การเปรียบเทียบวิธีการคัดเลือกตัวแปรสำหรับการถดถอยโลจิสติกในข้อมูลที่มีมิติสูง

นายรัชพงศ์ ปรัชญาเศรษฐ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

A COMPARISON OF FEATURE SELECTION METHODS FOR LOGISTIC REGRESSION IN HIGH DIMENSIONAL DATA

Mr. Ratchaphong Pratchayasait

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการคัดเลือกตัวแปรสำหรับการถดถอยโลจิสติกในข้อมูลที่มีมิติสูง
โดย	นายรัชพงศ์ ปรัชญาเศรษฐ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(รองศาสตราจารย์ ดร.วิฐูรา พึ่งพาพงศ์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.ดลชัย ละออนวล)

รัชพงศ์ ปรัชญาเศรษฐ : การเปรียบเทียบวิธีการคัดเลือกตัวแปรสำหรับการถดถอยโลจิสติกในข้อมูลที่มีมิติสูง. (A COMPARISON OF FEATURE SELECTION METHODS FOR LOGISTIC REGRESSION IN HIGH DIMENSIONAL DATA) อ.ที่ปรึกษาหลัก : รศ. ดร.เสกสรร เกียรติสุไพบูลย์

Regularization เป็นวิธีการป้องกันปัญหา overfitting ด้วยการเพิ่มฟังก์ชันการลงโทษไปในตัวแบบเพื่อให้เกิดการคัดกรองตัวแปรเข้าสู่ตัวแบบ งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของวิธีการคัดกรองตัวแปรสำหรับการวิเคราะห์การถดถอยโลจิสติกในข้อมูลที่มีมิติสูง ด้วยการใช้ฟังก์ชันการลงโทษในรูปแบบ (1) L_0-regularization (2) L_1-regularization (3) L_0L_2-regularization การวิจัยนี้ใช้การจำลองข้อมูลเพื่อทำการทดสอบ 18 กรณี โดยกำหนดค่าที่ต่างกันประกอบด้วย จำนวนตัวแปรอิสระมีจำนวน 200, 500 และ 1000 ตัวแปร ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0, 0.5 และ 0.9 อัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าเท่ากับ 1 และ 6 โดยจำลองข้อมูลแต่ละกรณีจำนวน 100 ชุด ในการศึกษานี้มุ่งเน้นที่การเปรียบเทียบประสิทธิภาพในการคัดกรองตัวแปรของตัวแบบ และประสิทธิภาพในการทำนายของตัวแบบ ซึ่งเปรียบเทียบประสิทธิภาพในแต่ละวิธีด้วย ความผิดพลาดในการตรวจจับเชิงบวก ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว และ พื้นที่ใต้เส้นโค้ง จากการศึกษาพบว่าวิธี L_0 มีความแม่นยำในการคัดกรองตัวแปรมากที่สุดเมื่อพิจารณาด้วยความผิดพลาดในการตรวจจับเชิงบวก เมื่อพิจารณาด้วยค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว พบว่าวิธี L_1 และ L_0L_2 มีประสิทธิภาพในการคัดกรองตัวแปรใกล้เคียงกัน แต่วิธี L_0L_2 จะมีประสิทธิภาพสูงกว่าเมื่อความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าสูง และเมื่อพิจารณาประสิทธิภาพในการทำนายของตัวแบบด้วยพื้นที่ใต้เส้นโค้ง พบว่าวิธี L_1 จะมีประสิทธิภาพสูงที่สุดในทุกกรณี

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6280267126 : MAJOR STATISTICS

KEYWORD: Logistic Regression, Feature Selection, Regularization, High Dimensional Data

Ratchaphong Pratchayasait : A COMPARISON OF FEATURE SELECTION METHODS FOR LOGISTIC REGRESSION IN HIGH DIMENSIONAL DATA. Advisor: Assoc. Prof. SEKSAN KIATSUPAIBUL, Ph.D.

Regularization is a method to circumvent the overfitting by adding penalty function to a model which results in a feature selection. This research aims to study and compare the performances of feature selection methods for binary logistic regression in high-dimensional data by using penalty function of the forms: (1) L_0-regularization (2) L_1-regularization and (3) L_0L_2-regularization. Simulated datasets are organized into 18 cases using various number of independent variables (features) (200, 500, 1000), correlation (0, 0.5, 0.9), and signal to noise ratio (1, 6), each with 100 simulated datasets. According to the performances, the study emphasizes on the accuracy of variables selection and predictive performance, which are compared in terms of False Positive, F_1-Score, and Area under the Curve (AUC). This paper shows that L_0-regularization yields the highest accuracy of the variables selection in terms of False Positive. For F_1-Score, L_1-regularization and L_0L_2-regularization, are comparable. However, L_0L_2-regularization tends to perform better when the correlations among independent variables are high. In addition, L_1-regularization outperforms other methods in terms of predictive performance measured by AUC.

Field of Study: Statistics Student's Signature .....

Academic Year: 2022 Advisor's Signature .....

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณรองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ เป็นอย่างยิ่งที่สละเวลาให้คำแนะนำ ให้ความรู้ที่เกี่ยวข้อง และให้คำปรึกษาที่มีประโยชน์ตลอดจนให้ความช่วยเหลือในการปรับปรุงแก้ไขข้อบกพร่องในวิทยานิพนธ์ฉบับนี้ จนทำให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

ผู้วิจัยขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช ประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์ และ ผู้ช่วยศาสตราจารย์ ดร.ดลชัย ละออนวล กรรมการสอบวิทยานิพนธ์ ที่ให้เกียรติสละเวลาเป็นคณะกรรมการสอบวิทยานิพนธ์ ตลอดจนให้คำแนะนำที่เป็นประโยชน์ ตรวจสอบและช่วยเหลือในการปรับปรุงแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

