อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท

นางสาวปฐมาภรณ์ สานุกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

POWER OF THE TEST OF THE TEST STATISTIC FOR AREA UNDER ROC CURVE ON A PROBIT MODEL

Miss Patamaporn Sanukool

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2009

หัวข้อวิทยานิพนธ์	อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท
โดย	นางสาวปฐมาภรณ์ สานุกูล
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)

..... กรรมการ

(รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา)

..... กรรมการ

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

..... กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

ปฐมาภรณ์ สานุกูล : อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท. (POWER OF THE TEST OF THE TEST STATISTIC FOR THE AREA UNDER ROC CURVE ON A PROBIT MODEL) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร. เสกสรร เกียรติสุไพบูลย์, 77 หน้า.

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติดังกล่าวบนตัวแบบโพรบิท โดยทำการศึกษา 2 ส่วน คือ ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve ซึ่งแบ่งเป็นกรณีตัวแปรอิสระจำนวน 1 และ 2 ตัวแปร โดยทำการศึกษาจากการพิสูจน์ทางคณิตศาสตร์ และส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เมื่อตัวแปรอิสระ 1 ตัวแปร โดยทำการศึกษาจากสถานการณ์จำลอง ซึ่งข้อมูลอยู่ภายใต้เงื่อนไขว่าตัวแปรอิสระ (X) มีการแจกแจงแบบปกติด้วย $\mu = 1$ และ $\sigma^2 = 1$ โดยที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง β_0 คงที่ ขนาดตัวอย่างเท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000 ในการทดลองซ้ำจำนวน 2,000 รอบ ผลการศึกษาสรุปได้ดังนี้

ส่วนที่ 1 กรณีตัวแปรอิสระจำนวน 1 ตัวแปร พบว่า เมื่อค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์จากวิธีการประมาณ 2 วิธี นั่นคือ $\hat{\beta}_1$ และ $\hat{\beta}_1'$ มีทิศทางเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองมีค่าเท่ากัน และกรณีตัวแปรอิสระจำนวน 2 ตัวแปร พบว่า เมื่อค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์จากวิธีการประมาณ 2 วิธี นั่นคือ $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ และ $(\hat{\beta}_0', \hat{\beta}_1', \hat{\beta}_2')$ โดยที่ $\hat{\beta}_0 = \hat{\beta}_0'$ และ $\hat{\beta}_1 = \hat{\beta}_1'$ จะมีช่วงเปิด (a,b) ซึ่งถ้าค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_2'$ ตกอยู่ในช่วงเปิดดังกล่าวแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองมีค่าเท่ากัน ในส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น และเมื่อขนาดตัวอย่างและระดับนัยสำคัญเพิ่มขึ้น ส่งผลให้ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้นเช่นกัน จนมีค่าอำนาจการทดสอบของตัวสถิติเข้าใกล้และเกือบเท่ากับ 1.000 ในเกือบทุกกรณี

ภาควิชา สถิติ ลายมือชื่อนิสิต ปฐมาภรณ์ สานุกูล

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก

ปีการศึกษา 2552

5081833126 : MAJOR STATISTICS

KEYWORDS: ROC CURVE / AREA UNDER THE ROC CURVE / PROBIT MODEL / POWER OF THE TEST.

PATAMAPORN SANUKOOL: POWER OF THE TEST OF THE TEST STATISTIC FOR THE AREA UNDER ROC CURVE ON A PROBIT MODEL. THESIS ADVISOR: ASST. PROF. SEKSAN KIATSUPAIBUL, Ph.D. 77 pp.

The objective of this research is to evaluate the effect of an invariance property of ROC curve and the power of the test of the test statistic on a probit model. The research is divided into two parts. Part 1: Some invariance properties of ROC curve are proved. The probit models under investigation are the univariate probit model and bivariate probit model. Part 2: The power of the test of the test statistic for the area under ROC curve is studied via simulation. We study the case when the regression coefficient β_1 varies, but β_0 is fixed. The experiment is done under the following conditions. The model assumes 1 independent variable (X), where X has normal distribution with mean 1 and variance 1. The number of sample size varies from 50 to 1,000. In each case study, the power of the test is estimated from 2,000 simulation runs. The conclusions of this research are as follows

Part 1: For the case of the probit model with 1 independent variable, given 2 estimates of the regression coefficient $\hat{\beta}_1$ and $\hat{\beta}_1'$, if the signs of $\hat{\beta}_1$ and $\hat{\beta}_1'$ are the same, the area under ROC curves of the two predicted models are equal. For the case of the probit model with 2 independent variables, given 2 estimates of the regression coefficients $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ and $(\hat{\beta}_0', \hat{\beta}_1', \hat{\beta}_2')$, where $\hat{\beta}_0 = \hat{\beta}_0'$ and $\hat{\beta}_1 = \hat{\beta}_1'$, there exists an open interval such that if $\hat{\beta}_2'$ is in the open interval, the area under ROC curves of the two predicted models are equal. Part 2: The power of the test statistic is an increasing function of the absolute value of β_1. When either the sample size or the significance level is higher, the power of the test statistic approach 1.000.

Department : Statistics Student's Signature ปรมาภรณ์ สานุกูล

Field of Study : Statistics Advisor's Signature

Academic Year : 2009

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยดีจากความอนุเคราะห์ของบุคคลหลายฝ่ายด้วยกัน ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร . เสกสรร เกียรติสุไพบูลย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาสละเวลาให้คำแนะนำ ปรึกษา ตลอดจนแก้ไขข้อบกพร่องต่าง ๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณ รองศาส ตราจารย์ ดร . ธีระพร วีระถาวร รองศาสตราจารย์ ดร . สุพล ดุรงค์วัฒนา รองศาสตราจารย์ ดร . กัลยา วานิชย์บัญชา และ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิชย์ ในฐานะประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาตรวจแก้ให้ข้อคิด และแนะแนวทางที่ทำให้วิทยานิพนธ์ฉบับนี้สมบู รณ์ยิ่งขึ้น ทั้งนี้ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติที่ให้โอกาสทางการศึกษา และประสิทธิประสาทความรู้ให้แก่ผู้วิจัยจนกระทั่งสำเร็จการศึกษา

ท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา ซึ่งให้การสนับสนุน และเพื่อน ๆ ที่ให้กำลังใจแก่ผู้วิจัยเสมอมาจนกระทั่งสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันงานวิจัยด้านต่าง ๆ เกือบทุกแขนงทั้งทางด้านเศรษฐศาสตร์ สังคมศาสตร์ การแพทย์ วิทยาศาสตร์ ธุรกิจ การเงิน ได้มีการนำเทคนิคการพยากรณ์มาใช้ซึ่งมีด้วยกันหลายวิธี โดยการเลือกใช้นั้นขึ้นอยู่กับลักษณะของข้อมูลที่นำมาศึกษาและจุดประสงค์ของการพยากรณ์ หากการพยากรณ์มุ่งหวังให้เกิดการจำแนกประเภท การแบ่งกลุ่ม วิธีการหนึ่งที่เป็นที่นิยมในปัจจุบัน คือ การพยากรณ์ด้วยตัวแบบโลจิท (Logit Model) ตัวแบบโพรบิท (Probit Model) และเทคนิคการทำเหมืองข้อมูลอื่น ๆ เช่น ตัวแบบต้นไม้ตัดสินใจ (decision tree) และ นิวรัลเน็ทเวอร์ค (Neural network) ซึ่งเป็นการวิเคราะห์ข้อมูลโดยการหาตัวแบบความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระเพื่อนำตัวแบบที่ได้ไปใช้ในการพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยที่ตัวแปรตามเป็นตัวแปรจำแนกพวกหรือข้อมูลเชิงคุณภาพ เช่ น ในด้านการแพทย์ต้องการศึกษาปัจจัยที่มีผลต่อการเกิดโรคเลปโตสไปโรลิส ซึ่งตัวแปรตามเป็นผู้ป่วยที่เป็นโรคเลปโตสไปโรลิสและไม่เป็นโรคเลปโตสไปโรลิส ทางด้านธุรกิจ เช่น การเช่าซื้อ ต้องการหาปัจจัยที่มีผลต่อการเกิดหนี้เสียของลูกหนี้จึงจำแนกลูกหนี้ออกเป็นลูกหนี้ที่มีโอกาสเกิดหนี้เสียและไม่มีโอกาสเกิดหนี้เสีย หรือแม้กระทั่งในการพยากรณ์อากาศเพื่อศึกษาปัจจัยที่ส่งผลต่อการตกของหิมะซึ่งจำแนกสภาพอากาศออกเป็นการตกของหิมะและไม่ตกของหิมะ เป็นต้น ส่วนตัวแปรอิสระอาจเป็นข้อมูลเชิงปริมาณหรือข้อมูลเชิงคุณภาพอย่างใดอย่างหนึ่งหรือทั้งสองอย่าง ด้วยความหลากหลายของตัวแบบพยากรณ์ ขั้นตอนที่สำคัญขั้นตอนหนึ่งในการพยากรณ์คือการเลือกตัวแบบที่มีประสิทธิภาพ การวัดประสิทธิภาพของตัวแบบสามารถทำได้โดยการวัดอัตราความถูกต้องของการพยากรณ์ ซึ่งการวัดที่เป็นที่นิยมวิธีหนึ่งคือการใช้ Receiver Operating Characteristic curve หรือ ROC curve

จากงานวิจัยของ Nancy A. Obuchowski (2003) ได้ศึกษาเกี่ยวกับ Receiver Operating Characteristic curve หรือ ROC curve โดยนำเสนอให้เห็นถึงเหตุผลและที่มาของการใช้ ROC curve เป็นเครื่องมือในการวัดอัตราความถูกต้องของการพยากรณ์ โดยค่าที่ถูกใช้เป็นตัวบอกความถูกต้องของการพยากรณ์คือ ค่าพื้นที่ใต้โค้ง ROC พร้อมทั้งยกตัวอย่างและวิธีการคำนวณหาพื้นที่ใต้โค้ง ROC และนอกจากนี้จากงานวิจัยของ Seong Ho Park (2004) ได้ศึกษาการใช้เส้นโค้ง ROC ในงานแพทย์รังสีวิทยา โดยนำเสนอให้เ ห็นว่า ROC curve สามารถใช้ในการคัดเลือกตัวแบบที่เหมาะสมที่สุดที่ใช้ในการพยากรณ์โดยพิจารณาจากค่าพื้นที่ใต้โค้ง ROC

นั่นคือ ตัวแบบใดให้ค่าพื้นที่ใต้โค้ง ROC มาก ตัวแบบดังกล่าวจะถือเป็นตัวแบบที่เหมาะสมที่ใช้ในการพยากรณ์

จากงานวิจัยที่กล่าวมาข้างต้นทำให้ผู้วิจัยต้องการศึกษาลักษณะความไม่ผันแปรของ ROC curve และ ประเมินอำนาจการทดสอบของพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เพื่อเป็นแนวทางในการนำความรู้ดังกล่าวไปประยุกต์ใช้ในการตัดสินใจเลือกตัวแบบที่เหมาะสมในการพยากรณ์โดยอาศัย ROC curve

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve ที่มีต่อตัวแบบพยากรณ์โพรบิท
2. หาอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC สำหรับการตรวจสอบความถูกต้องของตัวแบบพยากรณ์โพรบิท

ขอบเขตการวิจัย

ขอบเขตการวิจัยในวิทยานิพนธ์ฉบับนี้จะครอบคลุมถึงการศึกษาในเรื่องต่อไปนี้

1. ศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve ที่มีต่อตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระมีจำนวน 1 และ 2 ตัวแปร
2. หาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อมีตัวแปรอิสระตัวแปรเดียว ดังนี้
    1. คุณสมบัติของตัวแบบโพรบิทอย่างง่ายแบบ 2 กลุ่ม(Simple binary probit model) โดยกำหนด
        - ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพมีการแจกแจงแบบเบอร์นูลี และกำหนดค่าตัวแปรตามมีค่าเพียง 2 ค่า คือ 0 (เหตุการณ์ที่ไม่สนใจ) และ 1 (เหตุการณ์ที่สนใจ)
        - ตัวแปรอิสระ (X) เป็นข้อมูลเชิงปริมาณ โดยมีการแจกแจงแบบปกติ (Normal distribution) ด้วยพารามิเตอร์ μ และ σ นั่นคือ $X \sim N(\mu, \sigma^2)$ ในงานวิจัยครั้งนี้จะศึกษาที่ $\mu = 1$ และ $\sigma^2 = 1$
        - ค่าความคลาดเคลื่อน (e_i) โดยมีการแจกแจงแบบปกติ (Normal distribution) ด้วยพารามิเตอร์ μ และ σ นั่นคือ $e_i \sim N(\mu, \sigma^2)$ ในงานวิจัยครั้งนี้จะศึกษาที่ $\mu = 0$ และ $\sigma^2 = 1$

2. สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในการจำลองข้อมูล ที่ $\beta_1 = -1.00$ $\beta_1 = -0.80$ $\beta_1 = -0.60$ $\beta_1 = -0.40$ $\beta_1 = -0.20$ $\beta_1 = 0.20$ $\beta_1 = 0.40$ $\beta_1 = 0.60$ $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ โดยที่ β_0 คงที่ ($\beta_0 = 0.00$)
3. จำนวนขนาดตัวอย่าง (n) เป็น 50, 100, 200, 300, 400, 500 และ 1,000
4. กำหนดการกระทำซ้ำในแต่ละสถานการณ์เป็น 2,000 รอบ
5. กำหนดระดับนัยสำคัญ 0.01 และ 0.05

สมมติฐานการวิจัย

1. กรณีตัวแปรอิสระจำนวน 1 ตัวแปร เมื่อค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์จากวิธีการประมาณ 2 วิธี นั่นคือ $\hat{\beta}_1$ และ $\hat{\beta}_1'$ มีทิศทางเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองมีค่าเท่ากัน
2. กรณีตัวแปรอิสระจำนวน 2 ตัวแปร เมื่อค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์จากวิธีการประมาณ 2 วิธี นั่นคือ $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ และ $(\hat{\beta}_0', \hat{\beta}_1', \hat{\beta}_2')$ โดยที่ $\hat{\beta}_0 = \hat{\beta}_0'$ และ $\hat{\beta}_1 = \hat{\beta}_1'$ จะมีช่วงเปิด (a, b) ซึ่งถ้าค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_2'$ ตกอยู่ในช่วงเปิดดังกล่าวแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองมีค่าเท่ากัน
3. ค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยของตัวแบบโพรบิทในการจำลองข้อมูล จำนวนขนาดตัวอย่าง และ ระดับนัยสำคัญ ส่งผลให้อำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น

วิธีดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงทฤษฎี (Theorical Research) ได้กำหนดสถานการณ์ต่าง ๆ สำหรับการศึกษา Receiver Operating Characteristic curve หรือ ROC curve บนตัวแบบโพรบิท มีขั้นตอนการดำเนินการวิจัย ดังนี้

1. ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve ที่มีต่อตัวแบบพยากรณ์โพรบิท : พิสูจน์ทางคณิตศาสตร์ โดยแบ่งออกเป็น 2 กรณี นั่นคือ
   - กรณีตัวแปรอิสระจำนวน 1 ตัวแปร พร้อมยกตัวอย่าง
   - กรณีตัวแปรอิสระจำนวน 2 ตัวแปร พร้อมยกตัวอย่าง
2. หาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร: ศึกษาด้วยการจำลอง โดยมีขั้นตอนดังนี้

ขั้นที่ 1: คำนวณค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1

1. จำลองข้อมูลให้เป็นไปตาม H_0 : $AUC = 0.50$ โดยการจำลองข้อมูลตัวแปรตามให้มีการแจกแจงแบบเบอร์นูลีที่มีสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ที่ $\beta_1 = 0.00$ โดยที่ β_0 คงที่ ($\beta_0 = 0.00$) ตัวแปรอิสระจำนวน 1 ตัวแปรให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 1$ และ $\sigma^2 = 1$ นั่นคือ $X \sim N(1,1)$ และค่าความคลาดเคลื่อนให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 0$ และ $\sigma^2 = 1$ นั่นคือ $e_i \sim N(0,1)$ โดยกำหนดขนาดตัวอย่าง (n) เท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000
2. ทำการประมาณค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_0 และ β_1 เพื่อสร้างตัวแบบที่ใช้ในการพยากรณ์
3. พล็อต ROC curve โดยใช้ตัวแบบพยากรณ์ที่ได้จากข้อ 2 และคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC
4. นำค่าประมาณพื้นที่ใต้โค้ง ROC ดังกล่าวไปทำการคำนวณหาค่า p-value และทำการเปรียบเทียบค่า p-value ที่ระดับนัยสำคัญเป็น 0.01 และ 0.05 เพื่อที่จะได้ตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมติฐานว่าง นับจำนวนครั้งที่ปฏิเสธสมมติฐานว่าง H_0 : $AUC = 0.50$
5. ทำซ้ำในขั้นตอนที่ 1 – 4 จำนวน 2,000 ครั้ง สำหรับทุกขนาดตัวอย่าง
6. ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เท่ากับจำนวนครั้งที่ปฏิเสธสมมติฐานว่างหารด้วย 2,000
7. สรุปผลการวิจัย

ขั้นที่ 2: ทดสอบความสามารถในการควบคุมค่าคลาดเคลื่อนประเภทที่ 1 โดยใช้การทดสอบทวินามภายใต้ระดับนัยสำคัญ 0.05

ขั้นที่ 3: หาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร: ศึกษาด้วยการจำลอง

1. จำลองข้อมูลให้เป็นไปตาม H_1 : $AUC \neq 0.50$ โดยการจำลองข้อมูลตัวแปรตามให้มีการแจกแจงแบบเบอร์นูลีที่มีสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ที่ $\beta_1 = -1.00$ $\beta_1 = -0.80$ $\beta_1 = -0.60$ $\beta_1 = -0.40$ $\beta_1 = -0.20$ $\beta_1 = 0.20$ $\beta_1 = 0.40$ $\beta_1 = 0.60$ $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ โดยที่ β_0 คงที่ ($\beta_0 = 0.00$) ตัวแปรอิสระจำนวน 1 ตัวแปร ให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 1$ และ $\sigma^2 = 1$ นั่นคือ $X \sim N(1,1)$ และค่าความคลาดเคลื่อนให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 0$ และ $\sigma^2 = 1$ นั่นคือ $e_i \sim N(0,1)$ โดยกำหนดขนาดตัวอย่าง (n) เท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000
2. ทำการประมาณค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_0 และ β_1 เพื่อสร้างตัวแบบที่ใช้ในการพยากรณ์
3. พล็อต ROC curve โดยใช้ตัวแบบพยากรณ์ที่ได้จากข้อ 2 และคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC
4. นำค่าประมาณพื้นที่ใต้โค้ง ROC ดังกล่าวไปทำการคำนวณหาค่า p-value และทำการเปรียบเทียบค่า p-value ที่ระดับนัยสำคัญเป็น 0.01 และ 0.05 เพื่อที่จะได้ตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมติฐานว่าง นับจำนวนครั้งที่ปฏิเสธสมมติฐานว่าง H_0 : $AUC = 0.50$
5. ทำซ้ำในขั้นตอนที่ 1 – 4 จำนวน 2,000 ครั้ง สำหรับทุกสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 สำหรับทุกกรณี
6. ค่าอำนาจการทดสอบเท่ากับจำนวนครั้งที่ปฏิเสธสมมติฐานว่างหารด้วย2,000
7. สรุปผลการวิจัย

เกณฑ์การตัดสินใจ

อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท จะทำการศึกษาในกรณีที่ตัวสถิติสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ โดยมีหลักการพิจารณาดังต่อไปนี้

1. ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 หาได้จากสัดส่วนของเหตุการณ์ที่ปฏิเสธสมมติฐานว่าง $H_0 : AUC = 0.50$ เมื่อสมมติฐานว่างนั้นเป็นจริงและในการทดสอบว่าตัวสถิตินั้นสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้หรือไม่ จะใช้การทดสอบทวินาม (Binomail Test) โดยทำการทดสอบว่า ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของแต่ละตัวสถิติ มีค่าไม่เกินความน่าจะเป็นของความคลาดเคลื่อนที่กำหนด (α_0) หรือไม่ ภายใต้ระดับนัยสำคัญของการทดสอบทวินาม $\alpha^* = 0.10$ โดยมีรูปแบบการทดสอบเป็นดังนี้

$$H_0 : \alpha \le \alpha_0$$
$$H_1 : \alpha > \alpha_0$$

โดยทฤษฎีบทลิมิตสู่ส่วนกลาง จะได้ว่า

$$P\left(\frac{\hat{\alpha} - \alpha_0}{\sqrt{\alpha_0(1-\alpha_0)/n}} < z_{\alpha^*}\right) = 1-\alpha^*$$

หรือ

$$P\left(\hat{\alpha} < \alpha_0 + z_\alpha \cdot \sqrt{\alpha_0(1-\alpha_0)/n}\right) = 1-\alpha^*$$

โดยที่ α แทน ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของตัวสถิติ

$\hat{\alpha}$ แทน ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของตัวสถิติที่ได้จากการทดลอง

α_0 แทน ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ที่กำหนด

α^* แทน ระดับนัยสำคัญของการทดสอบทวินาม เท่ากับ 0.05

z_α แทน ค่า 1.645 ได้จากตารางการแจกแจงปกติมาตรฐานที่

ระดับนัยสำคัญ $\alpha^* = 0.10$

n แทน จำนวนการทำซ้ำของการทดลอง เท่ากับ 2,000 ครั้ง

- ถ้าค่า $\hat{\alpha}$ เปรียบเทียบกับ $\alpha_0 = 0.01$ ช่วงของการยอมรับ คือ $(0.006, 0.0137)$
- ถ้าค่า $\hat{\alpha}$ เปรียบเทียบกับ $\alpha_0 = 0.05$ ช่วงของการยอมรับ คือ $(0.042, 0.0580)$
- ถ้าค่า $\hat{\alpha}$ เปรียบเทียบกับ $\alpha_0 = 0.10$ ช่วงของการยอมรับ คือ $(0.089, 0.1110)$

ดังนั้น ถ้าค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 จากการทดลอง $(\hat{\alpha})$ อยู่ในช่วงของการยอมรับ กล่าวได้ว่า ตัวสถิตินั้นสามาร ถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้โดยที่ค่าอำนาจการทดสอบหาได้จากสัดส่วนของเหตุการณ์ที่ปฏิเสธสมมติฐานว่าง เมื่อสมมติฐานว่างนั้นไม่จริง

2. อำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC ในแต่ละสถานการณ์ พิจารณาได้จากค่าอำนาจการทดสอบของตัวสถิติที่สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ โดยที่ ค่าอำนาจการทดสอบหาได้จากสัดส่วนของเหตุการณ์ที่ปฏิเสธสมมติฐานว่างเมื่อสมมติฐานว่างนั้นไม่จริง

ค่าประมาณพื้นที่ใต้โค้ง ROC ($\hat{AUC}$) คือ ดัชนีในการบ่งชี้ความสามารถในการจำแนกกลุ่มหรือความเชื่อถือได้ของตัวแบบ

