นที่ใต้โค้ง ROC ที่น้อยกว่ากรณีที่ขนาดตัวอย่างมีจำนวนน้อยกว่า และจะให้อำนาจการทดสอบมากกว่าสำหรับความแตกต่างของพื้นที่ใต้โค้ง ROC เท่ากัน

แผนผังที่ 2 ขั้นตอนของการวิเคราะห์ข้อมูลจริง



ผลการวิเคราะห์

การวิเคราะห์ข้อมูลที่เกิดขึ้นจริง ซึ่งเป็นข้อมูลที่ได้จากการสุ่มตัวอย่างลูกค้าที่ทำการเช่าซื้อสินค้าชนิดหนึ่ง ที่เป็นข้อมูลระดับหน่วยย่อย ใช้จำนวนหน่วยตัวอย่างในการสร้างตัวแบบและการพยากรณ์ ชุดละ 5,000 ราย โดยเหตุการณ์ที่สนใจ คือ การเกิดหนี้ NPL ของลูกค้า และใช้ตัวแปรอิสระในแต่ละตัวแบบจำนวน 3 ตัวแปร ได้แก่ PAY LTV และ RTERM ได้ผลการวิเคราะห์ดังต่อไปนี้

1. ผลการวิเคราะห์ข้อมูลเบื้องต้น

ตาราง A แสดงตัวอย่างสำหรับการวิเคราะห์ข้อมูลจริง

	เวลาที่เริ่มเข้ามา	**จำนวนข้อมูลตัวอย่าง**	
		แบบสถิตย์	**แบบพลวัตร**
ตัวอย่างชุดที่ 1: ใช้สำหรับประมาณค่าพารามิเตอร์	พฤษภาคม 2543 - เมษายน 2546	5,000	19,313
ตัวอย่างชุดที่ 2: ใช้สำหรับทดสอบการพยากรณ์	พฤษภาคม 2546 - เมษายน 2549	5,000	19,227

ตาราง B แสดงสถิติเบื้องต้นของข้อมูลตัวอย่างชุดที่ 1

ตัวแปรตาม					
STATUS	ความถี่	ร้อยละ	ความถี่สะสม	ร้อยละสะสม	สัดส่วนหนี้ NPL เกิด : ไม่เกิด
ไม่เกิดหนี้ NPL	4,522	90.44	4,522	90.44	1 : 9
เกิดหนี้ NPL	478	9.56	5,000	100.00	

ตัวแปรอิสระ				
ตัวแปร	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ค่าต่ำสุด	ค่าสูงสุด
PAY	25,456	31,947	2,656	401,250
LTV	0.804	0.186	0.118	1
RTERM	40.825	11.829	10	60

ตาราง C แสดงสถิติเบื้องต้นของข้อมูลตัวอย่างชุดที่ 2

ตัวแปรตาม					
STATUS	ความถี่	ร้อยละ	ความถี่สะสม	ร้อยละสะสม	สัดส่วนหนี้ NPL เกิด : ไม่เกิด
ไม่เกิดหนี้ NPL	4,470	89.40	4,470	89.40	1 : 9
เกิดหนี้ NPL	530	10.60	5,000	100.00	

ตัวแปรอิสระ				
ตัวแปร	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ค่าต่ำสุด	ค่าสูงสุด
PAY	13,337	21,287	1200	812,250
LTV	0.762	0.145	0.126	1
RTERM	54.437	10.884	10	83

เมื่อพิจารณาสถิติเบื้องต้นของข้อมูลตัวอย่างสำหรับการวิเคราะห์ทั้ง 2 ชุด พบว่า มีลักษณะคล้ายคลึงกันทั้งในตัวแปรตามและตัวแปรอิสระ ซึ่งจะเห็นได้จากสัดส่วนของลูกค้าที่เกิดหนี้ NPL ต่อลูกค้าที่ไม่เกิดหนี้ NPL ในข้อมูลแต่ละชุดมีค่าใกล้เคียงกัน คือ ประมาณ 1 : 9 และลักษณะข้อมูลของตัวแปรอิสระแต่ละตัวแปรก็มีค่าค่อนข้างใกล้เคียงกันสำหรับตัวอย่างทั้ง 2 ชุด

คำนวณสัดส่วน Hazard function $\hat{\lambda}(t)$ และ Survival function $\hat{S}(t)$ ในแต่ละช่วงเวลาของข้อมูลที่ใช้สำหรับการวิเคราะห์แบบพลวัตร ซึ่งคำนวณได้จากสมการ (2.13) ได้ผลดังตารางที่ D

ตาราง D แสดง $\hat{\lambda}(t)$ และ $\hat{S}(t)$ ในแต่ละช่วงเวลาของข้อมูลตัวอย่างชุดที่ 1

ข้อมูลชุดที่ 1				
ช่วงเวลา (t)	จำนวนตัวอย่าง (n_t)	จำนวนที่เกิดหนี้ NPL (h_t)	$\hat{\lambda}(t)$	$\hat{S}(t)$
1	5,000	73	0.0146	0.9854
2	4,927	167	0.0339	0.9520
3	4,760	134	0.0282	0.9252
4	4,626	104	0.0225	0.9044

ตาราง E แสดง $\hat{\lambda}(t)$ และ $\hat{S}(t)$ ในแต่ละช่วงเวลาของข้อมูลตัวอย่างชุดที่ 2

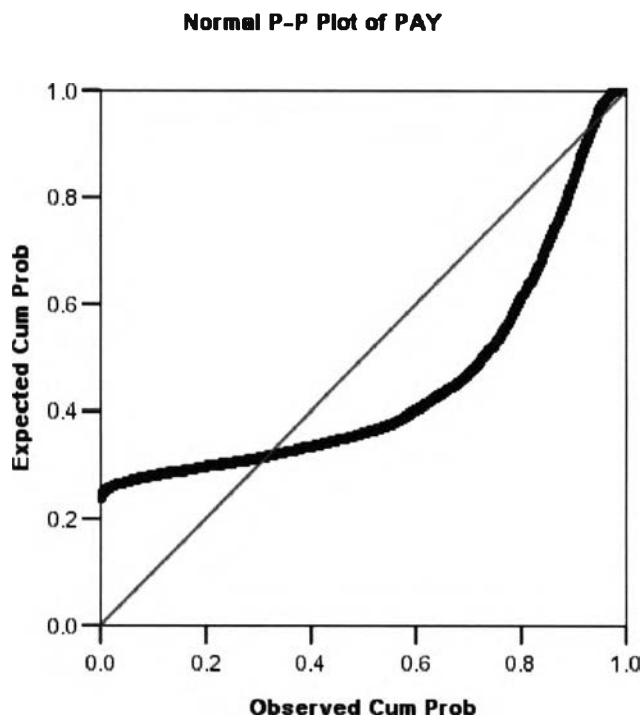
ข้อมูลชุดที่ 2				
ช่วงเวลา (t)	จำนวนตัวอย่าง (n_t)	จำนวนที่เกิดหนี้ NPL (h_t)	Hazard $\hat{\lambda}(t)$	Survivor $\hat{S}(t)$
1	5,000	95	0.0190	0.9810
2	4,905	167	0.0340	0.9476
3	4,738	154	0.0325	0.9168
4	4,584	114	0.0249	0.8940