นอกจากนี้ผู้วิจัยขอขอบพระคุณคณาจารย์ทุกท่านที่ให้ความรู้ทางด้านวิชาการ รวมไปถึงเจ้าหน้าที่ภาควิชาสถิติที่ช่วยอำนวยความสะดวกในด้านต่างๆ

สุดท้ายนี้ผู้วิจัยขอขอบพระคุณครอบครัวของผู้วิจัย ที่คอยสนับสนุน ผลักดัน และให้กำลังใจเสมอมา

รัชพงศ์ ปรัชญาเศรษฐ

สารบัญ

หน้า

สารบัญภาพ

หน้า

สารบัญตาราง

หน้า

บทที่ 1 บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เป็นที่น่าสังเกตว่าในปัจจุบันข้อมูลที่ถูกนำมาใช้เพื่อการวิเคราะห์ทางสถิติมีขนาดใหญ่ขึ้น เนื่องจากความสามารถในการจัดเก็บข้อมูลปริมาณมาก และเทคโนโลยีที่สามารถเข้าถึงข้อมูลได้อย่างสะดวกรวดเร็ว ทำให้เกิดข้อมูลชนิดที่มีจำนวนของตัวแปรมากกว่าจำนวนของตัวอย่าง ($p > n$) ซึ่งข้อมูลลักษณะนี้เรียกว่า ข้อมูลที่มีมิติสูง (High – Dimensional Data) วิธีการหนึ่งที่ใช้ในการวิเคราะห์ข้อมูลทางสถิติที่ใช้กันอย่างแพร่หลายคือ การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์การถดถอยแบบหนึ่งที่ตัวแปรตาม (Response Variable) มีลักษณะเป็นข้อมูลเชิงคุณภาพ ส่วนตัวแปรอิสระ (Explanatory Variable) เป็นได้ทั้งข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณ โดยมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ และสามารถนำสมการถดถอยที่ได้ไปใช้ในการพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจศึกษาได้อย่างไรก็ตาม การใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติกกับข้อมูลที่มีมิติสูง มักเกิดปัญหา โมเดลสามารถทำนายผลลัพธ์จากข้อมูลที่ใช้ในการเรียนรู้ได้อย่างแม่นยำ แต่ทำนายผลลัพธ์จากข้อมูลชุดอื่นได้ไม่แม่นยำ (Overfitting) เนื่องจากโมเดลตอบสนองต่อการรบกวนมากเกินไปและมีความซับซ้อนที่ไม่จำเป็นเกิดขึ้นจากการที่ข้อมูลมีจำนวนตัวแปรอิสระที่มากกว่าจำนวนตัวอย่าง จะสังเกตได้จากการที่ตัวโมเดลมีความเอนเอียงต่ำ (Low bias) แต่มีความแปรปรวนสูง (High variance) วิธีที่จะแก้ปัญหา Overfitting คือ การทำ Regularization และการคัดกรองตัวแปร (Feature Selection) เข้าสู่ตัวแบบ ด้วยการเพิ่มฟังก์ชันพีนอลตี้ (Penalty Function) เข้าไปในสมการที่ใช้ประมาณค่าสัมประสิทธิ์การถดถอย โดยฟังก์ชันนี้จะมีหลายรูปแบบ ในงานวิจัยนี้จะศึกษา 3 รูปแบบคือ L_0-norm penalty (Best Subset Selection) , L_1-norm penalty (Lasso) และ L_0L_2-norm penalty ทั้งสามรูปแบบมีความสามารถในการคัดกรองตัวแปรเข้าสู่ตัวแบบ พร้อมกับการประมาณค่าสัมประสิทธิ์การถดถอยในคราวเดียวกัน

การคัดกรองตัวแปรอิสระเข้าสู่ตัวแบบการวิเคราะห์การถดถอยโลจิสติกในข้อมูลที่มีมิติสูง มีวัตถุประสงค์เพื่อเพิ่มความสามารถในการพยากรณ์ และลดความซับซ้อนของตัวแบบ ทั้งนี้ยังสามารถแก้ปัญหา Overfitting ซึ่งมักเกิดในกรณีการวิเคราะห์การถดถอยโลจิสติกในข้อมูลที่มีมิติสูง โดยคัดกรองตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุด

วิธีเซตย่อยที่ดีที่สุด (Best Subset Selection : L_0) มีความสามารถในการคัดเลือกตัวแปรโดยทำการเลือกตัวแบบการถดถอยที่ดีที่สุดจากกลุ่มของตัวแบบ (Beale, Kendall, & Mann, 1967; Hocking & Leslie, 1967) ที่สร้างจากตัวแปร k ตัว เมื่อ k = {1, 2, ..., p} โดยพิจารณาจากค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of determination : R^2) และผลรวมความคลาดเคลื่อนกำลังสอง (Residual sum of squares : RSS) และนำตัวแบบทั้งหมด p ตัวแบบมาพิจารณาด้วย Cross-validation error เพื่อให้ได้ตัวแบบการถดถอยที่เหมาะสมที่สุด โดยวิธีนี้จะทำการสร้างตัวแบบทั้งหมดจำนวน 2^p ตัวแบบ ต่อมาถูกพัฒนาด้วย Mixed Integer Optimization (MIO) ด้วยการเพิ่ม L_0-norm penalty เข้าไปในสมการถดถอย (Bertsimas, King, & Mazumder, 2016)

วิธีแลสโซ่ นำเสนอโดย Tibshirani (Tibshirani, 1996) โดยเป็นวิธีที่สามารถประมาณค่าสัมประสิทธิ์ $\hat{\beta}$ ของการวิเคราะห์การถดถอยในข้อมูลที่มีมิติสูง เนื่องจากวิธีแลสโซ่มีการกำหนด L_1-norm penalty เข้าไปในสมการที่ใช้วิเคราะห์สัมประสิทธิ์การถดถอย จึงทำให้ค่าของสัมประสิทธิ์บางส่วนมีค่าเป็นศูนย์ จึงเป็นการคัดกรองตัวแปร (Feature Selection) เข้าตัวแบบไปพร้อมกับการประมาณค่าของสัมประสิทธิ์ ข้อจำกัดของวิธีแลสโซ่คือสามารถเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้มากที่สุดเพียง n ตัว หากตัวแปรอิสระมีจำนวนมากกว่าจำนวนตัวอย่างเป็นจำนวนมาก ตัวแบบที่ได้อาจไม่เหมาะสม และมักพบปัญหาการเลือกตัวแปรอิสระที่ไม่เหมาะสมเมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