ความคลาดเคลื่อนประเภทที่ 1 (Type I error) คือ ความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมติฐานว่าง เมื่อสมมติฐานว่างนั้นเป็นจริง

ความคลาดเคลื่อนประเภทที่ 2 (Type II error) คือ ความผิดพลาดที่เกิดจากการยอมรับสมมติฐานหลัก เมื่อสมมติฐานหลักนั้นเป็นเท็จ

อำนาจการทดสอบ (Power of the test) คือ ความน่าจะเป็นของการปฏิเสธสมมติฐานหลัก เมื่อสมมติฐานหลักนั้นเป็นเท็จ ซึ่งมีค่าเท่ากับ $1-\beta$ เมื่อ β คือ ความคลาดเคลื่อนประเภทที่ 2

ประโยชน์ที่คาดว่าจะได้รับ

1. ความเข้าใจในคุณสมบัติของ ROC curve เพื่อความสามารถในการประเมินคุณค่าของระเบียบวิธีวัดประสิทธิภาพการพยากรณ์ด้วย ROC curve
2. เพื่อเป็นแนวทางในการนำ ROC curve มาประยุกต์ใช้ในงานด้านการพยากรณ์เชิงสถิติ
3. เพื่อเป็นเอกสารค้นคว้าและข้อมูลอ้างอิงสำหรับผู้ที่สนใจศึกษาทฤษฎี หลักการ แนวทางในการนำงานวิจัยนี้ไปใช้หรือนำไปศึกษาในประเด็นอื่น ๆ

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงตัวแบบโพรบิท (Probit Model) การพยากรณ์ที่มีผลลัพธ์ของการพยากรณ์เพียง 2 ค่า (Binary Predictor) Receiver operating characteristic curve หรือ ROC curve และพื้นที่ใต้โค้ง ROC (Area under the ROC Curve หรือ (AUC))

2.1 ตัวแบบโพรบิท (Probit Model)

ตัวแบบโพรบิท มีรูปแบบเริ่มต้นดังนี้

กำหนดให้ Y_i^* เป็นตัวแปรลาเทนท์ (Latent variable) ของหน่วยสังเกตที่ $i\,;\, i=1,2,...,n$ ที่มีรูปแบบเป็นฟังก์ชันเชิงเส้น (Linear function) ของตัวแปรอิสระ $K+1$ ตัว ($1, X_{i1},...,X_{iK}$) ที่มีสัมประสิทธิ์ ($\beta_0, \beta_1,...,\beta_K$) และค่าความคลาดเคลื่อน (e_i) โดยที่ e_i มีรูปแบบการแจกแจงแบบปกติ โดยรูปแบบฟังก์ชันเชิงเส้น เป็นดังนี้

$$Y_i^* = \sum_{k=0}^{K} \beta_k X_{ik} + e_i \qquad (1)$$

และสามารถเขียนให้อยู่ในรูปเมทริกซ์ได้ดังนี้

$$Y^* = X\beta + e \qquad (2)$$

เมื่อ $X_{(n\times K)}$ เป็นเมทริกซ์ของตัวแปรอิสระ, $\beta_{(K\times 1)}$ เป็นเวกเตอร์พารามิเตอร์ของสัมประสิทธิ์การถดถอยและ $e_{(n\times 1)}$ เป็นเวกเตอร์สุ่มซึ่งแทนความคลาดเคลื่อน โดยมีรูปแบบการแจกแจงเป็นแบบปกติ

ในทางปฏิบัติตัวแปรลาเทนท์ Y_i^* ไม่สามารถเก็บข้อมูลหรือสังเกตค่าได้จริง ดังนั้น นผลลัพธ์ซึ่งเป็นสิ่งที่สังเกตได้ต้องนำมาปรับให้เป็นตัวแปรหุ่น (Dummy Variable) แทนด้วยค่า Y_i ดังนี้

$$Y_i = \begin{cases} 1 & \text{ถ้า } Y_i^* > 0 \\ 0 & \text{ถ้า } Y_i^* \le 0 \end{cases} \quad (3)$$

จากสมการ (2) และ (3) ข้างต้น จะได้ว่า

$$\begin{aligned} P(Y=1) &= P(e_i > -X_i'\beta) \\ &= 1-\Phi(-X_i'\beta) \\ &= \Phi(X_i'\beta) \end{aligned} \quad (4)$$

โดยที่ Φ คือ ฟังก์ชันการแจกแจงสะสม (Cumulative distribution function) ของค่าความคลาดเคลื่อน (e_i) ที่มีรูปแบบการแจกแจงแบบปกติ

<u>การประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบโพรบิท</u>

จากค่า Y_i ที่เก็บข้อมูลมาได้จะมีการแจกแจงแบบเบอร์นูลลี (Bernoulli distribution) ซึ่งความน่าจะเป็นถูกกำหนดโดยสมการ (4) (Green, 2000 : 811 – 895) ดังนี้

$$\Phi(X_i'\beta) = \int_{-\infty}^{X_i'\beta/\sigma} \left[\frac{1}{(2\pi)^{1/2}}\right] \exp\left[-\frac{t^2}{2}\right] dt \quad (5)$$

เมื่อ Φ เป็น การแจกแจงสะสมของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ
ดังนั้น

$$p_i = P(Y=1) = \int_{-\infty}^{\eta I} \phi(t)\,dt \quad (6)$$

เมื่อ ϕ เป็นฟังก์ชันความหนาแน่นของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ และพบว่า

$$\phi(t) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{t^2}{2}\right) \sim N(0,1)$$

และจะได้ว่า

$$\eta_i = \Phi^{-1}[p_i] = X_i'\beta \quad (7)$$

นั่นคือ การแปลงโพรบิท (Probit transformation) ซึ่งเป็นตัวผกผันของฟังก์ชันการแจกแจงสะสมของตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน (The inverse of the standard cumulative normal distribution function) สามารถเขียนได้เป็น

$$probit(p_i) = \Phi^{-1}(p_i) = \eta_i = X_i'\beta \quad (8)$$

โดยที่

$$p_i = \Phi(X_i'\beta) = \int_{-\infty}^{\eta_i}\left[\frac{1}{(2\pi)^{1/2}}\right]\exp\left[-\frac{t^2}{2}\right]dt$$

จากสมการที่ (6) จะพบว่าค่าของฟังก์ชันจะอยู่ระหว่าง 0 และ 1 สำหรับทุกค่าของ $X_i'\beta$ โดยฟังก์ชันดังกล่าวเป็นฟังก์ชันการแจกแจงสะสม (Cumulative distribution function) ของตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน (Standard normal distribution)

ในการประมาณค่าสัมประสิทธิ์การถดถอย (β) ในตัวแบบโพรบิทใช้การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) ซึ่งเป็นวิธีที่ให้ค่าประมาณของสัมประสิทธิ์การถดถอย $(\hat{\beta})$ มีค่าความน่าจะเป็นสูงสุดหรือใกล้เคียงกับข้อมูลมากที่สุด โดยเริ่มจากรูปแบบฟังก์ชันภาวะน่าจะเป็น (Likelihood function) ซึ่งมีขั้นตอนดังนี้

1. ระบุฟังก์ชันการแจกแจงความน่าจะเป็นของประชากร ซึ่งในกรณีนี้ตัวแปรที่ทำการศึกษาเป็นตัวแปรตาม (Y_I) ที่มีค่าเพียง 2 ค่า คือ 0 กับ 1 ดังนั้นจึงใช้ฟังก์ชันการแจกแจงความน่าจะเป็นแบบเบอร์นูลี คือ

$$g(Y_i) = P_i^{Y_i}(1-P_I)^{1-Y_i} \qquad \text{เมื่อ} \quad Y_i = 0,1 \quad (9)$$

2. สร้างฟังก์ชันของการแจกแจงความน่าจะเป็นร่วม (Joint probability density function) ของหน่วยตัวอย่างที่อิสระ n ค่า โดยการคูณฟังก์ชันการแจกแจงความน่าจะเป็นของทุกหน่วยตัวอย่าง $(g(Y_i))$ จะได้

$$g(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^{n} g(Y_i)$$

$$= \prod_{i=1}^{n} P_i^{Y_i} (1-P_I)^{1-Y_i}$$

$$= \prod_{i}^{n} \Phi(X_i'\beta)^{Y_i} [1-\Phi(X_i'\beta)]^{1-Y_i} \quad (10)$$

ดังนั้น

$$L = \prod_{i}^{n} \Phi(X_i'\beta)^{Y_i} [1-\Phi(X_i'\beta)]^{1-Y_i} \quad \text{เมื่อ } Y_i = 0,1 \quad (11)$$

ในการหาตัวประมาณค่าสัมประสิทธิ์การถดถอย $(\hat{\beta})$ ด้วยวิธีภาวะน่าจะเป็นสูงสุด คือ ต้องทำให้ L มีค่ามากที่สุดโดยทำการหาอนุพันธ์เทียบกับ β ต่าง ๆ ซึ่งในกรณีนี้การหาค่ามากที่สุดของ L เป็นปัญหาที่ซับซ้อน การประมาณค่าดังกล่าวสามารถประมาณค่าด้วยการใช้ล็อกธรรมชาติกับฟังก์ชันภาวะน่าเป็นโดยเรียกว่า ฟังก์ชันล็อกภาวะน่าจะเป็น (L) ซึ่งมีส่วนช่วยทำให้ลักษณะของฟังก์ชันภาวะน่าจะเป็นมีลักษณะง่ายขึ้น ดังนี้

$$\ln L = = \sum_{i} \{Y_i \ln \Phi(X_i'\beta) + (1-Y_i) \ln[1-\Phi(X_i'\beta)]\} \quad (12)$$

และสามารถหาค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอย $\beta' = (\beta_0, \beta_1, ..., \beta_K)$ โดยการหาค่า Frist order condition ของฟังก์ชันภาวะน่าจะเป็น โดยกำหนด

$$\frac{\partial \ln L(\beta)}{\partial \beta_k} = 0 \quad , \quad k = 0,1,...,K \quad (13)$$

และสามารถแก้สมการหาค่า Frist order condition ของแบบจำลองโพรบิท ซึ่งพบว่า

$$\frac{\partial \ln L(\beta)}{\partial \beta} = U(\beta) = \sum_{yi=0} -[\phi_i / (1-\Phi_i)] X_i + \sum_{yi=1} (\phi / \Phi_i) X_i = 0 \quad (14)$$

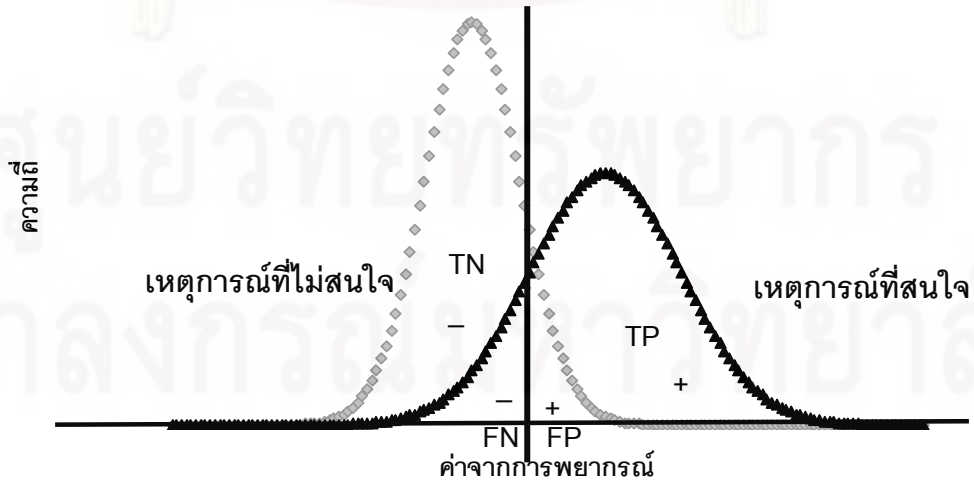
เมื่อ ϕ_i แทน ฟังก์ชันความหนาแน่นของตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน

Φ_i แทน ฟังก์ชันการแจกแจงสะสมของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

2.2 การพยากรณ์ที่มีผลลัพธ์ของการพยากรณ์เพียง 2 ค่า (Binary Predictor)

การพยากรณ์ที่มีค่าพยากรณ์เพียง 2 ค่า ถือเป็นหนึ่งในเหตุการณ์ที่ง่ายที่สุดสำหรับการพยากรณ์และเป็นส่วนสำคัญในการสร้าง Reciever Operating Characteristic curve หรือ ROC Curve โดยตัวแปรตามคุณภาพแบ่งออกเป็น 2 กรณีคือ ตัวแปรตามมีค่าเท่ากับ 1 หรือผลการทดสอบเป็นบวก (Positive) เมื่อเกิดเหตุการณ์ที่สนใจในหน่วยที่ i และตัวแปรตามมีค่าเท่ากับ 0 หรือผลการทดสอบเป็นลบ (Negative) เมื่อเกิดเหตุการณ์ที่ไม่สนใจในหน่วยที่ i เช่น การศึกษาว่าบุคคลผู้หนึ่งจะไปลงคะแนนเสียงเลือกตั้งหรือไม่ ในกรณีนี้ตัวแปรตามที่ศึกษามี 2 ค่า คือ การไปลงคะแนนเสียงเลือกตั้งซึ่งเป็นการเกิดเหตุการณ์ที่สนใจให้มีค่าตัวแปรตามเท่ากับ 1 และการไม่ไปลงคะแนนเสียงเลือกตั้งเป็นเหตุการณ์ที่ไม่สนใจให้มีค่าตัวแปรตามเท่ากับ 0 หรือการวินิจฉัยจำแนกผู้ป่วยโรคเลปโตสไปโรลิส ในกรณีนี้ตัวแปรตามที่ศึกษามี 2 ค่าเช่นกันคือ การป่วยเป็นโรคเลปโตสไปโรลิส (Disease) ซึ่งเป็นการเกิดเหตุการณ์ที่สนใจให้มีค่าตัวแปรตามเท่ากับ 1 และการไม่เป็นโรคเลปโตสไปโรลิส (Normal) เป็นเหตุการณ์ที่ไม่สนใจให้มีค่าตัวแปรตามเท่ากับ 0 เป็นต้น ซึ่งนั้นหมายถึง การพยากรณ์สามารถจำแนกประชากรออกเป็น 2 กลุ่ม โดยกลุ่มหนึ่งจะเป็นกลุ่มของเหตุการณ์ที่สนใจ และอีกกลุ่มก็จะเป็นกลุ่มของเหตุการณ์ที่ไม่สนใจ (จาก Thomas A. Lasko (2005) “ The use of receiver operating characteristic curves in biomedical informatics”) ดังในรูปที่ 1 ซึ่งเป็นตัวอย่างรูปที่แสดงถึงผลการพยากรณ์ที่จำแนกเหตุการณ์ออกเป็นกลุ่มของเหตุการณ์ที่ไม่สนใจและกลุ่มของเหตุการณ์ที่สนใจ

รูปที่ 2.1 แสดงผลการพยากรณ์จำแนกประชากรออกเป็นกลุ่มเหตุการณ์ที่สนใจและกลุ่มเหตุการณ์ที่ไม่สนใจ



จากรูปที่ 2.1 หากใช้จุดตัด ซึ่งเป็นตำแหน่งตรงเส้นตรงเป็นเกณฑ์ในการจำแนกเหตุการณ์ออกเป็นกลุ่มของเหตุการณ์ที่ไม่สนใจ และ กลุ่มของเหตุการณ์ที่สนใจ พบว่า

- ในเหตุการณ์ที่ไม่สนใจบางเหตุการณ์ซึ่งมีผลการพยากรณ์เป็นบวก ผลดังกล่าวเป็นผลบวกลวง หรือ เรียกว่า false positive (FP)
- ในเหตุการณ์ที่สนใจบางเหตุการณ์ซึ่งมีผลการพยากรณ์เป็นบวก ผลดังกล่าวเป็นผลบวกจริง หรือ เรียกว่า true positive (TP)
- ในเหตุการณ์ที่สนใจบางเหตุการณ์อาจมีผลการพยากรณ์เป็นลบ ผลดังกล่าวเป็นผลลบลวง หรือ เรียกว่า false negatives (FN)
- ในเหตุการณ์ที่ไม่สนใจบางเหตุการณ์ซึ่งมีผลการพยากรณ์เป็นลบ ผลดังกล่าวเป็นผลลบจริง หรือ เรียกว่า true negatives (TN)

ดังตารางที่ 2.1

ตารางที่ 2.1 แสดงค่า Sensitivity (SN) และ Specificity (SP)

	จำนวนเหตุการณ์ที่ให้ผลบวก	จำนวนเหตุการณ์ที่ให้ผลลบ	รวม
จำนวนเหตุการณ์ที่สนใจ	TP	FN	$TP+FN$
จำนวนเหตุการณ์ที่ไม่สนใจ	FP	TN	$FP+TN$

และจากผลการพยากรณ์ดังกล่าวมาข้างต้น เครื่องมือที่ใช้วัดความถูกต้องของการพยากรณ์ซึ่งผลที่เกิดจากการพยากรณ์มีเพียง 2 ค่า (ผศ. พิศิษฐ์ นามจันทรา "เอกสารประกอบการสอนวิชาเคมีคลินิก 3") คือ

- Sensitivity (SN) คือ สัดส่วนของเหตุการณ์ที่ให้ผลการพยากรณ์เป็นผลบวกจริง (TP) ต่อจำนวนเหตุการณ์ที่สนใจทั้งหมด นั่นคือ

$$SN = \frac{TP}{TP+FN} \tag{15}$$

- Specificity (SP) คือ สัดส่วนของเหตุการ ณ์ที่ให้ผลการพยากรณ์เป็นผลลบจริง (TN) ต่อจำนวนเหตุการณ์ที่ไม่สนใจทั้งหมด นั่นคือ

$$SP = \frac{TN}{FP + TN} \tag{16}$$

จากสมการที่ (15) และ (16) เราสามารถสังเกตได้ว่า ค่า Sensitivity จะขึ้นอยู่กับจำนวนเหตุการณ์ที่สนใจทั้งหมด ส่วนค่า Specificity ก็จะขึ้นอยู่กับจำนวนเหตุการณ์ที่ไม่สนใจทั้งหมด โดยไม่มีค่าใดขึ้นอยู่กับค่าการกระจายของเหตุการณ์ที่สนใจในกลุ่มตัวอย่างทั้งหมด ซึ่งนั้นเป็นเหตุผลที่ทำให้ ค่า Sensitivity และ ค่า Specificity กลายเป็นเครื่องมือที่ใช้วัดความถูกต้องของการพยากรณ์ในการพยากรณ์ที่มีผลลัพธ์เพียง 2 ค่าที่ได้รับความนิยมอย่างแพร่หลาย

2.3 Receiver Operating Characteristic curve หรือ ROC curve

ROC curve ถูกนำมาใช้ครั้งแรกในสมัยสงครามโลกครั้งที่สองสำหรับการตรวจจับและวิเคราะห์สัญญาณ ก่อนหน้าที่จะถูกนำมาใช้อย่างแพร่หลายใ นทฤษฎีการตรวจจับสัญญาณ (signal detection theory) หลังจากนั้นไม่นานในเหตุการณ์ที่ญี่ปุ่นโจมตีฐานทัพเรือสหรัฐที่อ่าวเพริ์ลฮาร์เบอร์ ในวันที่ 7 เดือน ธันวาคม ค.ศ. 1941 (พ.ศ. 2484) กองทัพสหรัฐได้นำ ROC curve มาใช้เพิ่มความถูกต้องของการพยากรณ์ในการตรวจจับหาเครื่องบินรบของประเทศญี่ปุ่น โดยใช้สัญญาณเรดาร์

และหลังจากในปี ค.ศ. 1950 เป็นต้นมา ROC curve ก็เริ่มเป็นที่รู้จักและได้รับการยอมรับอย่างกว้างขวาง โดยถูกนำมาประยุกต์ใช้กับศาสตร์ด้านต่าง ๆ มากมาย เช่น ด้านจิตวิทยา ด้านการแพทย์ ด้านรังสีเอ็กซเรย์ รวมถึงด้านวิศวกรรม

จาก David W. Hosmer. “Applied Logistic Regression” ให้ความหมายไว้ดังนี้

ROC curve เป็นกราฟที่พล็อตระหว่างค่าของ Sensitivity และ 1 - Specificity ซึ่งการพล็อตกราฟจะได้จากการกำหนดจุดตัด ที่ระดับต่างๆ เพื่อแบ่งผลลัพธ์ของการพยากรณ์ออกเป็น 2 กลุ่ม คือกลุ่มที่เกิดเหตุการณ์ $P(Y=1) \geq$ จุดตัด และกลุ่มที่ไม่เกิดเหตุการณ์ $P(Y=1) <$ จุดตัด ดังแสดงในตัวอย่าง

2.3.1 ตัวอย่างการคำนวณหาค่า Sensitivity และ ค่า Specificity ณ จุดตัดต่าง ๆ เพื่อใช้ในการพล็อตกราฟ

จากตัวอย่างเป็นผลจากการทดสอบความทนกลูโคส2-h oral glucose tolerance(OGTT) ในกลุ่มผู้ป่วยที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวาน (จาก Thomas A. Lasko (2005) “The use of receiver operating characteristic curves in biomedical informatics”) ดังนี้

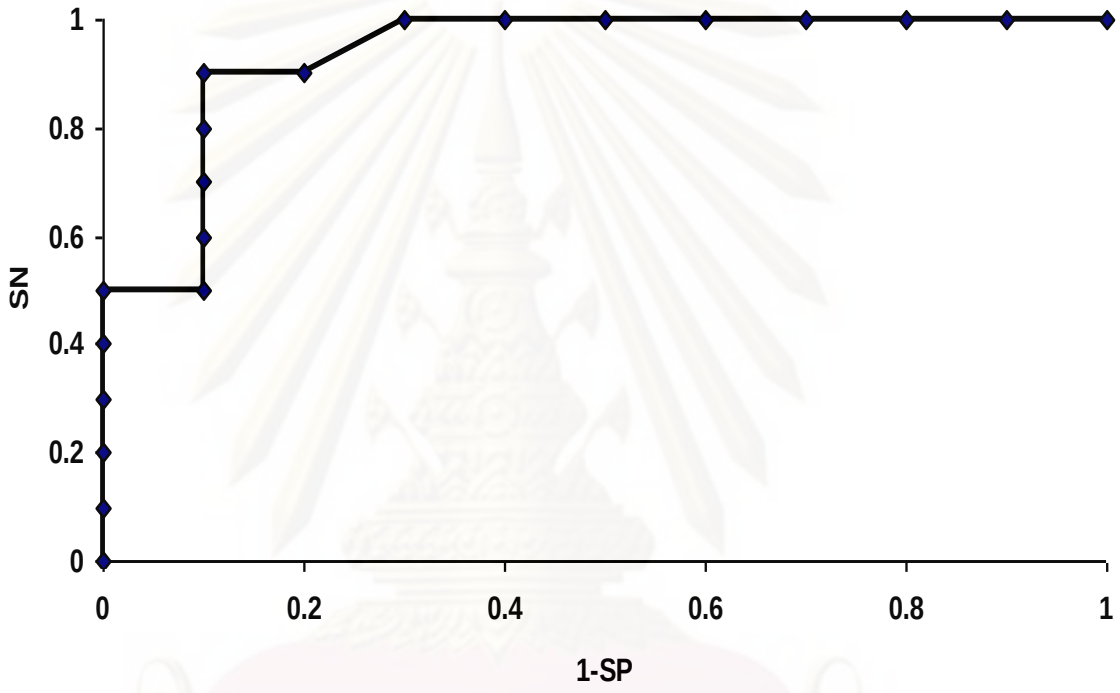
ตารางที่ 2.2 แสดงผลจากการทดสอบความทนกลูโคส 2- h oral glucose tolerance (OGTT)