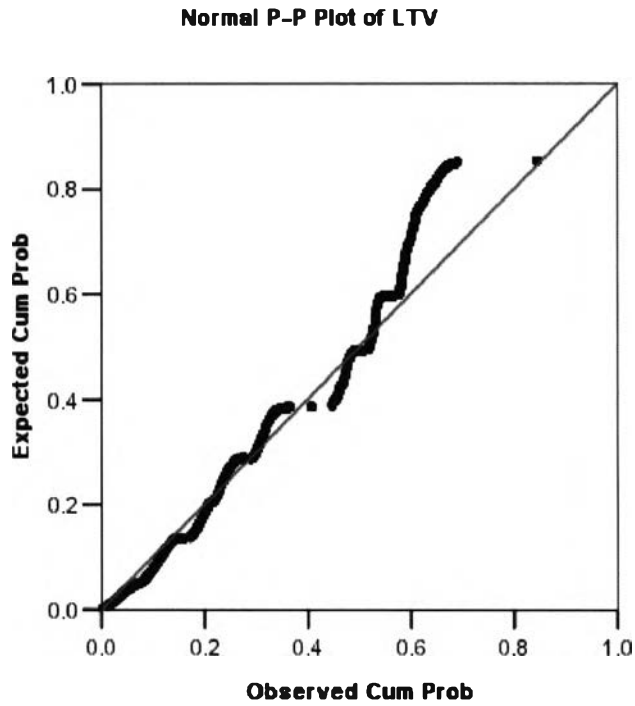
2. ตรวจสอบรูปแบบการแจกแจงของตัวแปรอิสระ

สำหรับข้อมูลข้อมูลจริงเราจะทำการตรวจสอบการแจกแจงของตัวแปรอิสระที่ใช้ในการวิเคราะห์ครั้งนี้ว่ามีการแจกแจงในลักษณะเดียวกันกับข้อมูลจำลองหรือไม่ ทั้งนี้ถ้าหากข้อมูลของตัวแปรอิสระในข้อมูลจริงมีรูปแบบการแจกแจงเช่นเดียวกันหรือคล้ายคลึงกันกับข้อมูลจำลองแล้ว การเปรียบเทียบประสิทธิภาพระหว่างตัวแบบทั้งสองก็น่าจะให้ผลลัพธ์ที่สอดคล้องกัน ซึ่งข้อมูลของตัวแปรอิสระ 3 ตัวแปร ที่ใช้ในข้อมูลจริงได้มาจากการแจกแจง 3 แบบ ได้แก่ การแจกแจงแบบปกติ แบบเบอร์นูลี และแบบเอ็กโพเนนเชียน ตามลำดับ ดังนั้น เราจะทำการทดสอบข้อมูลจริงว่ามีการแจกแจงใน 3 ลักษณะนี้หรือใกล้เคียงหรือไม่ ซึ่งสมมติฐานของการทดสอบ คือ

H_0 : ข้อมูลของตัวแปร X มาจากการแจกแจงตามที่กำหนด
H_1 : ข้อมูลของตัวแปร X ไม่ได้มาจากการแจกแจงตามที่กำหนด

- ทดสอบว่าตัวแปร PAY และ LTV มีการแจกแจงเป็นแบบปกติหรือไม่





One-Sample Kolmogorov-Smirnov Test

		PAY	LTV
N		5000	5000
Normal Parameters[a,b]	Mean	25456	.804
	Std. Deviation	31947	.186
Most Extreme Differences	Absolute	.245	.174
	Positive	.232	.146
	Negative	-.245	-.174
Kolmogorov-Smirnov Z		17.350	12.325
Asymp. Sig. (2-tailed)		.000	.000

a. Test distribution is Normal.

b. Calculated from data.

เมื่อพิจารณาจากกราฟ Normal P-P Plot พบว่า ตัวแปร PLY และ LTV มีแนวโน้มที่จะไม่ได้มาจากการแจกแจงแบบปกติ เนื่องจากกราฟที่ได้ไม่มีลักษณะเป็นเส้นตรงแนวทะแยงมุมจากจุด (0,0) ถึง (1,1) และจากสถิติทดสอบ Kolmogorov-Smirnov Z ของตัวแปรทั้งสองได้ค่าเท่ากับ 17.350 และ 12.325 ตามลำดับ ซึ่งมีค่า $p-value < 0.001$ ดังนั้น ข้อมูลของตัวแปร PLY และ LTV จึงไม่ได้มาจากการแจกแจงแบบปกติ ที่ระดับนัยสำคัญ 0.05

- ทดสอบว่าตัวแปร LTV มีการแจกแจงเป็นแบบทวินามหรือเบอร์นูลีหรือไม่

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
LTV	5000	.804	.186	.118	1

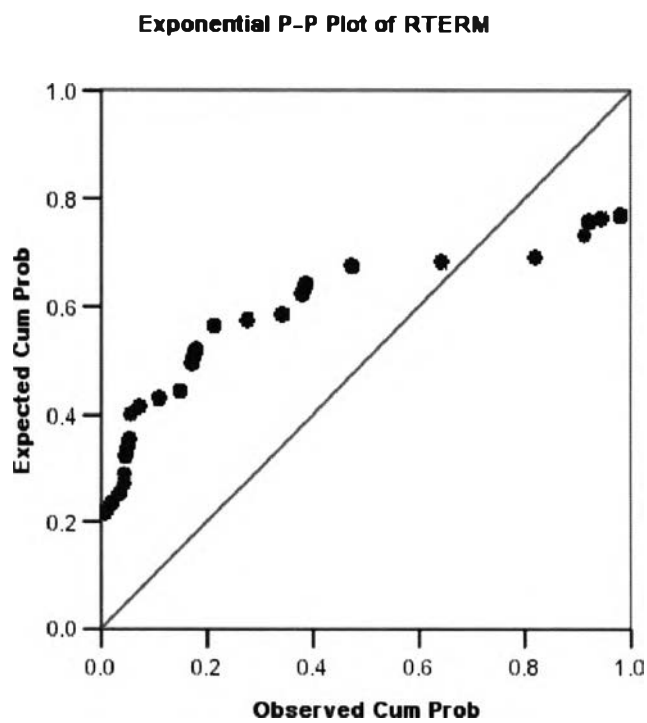
Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
LTV	Group 1	<= .8	2589	.52	.50	.012[a]
	Group 2	> .8	2411	.48		
	Total		5000	1.00		

a. Based on Z Approximation.

เนื่องจาก LTV มีค่าเฉลี่ยเท่ากับ 0.804 เราจึงทำการทดสอบสัดส่วนของจำนวนข้อมูลที่อยู่ในกลุ่มที่มีค่า ≥ 0.8 และกลุ่มที่มีค่า < 0.8 มีสัดส่วนที่เท่ากันหรือไม่ ซึ่งจากการทดสอบได้ค่า $p-value = 0.012$ นั่นแสดงว่า ข้อมูลของตัวแปร LTV ไม่ได้มีการแจกแจงแบบทวินามที่ระดับนัยสำคัญ 0.05

- ทดสอบว่าตัวแปร LTV มีการแจกแจงเป็นแบบทวินามหรือเบอร์นูลีหรือไม่



One-Sample Kolmogorov-Smirnov Test

		RTERM
N		5000
Exponential parameter.[a,b]	Mean	40.82
Most Extreme Differences	Absolute	.384
	Positive	.230
	Negative	-.384
Kolmogorov-Smirnov Z		27.166
Asymp. Sig. (2-tailed)		.000

a. Test Distribution is Exponential.

b. Calculated from data.