วิธี L_0L_2 ทำการเพิ่ม Penalty function เข้าไปในสมการการวิเคราะห์การถดถอยโลจิสติกด้วย L_0-norm penalty (Best Subset Selection) และ L_2-norm penalty (Ridge Regression) ศึกษาโดย Hazimeh และ Mazumder (Hazimeh & Mazumder, 2020) ซึ่งทำให้ตัวแบบมีคุณสมบัติคัดกรองตัวแปรจากวิธี L_0 และค่าประมาณสัมประสิทธิ์ที่ได้จะมีความเสถียร เนื่องจากวิธี L_2 สามารถแก้ไขปัญหาความสัมพันธ์ของตัวแปรอิสระที่สูงได้

ในการศึกษาครั้งนี้ มีวัตถุประสงค์ในการนำเสนอวิธีการคัดกรองตัวแปรอิสระจากการวิเคราะห์การถดถอยโลจิสติกแบบสองกลุ่ม กรณีข้อมูลที่มีมิติสูง จากวิธี L_0-norm penalty, L_1-norm penalty และ L_0L_2-norm penalty และนำมาเปรียบเทียบประสิทธิภาพโดยพิจารณาจากความผิดพลาดในการตรวจจับเชิงบวก(False Positive : FP), F1 Score และ AUC เพื่อเป็นแนวทางในการเลือกใช้วิธีสำหรับการคัดกรองตัวแปรที่มีประสิทธิภาพและเหมาะสมที่สุด

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของวิธีการคัดกรองตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยโลจิสติกที่มีลักษณะของข้อมูลที่มีมิติสูง และสัมประสิทธิ์การถดถอยบางตัวเป็นศูนย์ ด้วยวิธีการเพิ่ม Penalty Fucntion ด้วย L_0-norm penalty, L_1-norm penalty และ L_0L_2-norm penalty

1.3 ข้อตกลงเบื้องต้น

ในการศึกษาครั้งนี้สนใจศึกษา กรณีที่ตัวแปรอิสระ และตัวแปรตาม มีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) มีรูปแบบคือ

$$P(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

โดยที่ $P(y)$ คือ ความน่าจะเป็นเมื่อเกิดเหตุการณ์ที่สนใจ

y คือ เวกเตอร์ของตัวแปรตามขนาด n x 1

x คือ เมตริกซ์ตัวแปรอิสระขนาด $n \times p$

β คือ เวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

e คือ ค่าคงที่ทางคณิตศาสตร์ มีค่าประมาณ 2.71828

เมื่อ n คือ ขนาดตัวอย่าง

p คือ จำนวนตัวแปรอิสระ

1.4 ขอบเขตของการศึกษา

ในการศึกษานี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ $(n : p)$ ที่ 100 : 200, 100 : 500 และ 100 : 1000
2. ตัวแปรอิสระ (x) มีการแจกแจกปกติหลายตัวแปร โดยมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1
3. กำหนดค่าสัมประสิทธิ์การถดถอย β_j ที่ไม่เท่ากับ 0 จำนวน 20 ตัว โดยให้มีค่าเป็น 1 กับ 2 (Pungpapong, Zhang, & Zhang, 2015) ดังนี้

   $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 2$

   $\beta_{101} = \beta_{102} = \beta_{103} = \beta_{104} = \beta_{105} = \beta_{106} = \beta_{107} = \beta_{108} = \beta_{109} = \beta_{110} = 1$

   ส่วนสัมประสิทธิ์การถดถอยที่เหลือให้มีค่าเป็น 0

4. ตัวแปรตาม (y) มีค่าเป็น 0 และ 1 โดยใช้ latent variable model

$$y = \begin{cases} 0 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \leq 0 \\ 1 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon > 0 \end{cases}$$

โดยที่ $\varepsilon \sim Logistic(0, s)$

5. ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) ซึ่งเขียนในรูปสมการดังนี้

$$P(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

6. ศึกษาภายใต้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ

ระดับที่ 1 : $\rho = 0$

ระดับที่ 2 : $\rho = 0.5$

ระดับที่ 3 : $\rho = 0.9$

กล่าวคือสำหรับ $i = 1, 2, \ldots, n \quad x_i \sim N(0, \Sigma)$

เมตริกซ์ความแปรปรวนร่วม $\mathbf{\Sigma} = \begin{bmatrix} \boldsymbol{\rho_{11}} & \boldsymbol{\rho_{12}} & \cdots & \boldsymbol{\rho_{1p}} \\ \boldsymbol{\rho_{21}} & \boldsymbol{\rho_{22}} & \cdots & \boldsymbol{\rho_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\rho_{p1}} & \boldsymbol{\rho_{p2}} & \cdots & \boldsymbol{\rho_{pp}} \end{bmatrix}_{\boldsymbol{pxp}}$

โดยที่ $\rho_{ij} = \begin{cases} 1 & ; i = j \\ \rho^{|i-j|} & ; i \neq j \end{cases}$

7. กำหนดอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio (SNR)) ที่แตกต่างกันเพื่อให้เห็นถึงข้อแตกต่างของประสิทธิภาพในการพยากรณ์ของตัวแบบ (Hastie, Tibshirani, & Tibshirani, 2017; Mazumder, Radchenko, & Dedieu, 2017) โดยกำหนดให้มีค่าเป็น 1 และ 6