ระดับกลูโคสในพลาสม่า (mmol /L)		จุดตัด *					
กลุ่มผู้ป่วยไม่เป็นโรคเบาหวาน	กลุ่มผู้ป่วยที่เป็นโรคเบาหวาน	TP	TN	FP	FN	$1-SP$	SN
		10	0	10	0	1.00	1.00
4.86		10	1	9	0	0.90	1.00
5.69		10	2	8	0	0.80	1.00
6.01		10	3	7	0	0.70	1.00
6.06		10	4	6	0	0.60	1.00
6.27		10	5	5	0	0.50	1.00
6.37		10	6	4	0	0.40	1.00
6.55		10	7	3	0	0.30	1.00
7.29	7.29	9	8	2	1	0.20	0.90
7.82		9	9	1	1	0.10	0.90
	9.22	8	9	1	2	0.10	0.80
	9.79	7	9	1	3	0.10	0.70
	11.28	6	9	1	4	0.10	0.60
	11.83	5	9	1	5	0.10	0.50
12.06		5	10	0	5	0.00	0.50
	18.48	4	10	0	6	0.00	0.40
	18.50	3	10	0	7	0.00	0.30
	20.49	2	10	0	8	0.00	0.20
	22.66	1	10	0	9	0.00	0.10
	26.01	0	10	0	10	0.00	0.00

* ระดับกลูโคสในพลาสม่าแต่ละค่าเป็นจุดตัดในแต่ละบรรทัด

เมื่อได้ค่า Sensitivity และ 1 - Specificity จากการกำหนดจุดตัด ที่ระดับต่างๆ นำค่าดังกล่าวมาพล็อตกราฟระหว่างค่าของ Sensitivity และ 1 - Specificity ดังแสดงในรูปโดยจุดแต่ละจุดในกราฟคือค่า Sensitivity และ 1 - Specificity ในแต่ละบรรทัด ดังนี้

รูปที่ 2.2 แสดง ROC curve จากผลการทดสอบความทนกลูโคส 2- h oral glucose tolerance (ตารางที่ 2.2)



จาก ROC curve ในรูปที่ 2.2 ทำให้เห็นข้อได้เปรียบของการใช้ ROC curve ในการตรวจสอบความถูกต้องของตัวแบบอีกประการหนึ่ง นั่นคือ เป็นการง่ายมากที่เราสามารถมองเห็นการเปลี่ยนแปลงระหว่างค่า Sensitivity และ Specificity สำหรับทุก ๆ จุดตัด ซึ่งส่งผลให้เห็นภาพรวมทั้งหมดของความถูกต้องหรือความเชื่อถือได้ของตัวแบบ รวมถึงสามารถคำนวณหาร้อยละของการพยากรณ์ที่ถูกต้องในแต่ละจุดตัดได้อีกด้วย

2.4 พื้นที่ใต้โค้ง ROC (Area under the Curve (AUC))

จากที่ได้กล่าวไปแล้วข้างต้น เนื่องจาก ROC curve มักถูกนำไปใช้ในการตรวจสอบความถูกต้องของตัวแบบในการพยากรณ์ โดยค่าที่ใช้เป็นดัชนีในการบ่งชี้ความถูกต้องหรือความเชื่อถือได้ของตัวแบบ คือ ค่าประมาณพื้นที่ใต้โค้ง ROC (Area under the Curve หรือ ($\hat{AUC}$)) ซึ่งในการใช้ค่าประมาณพื้นที่ใต้โค้ง ROC สำหรับบ่งชี้ความถูกต้องของตัวแบบนั้น อาศัยแนวความคิดในการตีความได้หลายรูปแบบ

1. ค่าประมาณพื้นที่ใต้โค้ง ROC คือ ความน่าจะเป็นที่ตัวแบบสามารถสร้างค่าพยากรณ์ในกลุ่มเหตุการณ์ที่สนใจให้มีค่ามากกว่าค่าพยากรณ์ในกลุ่มเหตุการณ์ที่ไม่สนใจ ซึ่งนั้นส่งผลให้สามารถประเมินได้ว่าตัวแบบดังกล่าวมีความสามารถในการจำแนกระหว่างกลุ่มเหตุการณ์ที่สนใจกับเหตุการณ์ที่ไม่สนใจออกจากกันได้ดีเพียงใด

2. ค่าประมาณพื้นที่ใต้โค้ง ROC เป็นการรวมเครื่องมือพื้นฐานที่ใช้วัดความถูกต้องของการพยากรณ์ที่มีค่าพยากรณ์เพียง 2 ค่า ระหว่าง Sensitivity และ Specificity สำหรับทุก ๆ จุดตัด ให้เป็นดัชนีบ่งชี้แค่เพียงค่าเดียว คื อ ค่าเฉลี่ย Sensitivity สำหรับทุก Specificity หรือ ค่าเฉลี่ย Specificity สำหรับทุก Sensitivity

โดยในการคำนวณค่าพื้นที่ใต้โค้ง ROC สามารถคำนวณได้หลายวิธี สำหรับในงานวิจัยนี้ขอกล่าวอ้างถึงระเบียบวิธีที่ไม่ใช้พารามิเตอร์ (Nonparametric method) (จาก Thomas A. Lasko (2005) “The use of receiver operating characteristic curves in biomedical informatics”) ดังรายละเอียดต่อไปนี้

2.4.1 ระเบียบวิธีที่ไม่ใช้พารามิเตอร์ (Nonparametric method)

ในการหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้ค่าที่ได้จริงจากการพยากรณ์ (The empirical method) เป็นการสร้าง ROC curve โดยการเชื่อมต่อแต่ละจุดของข้อมูลซึ่งมีค่า Sensitivity (SN) อยู่ในแนวแกน y และ 1 - Specificity (SP) อยู่ในแนวแกน x $(1-SP, SN)$ เป็นเส้นตรง หลังจากนั้นทำการประมาณค่าพื้ นที่ใต้โค้ง ROC โดยอาศัยหลักเกณฑ์เชิงสี่เหลี่ยมคางหมู (trapezoidal rule) ซึ่งวิธีการดังกล่าวเป็นการอ้างระเบียบขั้นตอนทางสถิติที่ไม่ใช้พารามิเตอร์ โดยค่าประมาณของพื้นที่ใต้โค้ง ROC จากวิธีการดังกล่าวมีค่าเท่ากับผลการทดสอบข้อมูลชุดเดิมด้วยการทดสอบของแมนวิทนี (Mann–Whitney U Test) ซึ่งเป็นการทดสอบโดยการคำนวณค่าจากจำนวนคู่ที่มีความเป็นไปได้ของเหตุการณ์ที่สนใจกับเหตุการณ์ที่ไม่สนใจ

ซึ่งเป็นแนวความคิดเดียวกับการทดสอบผลรวมอันดับของวิลคอคสัน (Wilcoxon – rank – sum Test) และ การทดสอบ C-index

โดยกำหนดให้ $d_1, d_2, ..., d_{n_D}$ เป็นค่าพยากรณ์สำหรับกลุ่มเหตุการณ์ที่สนใจ

$h_1, h_2, ..., h_{n_H}$ เป็นค่าพยากรณ์สำหรับเหตุการณ์ที่ไม่สนใจ

n_D จำนวนค่าพยากรณ์ตัวอย่างจากกลุ่มเหตุการณ์ที่สนใจ

n_H จำนวนค่าพยากรณ์ตัวอย่างจากกลุ่มเหตุการณ์ที่ไม่สนใจ

และ $C(d_i, h_j)$ เป็นฟังก์ชันที่ใช้ในการเปรียบเทียบ โดยที่ $i = 1,2,...,n_D$ และ $j = 1,2,...,n_H$

เมื่อ

$$C(d_i, h_j) = \begin{cases} 1 & \text{ถ้า} \quad d_i > h_j \\ 0.5 & \text{ถ้า} \quad d_i = h_j \\ 0 & \text{ถ้า} \quad d_i < h_j \end{cases} \tag{18}$$

ดังนั้น ค่าประมาณของพื้นที่ใต้โค้ง ROC คือ ค่าเฉลี่ยของผลรวมที่เกิดจากฟังก์ชันที่ใช้ในการเปรียบเทียบ ($C(d_i, h_j)$) ของคู่เหตุการณ์ที่สนใจกับเหตุการณ์ที่ไม่สนใจทุกคู่ ดังนี้

$$\hat{AUC} = \frac{1}{n_D n_H} \sum_{i=1}^{n_D} \sum_{j=1}^{n_H} C(d_i, h_j) \tag{19}$$

จากที่ได้กล่าวไปแล้วว่า แนวความคิดในการหาค่าประมาณของพื้นที่ใต้โค้ง ROC ด้วยวิธีการข้างต้นเป็นแนวความคิดเดียวกับการทดสอบผลรวมอันดับของวิลคอคสัน ซึ่งเป็นการทดสอบโดยพิจารณาอันดับของข้อมูลเป็นสำคัญ หากจะใช้อันดับของข้อมูลในการหาค่าประมาณของพื้นที่ใต้โค้ง ROC สามารถประมาณค่าได้ดังรายละเอียด ดังนี้

โดยกำหนดให้ $d_1, d_2, ..., d_{n_D}$ เป็นค่าพยากรณ์สำหรับกลุ่มเหตุการณ์ที่สนใจ เมื่อ $d_1 \geq ... \geq d_{n_D}$

$h_1, h_2, ..., h_{n_H}$ เป็นค่าพยากรณ์สำหรับกลุ่มเหตุการณ์ที่ไม่สนใจ เมื่อ $h_1 \leq ... \leq h_{n_H}$

n_D จำนวนค่าพยากรณ์ตัวอย่างจากกลุ่มเหตุการณ์ที่สนใจ

n_H จำนวนค่าพยากรณ์ตัวอย่างจากกลุ่มเหตุการณ์ที่ไม่สนใจ

Z_a เป็นลำดับที่เกิดจากการรวม $d_1, d_2, ..., d_{nD}$ และ $h_1, h_2, ..., h_{nH}$ เข้าด้วยกันและทำการเรียงอันดับค่าพยากรณ์จากน้อยไปหามาก

โดยที่ $i = 1,2,...,n_D$ และ $j = 1,2,...,n_H$

ดังนั้น ค่าประมาณของพื้นที่ใต้โค้ง ROC สามารถเขียนให้อยู่ในรูปของการคำนวณโดยอาศัยอันดับ ดังนี้

$$\hat{AUC} = \frac{1}{n_D n_H}\left(\sum_{i=1}^{nD} r_i - \frac{n_D(n_D+1)}{2}\right) \quad (20)$$

เมื่อ r_i เป็นอันดับของ d_i ในอันดับ Z_a

ในกรณีที่ค่าพยากรณ์มีค่าเท่ากันหลายค่าซึ่งทำให้ค่าอันดับเท่ากันหลายค่าด้วย ให้ใช้อันดับเฉลี่ยของข้อมูลที่เท่ากันนั้น เช่น ถ้าค่าพยากรณ์ต่ำที่สุดของข้อมูลมี 2 ค่าเท่ากัน คือ 7 นั่นคือมี 7 สองค่า และเป็นค่าที่ต่ำที่สุดด้วย ดังนั้นอันดับที่ของ 7 คือ $\frac{(1+2)}{2} = 1.5$

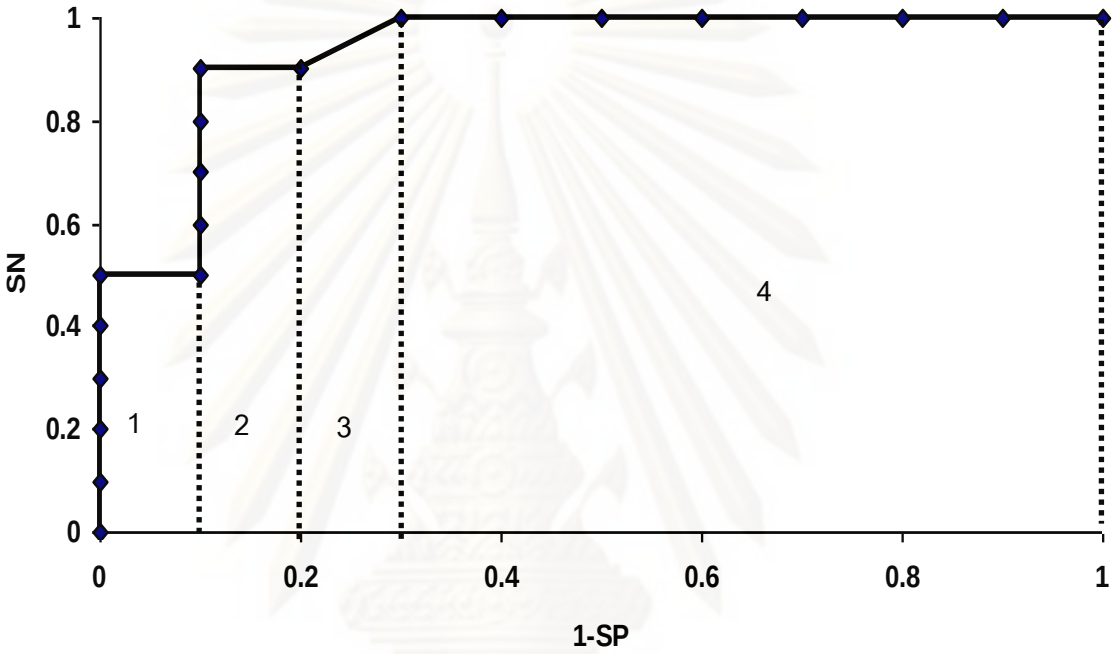
ทั้งสามวิธีที่กล่าวมาข้างต้นล้วนเป็นระเบียบวิธีที่ไม่ใช้พารามิเตอร์ ซึ่งเป็นระเบียบวิธีที่ไม่จำเป็นต้องมีข้อสมมติเกี่ยวกับข้อมูล ในการคำนวณค่าสถิติที่ใช้ในการวิเคราะห์ก็ทำได้ง่าย ไม่ยุ่งยากจึงส่งผลให้ระเบียบวิธีการนี้เป็นที่นิยมและถูกนำไปประยุกต์ใช้ในงานด้านต่าง ๆ อย่างกว้างขวาง ต่อไปเป็นตัวอย่างในการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC ด้วยวิธีการต่าง ๆ

2.4.2 ตัวอย่างในการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยอาศัยหลักเกณฑ์เชิงสี่เหลี่ยมคางหมู (trapezoidal rule)

ตัวอย่างการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยอาศัยหลักเกณฑ์เชิงสี่เหลี่ยมคางหมูจะแสดงให้เห็นโดยอาศัยผลจากการทดสอบความทนกลูโคส 2- h oral glucose tolerance (OGTT) ในกลุ่มผู้ป่วยที่เป็นโรคเบาหวานและไม่เป็นโรคเบาหวาน จากตารางที่ 2.2 และรูปที่ 2.2 ดังนี้

ขั้นที่ 1 ทำการเชื่อมต่อแต่ละจุดของข้อมูลเป็นเส้นตรง จนได้ ROC curve หลังจากนั้นแบ่งพื้นที่ใต้โค้งออกเป็นรูปสี่เหลี่ยมเล็ก ๆ ดังแสดงในภาพด้านล่าง

รูปที่ 2.3 แสดง ROC curve หลังจากทำการเชื่อมต่อจุดและแบ่งพื้นที่ใต้โค้งออกเป็นสี่เหลี่ยมเล็ก ๆ



ขั้นที่ 2 คำนวณพื้นที่ของสี่เหลี่ยมทั้ง 4 รูป

รูปที่ 1 = (0.1×0.5) = 0.050

รูปที่ 2 = $(0.2-0.1)\times0.9$ = 0.090

รูปที่ 3 = $\frac{1}{2}\times(0.3-0.2)\times(0.9+1.0)$ = 0.095

รูปที่ 4 = $(1.0-0.3)\times1.00$ = 0.700

ขั้นที่ 3 ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากับผลรวมของพื้นที่รูปสี่เหลี่ยมเล็ก ๆ ใต้โค้งทั้งหมด นั่นคือ

$\hat{AUC}$ = พท. รูปที่ 1 + พท.รูปที่ 2 + พท.รูปที่ 3 + พท.รูปที่ 4

= $0.050 + 0.090 + 0.095 + 0.700$

= 0.935

ดังนั้น ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากับ 0.935

2.4.3 ตัวอย่างในการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยการใช้ฟังก์ชันเปรียบเทียบ ($C(d_i, h_j)$)

ตัวอย่างการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยการใช้ฟังก์ชันเปรียบเทียบจะแสดงให้เห็นโดยอาศัยผลจากตารางที่ 2.2 เช่นเดียวกับในการคำนวณโดยใช้กฎสี่เหลี่ยมคางหมู **ตารางที่** 2.3 แสดงผลจากการทดสอบความทนกลูโคส 2- h oral glucose tolerance (OGTT) และการคำนวณหาค่า $\hat{AUC}$ โดยการใช้ $C(d_i, h_j)$

ระดับกลูโคสในพลาสม่า (mmol / L)		Partial Sums	
กลุ่มผู้ป่วยไม่เป็นโรคเบาหวาน (h_j)	กลุ่มผู้ป่วยที่เป็นโรคเบาหวาน (d_i)	$\sum_{i=1}^{n_D} C(d_i, h_j)$	$\sum_{j=1}^{n_H} C(d_i, h_j)$
4.86		(10)(1) = 10	
5.69		(10)(1) = 10	
6.01		(10)(1) = 10	
6.06		(10)(1) = 10	
6.27		(10)(1) = 10	
6.37		(10)(1) = 10	
6.55		(10)(1) = 10	
7.29	7.29	(9)(1)+(9)(0.5) = 9.5	(7)(1)+(1)(0.5) = 7.5
7.82		(9)(1) = 9	
	9.22		(9)(1) = 9
	9.79		(9)(1) = 9
	11.28		(9)(1) = 9
	11.83		(9)(1) = 9
12.06		(5)(1) = 5	
	18.48		(10)(1) = 10
	18.50		(10)(1) = 10
	20.49		(10)(1) = 10
	22.66		(10)(1) = 10
	26.01		(10)(1) = 10
ผลรวม (column total)		93.5	93.5

นั่นคือ $\hat{AUC}$ = column total / $n_D n_H$ = $93.5/(10\times10)$ = 0.935

ดังนั้น ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากับ 0.935

2.4.4 ตัวอย่างในการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้ผลรวมอันดับของวิลคอคสัน

ตัวอย่างการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้ผลรวมอันดับของวิลคอคสัน จะแสดงให้เห็นโดยอาศัยผลจากตารางที่ 2.2 เช่นเดียวกับในการคำนวณโดยใช้หลักเกณฑ์เชิงสี่เหลี่ยมคางหมูและการใช้ฟังก์ชันเปรียบเทียบ ดังนี้

ขั้นที่ 1 จัดเรียงอันดับข้อมูลทั้งหมดร่วมกันจากน้อยไปมาก ดังตาราง

ตารางที่ 2.4 แสดงผลจากการทดสอบความทนกลูโคส 2- h oral glucose tolerance (OGTT) และการคำนวณหาค่า $\hat{AUC}$ โดยใช้ผลรวมอันดับของวิลคอคสัน

ระดับกลูโคสในพลาสม่า (mmol / L)			
กลุ่มผู้ป่วยไม่เป็นโรคเบาหวาน (h_j)	อันดับ (r_j)	กลุ่มผู้ป่วยที่เป็นโรคเบาหวาน (d_i)	อันดับ (r_i)
4.86	1		
5.69	2		
6.01	3		
6.06	4		
6.27	5		
6.37	6		
6.55	7		
7.29	8.5	7.29	8.5
7.82	10		
		9.22	11
		9.79	12
		11.28	13
		11.83	14
12.06	15		
		18.48	16
		18.50	17
		20.49	18
		22.66	19
		26.01	20
ผลรวม (column total)			148.5

ขั้นที่ 2 คำนวณค่าประมาณพื้นที่ใต้โค้ง ROC ดังนี้

$$\text{จาก} \quad \hat{AUC} = \frac{1}{n_D n_H}\left(\sum_{i=1}^{n_D} r_i - \frac{n_D(n_D+1)}{2}\right)$$

$$= \frac{1}{10\times 10}\left(148.5 - \frac{10\times(10+1)}{2}\right)$$

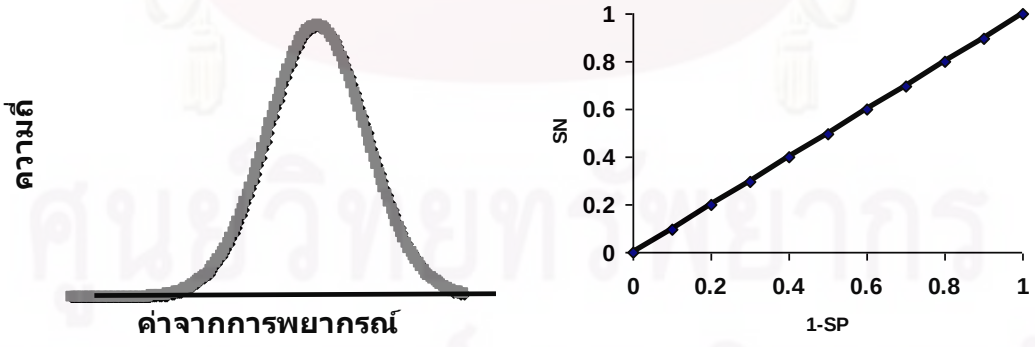
$$= 0.935$$

ดังนั้น ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากับ 0.935

จากข้างต้นเนื่องจากค่าประมาณพื้นที่ใต้โค้ง ROC เป็นค่าที่บ่งบอกถึงความสามารถในแบ่งกลุ่มได้ถูกต้องหรือความเชื่อถือได้ของตัวแบบ ซึ่งมีพิสัยอยู่ระหว่าง 0 ถึง 1 โดยที่เกณฑ์ทั่วไปของการวัดความสามารถหรือความเชื่อถือได้ของตัวแบบ สามารถตีความจากค่าดังกล่าวดังนี้

