การเปรียบเทียบตัวแบบความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร

นางสาววรุณี มุริกา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2551
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

A COMPARISON BETWEEN A STATIC LOGISTIC REGRESSION MODEL AND A DYNAMIC LOGISTIC REGRESSION MODEL

Miss Warunee Muriga

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Statistics Program in Statistics

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2008

Copyright of Chulalongkorn University

511616

หัวข้อวิทยานิพนธ์	การเปรียบเทียบตัวแบบความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร
โดย	นางสาววรุณี มุริกา
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร. โอวาท สุนันท์)

.....กรรมการ

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.....กรรมการ

(รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา)

วรุณี มุริกา : การเปรียบเทียบตัวแบบความถดถอยโลจิสติคแบบสถิตย์และแบบพลวัตร. (A Comparison between A Static Logistic Regression Model and A Dynamic Logistic Regression Model) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร. เสกสรร เกียรติสุไพบูลย์, 78 หน้า.

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบถดถอยโลจิสติคแบบสถิตย์ (Static model) และตัวแบบพลวัตร (Dynamic model) โดยวิธีการวิเคราะห์แบบสถิตย์จะพิจารณาช่วงเวลาเป็นเพียงหนึ่งช่วงเวลาเท่านั้น แต่วิธีการวิเคราะห์แบบพลวัตรจะทำการแบ่งช่วงเวลาออกเป็นหลายช่วงเวลาย่อยและพิจารณาข้อมูลของแต่ละหน่วยตัวอย่างในแต่ละช่วงเวลานั้น ซึ่งเป็นลักษณะของ Hazard model ในกรณีที่เวลาเป็นแบบไม่ต่อเนื่อง และทำการศึกษากับข้อมูล 2 ประเภท ได้แก่ ข้อมูลจำลองและข้อมูลจริง โดยข้อมูลจำลองอยู่ภายใต้เงื่อนไขว่า จำนวนตัวแปรอิสระที่ใช้ในแต่ละตัวแบบเท่ากับ 3 ตัวแปร (X_1, X_2, X_3) เมื่อ X_1 มีการแจกแจงแบบปกติด้วย $\mu = 0$ และ $\sigma^2 = 1$ X_2 มีการแจกแจงแบบเบอร์นูลีด้วย $p = 1/2$ และ X_3 เป็นตัวแปรที่มีค่าเปลี่ยนแปลงไปตามเวลาและมาจากการแจกแจงแบบเอกซโพเนนเชียลด้วยพารามิเตอร์ $\lambda = 1/36$ แบ่งช่วงเวลาออกเป็น 24 ช่วง ขนาดตัวอย่างที่ใช้เท่ากับ 10,000 และกระทำซ้ำจำนวน 1,000 ครั้ง และข้อมูลจริงเป็นข้อมูลทุติยภูมิเกี่ยวกับการเช่าซื้อสินค้าชนิดหนึ่ง โดยเหตุการณ์ที่สนใจคือการเกิดหนี้ NPL (Non Performing Loan) ใช้ตัวแปรอิสระจำนวน 3 ตัวแปร พิจารณาข้อมูลของลูกค้าในช่วงเวลา 1 ปี โดยแบ่งออกเป็น 4 ช่วงๆ ละ 3 เดือน และใช้ขนาดตัวอย่างเท่ากับ 10,000 เกณฑ์ที่ใช้ในการเปรียบเทียบคือ วัดประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบทั้งสองด้วยโค้ง ROC ผลการศึกษาสรุปดังนี้

สำหรับข้อมูลจำลอง พบว่า ในทุกกรณีของการทดลองตัวแบบถดถอยโลจิสติคแบบพลวัตรมีพื้นที่ใต้โค้ง ROC มากกว่าตัวแบบสถิตย์อย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 สำหรับข้อมูลจริง พบว่า ทั้งสองตัวแบบมีพื้นที่ใต้โค้ง ROC ใกล้เคียงกัน ซึ่งตัวแบบพลวัตรมีพื้นที่ใต้โค้งมากกว่าเล็กน้อย แต่เราไม่สามารถสรุปได้ว่าตัวแบบพลวัตรมีพื้นที่ใต้โค้ง ROC มากกว่าตัวแบบสถิตย์ที่ระดับนัยสำคัญ 0.05

ภาควิชา สถิติ ลายมือชื่อนิสิต วรุณี มุริกา

สาขาวิชา สถิติ ลายมือชื่อ อ. ที่ปรึกษาวิทยานิพนธ์หลัก เสกสรร เกียรติสุไพบูลย์

ปีการศึกษา 2551

4982223026 : MAJOR STATISTICS

KEY WORD : HAZARD FUNCTION/ LOGISTIC REGRESSION MODEL/ CREDIT SCORING MODEL/ ROC CURVE.