โดยที่ $SNR = \frac{Var(x\beta)}{Var(\varepsilon)}$

จาก $\varepsilon \sim Logistic(0, s)$ และ $Var(\varepsilon)$ คือ $\frac{s^2\pi^2}{3}$

8. ทำการจำลองข้อมูลในแต่ละกรณีจำนวน 100 รอบ

1.5 คำจำกัดความที่ใช้ในงานวิจัย

ข้อมูลที่มีมิติสูง (High-Dimensional Data)

คือ ข้อมูลที่มีจำนวนตัวแปรอิสระ (p) มากกว่าขนาดตัวอย่าง (n)

ความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP)

คือ ความผิดพลาดที่เกิดจากการทำนายว่าจะเกิดเหตุการณ์ที่สนใจในขณะที่ข้อมูลจริงไม่เกิดเหตุการณ์ที่ สนใจ

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

1. ความผิดพลาดในการตรวจจับเชิงบวก (False Positive: FP)

การวัดจำนวนที่เกิดความผิดพลาดจากการประมาณค่าสัมประสิทธิ์การถดถอยที่ได้จากการคัดเลือกตัวแปรแบบต่างๆ โดยค่าสัมประสิทธิ์ที่ประมาณได้นั้นมีค่าไม่เท่ากับศูนย์ ในขณะที่ค่าสัมประสิทธิ์ที่แท้จริงมีค่าเท่ากับศูนย์ ซึ่งสามารถคำนวณได้ดังนี้

$$FP = \sum_{j=1}^{p} 1_{\{\beta_j = 0 \ and \ \hat{\beta}_j \neq 0\}}$$

2. ความผิดพลาดในการตรวจจับเชิงลบ (False Negative: FN)

การวัดจำนวนที่เกิดความผิดพลาดจากการประมาณค่าสัมประสิทธิ์การถดถอยที่ได้จากการคัดเลือกตัวแปรแบบต่างๆ โดยค่าสัมประสิทธิ์ที่ประมาณได้นั้นมีค่าเท่ากับศูนย์ ในขณะที่ค่าสัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์ ซึ่งสามารถคำนวณได้ดังนี้

$$FP = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0 \ and \ \hat{\beta}_j = 0\}}$$

2. ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F_1 Score) เป็นค่าที่ใช้วัดความสามารถของโมเดล จากการเอาค่าความแม่นยำ (Precision) และ ค่าความไว (Sensitivity) มาคำนวณรวมกัน

$$F_1\ Score\ =\ 2\ x\ \frac{Precision\ x\ Sensitivity}{Precision\ +\ Sensitivity}$$

เมื่อ กำหนด Confusion Matrix จากผลของการประมาณค่าสัมประสิทธิ์การถดถอย β_j ดังนี้

ตารางที่ 1.6. 1 ตารางแสดง Confusion Matrix กรณีของ F_1-score

		ค่าของสัมประสิทธิ์ที่แท้จริง (β_j)	
		มีค่าไม่เท่ากับศูนย์	มีค่าเท่ากับศูนย์
ค่าของสัมประสิทธิ์การถดถอยที่ตัวแบบประมาณได้ ($\hat{\beta}_j$)	มีค่าไม่เท่ากับศูนย์	True Positive (TP) ($\beta_j \neq 0\ and\ \hat{\beta}_j \neq 0$)	False Positive (FP) ($\beta_j = 0\ and\ \hat{\beta}_j \neq 0$)
	มีค่าเท่ากับศูนย์	False Negative (FN) ($\beta_j \neq 0\ and\ \hat{\beta}_j = 0$)	True Negative (TN) ($\beta_j = 0\ and\ \hat{\beta}_j = 0$)

True Positive ($\beta_j \neq 0\ and\ \hat{\beta}_j \neq 0$) คือ จำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าไม่เท่ากับศูนย์ และค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์

False Positive ($\beta_j = 0\ and\ \hat{\beta}_j \neq 0$) คือ จำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าไม่เท่ากับศูนย์ และ ค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์

False Negative ($\beta_j \neq 0\ and\ \hat{\beta}_j = 0$) คือ จำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีเค่าเท่ากับศูนย์ และ ค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์

True Negative ($\beta_j = 0\ and\ \hat{\beta}_j = 0$) คือ จำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าเท่ากับศูนย์ และ ค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์

ค่าความแม่นยำ (Precision) คือ อัตราส่วนของจำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์ที่ประมาณได้มีค่าไม่เท่ากับศูนย์ และค่าสัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์ ต่อจำนวนเหตุการณ์ทั้งหมดที่ค่าสัมประสิทธิ์การถดถอยที่ประมาณค่าได้มีค่าไม่เท่ากับศูนย์

$$Precision = \frac{TP}{TP + FP}$$

ค่าความไว (Sensitivity) คือ อัตราส่วนของจำนวนเหตุการณ์ที่ค่าสัมประสิทธิ์ที่ประมาณได้มีค่าไม่เท่ากับศูนย์ และค่าสัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์ ต่อจำนวนเหตุการณ์ทั้งหมดที่ค่าสัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์

$$Sensitivity = \frac{TP}{TP + FN}$$

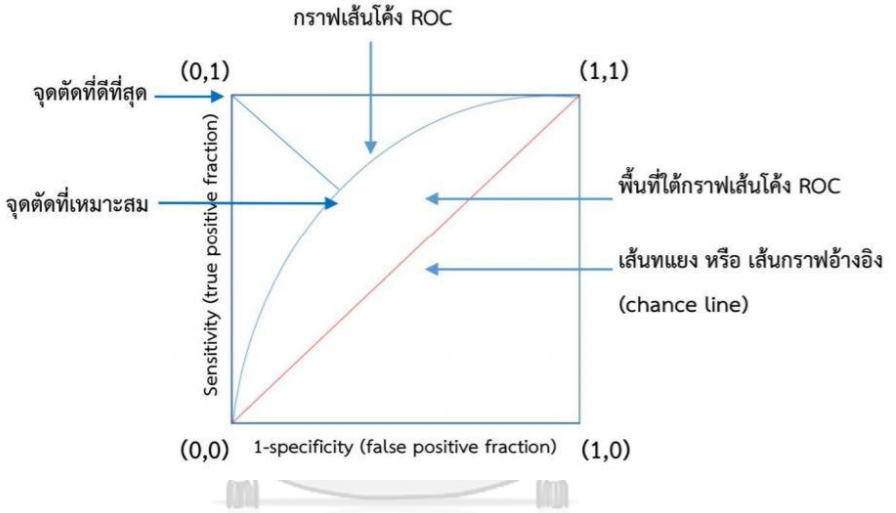
3. พื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ หาได้จากพื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) ที่สร้างจาก กราฟของ Sensitivity และ 1-Specificity ของ Confusion Matrix โดยจะแบ่งเป็นสองเกณฑ์ คือ