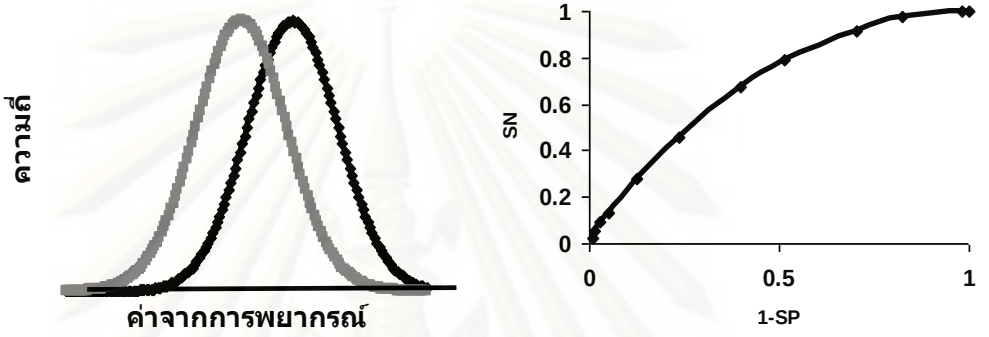
ถ้า $\hat{AUC} = 0.50$ ถือเป็นตัวแบบที่มีความเชื่อถือได้น้อย ไม่สามารถจำแนกเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้ ดังรูป 2.4 ดังต่อไปนี้

รูปที่ 2.4 แสดงผลการจำแนกประชากร 2 กลุ่ม และ ROC curve ที่ $\hat{AUC} = 0.50$



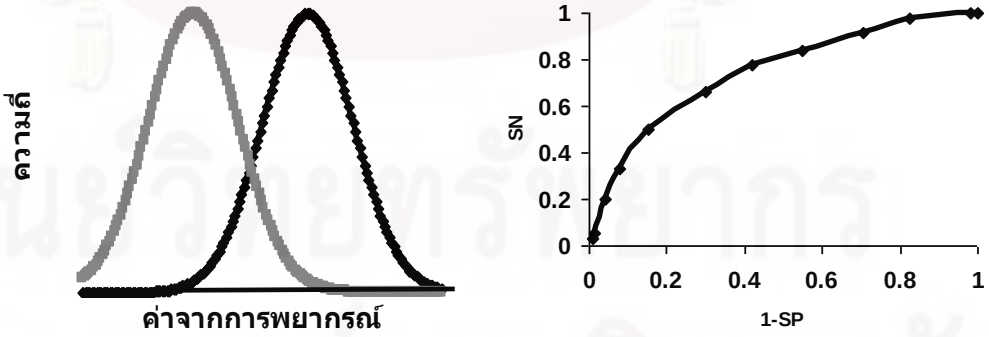
ถ้า $0.70 \le \hat{AUC} < 0.80$ ถือเป็นตัวแบบที่มีความเชื่อถือสามารถยอมรับได้ โดยสามารถจำแนกเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้พอใช้ ดังรูป 2.5 ดังต่อไปนี้

รูปที่ 2.5 แสดงผลการจำแนกประชากร 2 กลุ่ม และ ROC curve ที่ $0.70 \le \hat{AUC} < 0.80$



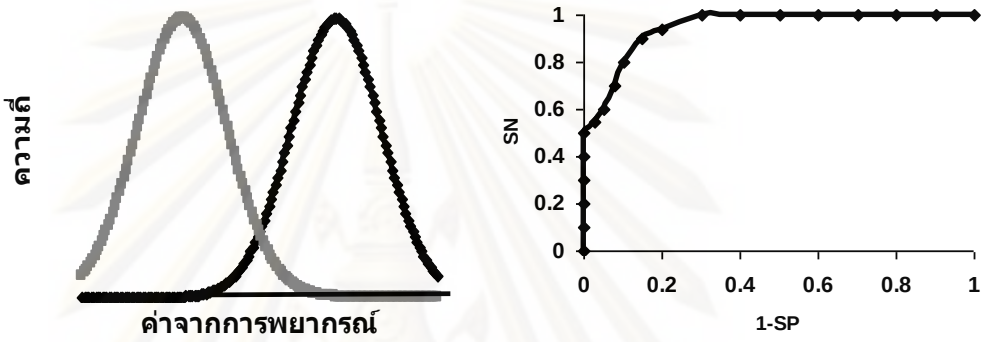
ถ้า $0.80 \le \hat{AUC} < 0.90$ ถือเป็นตัวแบบที่มีความเชื่อถือได้ในระดับดี โ ดยสามารถจำแนกเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้ดี ดังรูป 2.6 ดังต่อไปนี้

รูปที่ 2.6 แสดงผลการจำแนกประชากร 2 กลุ่ม และ ROC curve ที่ $0.80 \le \hat{AUC} < 0.90$



ถ้า $\hat{AUC} \geq 0.90$ ถือเป็นตัวแบบที่มีความเชื่อถือได้ในระดับดีมาก โด ยสามารถจำแนกเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้อย่างชัดเจน ดังรูป 2.7 ดังต่อไปนี้

รูปที่ 2.7 แสดงผลการจำแนกประชากร 2 กลุ่ม และ ROC curve ที่ $\hat{AUC} \geq 0.90$



2.4.5 การทดสอบสมมติฐาน

หากต้องการทดสอบสมมติฐานว่าตัวแบบที่ใช้ในการพย ากรณ์มีความถูกต้องและน่าเชื่อถือได้ภายใต้สมมติฐาน

H_0 : $AUC = 0.50$ แสดงว่า ค่าพยากรณ์ไม่สามารถ จำแนกเหตุการณ์ที่สนใจออกจากเหตุการณ์ที่ไม่สนใจได้

H_1 : $AUC \neq 0.50$ แสดงว่า ค่าพยากรณ์สา มารถจำแนกเหตุการณ์ที่สนใจออกจากเหตุการณ์ที่ไม่สนใจได้

สามารถทดสอบสมมติฐานได้ดังนี้ (จาก Hanley,James A. (1982) “The meaning and use of the area under a receiver operating characteristic curves”)

<u>ขั้นตอนในการทดสอบ</u>

1. สร้าง ROC curve จากตัวแบบที่ใช้ในการพยากรณ์
2. คำนวณตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC ตามสมการ (19)

<u>เกณฑ์ในการตัดสินใจ</u>

จะปฏิเสธสมมติฐานเมื่อค่า p-value ที่ได้จากการคำนวณมีค่าน้อยกว่าระดับนัยสำคัญที่กำหนดโดยค่า p-value สามารถคำนวณได้จาก

$$\Pr\left[|Z| > \left|\frac{A\hat{U}C - 0.5}{SD(A\hat{U}C)|_{AUC=0.5}}\right|\right] = 2\Pr\left[Z > \left|\frac{A\hat{U}C - 0.5}{SD(A\hat{U}C)|_{AUC=0.5}}\right|\right] \quad (21)$$

ในกรณีที่เป็นระเบียบวิธีที่ไม่ใช้พารามิเตอร์

$$SD(A\hat{U}C)|_{AUC=0.5} = \sqrt{\frac{AUC(1-AUC)+(n_D-1)(Q_1-AUC^2)+(n_H-1)(Q_2-AUC^2)}{n_D n_H}}\Bigg|_{AUC=0.5}$$

$$= \sqrt{\frac{0.5(1-0.5)+(n_D-1)(1/3-0.5^2)+(n_H-1)(1/3-0.5^2)}{n_D n_H}}$$

$$= \sqrt{\frac{n_D+n_H+1}{12 n_D n_H}}$$

$$= \frac{\sqrt{\frac{n_D n_H (n_D+n_H+1)}{12}}}{n_D n_H} \quad (22)$$

เมื่อเปรียบเทียบการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC จากทั้ง 3 วิธี พบว่า ผลการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้ฟังก์ชันเปรียบเทียบและผลรวมอันดับของวิลคอกซันจะให้ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากัน (จาก Yan, Lian.(2003) “Optimizing classifier performance via an approximation to the wilcoxon-mann-whiney statistic”) ส่วนผลการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้หลักเกณฑ์เชิงสี่เหลี่ยมคางหมู จะให้ค่าประมาณดังกล่าวแตกต่างจากการคำนวณโดยฟังก์ชันเปรียบเทียบและผลรวมอันดับของวิลคอกซันเพียงเล็กน้อย โดยในส่วนของงานวิจัยนี้ผู้วิจัยเลือกใช้วิธีการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC โดยอาศัยฟังก์ชันเปรียบเทียบและผลรวมอันดับของวิลคอกซัน

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท โดยพิจารณาจากความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ และแบ่งพิจารณากรณีศึกษาออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve แบ่งเป็นกรณีตัวแปรอิสระจำนวน 1 ตัวแปร และตัวแปรอิสระจำนวน 2 ตัวแปร

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร

สำหรับข้อมูลที่ใช้ในงานวิจัยนี้ เป็นข้อมูลจากการจำลองโดยใช้เทคนิคการจำลองโดยวิธีมอนติคาร์โล (Monte Carlo Simulation Method) และทำการวิเคราะห์ข้อมูลด้วยโปรแกรม R โดยมีแผนการทดลองและขั้นตอนในการวิจัย ดังต่อไปนี้

3.1 แผนการดำเนินงานวิจัย

การดำเนินงานวิจัยครั้งนี้แบ่งการดำเนินงานวิจัยออกเป็น 2 ส่วน โดยในส่วนที่ 1 จะทำการศึกษาโดยการพิสูจน์ทางคณิตศาสตร์ พร้อมยกตัวอย่างจากสถานการณ์จำลอง และในส่วนที่ 2 จะดำเนินงานวิจัยโดยกำหนดสถานการณ์จำลองต่าง ๆ แผนการดำเนินงานวิจัยเป็นดังต่อไปนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท แบ่งการพิสูจน์ทางคณิตศาสตร์ออกเป็น 3 กรณี

- กรณีไม่เจาะจงตัวแบบพยากรณ์
- กรณีตัวแปรอิสระจำนวน 1 ตัวแปรบนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง
- กรณีตัวแปรอิสระจำนวน 2 ตัวแปรบนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง

ส่วนที่ 2 หาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร โดยที่

1. คุณสมบัติของตัวแบบโพรบิทอย่างง่ายแบบ 2 กลุ่ม (Simple binary probit model) โดยกำหนด
   - ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพมีการแจกแจงแบบเบอร์นูลี และกำหนดค่าตัวแปรตามมีค่าเพียง 2 ค่า คือ 0 (เหตุการณ์ที่ไม่สนใจ) และ 1 (เหตุการณ์ที่สนใจ)
   - ตัวแปรอิสระ (X) เป็นข้อมูลเชิงปริมาณ โดยมีการแจกแจงแบบปกติ (Normal distribution) ด้วยพารามิเตอร์ μ และ σ นั่นคือ $X \sim N(\mu, \sigma^2)$ ในงานวิจัยครั้งนี้จะศึกษาที่ $\mu = 1$ และ $\sigma^2 = 1$
   - ค่าความคลาดเคลื่อน (e) โดยมีการแจกแจงแบบปกติ (Normal distribution) ด้วยพารามิเตอร์ μ และ σ นั่นคือ $e \sim N(\mu, \sigma^2)$ ในงานวิจัยครั้งนี้จะศึกษาที่ $\mu = 0$ และ $\sigma^2 = 1$
2. สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในการจำลองข้อมูล ที่ $\beta_1 = -1.00$ $\beta_1 = -0.80$ $\beta_1 = -0.60$ $\beta_1 = -0.40$ $\beta_1 = -0.20$ $\beta_1 = 0.20$ $\beta_1 = 0.40$ $\beta_1 = 0.60$ $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ โดยที่ β_0 คงที่ ($\beta_0 = 0.00$)
3. จำนวนขนาดตัวอย่าง (n) เป็น 50, 100, 200, 300, 400, 500 และ 1,000
4. กำหนดการกระทำซ้ำในแต่ละสถานการณ์เป็น 2,000 รอบ
5. กำหนดระดับนัยสำคัญ 0.01 และ 0.05

3.2 ขั้นตอนในการดำเนินงานวิจัย

ขั้นตอนในการดำเนินงานวิจัยในการศึกษาครั้งนี้ มีดังนี้

1. ศึกษาฟังก์ชันความน่าจะเป็นของตัวแปรตามในตัวแบบโพรบิท (Probit model)
2. ศึกษาการใช้ Receiver Operating Characteristic Curve หรือ ROC curve
3. ศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve
4. ทำการจำลองข้อมูลที่ใช้ในการวิจัยและหาอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท ในกรณีศึกษาข้างต้น
5. ทำการวิเคราะห์ผลการทดลอง และสรุปผลการวิจัย

ในขั้นตอนที่ 1 และ 2 เป็นขั้นตอนในการดำเนินงานวิจัยเป็นการศึกษาเชิงวิเคราะห์ ซึ่งรายละเอียดในการดำเนินงานวิจัยในขั้นตอนดังกล่าว ผู้วิจัยขอเสนอไปในบทที่ 2 ส่วนในขั้นตอนอื่น ๆ มีรายละเอียดการดำเนินการวิจัย ดังต่อไปนี้

การศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท

ทำการศึกษาโดยพิสูจน์ทางคณิตศาสตร์ พร้อมทั้งยกตัวอย่างจากทั้ง 3 กรณีที่กล่าวแล้วข้างต้น

การหาอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระ 1 ตัวแปร

แบ่งการทดลองเป็น 3 ขั้นตอนดังนี้

<u>คำนวณค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1</u>

1. กำหนดขนาดตัวอย่าง (n) เท่ากับ 50,100, 200, 300, 400, 500 และ 1,000
2. สร้างตัวแปรอิสระจำนวน 1 ตัวแปรให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 1$ และ $\sigma^2 = 1$ นั่นคือ $X_1 \sim N(1,1)$
3. สร้างค่าความคลาดเคลื่อนให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 0$ และ $\sigma^2 = 1$ นั่นคือ $e \sim N(0,1)$ โดยที่ $i = 1,2,...,n$
4. กำหนดสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_0 คงที่ ที่ $\beta_0 = 0.00$ และสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ที่ $\beta_1 = 0.00$
5. สร้างตัวแปรตาม (Y) ให้มีการแจกแจงแบบเบอร์นูลี โดยคำนวณจากข้อ 2, 3 และ 4 โดยตัวแปรตาม (Y) มีค่าได้เพียง 2 ค่า คือ 0 และ 1 ดังนี้

   จากสมการ (1) แทนค่าตามข้อกำหนดในข้อ 2, 3 และ 4 จะได้

$$Y_i^* = \beta_1 * x_{1i} + e_i$$

ทำการปรับ Y_i^* ให้เป็นตัวแปรหุ่น (Dummy Variable) แทนด้วยค่า Y_i ดังนี้

$Y_i = 1$ ถ้า $Y_i^* > 0$

$Y_i = 0$ ถ้า $Y_i^* \le 0$

6. นำข้อมูลตัวแปรตามจากข้อ 5 และตัวแปรอิสระมาทำการประมาณค่าพารามิเตอร์ด้วยการวิเคราะห์ความถดถอยโพรบิทเพื่อนำค่าพารามิเตอร์ดังกล่าวมาสร้างตัวแบบที่ใช้ในการพยากรณ์
7. สร้างค่าพยากรณ์ (y_i^*) จากตัวแบบที่ได้ในข้อ 6
8. พล็อต ROC curve โดยใช้ข้อมูลตัวแปรตามจากข้อ 5 และค่าพยากรณ์จากข้อ 7 หลังจากนั้นคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC
9. นำค่าประมาณพื้นที่ใต้โค้ง ROC ดังกล่าวไปทำการคำนวณหาค่า p-value ซึ่งได้อธิบายวิธีการคำนวณไว้ในบทที่ 2 และทำการเปรียบเทียบค่า p-value ที่ระดับนัยสำคัญที่กำหนด เพื่อที่จะได้ตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมติฐานว่าง นับจำนวนครั้งที่ปฏิเสธสมมติฐานว่าง $H_0 : AUC = 0.50$
10. ทำซ้ำขั้นตอนที่ 2 – 9 จำนวน 2,000 ครั้ง
11. ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เท่ากับจำนวนครั้งที่ปฏิเสธสมมติฐานว่างหารด้วย 2,000
12. ทำซ้ำขั้นตอนที่ 2 – 9 ทุกขนาดตัวอย่าง และระดับนัยสำคัญที่กำหนด

<u>ทดสอบความสามารถในการควบคุมค่าคลาดเคลื่อนประเภทที่ 1</u>

การทดสอบว่า ตัวสถิติสามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้หรือไม่ จะใช้การทดสอบทวินามภายใต้ระดับนัยสำคัญ 0.05 โดยตัวสถิติที่สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้จะมีค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วงดังต่อไปนี้

ที่ระดับนัยสำคัญของการทดสอบ 0.01 ช่วงของการยอมรับ คือ $(0.006, 0.0137)$

ที่ระดับนัยสำคัญของการทดสอบ 0.05 ช่วงของการยอมรับ คือ $(0.042, 0.0580)$

ที่ระดับนัยสำคัญของการทดสอบ 0.10 ช่วงของการยอมรับ คือ $(0.089, 0.1110)$

<u>หาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท</u>

1. กำหนดขนาดตัวอย่าง (n) เท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000
2. สร้างตัวแปรอิสระจำนวน 1 ตัวแปรให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 1$ และ $\sigma^2 = 1$ นั่นคือ $X_1 \sim N(1,1)$
3. สร้างค่าความคลาดเคลื่อนให้มีการแจกแจงแบบปกติด้วยพารามิเตอร์ $\mu = 0$ และ $\sigma^2 = 1$ นั่นคือ $e \sim N(0,1)$ โดยที่ $i = 1,2,...,n$

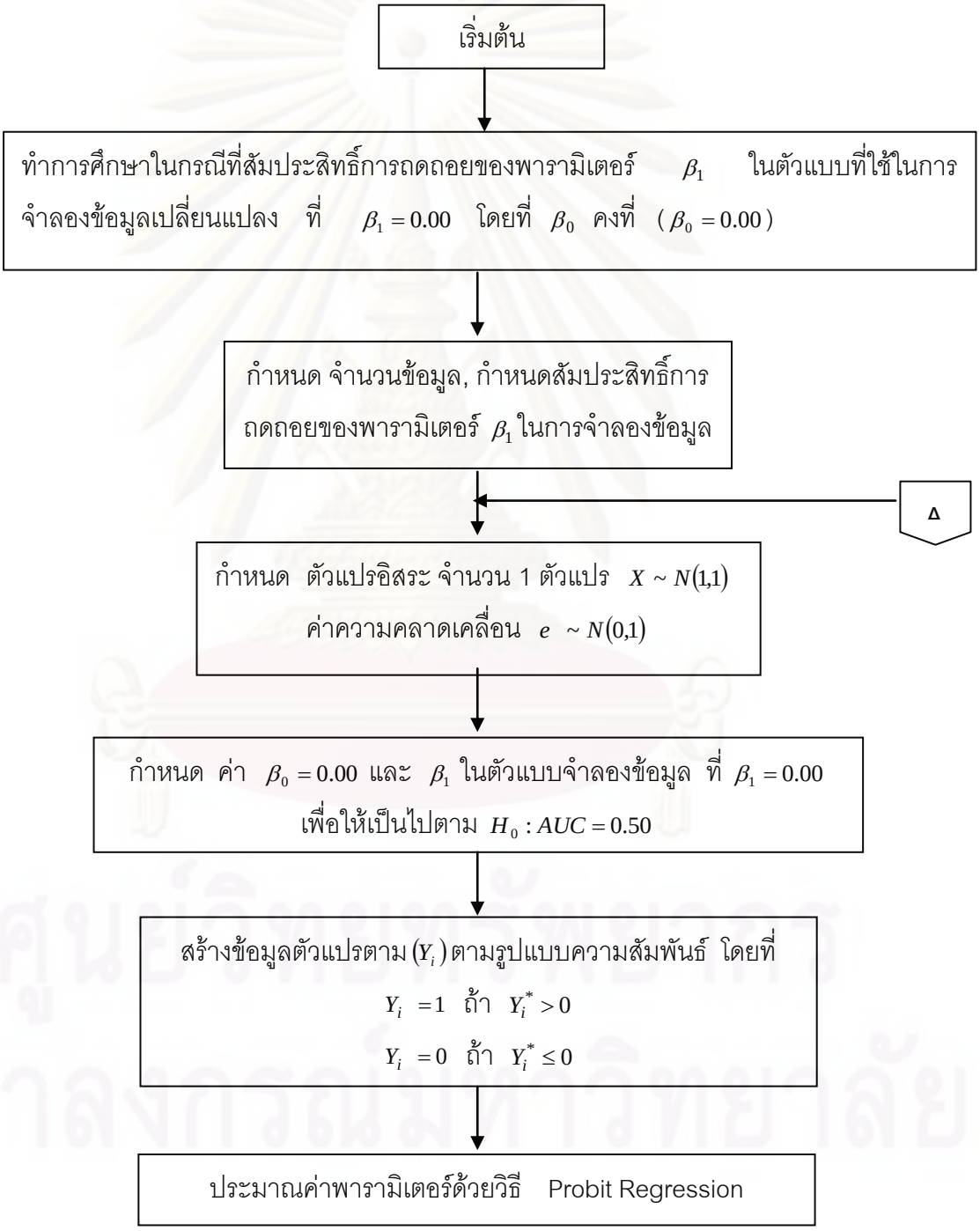
4. กำหนดสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_0 คงที่ ที่ $\beta_0 = 0.00$ และสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เปลี่ยนแปลง โดยที่ $\beta_1 = -1.00$ $\beta_1 = -0.80$ $\beta_1 = -0.60$ $\beta_1 = -0.40$ $\beta_1 = -0.20$ $\beta_1 = 0.20$ $\beta_1 = 0.40$ $\beta_1 = 0.60$ $\beta_1 = 0.80$ และ $\beta_1 = 1.00$
5. สร้างตัวแปรตาม (Y) ให้มีการแจกแจงแบบเบอร์นูลี โดยคำนวณจากข้อ 2, 3 และ 4 โดยตัวแปรตาม (Y) มีค่าได้เพียง 2 ค่า คือ 0 และ 1 ดังนี้
   จากสมการ (1) แทนค่าตามข้อกำหนดในข้อ 2, 3 และ 4 จะได้
   $$Y_i^* = \beta_1 * x_{1i} + e_i$$ โดยที่ β_1 เปลี่ยนแปลงตามกรณีในข้อ 4
   ทำการปรับ Y_i^* ให้เป็นตัวแปรหุ่น (Dummy Variable) แทนด้วยค่า Y_i ดังนี้
   $$Y_i = 1 \text{ ถ้า } Y_i^* > 0$$
   $$Y_i = 0 \text{ ถ้า } Y_i^* \le 0$$
6. นำข้อมูลตัวแปรตามจากข้อ 5 และตัวแปรอิสระมาทำการประมาณค่าพารามิเตอร์ด้วยการวิเคราะห์ความถดถอยโพรบิทเพื่อนำค่าพารามิเตอร์ดังกล่าวมา สร้างตัวแบบที่ใช้ในการพยากรณ์
7. สร้างค่าพยากรณ์ $\left(y_i^*\right)$ จากตัวแบบที่ได้ในข้อ 6
8. พล็อต ROC curve โดยใช้ข้อมูลตัวแปรตามจากข้อ 5 และค่าพยากรณ์จากข้อ 7 หลังจากนั้นคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC
9. นำค่าประมาณพื้นที่ใต้โค้ง ROC ดังกล่าวไปทำการคำนวณหาค่า p-value ซึ่งได้อธิบายวิธีการคำนวณไว้ในบทที่ 2 และทำการเปรียบเทียบค่า p-value ที่ระดับนัยสำคัญที่กำหนด เพื่อที่จะได้ตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมติฐานว่าง นับจำนวนครั้งที่ปฏิเสธสมมติฐานว่าง $H_0 : AUC = 0.50$
10. ทำซ้ำขั้นตอนที่ 2 – 9 จำนวน 2,000 ครั้ง
11. ค่าอำนาจการทดสอบเท่ากับจำนวนครั้งที่ปฏิเสธสมมติฐานว่างหารด้วย 2,000
12. ทำซ้ำขั้นตอนที่ 2 – 9 ทุกขนาดตัวอย่าง และระดับนัยสำคัญที่กำหนด