เมื่อพิจารณาจากกราฟ Exponential P-P Plot พบว่า ตัวแปร RTERM มีแนวโน้มที่จะไม่ได้มาจากการแจกแจงแบบเอกซโพเนนเชียล และจากสถิติทดสอบ Kolmogorov-Smirnov Z ได้ค่าเท่ากับ 27.166 $(p-value < 0.001)$ ดังนั้น ข้อมูลของตัวแปร PLY และ LTV จึงไม่ได้มาจากการแจกแจงแบบเอกซโพเนนเชียล ที่ระดับนัยสำคัญ 0.05

จากการตรวจสอบการแจกแจงของข้อมูลจำลอง พบว่า ตัวแปรอิสระทั้ง 3 ตัวแปรของข้อมูลจริง ไม่ได้มีการแจกแจงที่สอดคล้องกับตัวแปรอิสระของข้อมูลจำลองเลย ทั้งนี้อาจทำให้ผลลัพธ์ของการวิเคราะห์มีความเป็นไปได้ที่จะแตกต่างจากการวิเคราะห์ข้อมูลจำลอง

3. ผลการวิเคราะห์ความถดถอยโลจิสติคและการพยากรณ์

จากการพิจารณาค่าประมาณในตัวแบบทั้งสอง พบว่า ตัวแบบทั้งสองมีความเหมาะสมที่ระดับนัยสำคัญ 0.05 $(p-value < 0.01)$ และค่าสัมประสิทธิ์ความถดถอยของการประมาณของแต่ละตัวแปรอิสระมีค่าใกล้เคียงกันและมีทิศทางเดียวกันในตัวแบบทั้งสอง ซึ่งค่าประมาณในตัวแบบสถิตย์จะมีค่าสัมประสิทธิ์ความถดถอยสูงกว่าตัวแบบพลวัตรเล็กน้อยสำหรับตัวแปร LTV และ RTERM และมีทิศทางเป็นบวก ส่วนตัวแปร PAY มีค่าเท่ากันในทั้ง 2 ตัวแบบและมีทิศทางเป็นลบ ซึ่งสามารถอธิบายได้ว่า ลูกค้าที่มีจำนวนเงินที่จ่ายต่อเดือนมากจะมีโอกาสเป็นลูกหนี้ NPL น้อยลง และลูกค้าที่มีสัดส่วนเงินกู้ต่อมูลค่าของสินค้ามากหรือมีจำนวนงวดที่เหลืออยู่มากจะมีโอกาสเป็นลูกหนี้ NPL มากด้วย ดังตารางที่ F

ตาราง F แสดงค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบทั้งสองในข้อมูลจริง

ตัวแปร	ตัวแบบสถิตย์			ตัวแบบพลวัตร		
	$\hat{\beta}$	$SE(\hat{\beta})$	χ^2	$\hat{\beta}$	$SE(\hat{\beta})$	χ^{2*}
Intercept	-6.884	0.456	228.243	-7.610	0.402	92.731
PAY	-0.00001	0.000003	17.073	-0.00001	0.000003	5.177
LTV	4.286	0.363	139.808	3.993	0.342	35.288
RTERM	0.029	0.0054	28.209	0.021	0.0048	4.823

* ค่าสถิติที่ปรับแล้ว (n สำหรับตัวแบบสถิตย์และตัวแบบพลวัตร เท่ากับ 5,000 และ 19,313 ตามลำดับ)

ตาราง G แสดงค่าสถิติ R^2 และพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองสำหรับข้อมูลจริง

ตัวแบบ	ค่าสถิติ R^2	พื้นที่ใต้โค้ง ROC	95% ช่วงความเชื่อมั่นของพื้นที่ ROC	
			ขอบเขตล่าง (L)	ขอบเขตบน (U)
สถิตย์	0.047	0.6964	0.6705	0.7223
พลวัตร	0.046	0.7015	0.6757	0.7273

พิจารณาค่าสถิติ R^2 ของตัวแบบ พบว่า ตัวแบบทั้งสองมีค่า R^2 น้อยมาก แสดงว่าตัวแปรอิสระที่นำมาใช้อธิบายในครั้งนี้มีความสัมพันธ์กับตัวแปรตามหรือสถานะหนี้ของลูกค้าในระดับที่น้อย คือ ประมาณ 5% เท่านั้น ดังนั้น การใช้ตัวแปรอิสระทั้ง 3 ในตัวแบบอาจยังไม่เหมาะสมเท่าที่ควร หรืออีกนัยหนึ่งอาจมีตัวแปรอิสระอื่นที่มีความสำคัญกับการเกิดเหตุการณ์ที่สนใจมากกว่า แต่เราไม่ได้นำมาใช้ในการศึกษาในครั้งนี้ ด้วยเหตุนี้จึงอาจจะมีผลทำให้ประสิทธิภาพของการพยากรณ์อยู่ในระดับต่ำได้

จากการทดสอบการพยากรณ์และวัดความถูกต้องของการพยากรณ์ของตัวแบบทั้งสองด้วยโค้ง ROC พบว่า เส้นโค้ง ROC ของตัวแบบพลวัตรจะอยู่เหนือเส้นโค้ง ROC ของตัวแบบสถิตย์เล็กน้อย ดังรูป A ซึ่งทำให้ตัวแบบพลวัตรมีพื้นที่ใต้โค้งมากกว่าตัวแบบสถิตย์เล็กน้อย คือ ตัวแบบ

พลวัตรมีพื้นที่ใต้โค้ง ROC เท่ากับ 0.7015 และตัวแบบสถิตย์มีพื้นที่ใต้โค้ง ROC เท่ากับ 0.6964 และมีความแตกต่างของพื้นที่ทั้งสองเท่ากับ 0.0051 เมื่อทำการทดสอบสมมติฐานของความแตกต่างระหว่างพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสอง โดยใช้สมการ (a) ซึ่งมีสมมติฐานของการทดสอบ คือ