3.1 การคัดเลือกตัวแปร โดยทำการเปรียบเทียบค่า AUC ของการประมาณค่าสัมประสิทธิ์ β_j ที่ได้จากสมการถดถอย และค่าสัมประสิทธิ์ที่แท้จริงของข้อมูล ในตัวแบบที่ใช้พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน เพื่อวัดความถูกต้องในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ กราฟเส้น ROC ที่ได้จะมาจากค่าของ Sensitivity และ 1-Specificity ด้วยการกำหนดพารามิเตอร์ Lambda ที่ต่างกันซึ่งหมายถึง จำนวนของตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบที่แตกต่างกัน หากพารามิเตอร์ Lambda (λ) มีค่ามาก จะส่งผลให้จำนวนตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบมีจำนวนน้อย ในทางกลับกัน หากพารามิเตอร์ Lambda (λ) มีค่าน้อย จะส่งผลให้จำนวนตัวแปรที่ถูก

คัดเลือกเข้าสู่ตัวแบบมีจำนวนมาก และนำ AUC ที่ได้ในแต่ละชุดข้อมูลมาเปรียบเทียบกันด้วย Box plot ให้เห็นถึงลักษณะของข้อมูล

3.2 การคัดเลือกตัวแบบ โดยทำการเปรียบเทียบค่า AUC ของความสามารถในการพยากรณ์ระหว่างตัวแบบ จากตัวแบบวิเคราะห์การถดถอยโลจิสติก ด้วยการกำหนดจุดแบ่งแยกค่าของตัวแปรตาม (Cut-off) ที่ต่างกันเพื่อใช้จำแนกเหตุการณ์ที่สนใจศึกษากับเหตุการณ์ที่ไม่สนใจศึกษา หากเส้นโค้ง ROC เข้าใกล้จุดซ้ายบน จะทำให้มีพื้นที่ใต้เส้นโค้งมากขึ้น และจะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยหากค่ายิ่งเข้าใกล้ 1 ตัวแบบจะยิ่งมีความน่าเชื่อถือมากกว่า

ภาพที่ 1.6. 1 แสดงตัวอย่างกราฟ ROC



(ที่มา : https://li01.tci-thaijo.org/index.php/TBPS/article/download/248029/171563/)

ซึ่งค่าของ Sensitivity และ Specificity หาได้จาก Confusion Matrix ของการนำตัวแบบมาทำนายชุดข้อมูล เพื่อให้ได้ผลลัพธ์ของตัวแปรตาม

ตารางที่ 1.6. 2 ตารางแสดง Confusion Matrix กรณีของ AUC

		ค่าของข้อมูล (Actual)	
		Positive (1)	Negative (0)
ค่าพยากรณ์ (Predicted)	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

True Positive คือ จำนวนเหตุการณ์ที่ตัวแบบทำนายค่าของตัวแปรตามเป็น 1 และ ค่าของตัวแปรตามที่แท้จริงเป็น 1

False Positive คือ จำนวนเหตุการณ์ที่ตัวแบบทำนายค่าของตัวแปรตามเป็น 1 และ ค่าของตัวแปรตามที่แท้จริงเป็น 0

False Negative คือ จำนวนเหตุการณ์ที่ตัวแบบทำนายค่าของตัวแปรตามเป็น 0 และ ค่าของตัวแปรตามที่แท้จริงเป็น 1

True Negative คือ จำนวนเหตุการณ์ที่ตัวแบบทำนายค่าของตัวแปรตามเป็น 0 และ ค่าของตัวแปรตามที่แท้จริงเป็น 0

ค่าความจำเพราะ (Specificity) อัตราส่วนของจำนวนเหตุการณ์ที่ตัวแบบทำนายว่าไม่เกิดเหตุการณ์ที่สนใจได้ถูกต้อง กับ จำนวนเหตุการณ์ที่ไม่สนใจทั้งหมดที่เกิดขึ้นในข้อมูลจริง

$$Specificity = \frac{TN}{TN + FP}$$

ค่าความไว (Sensitivity) คือ อัตราส่วนของจำนวนเหตุการณ์ที่ตัวแบบทำนายว่าเกิดเหตุการณ์ที่สนใจได้ถูกต้อง กับ จำนวนเหตุการณ์ที่สนใจทั้งหมดที่เกิดขึ้นในข้อมูลจริง

$$Sensitivity = \frac{TP}{TP + FN}$$

1.7 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้วิธีการคัดกรองตัวแปรและประมาณค่าสัมประสิทธิ์ ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม และเพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบในแต่ละวิธี