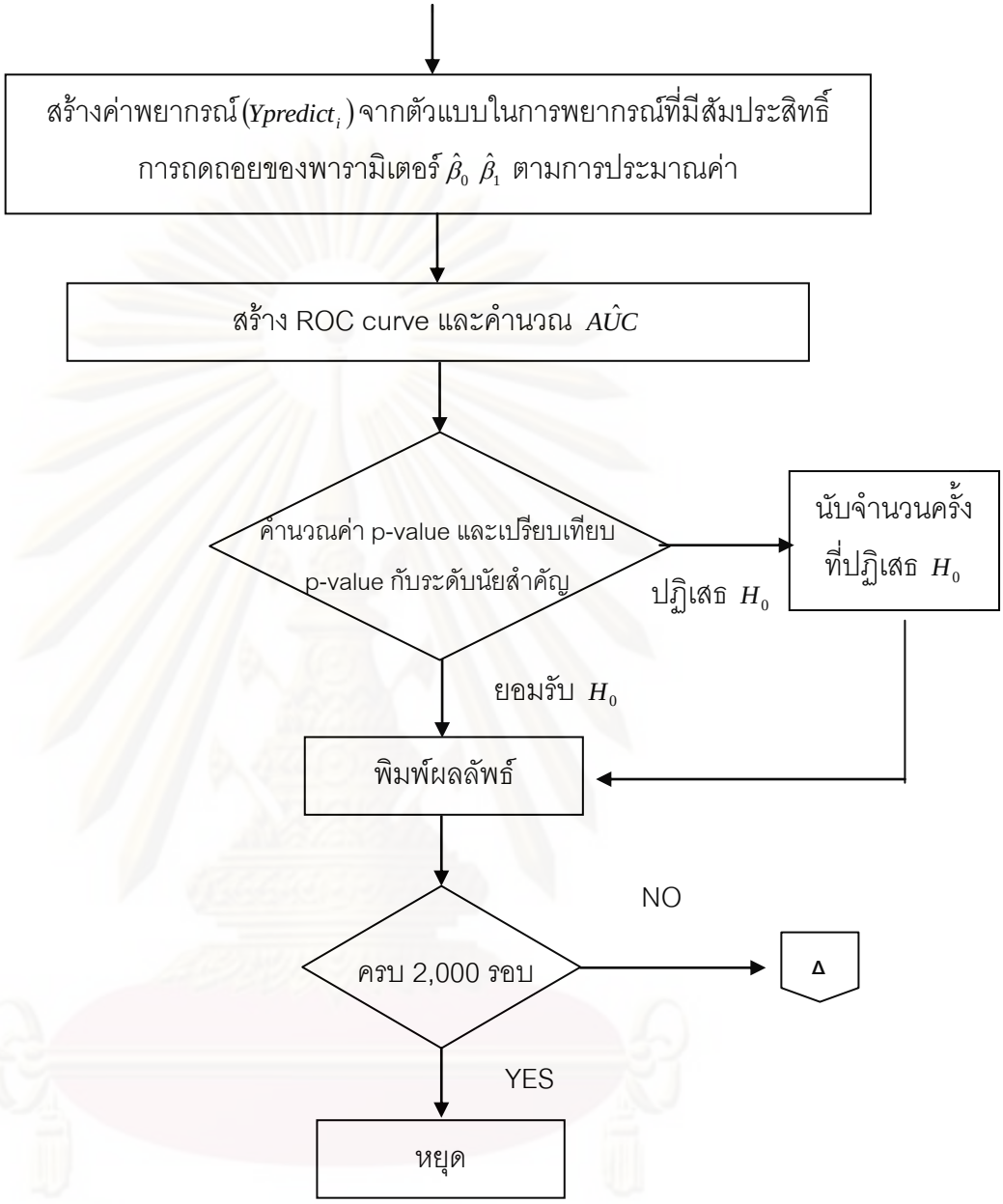
สรุปผลการทดลอง

เมื่อทำการศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และหาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิทในแต่ละกรณีศึกษาแล้ว นำผลการศึกษาโดยการพิสูจน์ทางคณิตศาสตร์และผลการทดลองที่ได้มาสรุปเป็นทฤษฎีและผลในรูปแบบตารางพร้อมทั้งรูปภาพเพื่อแสดงผลของการศึกษาในแต่ละกรณี

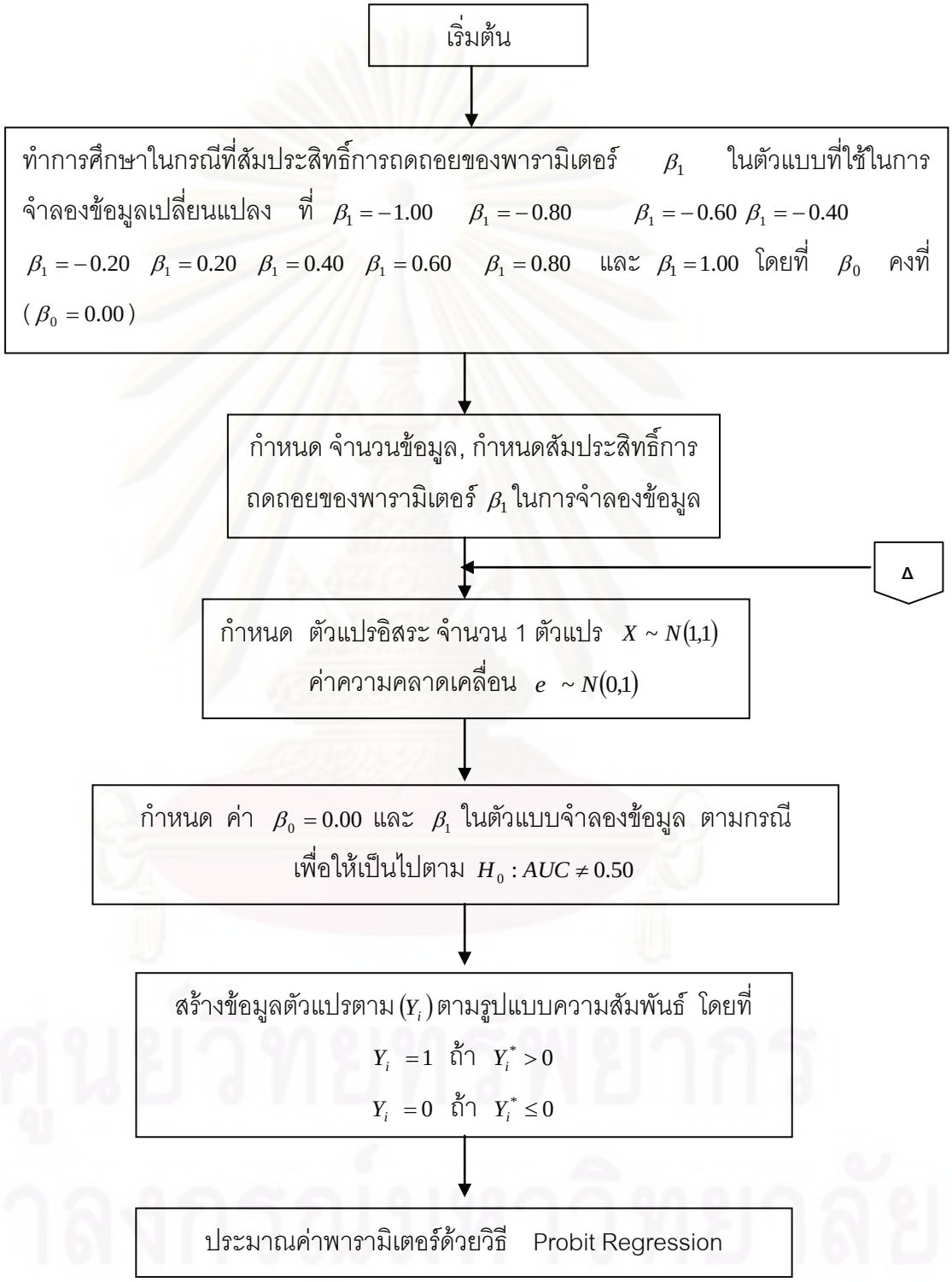
สำหรับขั้นตอนการจำลองในงานวิจัยข้างต้นนั้นได้แสดงเป็นแผนภาพแสดงขั้นตอนในงานวิจัยดังต่อไปนี้

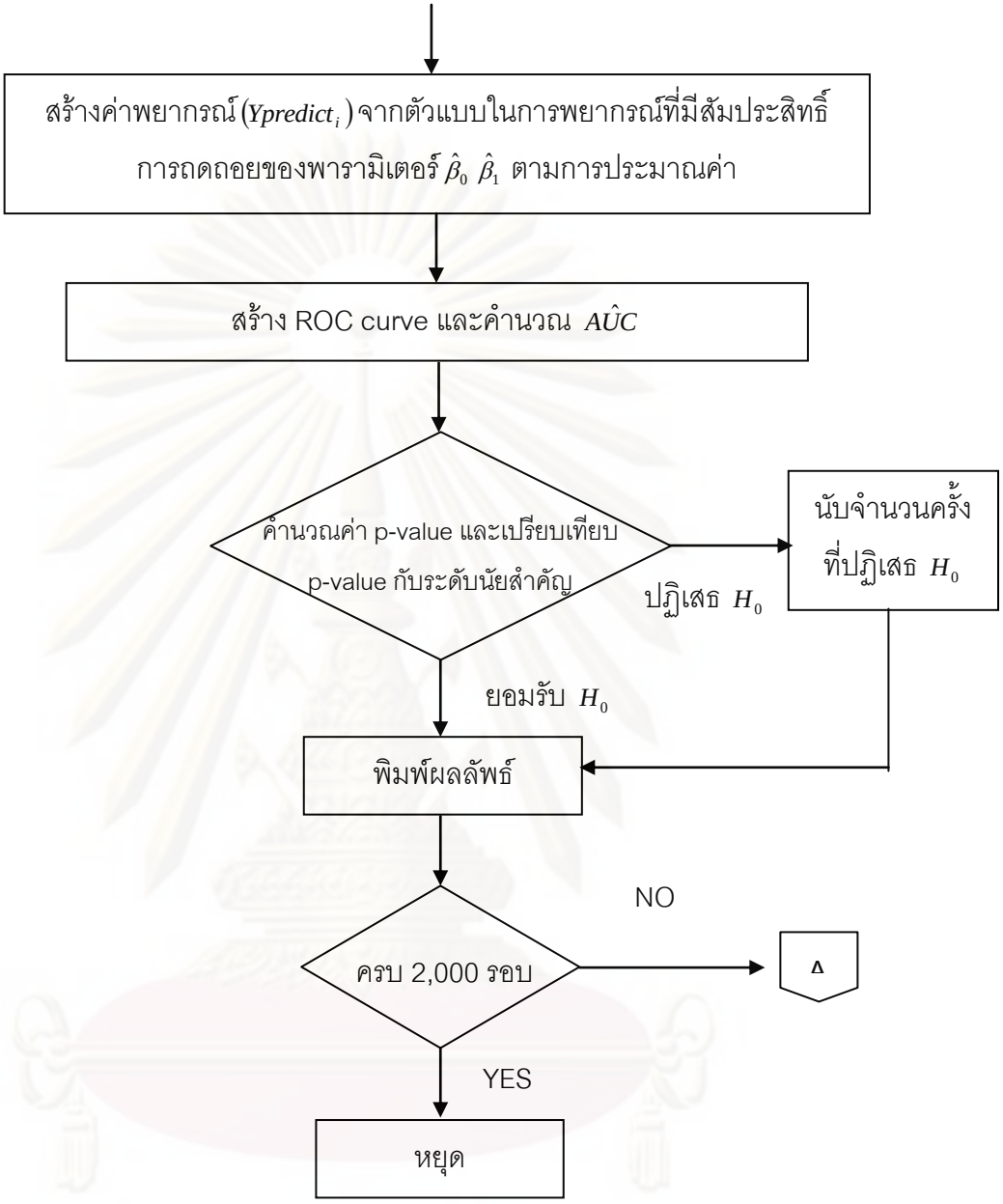
แผนภาพที่ 3.1แสดงขั้นตอนในการดำเนินงานวิจัยในส่วนการคำนวณค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1





แผนภาพที่ 3.2แสดงขั้นตอนในการดำเนินงานวิจัยในส่วนการหาอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC





บทที่ 4

ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท โดยพิจารณาจากความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ และทำการจำลองข้อมูลเพื่อศึกษาถึงกรณีดังกล่าว โดยแบ่งพิจารณากรณีศึกษาออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve แบ่งเป็นกรณีตัวแปรอิสระจำนวน 1 ตัวแปร และตัวแปรอิสระจำนวน 2 ตัวแปร

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร

การนำเสนอผลการวิจัยนี้ ได้นำเส นอเป็นทฤษฎีซึ่งได้จากการพิสูจน์ทางคณิตศาสตร์และรูปแบบตารางพร้อมทั้งรูปภาพ โดยมีการใช้สัญลักษณ์แทนความหมายต่าง ๆ ดังนี้

n แทน จำนวนขนาดตัวอย่าง

n_D แทน จำนวนขนาดตัวอย่างจากกลุ่มเหตุการณ์ที่สนใจ

n_H แทน จำนวนขนาดตัวอย่างจากกลุ่มเหตุการณ์ที่ไม่สนใจ

r_i แทน อันดับของ d_i ในการเรียงลำดับรวมของกลุ่มเหตุการณ์ที่สนใจกับกลุ่มเหตุการณ์ที่ไม่สนใจ

r_j แทน อันดับของ h_j ในการเรียงลำดับรวมของกลุ่มเหตุการณ์ที่สนใจกับกลุ่มเหตุการณ์ที่ไม่สนใจ

y_i^* แทน ค่าพยากรณ์

I_i แทน อันดับของค่าพยากรณ์จากการเรียงอันดับรวมจากน้อยไปหามาก

d_i แทน ค่าพยากรณ์สำหรับกลุ่มเหตุการณ์ที่สนใจ

h_j แทน ค่าพยากรณ์สำหรับกลุ่มเหตุการณ์ที่ไม่สนใจ

การนำเสนอผลการวิจัยในการศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท ได้จำแนกออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท แยกการพิสูจน์ทางคณิตศาสตร์เป็น 3 กรณี คือ

- กรณีไม่เจาะจงตัวแบบพยากรณ์
- กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง
- กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง

ส่วนที่ 2 ผลการวิจัยเปรียบเทียบอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระ 1 ตัวแปร ทำการศึกษาที่ขนาดตัวอย่างในการทดลองเท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000 และที่ระดับนัยสำคัญ 0.01 และ 0.05 แบ่งเป็น 3 ส่วน คือ

- ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC
- ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1
- อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท

การศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท โดยจะแสดงให้เห็นคุณสมบัติดังกล่าวโดยการพิสูจน์ทางคณิตศาสตร์จากกรณีตัวอย่าง 3 กรณี ดังนี้

- กรณีไม่เจาะจงตัวแบบพยากรณ์
- กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง
- กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง

กรณีไม่เจาะจงตัวแบบพยากรณ์

จากบทที่ 2 ทฤษฎีและสถิติที่เกี่ยวข้อง ได้นำเสนอถึงการคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC โดยผู้วิจัยเลือกใช้วิธีการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้ผลรวมอันดับของวิลคอกซันในการพิสูจน์ทางคณิตศาสตร์ได้ผลการศึกษา ดังนี้

ทฤษฎีบทที่ 1 กรณีข้อมูล 1 ชุด ถ้าตัวแบบพยากรณ์ A และตัวแบบพยากรณ์ B ให้ค่าอันดับของค่าพยากรณ์ในอันดับเดียวกัน แล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC ของตัวแบบ A และ ตัวแบบ B มีค่าเท่ากัน นั่นคือ

$$A\hat{U}C_A - A\hat{U}C_B = 0$$

พิสูจน์ จากการคำนวณค่าประมาณพื้นที่ใต้โค้ง ROC โดยใช้วิธีผลรวมอันดับของวิลคอคสัน คือ

$$A\hat{U}C = \frac{1}{n_D n_H}\left(\sum_{i=1}^{nD} r_i - \frac{n_D(n_D+1)}{2}\right)$$

ดังนั้น

$$A\hat{U}C_A - A\hat{U}C_B = \frac{1}{n_{AD}n_{AH}}\left(\sum_{i=1}^{nAD} r_{Ai} - \frac{n_{AD}(n_{AD}+1)}{2}\right) - \frac{1}{n_{BD}n_{BH}}\left(\sum_{i=1}^{nBD} r_{BI} - \frac{n_{BD}(n_{BD}+1)}{2}\right)$$

$$= \frac{1}{n_{AD}n_{AH}} \cdot \sum_{i=1}^{nAD} r_{Ai} - \frac{1}{n_{AD}n_{AH}} \cdot \frac{n_{AD}(n_{AD}+1)}{2} - \frac{1}{n_{BD}n_{BH}} \cdot \sum_{i=1}^{nBD} r_{Bi}$$

$$+ \frac{1}{n_{BD}n_{BH}} \cdot \frac{n_{BD}(n_{BD}+1)}{2}$$

$=$ 0 (เนื่องข้อมูลเป็นข้อมูลชุดเดียวกัน ดังนั้น $n_{AD} = n_{BD}$, $n_{AH} = n_{BH}$ และ ค่าอันดับของค่าพยากรณ์เป็นอันดับเดียวกัน นั่นคือ $r_{Ai} = r_{Bi}$)

กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท

จากทฤษฎีบทที่ 1 เป็นการพิสูจน์ทางคณิตศาสตร์ให้เห็นถึงคุณสมบัติการไม่ผันแปรของ ROC curve ในกรณีไม่เจาะจงตัวแบบพยากรณ์ ผู้วิจัยจะอาศัยทฤษฎีดังกล่าวในการอ้างอิงถึงเพื่อบทตามที่จะเกิดขึ้นในกรณีตัวแปรอิสระจำนวน 1 ตัวแปรบนตัวแบบพยากรณ์โพรบิท พร้อมยกตัวอย่างสถานการณ์จำลองตามบทตาม ดังนี้

บทตามที่ 1 กำหนดข้อมูล 1 ชุด และ ตัวแบบพยากรณ์โพรบิทสองตัวแบบ ได้แก่

ตัวแบบ A: $P(Y=1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_1)$ และ

ตัวแบบ B: $P(Y=1) = \Phi(\hat{\beta}'_0 + \hat{\beta}'_1 X_1)$

เมื่อ $\hat{\beta}_0 = \hat{\beta}'_0$ แล้ว ถ้าเครื่องหมายของ $\hat{\beta}_1$ กับ เครื่องหมายของ $\hat{\beta}'_1$ เป็นเครื่องหมายชนิดเดียวกัน ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองเท่ากัน

พิสูจน์ เมื่อกำหนดข้อมูล 1 ชุด

จากตัวแบบพยากรณ์โพรบิท A $P(Y=1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_1)$

จะทำให้ได้ $y_i^* = P(Y=1) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_1)$ เมื่อ $i = 1,2,...,n$ (23)

นำ y_i^* มาเรียงอันดับโดยเรียงจากน้อยไปหามาก จะได้

$I_1, I_2, I_3, ..., I_n$ เป็นอันดับที่เกิดจากการเรียง y_i^*

นั่นแสดงว่า

$$y_{I_1}^* < y_{I_2}^* < y_{I_3}^* , ..., < y_{I_n}^*$$

สมมติว่า

$$y^{*}_{I_l} \quad < \quad y^{*}_{I_{l+1}} \qquad \text{เมื่อ } l = 1,2,\ldots,n \text{ แทนอันดับที่เกิดขึ้น}$$

ดังนั้น $$\Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_l}}\right) \quad < \quad \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_{l+1}}}\right)$$

เนื่องจาก Φ เป็นฟังก์ชันเพิ่มสม่ำเสมอ (Strictly increasing function) ดังนั้น

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_l}} \quad &< \quad \hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_{l+1}}} \\ \hat{\beta}_1 x_{1_{I_l}} \quad &< \quad \hat{\beta}_1 x_{1_{I_{l+1}}} \end{aligned} \qquad (24)$$

นำ $\frac{\hat{\beta}'_1}{\hat{\beta}_1}$ คูณ (24)

$$\begin{aligned} \frac{\hat{\beta}'_1}{\hat{\beta}_1} \times \hat{\beta}_1 x_{1_{I_l}} \quad &< \quad \frac{\hat{\beta}'_1}{\hat{\beta}_1} \times \hat{\beta}_1 x_{1_{I_{l+1}}} \\ \Phi\left(\hat{\beta}'_0 + \hat{\beta}'_1 x_{1_{I_l}}\right) \quad &< \quad \Phi\left(\hat{\beta}'_0 + \hat{\beta}'_1 x_{1_{I_{l+1}}}\right) \longrightarrow \text{ตัวแบบ } B \\ y'^{*}_{I_l} \quad &< \quad y'^{*}_{I_{l+1}} \end{aligned}$$

เพราะฉะนั้น

$$I_1 = I'_1 \quad , \quad I_2 = I'_2 \quad , \quad I_3 = I'_3 \quad ,\ldots, \quad I_n = I'_n$$

ดังนั้น จากทฤษฎีบทที่ 1 เมื่อค่าอันดับของการพยากรณ์เป็นอันดับเดียวกัน ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองจะมีค่าเท่ากัน

ต่อไปจะเป็นตัวอย่างสถานการณ์จำลองตามบทตามที่ 1 ดังนี้

ตัวอย่างที่ 4.1 สถานการณ์จำลองกรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์โพรบิทเปลี่ยนแปลง โดยที่ $\hat{\beta}_0$ คงที่

ตารางที่ 4.1 แสดงค่าพยากรณ์, ค่าอันดับของค่าพยากรณ์ และพื้นที่ใต้โค้ง ROC จากการวิเคราะห์ข้อมูลใน 1 ครั้ง ที่ $\hat{\beta}_1 = 0.20$ และ $\hat{\beta}_1 = 0.40$ โดยที่ขนาดตัวอย่างในการทดลองเท่ากับ 50