WARUNEE MURIGA : A COMPARISON BETWEEN A STATIC LOGISTIC REGRESSION MODEL AND A DYNAMIC LOGISTIC REGRESSION MODEL. THESIS PRINCIPAL ADVISOR : ASST. PROF. SEKSAN KAITSUPAIBUL, Ph.D., 78 pp.

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 for each sample unit. A dynamic model, or a hazard model, divides time into several periods and consider data of explanatory variables in several periods for each sample unit. We consider a discrete time hazard model for the dynamic model. We perform the comparisons on two types of data: simulated data and real data .The simulated data are generated by R program using Monte Carlo Simulation techniques. The model assumes 3 independent variables, X_1, X_2, X_3, where X_1 has normal distribution with mean 0 and variance 1, X_2 has bernoulli distribution with probability of success $1/2$ and X_3 has exponential distribution with parameter $1/36$. The simulation is divided into 24 periods with sample size equal to 10,000 and the simulation is repeated for 1,000 times. The real data is a secondary data from a credit activity. The interested event of dependent variable is non performing loan (NPL) of customers. The number of independent variables is equal to 3, and the sample size equals 10,000. The data considered for each customer is within one-year period divided into four 3-months long subperiods. The criteria employed for the comparison is the area under ROC curve for a prediction purpose. The conclusions of this study are as follows.

For simulated data, in most simulation runs, the area under ROC curve from the dynamic model is greater than the area under ROC curve of the static model at significance level of 0.05. Therefore the dynamic model performs better than the static model. For the real data set, the area under ROC curve from the dynamic model is only slightly more than that of the static model, but not significant at significance level of 0.05. Therefore, we cannot conclude that the dynamic model performs better than the static model.

Department : Statistics Student's signature : Warunee Muriga

Field of study : Statistics Principal Advisor's signature :

Academic year : 2008

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดีจากความอนุเคราะห์ของบุคคลหลายฝ่ายด้วยกัน ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาสละเวลาให้คำแนะนำ ปรึกษา ตลอดจนแก้ไขข้อบกพร่องต่างๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. ธีระพร วีระถาวร รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา และ ดร.โอวาท สุนันท์ ในฐานะประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาตรวจแก้ ให้ข้อคิด และแนะแนวทางที่ทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น ทั้งนี้ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติที่ให้โอกาสทางการศึกษา และประสิทธิประสาทความรู้ให้แก่ผู้วิจัยจนกระทั่งสำเร็จการศึกษา

ท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา ซึ่งให้การสนับสนุน และเพื่อนๆ ที่ให้กำลังใจแก่ผู้วิจัยเสมอมาจนกระทั่งสำเร็จการศึกษา