บทที่ 2 ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

2.1 ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การวิจัยครั้งนี้ได้ศึกษาเกี่ยวกับการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) ด้วยการประมาณค่าสัมประสิทธิ์การถดถอย (β) โดย การวิเคราะห์การถดถอยโลจิสติกจะใช้วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) ในการประมาณค่าสัมประสิทธิ์ที่จะทำให้เกิดความน่าจะเป็นสูงสุดเพื่อที่จะสามารถทำนายค่าของตัวแปรตามได้ใกล้เคียงกับกับข้อมูลจริงมากที่สุด แต่หากข้อมูลเป็นข้อมูลที่มีมิติสูง คือมีจำนวนของตัวแปรอิสระมากกว่าจำนวนตัวอย่าง จะส่งผลให้ตัวแบบที่ได้จากการวิเคราะห์การถดถอยนั้นตอบสนองต่อการรบกวนมากเกินไป (Overfitting) เนื่องจากตัวแปรอิสระที่นำเข้าสู่ตัวแบบมีจำนวนเยอะรวมถึงค่าสัมประสิทธิ์ที่ประมาณได้ของตัวแปรอิสระหลายตัวที่นำเข้าสู่ตัวแบบนั้นมีค่าสูง ทำให้การเปลี่ยนแปลงค่าของตัวแปรอิสระส่งผลกระทบต่อการทำนายค่าตัวแปรตามของตัวแบบมากเกินไป ดังนั้นการศึกษาในครั้งนี้จะกล่าวถึงวิธีแก้ปัญหาดังกล่าว ด้วยการเพิ่ม Penalty Function 3 รูปแบบ คือ L_0-norm penalty, L_1-norm penalty และ L_0L_2-norm penalty เข้าไปในสมการการประมาณค่าสัมประสิทธิ์ เพื่อหลีกเลี่ยงการประมาณค่าสัมประสิทธิ์ที่มีค่าสูงเกินไป และเป็นการคัดกรองตัวแปรเข้าสู่ตัวแบบ

2.1.1 การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis)

การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม เป็นการวิเคราะห์เพื่อศึกษาว่าตัวแปรอิสระใดบ้างที่สามารถอธิบายโอกาสการเกิดเหตุการณ์หรือไม่เกิดเหตุการณ์ที่สนใจตามตัวแปรตามได้ และเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจจากสมการการถดถอยโลจิสติกที่เหมาะสม โดยตัวแปรตาม (y) มีค่าได้สองค่าคือ 1 เมื่อเกิดเหตุการณ์ที่สนใจศึกษา และ 0 เมื่อไม่เกิดเหตุการณ์ที่สนใจศึกษา (Agresti, 2003) ดังนั้น ตัวแปรตามมีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution)

โดย $$y = \begin{cases} 1 & \text{เมื่อเกิดเหตุการณ์ที่สนใจด้วยความน่าจะเป็น } P(y) \\ 0 & \text{ไม่เกิดเหตุการณ์ที่สนใจด้วยความน่าจะเป็น } 1 - P(y) \end{cases}$$

และตัวแปรอิสระเป็นได้ทั้งข้อมูลเชิงปริมาณและเชิงคุณภาพ โดยความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เขียนแทนด้วยสมการดังนี้

$$P(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}} \text{ หรือ } \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}}$$

โดยที่	$P(y)$	คือ ความน่าจะเป็นเมื่อเกิดเหตุการณ์ที่สนใจ
	$1 - P(y)$	คือ ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ที่สนใจ
	y	คือ เวกเตอร์ของตัวแปรตามขนาด n x 1
	x	คือ เมตริกซ์ตัวแปรอิสระขนาด $n \times p$
	β	คือ เวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1
	e	คือ ค่าคงที่ทางคณิตศาสตร์ มีค่าประมาณ 2.71828
เมื่อ	n	คือ ขนาดตัวอย่าง
	p	คือ จำนวนตัวแปรอิสระ

จากความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามของการวิเคราะห์การถดถอยโลจิสติกไม่อยู่ในรูปเชิงเส้น จึงต้องมีการปรับให้ความสัมพันธ์อยู่ในรูปเชิงเส้น ด้วยการเขียนในรูปแบบของ Odds หรือ Odd ratio

Odds หมายถึง อัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจ $P(y)$ กับโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ $1 - P(y)$ จะได้

$$Odds = \frac{P(y)}{1 - P(y)}$$

ค่าของ Odds แสดงถึงโอกาสที่จะเกิดเหตุการณ์ที่สนใจ เป็นกี่เท่าของโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ นั่นคือ เมื่อ Odds มีค่ามากกว่า 1 หมายถึง โอกาสที่จะเกิดเหตุการณ์ที่สนใจนั้นมากกว่าโอกาสที่จะไม่เกิดเหตุการณ์ที่สนใจ

การเขียนโมเดลโลจิสติกจะอยู่ในรูป log ของ Odds เรียกว่า logit

$$logit \ = \ log(odds) \ = \ log\left(\frac{P(y)}{1-P(y)}\right) \ = \ \beta_0 \ + \ \beta_1 x_1 + \ldots + \beta_p x_p$$

ซึ่งวิธีการถดถอยโลจิสติกมีวิธีการประมาณค่าสัมประสิทธิ์ (Parameter Estimation) ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) จากสมการ ดังนี้

$$L(\beta) \ = \ \prod_{i=1}^{n}\left(\frac{1}{1+e^{-(\beta_0 \ + \ \beta_1 x_1 \ + \ldots + \ \beta_p x_p)}}\right)^{y_i}\left(\frac{e^{-(\beta_0 \ + \ \beta_1 x_1 \ + \ldots + \ \beta_p x_p)}}{1+e^{-(\beta_0 \ + \ \beta_1 x_1 \ + \ldots + \ \beta_p x)}}\right)^{1-y_i}$$

จะได้ log-likelihood function คือ

$$l(\beta) \ = \ \sum_{i=1}^{n}\left(y_i(\beta_0 \ + \ \beta_1 x_1 + \ldots + \beta_p x_p) - log(1 + e^{\beta_0 \ + \ \beta_1 x_1 \ + \ldots + \ \beta_p x_p})\right)$$

2.1.2 วิธี L_0-norm penalty (Best Subset Selection)