ลำดับ	สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$							
	$\hat{\beta}_1 = 0.20$				$\hat{\beta}_1 = 0.40$			
	h_j	r_j	d_i	r_i	h_j	r_j	d_i	r_i
1	0.2250	1	0.2358	3	0.1560	1	0.1738	3
2	0.2263	2	0.2692	4	0.1581	2	0.2326	4
3	0.2824	5	0.3040	7	0.2573	5	0.2995	7
4	0.2943	6	0.3161	9	0.2803	6	0.3237	9
5	0.3150	8	0.3387	11	0.3215	8	0.3700	11
6	0.3261	10	0.3468	15	0.3441	10	0.3866	15
7	0.3419	12	0.3516	16	0.3765	12	0.3967	16
8	0.3432	13	0.3547	18	0.3793	13	0.4032	18
9	0.3456	14	0.3652	20	0.3843	14	0.4250	20
10	0.3532	17	0.3660	21	0.3999	17	0.4266	21
11	0.3581	19	0.3757	23	0.4103	19	0.4469	23
12	0.3731	22	0.3858	26	0.4414	22	0.4680	26
13	0.3846	24	0.3866	28	0.4655	24	0.4696	28
14	0.3848	25	0.4012	32	0.4659	25	0.4998	32
15	0.3860	27	0.4014	33	0.4682	27	0.5003	33
16	0.3958	29	0.4194	35	0.4886	29	0.5370	35
17	0.3982	30	0.4236	37	0.4935	30	0.5456	37
18	0.3996	31	0.4264	38	0.4965	31	0.5513	38
19	0.4091	34	0.4412	41	0.5161	34	0.5808	41
20	0.4233	36	0.4421	42	0.5451	36	0.5826	42
21	0.4346	39	0.4444	43	0.5678	39	0.5872	43
22	0.4398	40	0.4565	44	0.5781	40	0.6109	44
23	0.4662	45	0.4683	47	0.6294	45	0.6333	47
24	0.4678	46	0.4683	48	0.6324	46	0.6334	48
25	0.4736	49	0.4795	50	0.6434	49	0.6544	50
ผลรวมอันดับ				691				691
พื้นที่ใต้โค้ง				0. 5856				0. 5856

ผลการวิจัยเปรียบเทียบค่าพยากรณ์และค่าอันดับของค่าพยากรณ์ที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์มีการเปลี่ยนแปลง โดยที่ $\hat{\beta}_1 = 0.20$ และ $\hat{\beta}_1 = 0.40$ เมื่อ $\hat{\beta}_0$ คงที่ และขนาดตัวอย่างในการทดลองเท่ากับ 50 จากการวิเคราะห์ข้อมูลใน 1 ครั้ง พบว่าค่าพยากรณ์ทั้งในกลุ่มเหตุการณ์ที่สนใจและไม่สนใจมีความแตกต่างกันในทุกตัวอย่าง แต่ค่าอันดับของค่าพยากรณ์ r_j และ r_i ของทั้งสองตัวแบบพยากรณ์กลับมีอันดับเป็นอันดับเดียวกันทุกตัวอย่าง สอดคล้องกับทฤษฎีบทที่ 1 และบทตามที่ 1 จึงส่งผลให้ค่าประมาณพื้นที่ใต้โค้ง ROC ของตัวแบบพยากรณ์ทั้งสองมีค่าเท่ากัน โดยมีค่าเท่ากับ 0.5856 ซึ่งเป็นจริงตามการพิสูจน์ทฤษฎีบทที่ 1 และบทตามที่ 1

นอกจากนี้ในกรณีที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์เป็นกรณีอื่น ๆ เช่น $\hat{\beta}_1 = 0.60$, $\hat{\beta}_1 = 0.80$ และ $\hat{\beta}_1 = 1.00$ เมื่อขนาดตัวอย่างในการทดลองเท่ากับ 50 ผลการวิจัยเปรียบเทียบค่าพยากรณ์และค่าอันดับของค่าพยากรณ์ที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์มีการเปลี่ยนแปลงและค่าประมาณพื้นที่ใต้โค้ง ROC ก็มีลักษณะสอดคล้องและเป็นจริงตามทฤษฎีที่ 1 และบทตามที่ 1 เช่นเดียวกัน โดยผลการวิจัยแสดงให้เห็นในภาคผนวก

สรุปผลการศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท กรณีตัวแปรอิสระจำนวน 1 ตัวแปร

กรณีข้อมูล 1 ชุด และ ตัวแบบพยากรณ์โพรบิ ทสองตัวแบบ เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์โพรบิทมีค่าไม่เท่ากันแต่มีทิศทางเดียวกัน โดยที่ $\hat{\beta}_0$ คงที่ ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองยังคงมีค่าเท่ากัน เนื่องจากตัวแบบพยากรณ์โพรบิทแต่ละตัวแบบยังคงให้ค่าอันดับของค่าพยากรณ์เป็นอันดับเดียวกัน โดยหากมองในมุมของการที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์มาจากวิธีการประมาณ 2 วิธี นั่นคือ $\hat{\beta}_1$ และ $\hat{\beta}_1'$ โดยที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ดังกล่าวมีทิศทางเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองจึงมีค่าเท่ากัน

กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท

เช่นเดียวกับกรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท ผู้วิจัยจะอาศัยทฤษฎีบทที่ 1 ในการอ้างอิงถึงเพื่อบทตามที่จะเกิดขึ้นในกรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมยกตัวอย่างสถานการณ์จำลองตามบทตาม ดังนี้

บทตามที่ 2 กำหนดข้อมูล 1 ชุด และ ตัวแบบพยากรณ์โพรบิทสองตัวแบบ ได้แก่

ตัวแบบ A: $P(Y=1) = \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2\right)$ และ

ตัวแบบ B: $P(Y=1) = \Phi\left(\hat{\beta}'_0 + \hat{\beta}'_1 X_1 + \hat{\beta}'_2 X_2\right)$

สมมติว่า $\hat{\beta}_0 = \hat{\beta}'_0$ และ $\hat{\beta}_1 = \hat{\beta}'_1$ และ ค่าพยากรณ์ที่ได้จากตัวแบบ A มีค่าไม่ซ้ำกัน จะมีช่วงเปิด (a,b)ซึ่งถ้า $\hat{\beta}'_2 \in (a,b)$ แล้วค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองเท่ากัน

พิสูจน์ กำหนดข้อมูล 1 ชุด เมื่อค่าพยากรณ์จากตัวแบบพยากรณ์โพรบิทมีค่าไม่เท่ากัน

จากตัวแบบพยากรณ์โพรบิท A $\quad P(Y=1) = \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2\right)$

จะทำให้ได้ $\quad y_i^* = P(Y=1) = \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2\right)$เมื่อ $i = 1,2,...,n$ (25)

นำ y_i^* มาเรียงอันดับโดยเรียงจากน้อยไปหามาก จะได้

$I_1, I_2, I_3, ..., I_n$ เป็นอันดับที่เกิดจากการเรียง y_i^*

เนื่องจากค่าพยากรณ์ไม่เท่ากัน

$$y_{I1}^* < y_{I2}^* < y_{I3}^* , ..., < y_{In}^*$$

จาก $\quad y_{Il}^* < y_{Il+1}^*$

กำหนด $l = 1$

$$\left.\begin{array}{l} y_{I1}^* < y_{I2}^* \\ \text{ดังนั้น} \quad \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_1}} + \hat{\beta}_2 x_{2_{I_1}}\right) < \Phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_2}} + \hat{\beta}_2 x_{2_{I_2}}\right) \\ \text{เนื่องจาก } \Phi \text{ เป็นฟังก์ชันเพิ่มสม่ำเสมอ (Strictly increasing function)} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_1}} + \hat{\beta}_2 x_{2_{I_1}} < \hat{\beta}_0 + \hat{\beta}_1 x_{1_{I_2}} + \hat{\beta}_2 x_{2_{I_2}} \end{array}\right\} (a_1, b_1)$$

กำหนด $l=2$ (เมื่อ l แทนอันดับที่เกิดขึ้น)

$$y^*_{I_2} < y^*_{I_3}$$

ดังนั้น $\Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_2}}+\hat{\beta}_2 x_{2_{I_2}}\right) < \Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_3}}+\hat{\beta}_2 x_{2_{I_3}}\right)$ $\Big\}\ (a_2, b_2)$

เนื่องจาก Φ เป็นฟังก์ชันเพิ่มสม่ำเสมอ (Strictly increasing function) ดังนั้น

$$\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_1}}+\hat{\beta}_2 x_{2_{I_2}} < \hat{\beta}_0+\hat{\beta}_1 x_{1_{I_2}}+\hat{\beta}_2 x_{2_{I_3}}$$

กำหนด $l=3$ (เมื่อ l แทนอันดับที่เกิดขึ้น)

$$y^*_{I_3} < y^*_{I_4}$$

ดังนั้น $\Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_3}}+\hat{\beta}_2 x_{2_{I_3}}\right) < \Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_4}}+\hat{\beta}_2 x_{2_{I_4}}\right)$ $\Big\}\ (a_3, b_3)$

เนื่องจาก Φ เป็นฟังก์ชันเพิ่มสม่ำเสมอ (Strictly increasing function) ดังนั้น

$$\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_3}}+\hat{\beta}_2 x_{2_{I_3}} < \hat{\beta}_0+\hat{\beta}_1 x_{1_{I_4}}+\hat{\beta}_2 x_{2_{I_4}}$$

$$\vdots$$

กำหนด $l=n-1$ (เมื่อ l แทนอันดับที่เกิดขึ้น)

$$y^*_{I_{n-1}} < y^*_{I_n}$$

ดังนั้น $\Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_{n-1}}}+\hat{\beta}_2 x_{2_{I_{n-1}}}\right) < \Phi\left(\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_n}}+\hat{\beta}_2 x_{2_{I_n}}\right)$ $\Big\}\ (a_{n-1}, b_{n-1})$

เนื่องจาก Φ เป็นฟังก์ชันเพิ่มสม่ำเสมอ (Strictly increasing function) ดังนั้น

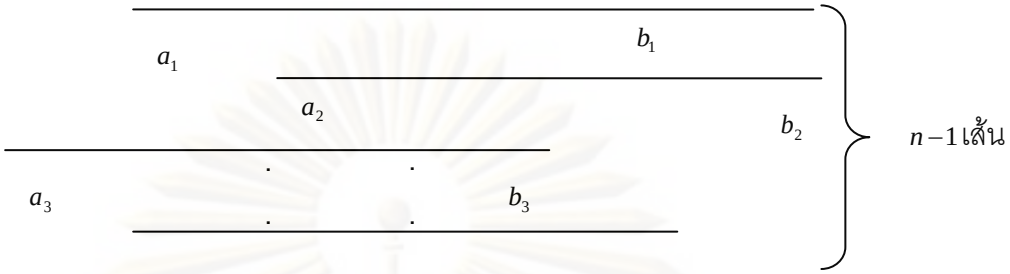
$$\hat{\beta}_0+\hat{\beta}_1 x_{1_{I_{n-1}}}+\hat{\beta}_2 x_{2_{I_{n-1}}} < \hat{\beta}_0+\hat{\beta}_1 x_{1_{I_n}}+\hat{\beta}_2 x_{2_{I_n}}$$

ตามเงื่อนไขเบื้องต้นที่กำหนด เมื่ออสมการ $n-1$ อสมการ เป็นจริง แสดงว่า $(a,b)=\bigcap_{l=1}^{n-1}(a_l, b_l)$

ดังนั้นจากตัวแบบ B $\quad P(Y=1) = \Phi\left(\hat{\beta}_0'+\hat{\beta}_1' X_1+\hat{\beta}_2' X_2\right)$

สมมติว่า $\hat{\beta}_0=\hat{\beta}_0'$ และ $\hat{\beta}_1=\hat{\beta}_1'$ นั่นคือ $P(Y=1) = \Phi\left(\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2' X_2\right)$

รูปที่ 4.1 ช่วงของค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_2'$ ที่เกิดจากอสมการ $n-1$ อสมการ และเกิดการทับซ้อนเพื่อให้ทุกอสมการเป็นจริง



ดังนั้น ถ้า $\hat{\beta}_2' \in (a,b)$ จะทำให้ค่าอันดับของค่าพยากรณ์จากตัวแบบ B ให้อันดับเดียวกับอันดับของค่าพยากรณ์จากตัวแบบ A และจากทฤษฎีบทที่ 1 เมื่อค่าอันดับของการพยากรณ์เป็นอันดับเดียวกัน ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองเท่ากัน

ต่อไปจะเป็นตัวอย่างสถานการณ์จำลองตามบทตามที่ 2 ดังนี้

ตัวอย่างที่ 4.2 สถานการณ์จำลองกรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ในตัวแบบพยากรณ์คงที่ โดยที่ $X_2 \sim Ber(0.50)$

ตารางที่ 4.2 แสดงค่าพยากรณ์, ค่าอันดับของค่าพยากรณ์ และพื้นที่ใต้โค้ง ROC จากการวิเคราะห์ข้อมูลใน 1 ครั้ง ที่ $\hat{\beta}_2' = 0.7610$ และ $\hat{\beta}_2' = 0.7700$ เมื่อ $X_2 \sim Ber(0.50)$ โดยที่ขนาดตัวอย่างในการทดลองเท่ากับ 50

ลำดับ	สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_2'$							
	$\hat{\beta}_2' = 0.7610$				$\hat{\beta}_2' = 0.7700$			
	h_j	r_j	d_i	r_i	h_j	r_j	d_i	r_i
1	0.1154	1	0.2853	2	0.1154	1	0.2853	2
2	0.2878	3	0.3288	5	0.2878	3	0.3288	5
3	0.3243	4	0.3439	6	0.3243	4	0.3439	6
4	0.3767	8	0.3588	7	0.3767	8	0.3588	7
5	0.3862	9	0.3981	10	0.3862	9	0.3981	10
6	0.4197	12	0.4075	11	0.4197	12	0.4075	11
7	0.4314	13	0.4587	15	0.4314	13	0.4587	15
8	0.4390	14	0.4671	16	0.4390	14	0.4671	16
9	0.4706	17	0.4717	19	0.4706	17	0.4717	19
10	0.4715	18	0.4888	20	0.4715	18	0.4888	20
11	0.4963	21	0.5360	23	0.4963	21	0.5396	23
12	0.5155	22	0.5539	25	0.5155	22	0.5575	25
13	0.5537	24	0.5933	29	0.5537	24	0.5933	29
14	0.5578	26	0.6657	33	0.5578	26	0.6690	33
15	0.5676	27	0.6971	35	0.5711	27	0.7003	35
16	0.5765	28	0.7133	37	0.5765	28	0.7163	37
17	0.6019	30	0.7225	39	0.6054	30	0.7255	39
18	0.6315	31	0.7243	40	0.6348	31	0.7273	40
19	0.6560	32	0.7288	41	0.6593	32	0.7318	41
20	0.6751	34	0.7459	42	0.6783	34	0.7487	42
21	0.6996	36	0.7709	43	0.7027	36	0.7736	43
22	0.7146	38	0.7759	44	0.7177	38	0.7786	44
23			0.7793	45			0.7820	45
24			0.7883	46			0.7909	46
25			0.8021	47			0.8046	47
26			0.8299	48			0.8322	48
27			0.8665	49			0.8684	49
28			0.8822	50			0.8839	50
ผลรวมอันดับ				827				827
พื้นที่ใต้โค้ง				0.6834				0.6834

ในกรณีเมื่อตัวแปรอิสระ X_2 มีการแจกแจงแบบเบอร์นูลลี ผลการวิจัยเปรียบเทียบค่าพยากรณ์และค่าอันดับของค่าพยากรณ์ที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ โดยที่ $\hat{\beta}_0 = \hat{\beta}_0'$ และ $\hat{\beta}_1 = \hat{\beta}_1'$ คงที่ จะได้ $\hat{\beta}_2' \in (0.7604, 0.7708)$ โดยเลือกพิจารณาในกรณี $\hat{\beta}_2' = 0.7610$ และ $\hat{\beta}_2' = 0.7700$ ซึ่งตกอยู่ในช่วงดังกล่าว เมื่อขนาดตัวอย่างในการทดลองเท่ากับ 50 จากการวิเคราะห์ข้อมูล 1 ครั้ง พบว่า ค่าพยากรณ์ทั้งในกลุ่มเหตุการณ์ที่สนใจและไม่สนใจมีความแตกต่างกัน แต่ค่าอันดับของค่าพยากรณ์ r_j และ r_i ของทั้งสองตัวแบบพยากรณ์มีอันดับเป็นลำดับเดียวกันทุกตัวอย่าง สอดคล้องกับทฤษฎีที่ 1 และบทตามที่ 2 ดังนั้นจึงส่งผลให้ค่าประมาณพื้นที่ใต้โค้ง ROC มีค่าเท่ากัน โดยมีค่าเท่ากับ 0.6834 ซึ่งเป็นจริงตามการพิสูจน์ทฤษฎีบทที่ 1 และบทตามที่ 2

สรุปผลการวิจัยเกี่ยวกับปัจจัยและผลกระทบที่มีต่อพื้นที่ใต้โค้ง ROC กรณีตัวแปรอิสระจำนวน 2 ตัวแปร

กรณีข้อมูล 1 ชุด และ ตัวแบบพยากรณ์โพรบิทสองตัวแบบ เมื่อค่าพยากรณ์จากตัวแบบพยากรณ์มีค่าไม่เท่ากัน โดยที่ค่าสัมประสิทธิ์การถดถอยของ พารามิเตอร์ $\hat{\beta}_0$ และ $\hat{\beta}_1$ คงที่ พบว่าจะมี (a,b) ซึ่งถ้า $\hat{\beta}_2' \in (a,b)$ ซึ่งทำให้ค่าอันดับของค่าพยากรณ์จากทั้งสองตัวแบบพยากรณ์เป็นค่าอันดับเดียวกัน และเมื่อได้ค่าอันดับของค่าพยากรณ์เป็นค่าอันดับเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC ก็จะมีค่าเท่ากัน และหากมองในมุมของการที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์มาจากวิธีการประมาณ 2 วิธี นั่นคือ $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ และ $(\hat{\beta}_0', \hat{\beta}_1', \hat{\beta}_2')$ โดยที่ $\hat{\beta}_0 = \hat{\beta}_0'$ และ $\hat{\beta}_1 = \hat{\beta}_1'$ จะมีช่วงเปิด (a,b) ซึ่งถ้าค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_2'$ ตกอยู่ในช่วงเปิดดังกล่าวแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองจะมีค่าเท่ากัน

ส่วนที่ 2 ผลการวิจัยเปรียบเทียบอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC

การวิจัยในกรณีนี้ ได้ทำการศึกษากรณีที่จำนวนขนาดตัวอย่างในการทดลองเป็น 50, 100, 200, 300, 400, 500 และ 1,000 โดยในแต่ละขนาดตัวอย่าง ได้ทำการศึกษาพื้นที่ใต้โค้ง ROC ค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC ที่ระดับนัยสำคัญ 0.01 และ 0.05 ซึ่งผลการวิจัยส่วนนี้นำเสนอในตารางที่ 4.3-4.5 และรูปภาพที่ 4.2-4.3 ดังต่อไปนี้

ตารางที่ 4.3 แสดงค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีการเปลี่ยนแปลง โดยที่ β_0 คงที่ เมื่อขนาดตัวอย่างในการทดลองเท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000

สัมประสิทธิ์การถดถอยของ β_1	พื้นที่ใต้โค้ง ROC เฉลี่ย ขนาดตัวอย่าง						
	$n=50$	$n=100$	$n=200$	$n=300$	$n=400$	$n=500$	$n=1{,}000$
-1.00	0.8474	0.8488	0.8481	0.8463	0.8468	0.8471	0.8473
-0.80	0.8031	0.8047	0.8035	0.8016	0.8020	0.8020	0.8023
-0.60	0.7450	0.7448	0.7449	0.7426	0.7443	0.7437	0.7439
-0.40	0.6727	0.6735	0.6738	0.6713	0.6721	0.6715	0.6723
-0.20	0.6011	0.5930	0.5909	0.5889	0.5889	0.5885	0.5890
0.00	0.5657	0.5450	0.5321	0.5264	0.5229	0.5203	0.5136
0.20	0.6014	0.5917	0.5872	0.5896	0.5890	0.5889	0.5887
0.40	0.6737	0.6717	0.6705	0.6725	0.6720	0.6715	0.6714
0.60	0.7448	0.7445	0.7425	0.7432	0.7433	0.7441	0.7438
0.80	0.8026	0.8023	0.8013	0.8021	0.8010	0.8024	0.8019
1.00	0.8462	0.8475	0.8466	0.8468	0.8459	0.8469	0.8471

ผลการวิจัยเปรียบเทียบค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีการเปลี่ยนแปลง โดยที่ β_0 คงที่ จากตารางที่ 4.3 พบว่า

ขนาดตัวอย่างเท่ากับ 50 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8474 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8462และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5657

ขนาดตัวอย่างเท่ากับ 100 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8488 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8475 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5450

ขนาดตัวอย่างเท่ากับ 200 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8481 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8466และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5321

ขนาดตัวอย่างเท่ากับ 300 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8467 อันดับสองที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8463 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5263

ขนาดตัวอย่างเท่ากับ 400 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8468 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8459และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5229

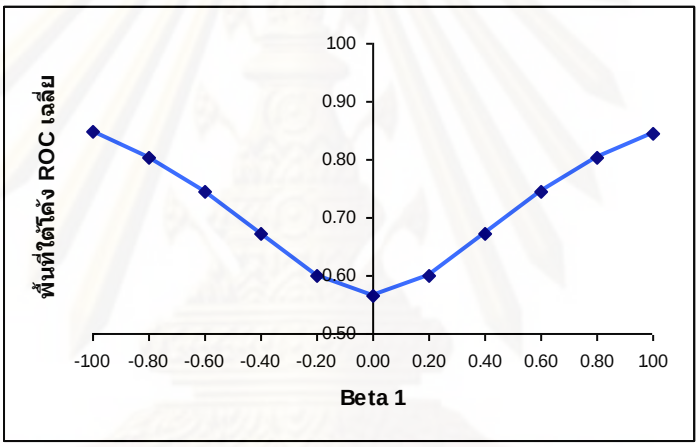
ขนาดตัวอย่างเท่ากับ 500 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8471 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.8469และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ 0.5203

ขนาดตัวอย่างเท่ากับ 1,000 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุดเท่ากับ 0.8473 อันดับสองที่ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROCเท่ากับ

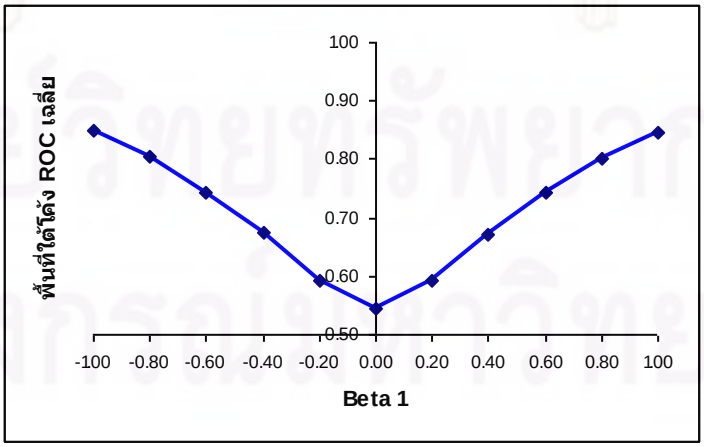
0.8471และมีค่าน้อยที่สุดที่ $\beta_1 = 0.00$ โดยมีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เท่ากับ 0.5136 และสามารถแสดงให้เห็นถึงการเปลี่ยนแปลงของค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง ในกรณีศึกษาทุกกรณี ดังรูปที่ 4.2