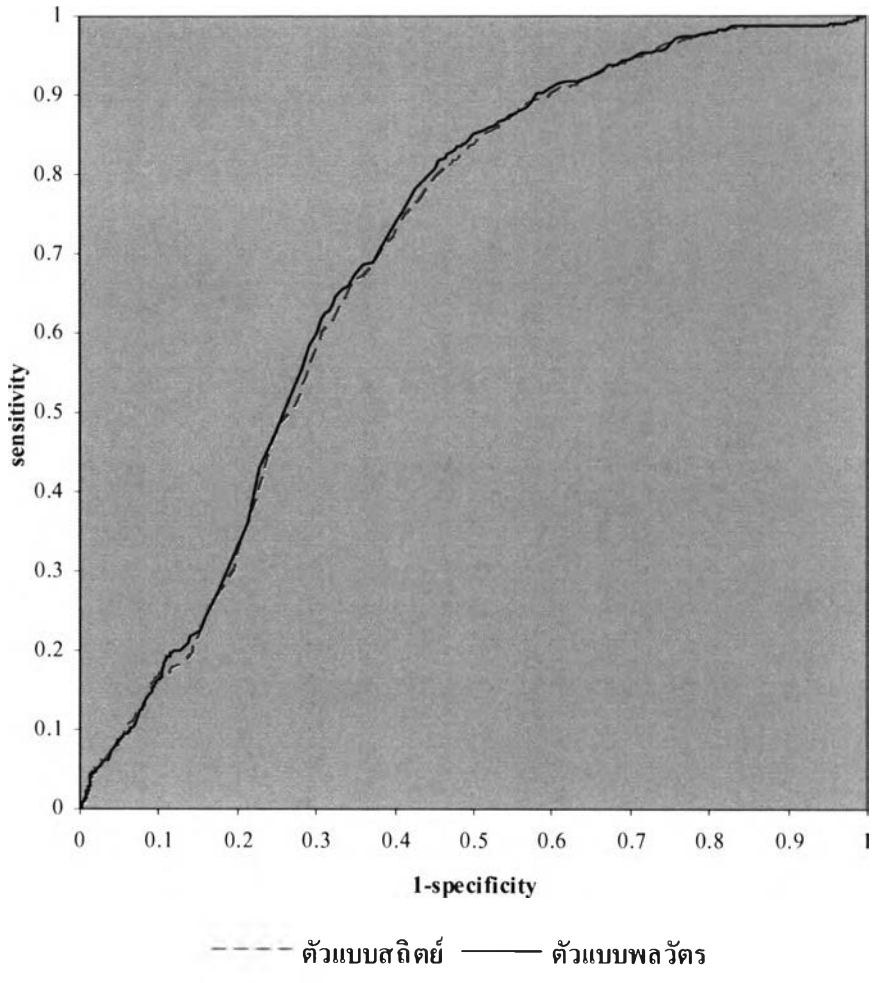
$$H_0 : \theta_2 \leq \theta_1 \ \ vs. \ \ H_1 : \theta_2 > \theta_1$$

เมื่อ θ_1 คือ พื้นที่ใต้โค้ง ROC ของตัวแบบสถิตย์

θ_2 คือ พื้นที่ใต้โค้ง ROC ของตัวแบบพลวัตร

จากการคำนวณได้ค่า $n = 64,467$ ซึ่งหมายความว่า สำหรับตัวแบบทั้งสองที่ให้ความแตกต่างของพื้นที่ใต้โค้ง ROC เท่ากับ 0.0051 จะต้องมีขนาดตัวอย่างในแต่ละค่าของตัวแปรตาม STATUS (0,1) เท่ากับ 64,467 หน่วยตัวอย่างขึ้นไป จึงจะเกิดความแตกต่างอย่างมีนัยสำคัญของพื้นที่ใต้โค้ง ROC ที่ระดับนัยสำคัญ 0.05 และมีอำนาจการทดสอบ 80% แต่เนื่องจากจำนวนตัวอย่างสำหรับแต่ละค่าของตัวแปรตาม STATUS (0, 1) ที่ใช้ในการวิเคราะห์ครั้งนี้ มีค่าเท่ากับ 4,470 และ 530 สำหรับ STATUS = 0 และ STATUS = 1 ตามลำดับ ซึ่งน้อยกว่าค่า n ที่ได้จากการคำนวณ ดังนั้น เราจึงไม่สามารถสรุปได้ว่าพื้นที่ใต้โค้ง ROC ของตัวแบบพลวัตรมากกว่าพื้นที่ใต้โค้ง ROC ของตัวแบบสถิตย์ที่ระดับนัยสำคัญ 0.05 นั่นคือ ตัวแบบทั้งสองมีประสิทธิภาพในการพยากรณ์ไม่แตกต่างกัน

รูปที่ A แสดงโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรสำหรับข้อมูลจริง



สรุปและอภิปรายผลสำหรับข้อมูลจริง

จากการนำวิธีการวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตรมาประยุกต์ใช้กับข้อมูลที่เกิดขึ้นจริง พบว่า ตัวแบบพลวัตรยังเป็นตัวแบบที่ให้พื้นที่ใต้โค้ง ROC ของการพยากรณ์มากกว่าพื้นที่ใต้โค้ง ROC ของตัวแบบสถิตย์ แต่มีค่ามากกว่าเพียงเล็กน้อย จึงทำให้ผลการทดสอบไม่แสดงนัยสำคัญทางสถิติที่ระดับ 0.05 ดังนั้น เราจึงไม่สามารถกล่าวได้ว่าตัวแบบพลวัตรมีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบสถิตย์ ทั้งนี้อาจเนื่องจาก ข้อมูลจริงที่ใช้ในการศึกษาครั้งนี้ จะใช้ตัวแปรอิสระในแต่ละตัวแบบจำนวน 3 ตัวแปร แต่มีเพียง 1 ตัวแปรเท่านั้น คือ เวลาที่เหลืออยู่ของลูกหนี้ ที่มีค่าผันแปรกับเวลาและมีอัตราผันแปรค่อนข้างน้อย จึงทำให้ข้อมูลที่ใช้ในการวิเคราะห์แบบพลวัตรไม่แตกต่างจากข้อมูลที่ใช้ในการวิเคราะห์แบบสถิตย์ที่เป็นเสมือนการใช้

ค่าเฉลี่ยของข้อมูล และเนื่องจากตัวแปรที่มีค่าผันแปรตามเวลานี้มีความสัมพันธ์กับเหตุการณ์ที่สนใจค่อนข้างน้อย ซึ่งจะมีผลต่อค่าความน่าจะเป็นของเหตุการณ์ที่สนใจของการพยากรณ์น้อยด้วย และจากการตรวจสอบรูปแบบการแจกแจงของตัวแปรอิสระที่ใช้ในตัวแบบประมาณ พบว่ามีรูปแบบของการแจกแจงที่แตกต่างกับข้อมูลจริงในทุกตัวแปร และข้อมูลมีการกระจายค่อนข้างมาก ซึ่งอาจเป็นสาเหตุให้ผลของการเปรียบเทียบประสิทธิภาพระหว่างตัวแบบทั้งสองไม่สอดคล้องกับการวิเคราะห์ข้อมูลจำลอง และอีกเหตุผลหนึ่ง คือ ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลทุติยภูมิที่มีการเก็บข้อมูลโดยมุ่งตามแบบสถิตย์มาตั้งแต่ต้น การปรับข้อมูลสำหรับการวิเคราะห์แบบพลวัตรจึงทำได้ในขอบเขตที่จำกัดและอาจยังไม่เพียงพอที่จะเห็นประสิทธิภาพที่เหนือกว่าของตัวแบบพลวัตรเมื่อเทียบกับตัวแบบสถิตย์