วิธี L_0-norm penalty (Best Subset Selection) มี Penalty function ดังนี้ $\sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$ คือ จำนวนของสัมประสิทธิ์ β ที่ไม่เป็นศูนย์ (Miller, 2002) จะได้ log-likelihood function คือ

$$l(\beta) = \sum_{i=1}^{n} \left(y_i(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) - log(1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}) \right) - \lambda \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$$

2.1.3 วิธีแลสโซ่ (LASSO)

วิธีแลสโซ่ นำเสนอโดย Tibshirani (1996) โดยเป็นวิธีที่สามารถคัดกรองตัวแปรเข้าสู่ตัวแบบพร้อมกับการประมาณค่าสัมประสิทธิ์ ด้วยการเพิ่ม Penalty Function เข้าไปในสมการถดถอย ซึ่งวิธีแลสโซ่ใช้ L_1-norm penalty ดังนี้ $\sum_{j=1}^{p} |\beta_j|$ คือ ผลรวมของค่าสัมบูรณ์ของสัมประสิทธิ์ β จะได้ log-likelihood function (Hastie, Tibshirani, & Friedman, 2009) คือ

$$l(\beta) = \sum_{i=1}^{n} \left(y_i(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) - log(1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}) \right) - \lambda \sum_{j=1}^{p} |\beta_j|$$

และมีพารามิเตอร์ แลมบ์ดา (λ) ซึ่งจะมีค่ามากกว่าหรือเท่ากับศูนย์ ใช้ในการกำหนดน้ำหนักของ Penalty function โดยค่าของแลมบ์ดาที่เหมาะสมจะหาได้จากวิธี Cross-validation หากแลมบ์ดามีค่ามาก จะส่งผลให้ค่าสัมประสิทธิ์ส่วนใหญ่เป็นศูนย์ และบางส่วนไม่เท่ากับศูนย์ (Gareth, Daniela, Trevor, & Robert, 2013) ดังนั้นวิธีแลสโซ่จึงเป็นวิธีที่สามารถประมาณค่าสัมประสิทธิ์พร้อมกับการคัดกรองตัวแปรได้โดยอัตโนมัติ

อย่างไรก็ตาม วิธีแลสโซ่มีข้อจำกัดคือ สามารถเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้มากที่สุดเพียง n ตัว นั่นคือ หากข้อมูลมีจำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่างเป็นจำนวนมาก ตัวแบบที่ได้อาจไม่เหมาะสม และการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของวิธีแลสโซ่มีแนวโน้มที่จะเลือกตัวแปรอิสระเพียงตัวเดียวจากกลุ่มของตัวแปรอิสระที่มีความสัมพันธ์กันสูงโดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม ดังนั้นวิธีแลสโซ่จะมีประสิทธิภาพสูงเมื่อใช้กับข้อมูลที่ ตัวแปรอิสระมีความสัมพันธ์กันไม่สูงมากนัก (วิฐรา พึ่งพาพงศ์, 2558)

2.1.4 วิธี L_0L_2-norm penalty

วิธี L_0L_2-norm penalty เป็นการนำ Penalty function ของทั้ง L_0-norm penalty function ดังนี้ $\sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$ คือ จำนวนของสัมประสิทธิ์ β ที่ไม่เป็นศูนย์ (Miller, 2002) และ L_2-norm penalty function (Ridge Regression) ดังนี้ $\sum_{j=1}^{p} \beta_j^2$ คือ ผลรวมของสัมประสิทธิ์ β^2 (Duffy & Santner, 1989; Le Cessie & Van Houwelingen, 1992) จะได้ log-likelihood function

$$l(\beta) = \sum_{i=1}^{n} \left(y_i(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) - log(1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}) \right) - \left(\lambda \sum_{j=1}^{p} 1_{\beta_j \neq 0} + \gamma \sum_{j=1}^{p} \beta_j^2 \right)$$

ทุกวิธีการในการวิเคราะห์การถดถอยโลจิสติกทั้งในวิธีการเพิ่ม L_0-norm penalty, L_1-norm penalty และ L_0L_2-norm penalty เข้าไปในสมการถดถอย จะต้องมีการเลือก พารามิเตอร์ปรับ (Tuning Parameter) (λ) ซึ่งในงานวิจัยนี้ใช้วิธีการ 5 Fold Cross-Validation ในการปรับค่าพารามิเตอร์ด้วยการหาพารามิเตอร์ที่ทำให้ค่าของ ค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error : MSE) ของตัวแบบน้อยที่สุด เพื่อเป็นการเลือกตัวแบบที่ดีที่สุดก่อนจะนำไปเปรียบเทียบเพื่อหาความสามารถตามเกณฑ์การตัดสินใจ

บทที่ 3 วิธีการดำเนินการศึกษา

3.1 วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ และเปรียบเทียบวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) ด้วยการประมาณค่าสัมประสิทธิ์การถดถอย (β) จากวิธีการใช้ Penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection), L_1-norm penalty (LASSO) และ L_0L_2-norm penalty ซึ่งทำการศึกษากับข้อมูลที่จำลองขึ้นมาและข้อมูลมีลักษณะเป็นข้อมูลที่มีมิติสูง โดยพิจารณาเปรียบเทียบประสิทธิภาพของแต่ละวิธีจาก ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score) และ พื้นที่ใต้เส้นโค้ง ROC (AUC) โดยทำการวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชัน 3.6.1 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.2 ขอบเขตของการศึกษา

ในการศึกษานี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ($n : p$) ที่ 100 : 200, 100 : 500 และ 100 : 1000
2. ตัวแปรอิสระ (x) มีการแจกแจกปกติหลายตัวแปร โดยมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1
3. กำหนดค่าสัมประสิทธิ์การถดถอย β_j ที่ไม่เท่ากับ 0 จำนวน 20 ตัว โดยให้มีค่าเป็น 1 กับ 2 (Pungpapong et al., 2015) ดังนี้

   $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 2$

   $\beta_{101} = \beta_{102} = \beta_{103} = \beta_{104} = \beta_{105} = \beta_{106} = \beta_{107} = \beta_{108} = \beta_{109} = \beta_{110} = 1$