รูปที่ 4.2 แสดงการเปลี่ยนแปลงของค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง และขนาดตัวอย่างในการทดลองเท่ากับ 50, 100, 200, 300, 400, 500 และ 1,000

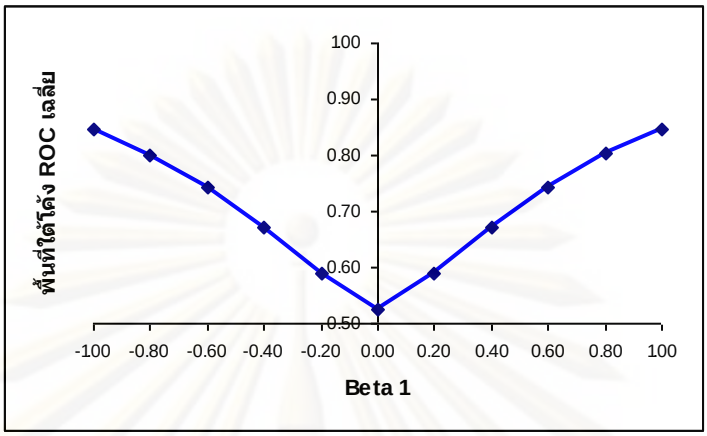
ก) $n = 50$



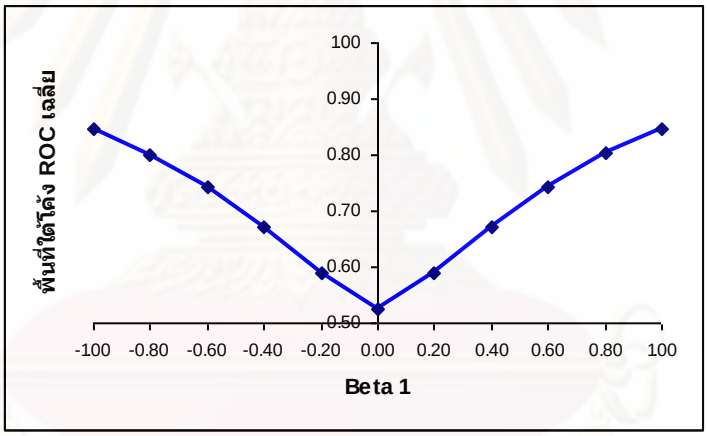
ข) $n = 100$



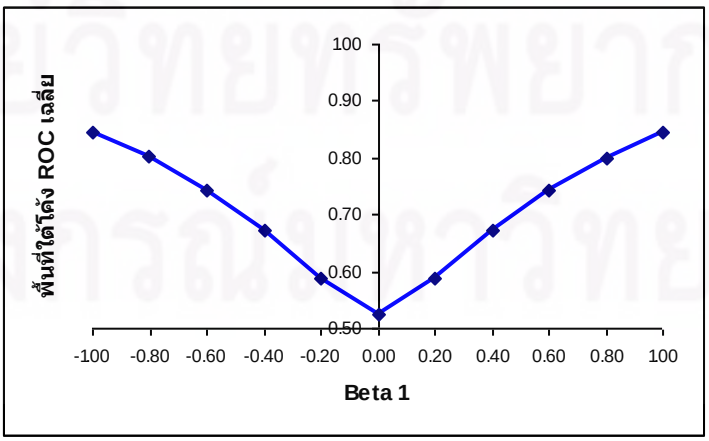
ค) $n = 200$



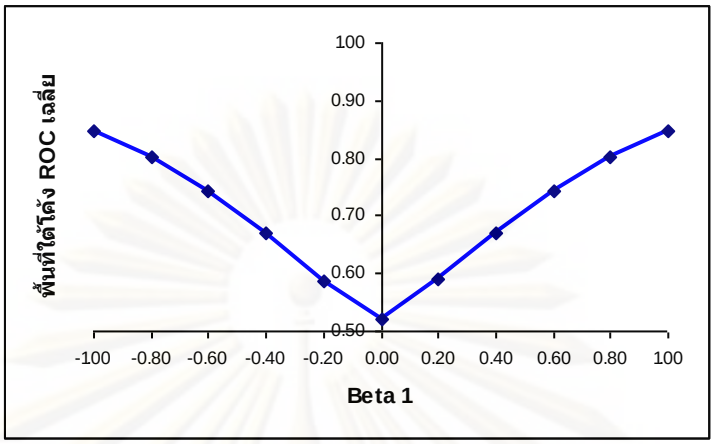
ง) $n = 300$



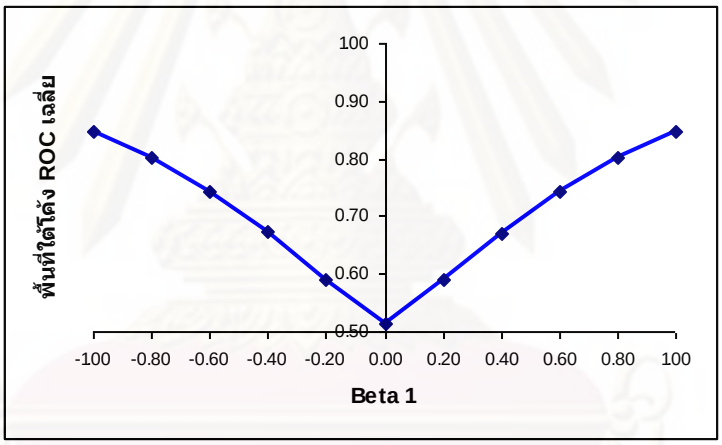
จ) $n = 400$



ฉ) $n = 500$



ช) $n = 1{,}000$



จากตารางที่ 4.3 และรูปที่ 4.2 สามารถสรุปผลการวิจัยเกี่ยวกับพื้นที่ใต้โค้ง ROC ดังนี้

1. <u>เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง</u> เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 มีค่าเท่ากับ $\beta_1 = -1.00$ และ $\beta_1 = 1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC ค่อนข้างมากในทุกขนาดตัวอย่าง และที่ สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 มีค่าเท่ากับ $\beta_1 = 0.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC น้อยมากในทุกขนาดตัวอย่าง

2. <u>เมื่อขนาดตัวอย่างเพิ่มขึ้น</u> ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีแนวโน้มเกือบจะเท่ากันในทุกขนาดตัวอย่างที่มีการเปลี่ยนแปลง แม้ขนาดตัวอย่างจะมีความแตกต่างกันมากก็ตาม ซึ่งน่าจะเป็นข้อสังเกตได้ว่าขนาดตัวอย่างไม่เป็นปัจจัยที่ส่งผลกระทบต่อพื้นที่ใต้โค้ง ROC

จากผลการวิจัยในตัวแบบจำลองข้อมูลข้างต้นพบว่า ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC แปรผันตรงกับค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 กล่าวคือ เ มื่อระดับความสัมพันธ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 กับค่าพยากรณ์เพิ่มขึ้น ส่งผลให้ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น นอกจากนี้สามารถตั้งข้อสังเกตได้ว่า ในระดับความสัมพันธ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 กับค่าพยากรณ์ที่เท่ากัน แม้ขนาดตัวอย่างจะมีการเปลี่ยนแปลง ก็ไม่ส่งผลกระทบต่อค่าเฉลี่ยพื้นที่ใต้โค้ง ROC

ตารางที่ 4.4 แสดงค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 กำหนดความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ที่ระดับนัยสำคัญ 0.01 และ 0.05 จำแนกตามขนาดตัวอย่าง เมื่อสัมประสิทธิ์ความถดถอยของพารามิเตอร์ β_1 ของตัวแบบจำลองข้อมูล ที่ $\beta_1 = 0.00$

ขนาดตัวอย่าง (n)	ระดับนัยสำคัญ	
	$\alpha = 0.01$	$\alpha = 0.05$
50	0.0085	0.0580
100	0.0105	0.0500
200	0.0135	0.0570
300	0.0100	0.0535
400	0.0130	0.0530
500	0.0125	0.0495
1,000	0.0080	0.0500

ผลการวิจัยค่าประมาณความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 โดยกำหนดความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ที่ระดับนัยสำคัญ 0.01 และ 0.05 ตามลำดับ จำแนกตามขนาดตัวอย่าง พบว่า

จากการทดสอบความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ สำหรับทุกขนาดตัวอย่างและทุกระดับนัยสำคัญ

ตารางที่ 4.5 แสดงค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีการเปลี่ยนแปลง โดยที่ β_0 คงที่ ที่ระดับนัยสำคัญ 0.01 และ 0.05

สัมประสิทธิ์การถดถอยของ	ระดับนัยสำคัญ 0.01							ระดับนัยสำคัญ 0.05						
β_1	$n=50$	$n=100$	$n=200$	$n=300$	$n=400$	$n=500$	$n=1{,}000$	$n=50$	$n=100$	$n=200$	$n=300$	$n=400$	$n=500$	$n=1{,}000$
-1.00	0.9020	0.9985	1.0000	1.0000	1.0000	1.0000	1.0000	0.9730	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
-0.80	0.7930	0.9930	0.9995	1.0000	1.0000	1.0000	1.0000	0.9340	0.9980	1.0000	1.0000	1.0000	1.0000	1.0000
-0.60	0.5795	0.9275	0.9995	1.0000	1.0000	1.0000	1.0000	0.8055	0.9800	1.0000	1.0000	1.0000	1.0000	1.0000
-0.40	0.2760	0.6210	0.9325	0.9895	0.9990	1.0000	1.0000	0.5095	0.8310	0.9845	0.9995	1.0000	1.0000	1.0000
-0.20	0.0620	0.1405	0.3315	0.5195	0.6845	0.7875	0.9910	0.1810	0.3395	0.5880	0.7535	0.8525	0.9210	1.0000
0.20	0.0635	0.1400	0.3105	0.5075	0.6770	0.8020	0.9890	0.1880	0.3320	0.5415	0.7585	0.8535	0.9270	1.0000
0.40	0.2785	0.6055	0.9345	0.9925	0.9995	1.0000	1.0000	0.5270	0.8235	0.9800	0.9990	1.0000	1.0000	1.0000
0.60	0.5705	0.9230	0.9990	1.0000	1.0000	1.0000	1.0000	0.7970	0.9865	1.0000	1.0000	1.0000	1.0000	1.0000
0.80	0.7930	0.9910	1.0000	1.0000	1.0000	1.0000	1.0000	0.9360	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000
1.00	0.9020	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	0.9765	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000

ผลการวิจัยเปรียบเทียบค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีการเปลี่ยนแปลง โดยที่ β_0 คงที่ ที่ระดับนัยสำคัญ 0.01 และ 0.05 จากตารางที่ 4.5 พบว่า

ที่ระดับนัยสำคัญ 0.01

ขนาดตัวอย่างเท่ากับ 50 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 0.902 และมีค่าน้อยที่สุดที่ $\beta_1 = -0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.062

ขนาดตัวอย่างเท่ากับ 100 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 0.999 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.140

ขนาดตัวอย่างเท่ากับ 200 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$, $-\beta_1 = 0.80$, $-\beta_1 = 0.60$, $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.000 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.311

ขนาดตัวอย่างเท่ากับ 300 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$, $\beta_1 = -0.80$, $\beta_1 = -0.60$, $\beta_1 = 0.60$, $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.00 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.508

ขนาดตัวอย่างเท่ากับ 400 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$, $\beta_1 = -0.80$, $\beta_1 = -0.60$, $\beta_1 = -0.40$, $\beta_1 = 0.40$, $\beta_1 = 0.60$, $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.00 และมีค่าน้อยที่สุดที่

ขนาดตัวอย่างเท่ากับ 500 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยมีค่าอำนาจการทดสอบเท่ากับ 1.00 ในเกือบทุกกรณี โดยยังคงมีเพียงกรณีที่ $\beta_1 = -0.20$ และ $\beta_1 = 0.20$ ซึ่งมีค่าอำนาจการทดสอบเท่ากับ 0.788 และ 0.802 ตามลำดับ

ขนาดตัวอย่างเท่ากับ 1,000 พบว่า เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยมีค่าอำนาจการทดสอบเท่ากับ 1.00 ในเกือบทุกกรณี โดยยังคงมีเพียงกรณีที่ $\beta_1 = -0.20$ และ $\beta_1 = 0.20$ ซึ่งมีค่าอำนาจการทดสอบเท่ากับ 0.991 และ 0.989 ตามลำดับ

ที่ระดับนัยสำคัญ 0.05

ขนาดตัวอย่างเท่ากับ 50 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 0.977 อันดับสองที่ $\beta_1 = -1.00$ มีค่าอำนาจการทดสอบเท่ากับ 0.973 และมีค่าน้อยที่สุดที่ $\beta_1 = -0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.181

ขนาดตัวอย่างเท่ากับ 100 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = 1.00$ และ $\beta_1 = -1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.000 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.332

ขนาดตัวอย่างเท่ากับ 200 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$, $\beta_1 = -0.80$, $\beta_1 = -0.60$, $\beta_1 = 0.60$, $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.000 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.542

ขนาดตัวอย่างเท่ากับ 300 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยที่ $\beta_1 = -1.00$, $\beta_1 = -0.80$, $\beta_1 = -0.60$, $\beta_1 = -0.40$, $\beta_1 = 0.60$, $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ มีค่าอำนาจการทดสอบมากที่สุดเท่ากับ 1.000 และมีค่าน้อยที่สุดที่ $\beta_1 = 0.20$ โดยมีค่าอำนาจการทดสอบเท่ากับ 0.754

ขนาดตัวอย่างเท่ากับ 400 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยมีค่าอำนาจการทดสอบเท่ากับ 1.000 ในเกือบทุกกรณี ยกเว้นเพียงกรณีที่ $\beta_1 = -0.20$ และ $\beta_1 = 0.20$ ซึ่งมีค่าอำนาจการทดสอบเท่ากับ 0.853 และ 0.854 ตามลำดับ

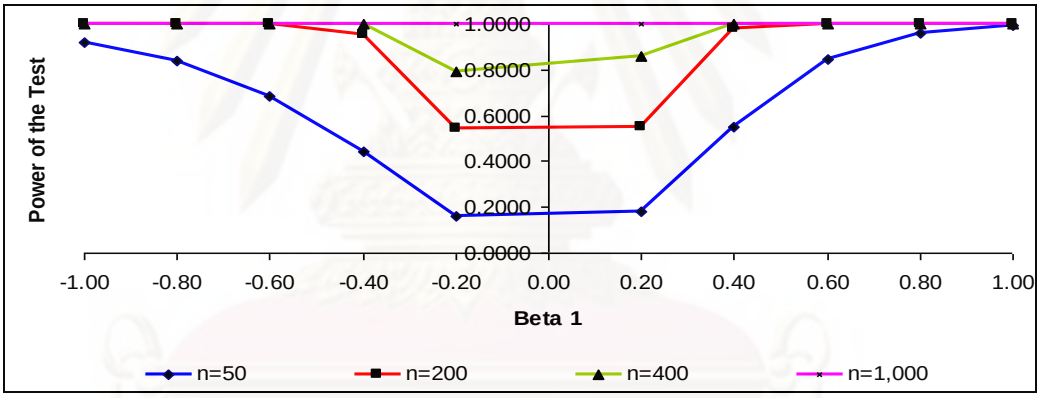
ขนาดตัวอย่างเท่ากับ 500 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น

โดยมีค่าอำนาจการทดสอบเท่ากับ 1.000 ในเกือบทุกกรณี ยกเว้นเพียงกรณีที่ $\beta_1 = -0.20$ และ $\beta_1 = 0.20$ ซึ่งมีค่าอำนาจการทดสอบเท่ากับ 0.921 และ 0.927 ตามลำดับ

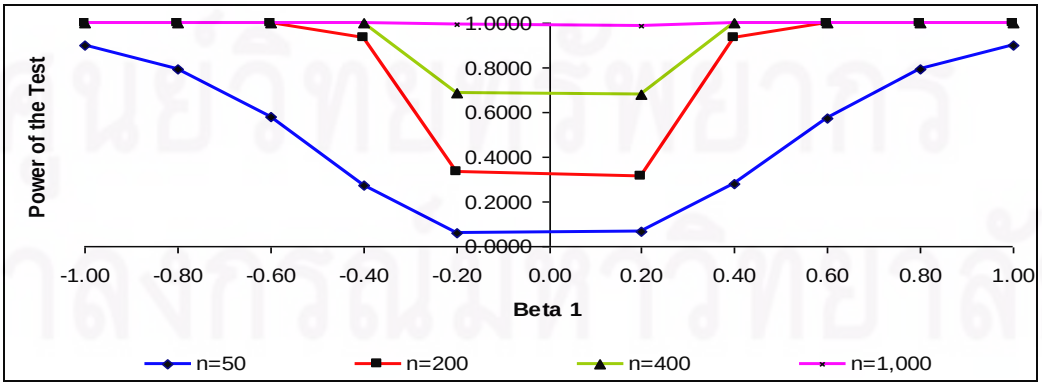
ขนาดตัวอย่างเท่ากับ 1,000 พบว่า เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น โดยมีค่าอำนาจการทดสอบเท่ากับ 1.000 ในทุกกรณี และสามารถแสดงให้เห็นถึงการเปลี่ยนแปลงของอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC ที่ระดับนัยสำคัญ 0.01 และ 0.05 ดังตัวอย่างในรูปที่ 4.3 ซึ่งแสดงในกรณีที่ขนาดตัวอย่างเท่ากับ 50, 200, 400 และ 1,000

รูปที่ 4.3 แสดงการเปลี่ยนแปลงของค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC ที่ขนาดตัวอย่าง 50, 200, 400 และ 1,000 ที่ระดับนัยสำคัญ 0.01 และ 0.05

ก) ระดับนัยสำคัญ 0.01



ข) ระดับนัยสำคัญ 0.05



จากตารางที่ 4.5 และรูปที่ 4.3 สามารถสรุปผลการวิจัยเกี่ยวกับค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC ดังนี้

1. <u>เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง</u> เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น โดยที่เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 มีค่าเท่ากับ $\beta_1 = -1.00$ และ $\beta_1 = 1.00$ ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเข้าใกล้และเกือบเท่ากับ 1.000 ในทุกขนาดตัวอย่าง และเมื่อ สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ตกอยู่ในช่วง $(0.20,0.40)$ เส้นกราฟของค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีความชันมาก นั่นคือ ในช่วงดังกล่าวค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้ นที่ใต้โค้ง ROC มีความไวค่อนข้างสูง

2. <u>เมื่อขนาดตัวอย่างเพิ่มขึ้น</u> ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มขึ้น จนมีค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเข้าใกล้และเกือบเท่ากับ 1.000 ในเกือบทุกกรณี

3. <u>ระดับนัยสำคัญ</u> ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น เมื่อระดับนัยสำคัญเพิ่มขึ้น

จากผลการวิจัยในตัวแบบจำลองข้อมูลข้างต้นพบว่า ค่า อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC แปรผันตรงกับค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ขนาดตัวอย่าง และระดับนัยสำคัญ กล่าวคือ เ มื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ขนาดตัวอย่าง และระดับนัยสำคัญเพิ่มขึ้น ส่งผลให้ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น

สรุปผลการวิจัยเกี่ยวกับพื้นที่ใต้โค้ง ROC ความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท

สำหรับทุกขนาดตัวอย่าง เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีการเปลี่ยนแปลง โดยที่ β_0 คงที่ ได้ข้อสังเกตว่า ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีแนวโน้มเกือบจะเท่ากัน ณ สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเดียวกัน โดยที่ในเกือบทุกกรณี $\beta_1 = -1.00$ มีค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มากที่สุด อันดับสองที่ $\beta_1 = 1.00$ และมีค่าน้อยที่สุดที่ $\beta_1 = -0.20$ และ $\beta_1 = 0.20$ และจากการทดสอบความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 สามารถสรุปผลได้ว่า ตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC

สามารถควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ สำหรับทุกสถานการณ์ที่ศึกษา ส่วนปัจจัยที่มีผลกระทบต่อค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท คือ สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ขนาดตัวอย่าง และระดับนัยสำคัญ กล่าวคือ เ มื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ขนาดตัวอย่าง และระดับนัยสำคัญเพิ่มขึ้น ส่งผลให้ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท โดยแบ่งพิจารณากรณีศึกษาออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร และตัวแปรอิสระจำนวน 2 ตัวแปร

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน1 ตัวแปร

โดยในงานวิจัยนี้ได้ทำการพิสูจน์ทางคณิตศาสตร์และศึกษาผลการวิจัยในสถานการณ์ต่าง ๆ ดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท แยกการพิสูจน์ทางคณิตศาสตร์เป็น 3 กรณี คือ

- กรณีไม่เจาะจงตัวแบบพยากรณ์
- กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง
- กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท พร้อมตัวอย่าง

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร

- ตัวแบบโพรบิทอย่างง่ายแบบ 2 กลุ่ม (Simple binary probit model)
- ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพมีการแจกแจงแบบเบอร์นูลี และกำหนดค่าตัวแปรตามมีค่าเพียง 2 ค่า คือ 0 และ 1
- ในกรณีตัวแปรอิสระจำนวน 1 ตัวแปร ตัวแปรอิสระ (X) เป็นข้อมูลเชิงปริมาณ โดยมีการแจกแจงแบบปกติ ด้วยพารามิเตอร์ μ และ σ นั่นคือ $X \sim N(\mu, \sigma^2)$ ในงานวิจัยครั้งนี้จะศึกษาที่ $\mu = 1$ และ $\sigma^2 = 1$

- สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในการจำลองข้อมูล ที่ $\beta_1 = -1.00$ $\beta_1 = -0.80$ $\beta_1 = -0.60$ $\beta_1 = -0.40$ $\beta_1 = -0.20$ $\beta_1 = 0.00$ $\beta_1 = 0.20$ $\beta_1 = 0.40$ $\beta_1 = 0.60$ $\beta_1 = 0.80$ และ $\beta_1 = 1.00$ โดยที่ β_0 คงที่ ($\beta_0 = 0.00$)
- จำนวนขนาดตัวอย่าง (n) เป็น 50, 100, 200, 300, 400, 500 และ 1,000
- กำหนดระดับนัยสำคัญ 0.01 และ 0.05
- กำหนดการกระทำซ้ำในแต่ละสถานการณ์เป็น 2,000 รอบ

สรุปผลการวิจัย

จากการศึกษาผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve และอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท ในแต่ละสถานการ ณ์ที่กำหนด ผลสรุปของการวิจัยเป็นดังนี้