ข้อเสนอแนะ

สำหรับการวิเคราะห์ข้อมูลจริงนี้เป็นเพียงตัวอย่างการนำวิธีการวิเคราะห์และการเปรียบเทียบประสิทธิภาพของตัวแบบถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตรมาประยุกต์ใช้กับงานจริง ซึ่งผลการเปรียบเทียบอาจไม่ได้แสดงให้เห็นประสิทธิภาพที่มากกว่าของตัวแบบพลวัตรเมื่อเทียบกับตัวแบบสถิตย์ แต่ก็ทำให้เราเห็นได้ว่า เมื่อข้อมูลของตัวแปรอิสระที่ใช้ในการวิเคราะห์เปลี่ยนไปไม่ได้มีการแจกแจงแบบปกติหรือแบบเบอร์นูลีย์ และตัวแปรที่ผันตามเวลาไม่ได้มีการแจกแจงแบบเอกซโพเนนเชียล เช่นในข้อมูลจำลองแล้ว ผลการเปรียบเทียบประสิทธิภาพระหว่างตัวแบบทั้งสองอาจเปลี่ยนไปได้ ทั้งนี้ การวิเคราะห์จะให้ผลที่ดีขึ้น จะต้องขึ้นอยู่กับข้อมูลที่ใช้ในการวิเคราะห์ด้วย ซึ่งจะต้องเป็นข้อมูลตัวอย่างที่ดี ตัวแปรที่ใช้มีความสำคัญกับเรื่องที่ต้องการศึกษาเป็นต้น ซึ่งถ้าเราสามารถควบคุมหรือจัดการข้อมูลที่ใช้ในการวิเคราะห์ให้มีความเหมาะสมแล้ว เชื่อได้ว่าตัวแบบถดถอยโลจิสติคแบบพลวัตรน่าจะเป็นตัวแบบที่ให้ประสิทธิภาพดีกว่าตัวแบบถดถอยโลจิสติคแบบสถิตย์สำหรับข้อมูลที่เกิดขึ้นจริง