สารบัญ

หน้า

บทคัดย่อภาษาไทย. ..... ง

บทคัดย่อภาษาอังกฤษ..... จ

กิตติกรรมประกาศ..... ฉ

สารบัญ..... ช

สารบัญตาราง..... ฌ

สารบัญภาพ..... ฐ

บทที่ 1 บทนำ..... 1

ความเป็นมาและความสำคัญของปัญหา..... 1

วัตถุประสงค์ของการวิจัย..... 2

สมมติฐานของการวิจัย..... 2

ขอบเขตของการวิจัย..... 3

ข้อตกลงเบื้องต้น..... 3

คำจำกัดความที่ใช้ในการวิจัย..... 4

เกณฑ์การตัดสินใจ..... 7

ประโยชน์ที่คาดว่าจะได้รับ..... 8

บทที่ 2 ทฤษฎีและสถิติที่เกี่ยวข้อง..... 9

ข้อสมมติทั่วไปของตัวแบบถดถอยโลจิสติค..... 9

การประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด..... 11

การวิเคราะห์ความถดถอยโลจิสติคแบบสถิตย์..... 12

การวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตร..... 13

Survival function หรือ ฟังก์ชันการอยู่รอด..... 14

Hazard function หรือ ฟังก์ชันอัตราภาวะภัย..... 14

บทที่ 3 วิธีดำเนินการวิจัย..... 16

แผนการดำเนินการวิจัย..... 16

ขั้นตอนการดำเนินงาน..... 18

หน้า

การจัดข้อมูลสำหรับการวิเคราะห์แบบสถิตย์และแบบพลวัตร..... 19
การคำนวณความน่าจะเป็นของเหตุการณ์ของหน่วยตัวอย่าง $(\hat{p}_i)$ สำหรับตัวแบบพลวัตร..... 20
บทที่ 4 ผลการวิเคราะห์ข้อมูล..... 22
สำหรับการแบ่งช่วงเวลาเป็น 12 ช่วงเวลาย่อย..... 23
กรณีที่ 1..... 23
กรณีที่ 2..... 26
กรณีที่ 3..... 29
สำหรับการแบ่งช่วงเวลาเป็น 24 ช่วงเวลาย่อย..... 32
กรณีที่ 4..... 32
กรณีที่ 5..... 35
กรณีที่ 6..... 38
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ..... 41
สรุปผลการวิจัย..... 41
อภิปรายผลการวิจัย..... 44
ข้อเสนอแนะ..... 45
รายการอ้างอิง..... 46
ภาคผนวก..... 48
ประวัติผู้เขียนวิทยานิพนธ์..... 78

สารบัญตาราง

ตารางที่ หน้า

4.1 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 1..... 23

4.2 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 1..... 24

4.3 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 1..... 25

4.4 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 2..... 26

4.5 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 2..... 26

4.6 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 2..... 27

4.7 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 3..... 29

4.8 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 3..... 29

4.9 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 3..... 30

4.10 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 4..... 32

4.11 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 4..... 33

4.12 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 4..... 33

4.13 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 5..... 35

4.14 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 5..... 35

4.15 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 5..... 36

ตารางที่ หน้า

4.16 แสดงตัวอย่างค่าสัมประสิทธิ์ความถดถอยของการประมาณของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง สำหรับกรณีที่ 6..... 38

4.17 แสดงพิสัยของค่าสถิติ R^2 ของตัวแบบทั้งสอง จากการจำลอง 1,000 ครั้ง กรณีที่ 6..... 38

4.18 แสดงพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสองและสถิติทดสอบความแตกต่างของพื้นที่ใต้โค้ง ROC กรณีที่ 6..... 39

5.1 สรุปค่าสถิติ R^2 ของตัวแบบสถิตย์และตัวแบบพลวัตรในการศึกษาทั้ง 6 กรณี..... 41

5.2 สรุปพื้นที่ใต้โค้ง ROC และผลการเปรียบเทียบของตัวแบบสถิตย์และตัวแบบพลวัตรในการศึกษาทั้ง 6 กรณี..... 42

5.3 เปรียบเทียบค่าพารามิเตอร์และค่าสัมประสิทธิ์ความถดถอยของการประมาณในตัวแบบสถิตย์และพลวัตร สำหรับข้อมูลจำลองทั้ง 6 กรณี..... 43

สารบัญภาพ

รูปภาพ หน้า

1.1 ตัวอย่างเส้นโค้งและพื้นที่ใต้โค้ง ROC..... 5

3.1 การจัดข้อมูลสำหรับการวิเคราะห์แบบสถิตย์และแบบพลวัตร..... 19

3.2 การคำนวณความน่าจะเป็นของเหตุการณ์ของหน่วยตัวอย่าง $(\hat{p}_t)$ สำหรับตัวแบบพลวัตร..... 20

4.1 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 1 25

4.2 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 2 28

4.3 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 3 31

4.4 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 4 34

4.5 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 5 37

4.6 แสดงตัวอย่างโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรในการทดลอง 1 ครั้ง กรณีที่ 6 40

A แสดงโค้ง ROC ของตัวแบบสถิตย์และตัวแบบพลวัตรสำหรับข้อมูลจริง 62