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve

<u>กรณีไม่เจาะจงตัวแบบพยากรณ์</u>

กรณีข้อมูล 1 ชุด ถ้าตัวแบบพยากรณ์ A และตัวแบบพยากรณ์ B ให้ค่าอันดับของค่าพยากรณ์ในอันดับเดียวกัน แล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC ของตัวแบบ A และ ตัวแบบ B มีค่าเท่ากัน

<u>กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท</u>

กรณีข้อมูล 1 ชุด บนตัวแบบพยากรณ์โพรบิทสองตัวแบบ เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์โพรบิทมีค่าไม่เท่ากันแต่มีทิศทางเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC มีค่าเท่ากัน เนื่องจากค่าอันดับของค่าพยากรณ์ r_j และ r_i ยังคงมีค่าอันดับเดียวกัน โดยหากมองในมุ มของการที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์มาจากวิธีการประมาณ 2 วิธี นั่นคือ $\hat{\beta}_1$ และ $\hat{\beta}_1'$ โดยที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ดังกล่าวมีทิศทางเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์โพรบิททั้งสองจึงมีค่าเท่ากัน

กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท

กรณีข้อมูล 1 ชุด บนตัวแบบพยากรณ์โพรบิทสองตัวแบบ เมื่อค่าพยากรณ์จากตัวแบบพยากรณ์มีค่าไม่เท่ากัน โดยที่ค่าสัมประสิทธิ์การถดถอยของ พารามิเตอร์ β_0 และ β_1 คงที่ พบว่าจะสามารถหาช่วงเปิดของค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_2 ในตัวแบบพยากรณ์ตัวใดตัวหนึ่ง ซึ่งทำให้ค่าอันดับของค่าพยากรณ์จากทั้งสองตัวแบบพยากรณ์เป็นค่าอันดับเดียวกัน และเมื่อได้ค่าอันดับของค่าพยากรณ์เป็นค่าอันดับเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC ก็จะมีค่าเท่ากัน และหากมองในมุมของการที่ค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์มาจากวิธีการประมาณ 2 วิธี นั่นคือ $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ และ $(\hat{\beta}'_0, \hat{\beta}'_1, \hat{\beta}'_2)$ โดยที่ $\hat{\beta}_0 = \hat{\beta}'_0$ และ $\hat{\beta}_1 = \hat{\beta}'_1$ จะมีช่วงเปิด (a,b) ซึ่งถ้าค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}'_2$ ตกอยู่ในช่วงเปิดดังกล่าวแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ทั้งสองจะมีค่าเท่ากัน

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบพยากรณ์โพรบิท เมื่อตัวแปรอิสระจำนวน 1 ตัวแปร

1. เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลเปลี่ยนแปลง

   - ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น ส่งผลให้ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น

   - อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 เพิ่มขึ้น อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น

2. เมื่อขนาดตัวอย่างเพิ่มขึ้น

   - ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC ในกรณีที่สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีค่าเท่ากัน แม้ว่าขนาดตัวอย่างเพิ่มขึ้นแต่ก็ได้ข้อสังเกตว่า ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีแนวโน้มเกือบจะเท่ากัน

   - อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เมื่อขนาดตัวอย่างเพิ่มขึ้นค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น จนมี

ค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC มีค่าเข้าใกล้และเกือบเท่ากับ 1.000 ในเกือบทุกกรณี

3. ระดับนัยสำคัญ

- อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC โดยอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC เพิ่มขึ้น เมื่อระดับนัยสำคัญเพิ่มขึ้น

อภิปรายผลการวิจัย

ส่วนที่ 1 ผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์โพรบิท

<u>กรณีไม่เจาะจงตัวแบบ</u>

ในการศึกษาถึงผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์ กรณีข้อมูล 1 ชุด หากตัวแบบพยากรณ์ใด ๆ ให้ค่าอันดับของการพยากรณ์เป็นอันดับเดียวกัน ตัวแบบพยากรณ์ดังกล่าวจะให้ค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากัน ซึ่ งสอดคล้องกับงานวิจัยของ James A. Hanley และ Barbara J. McNeil (1982) พบว่า ค่าประมาณพื้นที่ใต้โค้ง ROC ไม่ได้ขึ้นอยู่กับค่าจริงของการพยากรณ์แต่ขึ้นอยู่กับค่าอันดับของค่าพยากรณ์เป็นสำคัญ

<u>กรณีตัวแปรอิสระจำนวน 1 ตัวแปร บนตัวแบบพยากรณ์โพรบิท</u>

ในการศึกษาถึงผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve กรณีข้อมูล 1 ชุด บนตัวแบบพยากรณ์โพรบิทสองตัวแบบ เมื่อสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์มีค่าไม่เท่ากันแต่มีทิศทางเดียวกัน พบว่า ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ดังกล่าวมีค่าเท่ากันเสมอ แม้ว่าค่าพยากรณ์จากแต่ละตัวแบบพยากรณ์จะมีค่าแตกต่างกัน เนื่องจากการเปลี่ยนแปลงสัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$ ในตัวแบบพยากรณ์ เปรียบเสมือนการเปลี่ยนแปลงระดับความสัมพันธ์ที่ตัวแปรอิสระ X_1 มีกับค่าพยากรณ์ แต่ความแตกต่างของข้อมูลตัวอย่างซึ่งอยู่ในกลุ่มเหตุการณ์ที่สนใจและกลุ่มเหตุการณ์ที่ไม่สนใจยังคงอยู่ในตัวแปรอิสระ X_1 ดังนั้นเมื่อนำค่าพยากรณ์จากแต่ละตัวแบบมาทำการเรียงอันดับจึงยังคงให้ค่าอันดับเป็นค่าเดิมและเป็นค่าเดียวกัน ซึ่งแสดงให้เห็นว่า ค่าประมาณพื้นที่ใต้โค้ง ROC

ไม่ได้มีค่าขึ้นอยู่กับค่าจริงของค่าพยากรณ์ แต่กลับมีค่าขึ้นอยู่กับค่าอันดับเป็นสำคัญ ซึ่งสอดคล้องกับงานวิจัยของ James A. Hanley และ Barbara J. McNeil (1982) เช่นเดียวกัน และจากข้างต้นทำให้ผู้วิจัยสรุปได้ว่าแม้จะเกิดความผิดพลาดจากวิธีการในการประมาณค่าสัมประสิทธิ์ความถดถอยของพารามิเตอร์ในตัวแบบพยากรณ์โพรบิทซึ่งส่งผลให้ค่าประมาณของสัมประสิทธิ์ความถดถอยของพารามิเตอร์ดังกล่าวอาจมีค่าที่คลาดเคลื่อนไปจากค่าสัมประสิทธิ์ความถดถอยของพารามิเตอร์จริงที่ควรจะเป็น ก็จะไม่ ส่งผลให้ค่าประมาณพื้นที่ใต้โค้ง ROC จากตัวแบบพยากรณ์ดังกล่าวเปลี่ยนแปลงไป นั่นคือ ยังคงค่าตัวชี้วัดถึงความ สามารถในการจำแนกกลุ่มเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้เท่าเดิม

<u>กรณีตัวแปรอิสระจำนวน 2 ตัวแปร บนตัวแบบพยากรณ์โพรบิท</u>

ในการศึกษาถึงผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve กรณีข้อมูล 1 ชุด และตัวแบบพยากรณ์โพรบิทสองตัวแบบ เมื่อสัมประสิทธิ์การถดถอยของ พารามิเตอร์ $\hat{\beta}_0$ และ $\hat{\beta}_1$ ในตัวแบบพยากรณ์ทั้งสองคงที่ พบว่า จะมี (a,b) ซึ่งถ้า $\hat{\beta}_2' \in (a,b)$ ซึ่งทำให้ค่าอันดับของค่าพยากรณ์จากทั้งสองตัวแบบพยากรณ์เป็นค่าอันดับเดียวกัน และเมื่อได้ค่าอันดับของค่าพยากรณ์เป็นค่าอันดับเดียวกันแล้ว ค่าประมาณพื้นที่ใต้โค้ง ROC ก็จะมีค่าเท่ากัน นั่นแสดงให้เห็นว่าหาก (a,b) เป็นช่วงที่กว้างจะแสดงให้เห็นว่าตัวแบบพยากรณ์โพรบิทดังกล่าวมีความไวต่ำ ในทางกลับกันหาก (a,b) เป็นช่วงที่แคบจะแสดงให้เห็นว่าตัวแบบพยากรณ์โพรบิทดังกล่าวมีความไวสูง

ส่วนที่ 2 **อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง** ROC

- ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC เมื่อค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 ในตัวแบบจำลองข้อมูลมีความสัมพันธ์กับ Y เพิ่มขึ้น ส่งผลให้ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC มีค่าเพิ่มขึ้น นั้นคือ เมื่อ ค่าสัมบูรณ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 มีความสัมพันธ์กับ Y มากขึ้นเท่าไร ก็จะส่งผลให้ค่าพยากรณ์ในกลุ่มเหตุการณ์ที่สนใจกับค่าพยากรณ์ในกลุ่มเหตุการณ์ที่ไม่สนใจแตกต่างกันมากเท่านั้น ดังนั้นจึงส่งผลให้เมื่อ สัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 มีความสัมพันธ์กับค่าพยากรณ์เพิ่มขึ้น ค่าเฉลี่ยพื้นที่ใต้โค้ง ROC ก็จะมีค่าเพิ่มขึ้นด้วย สอดคล้องกับคำกล่าวในงานวิจัย Nancy A. Obuchowski (2003) ว่าค่าพื้นที่ใต้โค้ง ROC จะเป็นตัวชี้วัดถึงความสามารถในการจำแนกกลุ่มเหตุการณ์ที่สนใจออกจากกลุ่มของเหตุการณ์ที่ไม่สนใจ นั่นคือตัวแบบในการพยากรณ์ตัวแบบใด

สามารถสร้างค่าพยากรณ์ในทั้งสองกลุ่มให้มีความแตกต่างกันได้มาก ก็จะมีค่าพื้นที่ใต้โค้ง ROC มาก

- อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC สำหรับปัจจัยที่มีผลกระทบต่อค่าอำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท คือ ระดับความสัมพันธ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 กับค่าพยากรณ์ในตัวแบบจำลองข้อมูล เนื่องจากเมื่อ ระดับความสัมพันธ์ของสัมประสิทธิ์การถดถอยของพารามิเตอร์ β_1 กับค่าพยากรณ์ในตัวแบบจำลองข้อมูลมีค่าเพิ่มสูงขึ้น จะส่งผลให้ค่าพยากรณ์ทั้งในกลุ่มเหตุการณ์ที่สนใจและกลุ่มเหตุการณ์ที่ไม่สนใจมีค่าแตกต่างกันมากขึ้น นั่นคือ ตัวแบบพยากรณ์ดังกล่าวจะสามารถจำแนกกลุ่มเหตุการณ์ที่สนใจออกจากกลุ่มเหตุการณ์ที่ไม่สนใจได้ดียิ่งขึ้น ส่งผลให้ค่า ประมาณพื้นที่ใต้โค้ง ROC มีค่าเพิ่มสูงขึ้นและแตกต่างจากค่าประมาณพื้นที่ใต้โค้ง ROC เท่ากับ 0.50 มากยิ่งขึ้น และเนื่องจากอำนาจการทดสอบจะแปรตามขนาดของความแตกต่างระหว่างค่าประมาณพื้นที่ใต้โค้ง ROC ของกลุ่มประชากรภายใต้สมมติฐานศูนย์ (H_0) กับสมมติฐานทางเลือก (H_1) ถ้าค่าประมาณพื้นที่ใต้โค้ง ROC ของกลุ่มประชากรภายใต้สมมติฐานทางเลือกยิ่งแตกต่างจากค่าประมาณพื้นที่ใต้โค้ง ROC ของกลุ่มประชากรภายใต้สมมติฐานศูนย์มากเท่าไร จะทำให้มีอำนาจการทดสอบของตัวสถิติเพิ่มมากขึ้นเท่านั้น นอกจากนี้ เมื่อขนาดตัวอย่างและระดับนัยสำคัญเพิ่มขึ้นอำนาจการทดสอบของตัวสถิติก็มีค่าเพิ่มขึ้นเช่นกัน

ข้อเสนอแนะ

1. ในส่วนของผลกระทบของคุณสมบัติความไม่ผันแปรของ ROC curve บนตัวแบบพยากรณ์ อาจมีการศึกษาในกรณีที่ตัวแปรอิสระและจำนวนขนาดตัวอย่างในการทดลองมีจำนวนมากขึ้น โด ยหากต้องการพิจารณาช่วงเปิดของสัมประสิทธิ์การถดถอยของพารามิเตอร์ อาจมีการนำโปรแกรมเชิงเส้น (Linear Programming) เข้ามาช่วยในการคำนวณเพื่อความสะดวกและรวดเร็ว

2. ในงานวิจัยครั้งนี้ในส่วนของการหาอำนาจการทดสอบของตัวสถิติพื้นที่ใต้โค้ง ROC ได้ศึกษาในกรณีที่ประชากรมีการแจกแจงแบบปกติ ในงานวิจัยต่อไปอาจทำการศึกษาในกรณีที่ประชากรไม่ได้มีการแจกแจงปกติ

รายการอ้างอิง

ภาษาไทย

กัลยา วาณิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 1. กรุงเทพมหานคร : บริษัท ธรรมสาร จำกัด, 2548.

ธีระพร วีระถาวร. ความน่าจะเป็นกับการประยุกต์. พิมพ์ครั้งที่2. กรุงเทพมหานคร : นำอักษรการพิมพ์, 2539.

มนทกานติ หรรษวรวงศ์. การเปรียบเทียบอำนาจการทดสอบของตัวสถิติทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ 2 ประชากร. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2530.

เสกสรร เกียรติสุไพบูลย์. เอกสารประกอบการสอนวิชาการจำลอง. กรุงเทพมหานคร : จุฬาลงกรณ์มหาวิทยาลัย, 2550.

เสาวรส ใหญ่สว่าง. เอกสารคำสอนสถิติที่ไม่ใช้พารามิเตอร์. กรุงเทพมหานคร : จุฬาลงกรณ์มหาวิทยาลัย, 2551.

ภาษาอังกฤษ

Bewick, Viv., Cheek, Liz. and Ball, Jonathan. Receiver operating characteristic curves. Critical Care 8(December 2004) : 508-512.

D.M. Green and J.M. Swets. Signal detection theory and psychophysics. New York : John Wiley and Sons Inc, 1966.

Daniel A. Powers, Yu Xie. Statistic Methods for Categorical Data Analysis. USA : Academic Press, 2000.

Green, William H. Econometric Analysis. 4th edition. New Jersey, USA : Prentice Hall International Inc, 2000.

Hanley , Jame A. and McNeil, Barbara J. The Meaning and use of the Area under a ReceiverOperating Characteristic(ROC) curve. Radiology 143(April 1982) : 29-36.

Lasko , Thomas A., Bhagwat, Jui G., Zou, Nelly H. and Ohno-Machado, Lucila. The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics 38 (2005) :404-415.

Mithat Gonen. Receiver Operating Characteristic(ROC) Curves. Statistic and Data Analysis 211-31 : 1-18.

Obuchowski, Nancy A. Receiver Operating Characteristic Curves and Their Use in Radiology. Radiology 229(October 2003) : 3-8.

Park, Seong Ho., Goo, Jin MO. and Jo, Chan-Hee. Receiver Operating Characteristic(ROC) Curve:Practical Review for Radiologists. Korean J Radiol 5(March 2004) : 11-18.

Streiner, David L. and Cairney, John. What's Under the ROC? An Introduction to Receiver Operating Characteristics Curves. The Canadian Journal of Phychiatry 52(February 2007) : 121-128.

ภาคผนวก

ภาคผนวก ก ตัวอย่างการจำลองสถานการณ์จากบทตามที่ 1

ตารางที่ A ตารางแสดงค่าพยากรณ์, ค่าอันดับของค่าพยากรณ์ และพื้นที่ใต้โค้ง ROC จากการวิเคราะห์ข้อมูลใน 1 ครั้ง ที่ $\hat{\beta}_1 = 0.60$ และ $\hat{\beta}_1 = 0.80$ โดยที่ขนาดตัวอย่างในการทดลองเท่ากับ 50

ลำดับ	สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$							
	$\hat{\beta}_1 = 0.60$				$\hat{\beta}_1 = 0.80$			
	h_j	r_j	d_i	r_i	h_j	r_j	d_i	r_i
1	0.1027	1	0.1232	3	0.0640	1	0.0840	3
2	0.1050	2	0.1990	4	0.0662	2	0.1684	4
3	0.2335	5	0.2951	7	0.2109	5	0.2906	7
4	0.2667	6	0.3313	9	0.2534	6	0.3391	9
5	0.3281	8	0.4022	11	0.3347	8	0.4350	11
6	0.3624	10	0.4277	15	0.3810	10	0.4695	15
7	0.4122	12	0.4431	16	0.4486	12	0.4904	16
8	0.4165	13	0.4532	18	0.4544	13	0.5039	18
9	0.4241	14	0.4865	20	0.4647	14	0.5484	20
10	0.4481	17	0.4890	21	0.4971	17	0.5516	21
11	0.4640	19	0.5199	23	0.5184	19	0.5922	23
12	0.5116	22	0.5515	26	0.5814	22	0.6328	26
13	0.5478	24	0.5539	28	0.6281	24	0.6359	28
14	0.5484	25	0.5984	32	0.6288	25	0.6911	32
15	0.5519	27	0.5991	33	0.6333	27	0.6920	33
16	0.5821	29	0.6515	35	0.6711	29	0.7536	35
17	0.5893	30	0.6634	37	0.6800	30	0.7670	37
18	0.5936	31	0.6713	38	0.6853	31	0.7758	38
19	0.6220	34	0.7109	41	0.7193	34	0.8181	41
20	0.6627	36	0.7132	42	0.7663	36	0.8204	42
21	0.6937	39	0.7192	43	0.8000	39	0.8265	43
22	0.7073	40	0.7493	44	0.8143	40	0.8562	44
23	0.7720	45	0.7767	47	0.8771	45	0.8813	47
24	0.7756	46	0.7767	48	0.8803	46	0.8813	48
25	0.7886	49	0.8012	50	0.8916	49	0.9023	50
ผลรวมอันดับ				691				691
พื้นที่ใต้โค้ง				0. 5856				0. 5856

ตารางที่ B ตารางแสดงค่าพยากรณ์, ค่าอันดับของค่าพยากรณ์ และพื้นที่ใต้โค้ง ROC จากการวิเคราะห์ข้อมูลใน 1 ครั้ง ที่ $\hat{\beta}_1 = 1.00$ โดยที่ขนาดตัวอย่างในการทดลองเท่ากับ 50

ลำดับ	สัมประสิทธิ์การถดถอยของพารามิเตอร์ $\hat{\beta}_1$			
	$\hat{\beta}_1 = 1.00$			
	h_j	r_j	d_i	r_i
1	0.0377	1	0.0550	3
2	0.0395	2	0.1411	4
3	0.1897	5	0.2862	7
4	0.2405	6	0.3469	9
5	0.3414	8	0.4683	11
6	0.4000	10	0.5117	15
7	0.4854	12	0.5378	16
8	0.4927	13	0.5546	18
9	0.5057	14	0.6091	20
10	0.5462	17	0.6130	21
11	0.5725	19	0.6615	23
12	0.6487	22	0.7086	26
13	0.7031	24	0.7120	28
14	0.7040	25	0.7730	32
15	0.7091	27	0.7739	33
16	0.7513	29	0.8370	35
17	0.7610	30	0.8499	37
18	0.7667	31	0.8583	38
19	0.8026	34	0.8962	41
20	0.8493	36	0.8982	42
21	0.8804	39	0.9033	43
22	0.8930	40	0.9270	44
23	0.9424	45	0.9454	47
24	0.9447	46	0.9454	48
25	0.9525	49	0.9593	50
ผลรวมอันดับ				691
พื้นที่ใต้โค้ง				0. 5856

ภาคผนวก ข ตัวอย่างการใช้โปรแกรม R ในการดำเนินงานวิจัย

ส่วนที่ 2 อำนาจการทดสอบของตัวสถิติสำหรับพื้นที่ใต้โค้ง ROC บนตัวแบบโพรบิท โดยใช้ฟังก์ชันเปรียบเทียบ $C(d_i, h_j)$

```
AUC=c()

SD=c()

test=c()

p_value=c()

count=c()

z<-2000

alpha<-0.05                # α = 0.01 และ α = 0.05 #

for(n in 1:z)

{

        beta<-0.5

        beta0<- 0.00     #FIX Beta 0 = 0.00 #

        x<-rnorm(100,1,1)

        error<-rnorm(100,0,1)

        y<-beta0+beta*x+error

        y1<-ifelse(y>0,1,0)

                #test probit regression

                Data<-data.frame(y1,x)
```

```
        Probit<-glm(y1~x,family=binomial(link=probit),data=Data)
        summary(Probit)
# ROC CURVE and AUC #
    ypredict<-predict.glm(Probit,Data,type="response")
    H<-ifelse(y1==0,ypredict,NA)
    D<-ifelse(y1==1,ypredict,NA)
    h<-na.omit(H)
    d<-na.omit(D)
    tp<-c()
    tn<-c()
    for(i in 1:length(ypredict))
    {       tp[i]<-0
            for(j in 1:length(d))
            {
                ifelse(d[j]>ypredict[i],tp[i]<-tp[i]+1,0)
            }
            tn[i]<-0
            for(j in 1:length(h))
            {
                ifelse(h[j]<=ypredict[i],tn[i]<-tn[i]+1,0)
```

```
        }
}
sen<-tp/sum(y1)
spec<-tn/sum(1-y1)
plot(1-spec,sen)
sumch<-c()
for(j in 1:length(h))
{
        sumch[j]<-0
        for(i in 1:length(d))
        {
                if(h[j]<d[i])
                {
                        sumch[j]<-sumch[j]+1
                }
                else
                {
                        if(h[j]==d[i])
                        {
                                sumch[j]<-sumch[j]+0.5
```

```
                                        }

                                        # case > not to do

                                    }

                                }

                            }

AUC[n]<-sum(sumch)/(length(h)*length(d))

SD[n]<-sqrt((sum(y1)+ sum(1-y1)+1)/(12*sum(y1)*sum(1-y1)))

test[n]<-abs((AUC[n]-0.50))/SD[n]

p_value[n]<-2*(1-pnorm(test[n], mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE))

count[n]<-ifelse(p_value[n]<alpha,1,0)

}

DATA<-data.frame(count)

setwd("C:/seednum")

write.table(DATA,"DA15.txt",sep=",")
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวปฐมาภรณ์ สานุกูล เกิดวันที่ 7 เมษายน พ.ศ. 2525 สำเร็จการศึกษาปริญญาตรีจากสาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ประกาศนียบัตร (ป.บัณฑิต) สาขาวิชาชีพครู ภาควิชาศึกษาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ และได้เข้าศึกษาต่อในระดับปริญญาโท ที่ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปี พ.ศ. 2550