คำสั่งที่ใช้ในการวิเคราะห์ข้อมูลจำลองจากโปรแกรม R

```
Rocarea<-data.frame(StROC=0,HzROC=0)
Stattable<-data.frame(StAIC=1,StDev1=1,StDev2=1,Stn=1,HzAIC=1,HzDev1=1,HzDev2=1,Hzn=1)
R2table<-data.frame(StR2=1,HzR2=1)
Coeftable<-data.frame(St_coef=0,Hz_coef=0,r=0)

# Simulate data #
for(i in 1:1000){
N<-10000
X1<-rnorm(N,0,1)
X2<-1*(runif(N,min=0,max=1)>0.5)
X3<-round(rexp(N,1/36),0)
X3<-ifelse(X3==0,1,X3)
CLt0<-data.frame(ID=1:N,X1=X1,X2=X2,X3=X3,Y=0,t=0)
CL<-CLt0
        for(j in 1:24){
                CLtemp<-CL[CL$X3>0 & CL$Y==0 & CL$t==j-1,]
                n<-length(CLtemp$X3)
                U<-runif(n)
                CLtemp$X3<-CLtemp$X3-1
                p<-exp(-4.2+(1*CLtemp$X1)+(1*CLtemp$X2)+((1/12)*CLtemp$X3))/
                   (1+exp(-4.2+(1*CLtemp$X1)+(1*CLtemp$X2)+((1/12)*CLtemp$X3)))
                CLtemp$Y<-ifelse(U<p,1,0)
                CLtemp$t<-(CLtemp$t)+1
                CL<-rbind(CL,CLtemp)
        }
```

```
#Separate data : Trainning set and Validation set#
        CLData<-CL[CL$t>0,]
        CLData$X3<-CLData$X3+1
        CLModel<-CLData[CLData$t<=12,]
        CLTest<-CLData[CLData$t>12,]

#Set data for analysis#
        CLst1<-CLModel[CLModel$Y==1,]
        CLst2<-CLModel[CLModel$X3==1 & CLModel$Y==0,]
        CLst3<-CLModel[CLModel$X3>1 & CLModel$Y==0 & CLModel$t==12,]
        CLModelSt<-rbind(CLst1,CLst2,CLst3)
        CLModelSt<-
        CLModelSt[order(CLModelSt$ID,CLModelSt$X1,CLModelSt$X2,CLModelSt$X3,
CLModelSt$Y,CLModelSt$t),]
        CLsttest1<-CLTest[CLTest$Y==1,]
        CLsttest2<-CLTest[CLTest$X3==1 & CLTest$Y==0,]
        CLsttest3<-CLTest[CLTest$X3>1 & CLTest$Y==0 & CLTest$t==24,]
        CLTestSt<-rbind(CLsttest1,CLsttest2,CLsttest3)
        CLTestSt<-CLTestSt[order(CLTestSt$ID,CLTestSt$X1,CLTestSt$X2,CLTestSt$X3,
CLTestSt$Y,CLTestSt$t),]
        CLModelSt1<-CLModel[CLModel$t==1,]
        CLModelSt1<-
CLModelSt1[order(CLModelSt1$ID,CLModelSt1$X1,CLModelSt1$X2,CLModelSt1$X3,
CLModelSt1$Y,CLModelSt1$t),]
        CLModelSt1$Y<-CLModelSt$Y

#Estimate and predict Static model#
        Smodel<-glm(Y~X1+X2+X3, family=binomial(link=logit),data=CLModelSt1)
        summary(Smodel)
        CLTestSt13<-CLTest[CLTest$t==13,]
```

```
        CLTestSt13<-
CLTestSt13[order(CLTestSt13$ID,CLTestSt13$X1,CLTestSt13$X2,CLTestSt13$X3,CLTestSt1
3$Y,CLTestSt13$t),]
        CLTestSt13$Y<-CLTestSt$Y
        CLTestSt13$preStemp<-predict.glm(Smodel,newdata=CLTestSt13,type="response")
        CLTestSt13Y1<-CLTestSt13[CLTestSt13$Y==1,]
        CLTestSt13Y0<-CLTestSt13[CLTestSt13$Y==0,]

#Compute R-square of static model#
StG<-(Smodel$null.deviance-Smodel$deviance)
StRsq<-(1-exp(-StG/10000))

#Compute sensitivity and 1-specificity of static model#
        snumroc<-data.frame(stsen=1,stmspe=1)
        for(k in 1:100){
                Ys<-ifelse(CLTestSt13$preStemp>(k*0.01),1,0)
                scsen<-ifelse(CLTestSt13$Y==1&Ys==1,1,0)
                scspe<-ifelse(CLTestSt13$Y==0&Ys==0,1,0)
                stsen<-(sum(scsen))/nrow(CLTestSt13Y1)
                stspe<-(sum(scspe))/nrow(CLTestSt13Y0)
                stmspe<-(1-stspe)
                snumroctemp<-data.frame(stsen,stmspe)
                snumroc<-rbind(snumroc,snumroctemp)
        }
        CLTestSt13<-
data.frame(ID=CLTestSt13$ID,X1=CLTestSt13$X1,X2=CLTestSt13$X2,X3=CLTestSt13$X3,
Y=CLTestSt13$Y,t=CLTestSt13$t,psy1=CLTestSt13$preStemp)

#Estimate Dynamic or Hazard model#
        Hmodel<-glm(Y~X1+X2+X3,family=binomial(link=logit),data=CLModel)
        summary(Hmodel)
```

```
#Predict and compute probability of event for Dynamic model
        CLTesth13<-CLTest[CLTest$t==13,]
        CLTesth13<-
CLTesth13[order(CLTesth13$ID,CLTesth13$X1,CLTesth13$X2,CLTesth13$X3,CLTesth13$Y,
CLTesth13$t),]
        CLTesth13$Y<-CLTestSt$Y
        CLTesth13nex<-data.frame(CLTesth13,py0cum=1)
        for(d in 1:12){
                CLTesthnext<-data.frame(CLTesth13nex)
                CLTesthnext$X3<-(CLTesthnext$X3-d)
                py1ti<-predict.glm(Hmodel,newdata=CLTesthnext,type="response")
                CLTesthnext$py1t<-ifelse(CLTesthnext$X3>=0,py1ti,0)
                CLTesthnext$py0t<-(1-CLTesthnext$py1t)
                CLTesth13<-cbind(CLTesth13,CLTesthnext$py0t)
                CLTesth13nex$py0cum<-CLTesthnext$py0cum*CLTesthnext$py0t
        }
        CLTesth13nex<-
CLTesth13nex[order(CLTesth13nex$ID,CLTesth13nex$X1,CLTesth13nex$X2,CLTesth13nex$
X3,CLTesth13nex$Y,CLTesth13nex$t,CLTesth13nex$py0cum),]
        CLTestHa<-data.frame(CLTesth13nex)
        CLTestHa<-
data.frame(ID=CLTestHa$ID,X1=CLTestHa$X1,X2=CLTestHa$X2,X3=CLTestHa$X3,Y=CLTestHa
$Y,t=CLTestHa$t,py0cum=CLTestHa$py0cum)
        CLTestHa$phy1<-(1-CLTestHa$py0cum)

        CLTestHaY1<-CLTestHa[CLTestHa$Y==1,]
        CLTestHaY0<-CLTestHa[CLTestHa$Y==0,]

# Compute R-square of Dynamic model#
        HzG<-(Hmodel$null.deviance-Hmodel$deviance)
        HzRsq<-(1-exp(-HzG/10000))
```

```
# Compute sensitivity and 1-specificity of Dynamic model
        hnumroc<-data.frame(hasen=1,hamspe=1)
        for(c in 1:100){
                Yh<-ifelse(CLTestHa$phy1>(c*0.01),1,0)
                hcsen<-ifelse(CLTestHa$Y==1&Yh==1,1,0)
                hcspe<-ifelse(CLTestHa$Y==0&Yh==0,1,0)
                hasen<-(sum(hcsen))/nrow(CLTestHaY1)
                haspe<-(sum(hcspe))/nrow(CLTestHaY0)
                hamspe<-(1-haspe)
                hnumroctemp<-data.frame(hasen,hamspe)
                hnumroc<-rbind(hnumroc,hnumroctemp)
        }
        CLTestHa<-data.frame(CLTestHa)

#Plot 2 ROC curves#
        nf <- layout(matrix(c(2,1,1,2),2,2,byrow=TRUE), c(2,1), c(1,2), TRUE)
        layout.show(nf)
        sroc<-plot(snumroc$stmspe,snumroc$stsen,type="l",main="ROC",xlab="1-specificity",
ylab="sensitivity",lwd=2)
        hroc<-plot(hnumroc$hamspe,hnumroc$hasen,type="l",xlab="1-specificity",
ylab="sensitivity",lwd=2,col="blue",axes=FALSE)

#Compute area under ROC curve of 2 models#
        sarearoc<-0
        snumroc<-data.frame(stsen=snumroc$stsen,stmspe=snumroc$stmspe)
        for(a in 1:(length(snumroc$stsen)-1)){
                sarearoc<-sarearoc
                sdiffmspec<-snumroc$stmspe[a]-snumroc$stmspe[a+1]
                splussen<-snumroc$stsen[a]+snumroc$stsen[a+1]
                sareatemp<-((1/2)*sdiffmspec*splussen)
                sarearoc<-sarearoc+sareatemp
        }
```

```
        harearoc<-0
        hnumroc<-data.frame(hasen=hnumroc$hasen,hamspe=hnumroc$hamspe)
        for(a in 1:(length(hnumroc$hasen)-1)){
                harearoc<-harearoc
                diffmspec<-hnumroc$hamspe[a]-hnumroc$hamspe[a+1]
                plussen<-hnumroc$hasen[a]+hnumroc$hasen[a+1]
                hareatemp<-((1/2)*diffmspec*plussen)
                harearoc<-harearoc+hareatemp
        }

#Collect Statistics value#
        Roctemp<-data.frame(StROC=sarearoc,HzROC=harearoc)
        Rocarea<-rbind(Rocarea,Roctemp)
        Stattemp<-
data.frame(StAIC=Smodel$aic,StDev1=Smodel$null.deviance,StDev2=Smodel$deviance,
Stn=nrow(CLModelSt1),HzAIC=Hmodel$aic,HzDev1=Hmodel$null.deviance,
HzDev2=Hmodel$deviance,Hzn=nrow(CLModel))
        Stattable<-rbind(Stattable,Stattemp)
        R2temp<-data.frame(StR2=StRsq,HzR2=HzRsq)
        R2table<-rbind(R2table,R2temp)
        Coeftemp<-data.frame(St_coef=coef(Smodel),Hz_coef=coef(Hmodel),r=i)
        Coeftable<-rbind(Coeftable,Coeftemp)
}
Coeffall<-Coeftable[Coeftable$r!=0,]
Statall<-Stattable[Stattable$StAIC>1,]
Rocareaall<-Rocarea[Rocarea$StROC!=0,]
R2all<-R2table[R2table$StR2!=1 & R2table$HzR2!=1,]

#Test hypothesis of area under ROC for 1000 rounds#
CM<-t.test(Rocareaall$HzROC,Rocareaall$StROC,alternative="greater",paired=TRUE,
    var.equal=FALSE,conf.level=0.95)
```

คำสั่งที่ใช้ในการวิเคราะห์ข้อมูลจำลองจากโปรแกรม SAS

```
/*model data over 1 year */
data Credit.model1year_data ;
set Credit.model_data ;
if now_term <= 14;
run ;

proc sort data=credit.model1year_data ;
by t2cont date;
run ;

proc sort data=credit.stmodel_sample ;
by t2cont;
run ;

proc sql ;
      create table credit.ha1year_id as
      select t2cont from credit.stmodel_sample ;
run;

data credit.ha1year_id ;
set credit.ha1year_id ;
 select_id =1;
run;

data Credit.model1year_data ;
merge Credit.model1year_data Credit.ha1year_id ;
by T2CONT ;
run ;

data Credit.hamodel_sample1 ;
set Credit.model1year_data ;
if select_id =1 ;
run ;

data credit.hamodel_sample1 ;
set credit.hamodel_sample1 ;
format T2NRS1 $1. ;
format M1SEX $1. ;
format region $2. ;
informat region $2. ;
run;

data Credit.hamodel_sample1 ;
set Credit.hamodel_sample1 ;
drop select_id ;
run ;

/*set 5 periods time 1-5 */
proc sort data=credit.hamodel_sample1 ;
by t2cont date;
run ;
```

```
data Credit.hamodel_sample1 ;
set Credit.hamodel_sample1 ;
by t2cont ;
retain time_id ;
if first.t2cont then time_id=1 ;
else time_id=time_id+1;
run;

data Credit.time1 ;
set Credit.hamodel_sample1 ;
if time_id=1 ;
run ;

data Credit.time2 ;
set Credit.hamodel_sample1 ;
if time_id=2 ;
run ;

data Credit.time3 ;
set Credit.hamodel_sample1 ;
if time_id=3 ;
run ;

data Credit.time4 ;
set Credit.hamodel_sample1 ;
if time_id=4 ;
run ;

data Credit.time5 ;
set Credit.hamodel_sample1 ;
if time_id=5 ;
run ;

/*set X1,Y2 for time1 */
data Credit.time1x ;
set Credit.time1 ;
drop status status1 ;
run ;

proc sql ;
      create table credit.time2y as
      select t2cont, status, status1 from credit.time2 ;
run;

data Credit.time1x1y2 ;
merge Credit.time1x Credit.time2y ;
by T2CONT ;
run ;

/*set X2,Y3 for time2 */
data Credit.time2x ;
set Credit.time2 ;
if status=0 ;
run ;

data Credit.time2x ;
set Credit.time2x ;
drop status status1 ;
run ;
```

```
proc sql ;
      create table credit.time3y as
      select t2cont, status, status1 from credit.time3 ;
run;

data Credit.time2x2y3 ;
merge Credit.time2x Credit.time3y ;
by T2CONT ;
run ;

/*set X3,Y4 for time3 */
data Credit.time3x ;
set Credit.time3 ;
if status=0 ;
run ;

data Credit.time3x ;
set Credit.time3x ;
drop status status1 ;
run ;

proc sql ;
      create table credit.time4y as
      select t2cont, status, status1 from credit.time4 ;
run;

data Credit.time3x3y4 ;
merge Credit.time3x Credit.time4y ;
by T2CONT ;
run ;

/*set X4,Y5 for time4 */
data Credit.time4x ;
set Credit.time4 ;
if status=0 ;
run ;
data Credit.time4x ;
set Credit.time4x ;
drop status status1 ;
run ;
proc sql ;
      create table credit.time5y as
      select t2cont, status, status1 from credit.time5 ;
run;
data Credit.time4x4y5 ;
merge Credit.time4x Credit.time5y ;
by T2CONT ;
run ;

/*corr hazard sample for time1-4 */
PROC SQL;
CREATE TABLE CREDIT.hamodel_sample AS
SELECT * FROM CREDIT.time1x1y2
OUTER UNION CORR
SELECT * FROM CREDIT.time2x2y3
OUTER UNION CORR
SELECT * FROM CREDIT.time3x3y4
OUTER UNION CORR
SELECT * FROM CREDIT.time4x4y5;
run;
```

```
proc sort data=credit.hamodel_sample ;
by t2cont date;
run ;

data Credit.hamodel_sample ;
set Credit.hamodel_sample ;
drop time_id ;
run ;

/* new set static model sample*/
data Credit.stmodel_xf ;
set Credit.hamodel_sample ;
by t2cont ;
if first.t2cont ;
run ;

data Credit.stmodel_yl ;
set Credit.hamodel_sample ;
by t2cont ;
if last.t2cont ;
run ;

data Credit.stmodel_xf ;
set Credit.stmodel_xf ;
drop status status1 ;
run ;

proc sql ;
      create table credit.stmodel_ylyy as
      select t2cont, status, status1 from credit.stmodel_yl ;
run;

data Credit.stmodel_sample ;
merge Credit.stmodel_xf Credit.stmodel_ylyy ;
by T2CONT ;
run ;

/*estimate model and test with old data*/
proc logistic data=credit.stmodel14_sample outest=credit.stresult14
outmodel=credit.staticfix14_model SIMPLE;
      model status(event='1') =  payment ltv rem_term  / rsquare
ctable;
      score out=credit.stscore_old OUTROC=credit.stroc14;
quit;

/*from est_model, test with new data*/
proc logistic inmodel=credit.staticfix14_model;
      score data=credit.stpredict14_sample out=credit.stscore_new
outroc=credit.stdataroc;
quit;

proc sql ;
      create table credit.stresult_old as
      select t2cont,status,P_1 from credit.stscore_old ;
run;

proc sql ;
      create table credit.stresult_new as
      select t2cont,status,P_1 from credit.stscore_new ;
run;
```

```
/*estimate model and test with old data*/
proc logistic data=credit.hamodel14_sample outest=credit.haresult14
outmodel=credit.hazardfix14_model SIMPLE;
      model status(event='1') = payment ltv rem_term  / rsquare
ctable;
      score OUTROC=credit.haroc14;
quit;

/*compute p & 1-p in time5 */
data credit.time1_pre ;
set credit.hamodel14_sample ;
by t2cont ;
if first.t2cont ;
run;

data credit.time1_pre ;
set credit.time1_pre ;
 rem_term1=rem_term ;
 rem_term=rem_term-3 ;
 if rem_term<0 then rem_term=0 ;
run;

proc logistic inmodel=credit.hazardfix14_model;
      score data=credit.time1_pre out=credit.hascoret1 ;
quit;

/*compute p & 1-p in time6 */
data credit.time2_pre ;
set credit.time1_pre ;
 rem_term1=rem_term ;
 rem_term=rem_term-3 ;
run;

proc logistic inmodel=credit.hazardfix14_model;
      score data=credit.time2_pre out=credit.hascoret2 ;
quit;
data credit.hascoret2 ;
set credit.hascoret2 ;
if rem_term<=-3 then P_0=1;
if rem_term<=-3 then P_1=0;
run;

/*compute p & 1-p in time3 */
data credit.time3_pre ;
set credit.time2_pre ;
 rem_term1=rem_term ;
 rem_term=rem_term-3 ;
run;

proc logistic inmodel=credit.hazardfix14_model;
      score data=credit.time3_pre out=credit.hascoret3 ;
quit;

data credit.hascoret3 ;
set credit.hascoret3 ;
if rem_term<=-3 then P_0=1;
if rem_term<=-3 then P_1=0;
run;
```

```
/*compute p & 1-p in time4 */
data credit.time4_pre ;
set credit.time3_pre ;
 rem_term1=rem_term ;
 rem_term=rem_term-3 ;
run;

proc logistic inmodel=credit.hazardfix14_model;
      score data=credit.time4_pre out=credit.hascoret4 ;
quit;

data credit.hascoret4 ;
set credit.hascoret4 ;
if rem_term<=-3 then P_0=1;
if rem_term<=-3 then P_1=0;
run;

/*P_0 in time1-4 */
proc sql ;
      create table credit.time1_P0 as
      select t2cont, P_0 from credit.hascoret1 ;
run;

data credit.time1_P0 ;
set credit.time1_P0 ;
 P_01= P_0;
 drop P_0 ;
run;

proc sql ;
      create table credit.time2_P0 as
      select t2cont, P_0 from credit.hascoret2 ;
run;

data credit.time2_P0 ;
set credit.time2_P0 ;
 P_02= P_0;
 drop P_0 ;
run;

proc sql ;
      create table credit.time3_P0 as
      select t2cont, P_0 from credit.hascoret3 ;
run;

data credit.time3_P0 ;
set credit.time3_P0 ;
 P_03= P_0;
 drop P_0 ;
run;

proc sql ;
      create table credit.time4_P0 as
      select t2cont, P_0 from credit.hascoret4 ;
run;
```

```
data credit.time4_P0 ;
set credit.time4_P0 ;
 P_04= P_0;
 drop P_0 ;
run;

data Credit.time1to4_P0 ;
merge Credit.time1_P0 Credit.time2_P0 Credit.time3_P0 Credit.time4_P0
;
by T2CONT ;
run ;

data credit.time1to4_P0 ;
set credit.time1to4_P0 ;
 cumP_0= P_01*P_02*P_03*P_04 ;
 cumP_1=1-cumP_0 ;
run;

proc sort data=credit.time1to4_P0 ;
by t2cont;
run ;

proc sql ;
      create table credit.hastatus14_last as
      select t2cont, status from credit.stmodel14_sample ;
run;

data Credit.time1to4_P0 ;
merge Credit.time1to4_P0 credit.hastatus14_last ;
by T2CONT ;
run ;

proc sql ;
      create table credit.hascore_old as
      select t2cont, status,cumP_1 from credit.time1to4_P0 ;
run;

proc gplot data= credit.stroc14 ;
      plot  _1MSPEC_ * _SENSIT_ ;
run;

symbol1 v=dot i=join ;
proc gplot data= credit.stdataroc ;
      plot  _1MSPEC_ * _SENSIT_ ;
run;
```

ตัวอย่างค่าของพื้นที่ใต้โค้ง ROC จากการวิเคราะห์ข้อมูลจำลองทั้ง 6 กรณี จำนวน 30 รอบ

i	Case1		Case2		Case3		Case4		Case5		Case6	
	St	Dy	St	Dy	St	Dy	St	Dy	St	Dy	St	Dy
1	0.7936	0.8029	0.9329	0.9332	0.8089	0.8127	0.7959	0.8030	0.9448	0.9482	0.7783	0.7890
2	0.7998	0.8058	0.9267	0.9278	0.8027	0.8080	0.7774	0.8017	0.9358	0.9399	0.7830	0.8031
3	0.8086	0.8165	0.9325	0.9337	0.8090	0.8106	0.7726	0.7935	0.9407	0.9443	0.7812	0.7978
4	0.7883	0.7962	0.9333	0.9328	0.8167	0.8193	0.8302	0.8351	0.9372	0.9420	0.8088	0.8227
5	0.8031	0.8114	0.9282	0.9277	0.7888	0.7938	0.7658	0.7816	0.9408	0.9454	0.7722	0.7855
6	0.8123	0.8148	0.9316	0.9318	0.7897	0.7934	0.8106	0.8215	0.9430	0.9465	0.8164	0.8292
7	0.8093	0.8117	0.9298	0.9304	0.7960	0.7979	0.7884	0.8021	0.9387	0.9424	0.7859	0.8059
8	0.8112	0.8170	0.9305	0.9303	0.8006	0.8057	0.7841	0.7953	0.9419	0.9454	0.7993	0.8045
9	0.8166	0.8237	0.9325	0.9318	0.7991	0.8034	0.8159	0.8290	0.9412	0.9456	0.7769	0.7973
10	0.7950	0.8033	0.9281	0.9283	0.7775	0.7849	0.7945	0.8104	0.9402	0.9443	0.8053	0.8139
11	0.7879	0.7968	0.9323	0.9334	0.8119	0.8119	0.7968	0.8096	0.9422	0.9456	0.7796	0.7920
12	0.7950	0.8029	0.9320	0.9320	0.7922	0.7954	0.7714	0.7926	0.9360	0.9410	0.7863	0.8038
13	0.7899	0.7947	0.9275	0.9292	0.7938	0.7915	0.7870	0.8013	0.9388	0.9431	0.8015	0.8191
14	0.8041	0.8112	0.9319	0.9323	0.7992	0.8072	0.7788	0.8039	0.9361	0.9408	0.8015	0.8130
15	0.7948	0.8021	0.9291	0.9294	0.8020	0.8027	0.7697	0.7891	0.9429	0.9461	0.7795	0.7895
16	0.7946	0.8060	0.9313	0.9316	0.8038	0.8082	0.7922	0.8062	0.9411	0.9451	0.8079	0.8263
17	0.8005	0.8077	0.9335	0.9333	0.7904	0.7967	0.8012	0.8065	0.9412	0.9450	0.7656	0.7858
18	0.8053	0.8118	0.9309	0.9313	0.8088	0.8098	0.8024	0.8138	0.9384	0.9413	0.7797	0.7819
19	0.7974	0.8060	0.9297	0.9291	0.8137	0.8148	0.7694	0.7879	0.9402	0.9433	0.7905	0.7972
20	0.8125	0.8160	0.9276	0.9282	0.7957	0.8004	0.8210	0.8243	0.9420	0.9455	0.7710	0.7837
21	0.7771	0.7849	0.9286	0.9292	0.7975	0.8026	0.7959	0.8103	0.9407	0.9437	0.8038	0.8233
22	0.7927	0.7980	0.9299	0.9297	0.7924	0.7966	0.7830	0.8054	0.9399	0.9441	0.7971	0.8109
23	0.8123	0.8155	0.9385	0.9391	0.8220	0.8250	0.8105	0.8300	0.9387	0.9428	0.8048	0.8154
24	0.7806	0.7888	0.9362	0.9368	0.7992	0.7997	0.7736	0.7910	0.9383	0.9422	0.7773	0.7925
25	0.7928	0.7966	0.9360	0.9370	0.7832	0.7865	0.8002	0.8105	0.9407	0.9448	0.7831	0.7995
26	0.7928	0.8019	0.9323	0.9329	0.8004	0.8023	0.7790	0.7918	0.9446	0.9487	0.7924	0.8003
27	0.7938	0.7966	0.9302	0.9305	0.8096	0.8169	0.7857	0.7975	0.9384	0.9417	0.7908	0.8098
28	0.7978	0.8035	0.9347	0.9349	0.7967	0.7963	0.7994	0.8075	0.9451	0.9493	0.7919	0.8048
29	0.7992	0.8081	0.9349	0.9345	0.7960	0.7987	0.7730	0.7912	0.9433	0.9466	0.7971	0.8171
30	0.7831	0.7901	0.9345	0.9358	0.7794	0.7850	0.8142	0.8263	0.9402	0.9436	0.8199	0.8333

ประวัติผู้เขียนวิทยานิพนธ์

นางสาววรุณี มุริกา เกิดวันที่ 4 กรกฎาคม พ.ศ. 2521 สำเร็จการศึกษาปริญญาตรี จากคณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม และได้เข้าศึกษาต่อในระดับปริญญาโท ที่ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปี พ.ศ. 2549