   ส่วนสัมประสิทธิ์การถดถอยที่เหลือให้มีค่าเป็น 0

4. ตัวแปรตาม (y) มีค่าเป็น 0 และ 1 โดยใช้ latent variable model

$$y = \begin{cases} 0 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \leq 0 \\ 1 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon > 0 \end{cases}$$

โดยที่ $\varepsilon \sim Logistic(0, s)$

5. ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) ซึ่งเขียนในรูปสมการดังนี้

$$P(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

6. ศึกษาภายใต้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ

ระดับที่ 1 : $\rho = 0$

ระดับที่ 2 : $\rho = 0.5$

ระดับที่ 3 : $\rho = 0.9$

กล่าวคือสำหรับ $i = 1, 2, \ldots, n \quad x_i \sim N(0, \Sigma)$

เมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\rho_{11}} & \boldsymbol{\rho_{12}} & \cdots & \boldsymbol{\rho_{1p}} \\ \boldsymbol{\rho_{21}} & \boldsymbol{\rho_{22}} & \cdots & \boldsymbol{\rho_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\rho_{p1}} & \boldsymbol{\rho_{p2}} & \cdots & \boldsymbol{\rho_{pp}} \end{bmatrix}_{\boldsymbol{pxp}}$

โดยที่ $\rho_{ij} = \begin{cases} 1 & ; i = j \\ \rho^{|i-j|} & ; i \neq j \end{cases}$

7. กำหนดอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio (SNR)) ที่แตกต่างกันเพื่อให้เห็นถึงข้อแตกต่างของประสิทธิภาพในการพยากรณ์ของตัว

แบบ (Hastie et al., 2017; Mazumder et al., 2017) โดยกำหนดให้มีค่าเป็น 1 และ 6

$$\text{โดยที่} \quad SNR = \frac{Var(x\beta)}{Var(\varepsilon)}$$

จาก $\varepsilon \sim Logistic(0, s)$ และ $Var(\varepsilon)$ คือ $\frac{s^2\pi^2}{3}$

8. ทำการจำลองข้อมูลในแต่ละกรณีจำนวน 100 รอบ

3.3 ขั้นตอนการดำเนินการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง

2. กำหนดและจำลองข้อมูล

2.1 กำหนดค่าเริ่มต้นโดยการสร้างข้อมูลที่มีจำนวนของตัวอย่าง n และจำนวนตัวแปรอิสระ p ตัว โดยใช้อัตราส่วน $n : p$ ดังนี้ 100:200, 100:500, 100:1000

2.2 จำลองตัวแปรอิสระ (x) ให้มีการแจกแจงปกติ $x_i \sim N(0, \Sigma)$

2.3 กำหนดค่าสัมประสิทธิ์การถดถอย β_j ที่ไม่เท่ากับ 0 จำนวน 20 ตัว โดยให้มีค่าเป็น 1 กับ 2 (Pungpapong et al., 2015) ดังนี้

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 2$$
$$\beta_{101} = \beta_{102} = \beta_{103} = \beta_{104} = \beta_{105} = \beta_{106} = \beta_{107} = \beta_{108} = \beta_{109} = \beta_{110} = 1$$

ส่วนสัมประสิทธิ์การถดถอยที่เหลือให้มีค่าเป็น 0

2.4 กำหนดอัตราสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio (SNR)) = 1 และ 6

$$\text{โดยที่} \quad SNR = \frac{Var(x\beta)}{Var(\varepsilon)}$$

จาก $\varepsilon \sim Logistic(0, s)$ และ $Var(\varepsilon)$ คือ $\frac{s^2\pi^2}{3}$

2.5 จำลองตัวแปรตาม (y) ให้มีค่าเป็น 0 และ 1 โดยใช้ latent variable model

$$y = \begin{cases} 0 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \leq 0 \\ 1 \text{ เมื่อ } \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon > 0 \end{cases}$$

โดยที่ $\varepsilon \sim Logistic(0, s)$

2.6 กำหนดความสัมพันธ์ระหว่างตัวแปรอิสระ 3 ระดับ คือ

ระดับที่ 1 : $\rho = 0$

ระดับที่ 2 : $\rho = 0.5$

ระดับที่ 3 : $\rho = 0.9$

กล่าวคือสำหรับ $i = 1, 2, \ldots, n \quad x_i \sim N(0, \Sigma)$

เมตริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\rho_{11}} & \boldsymbol{\rho_{12}} & \cdots & \boldsymbol{\rho_{1p}} \\ \boldsymbol{\rho_{21}} & \boldsymbol{\rho_{22}} & \cdots & \boldsymbol{\rho_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\rho_{p1}} & \boldsymbol{\rho_{p2}} & \cdots & \boldsymbol{\rho_{pp}} \end{bmatrix}_{\boldsymbol{pxp}}$

โดยที่ $\rho_{ij} = \begin{cases} 1 & ; i = j \\ \rho^{|i-j|} & ; i \neq j \end{cases}$

2.7 จำลองข้อมูลภายใต้ความสัมพันธ์ตามตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม

$$P(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

โดยที่	$P(y)$	คือ ความน่าจะเป็นเมื่อเกิดเหตุการณ์ที่สนใจ
	$1-P(y)$	คือ ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ที่สนใจ
	y	คือ เวกเตอร์ของตัวแปรตามขนาด n x 1
	x	คือ เมตริกซ์ตัวแปรอิสระขนาด n x p
	β	คือ เวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1
	e	คือ ค่าคงที่ทางคณิตศาสตร์มีค่าประมาณ 2.71828
เมื่อ	n	คือ ขนาดตัวอย่าง
	p	คือ จำนวนตัวแปรอิสระ

3. นำข้อมูลที่ได้จากการจำลองมาประมาณค่าสัมประสิทธิ์การถดถอย $(\hat{\beta})$ ด้วยวิธีการเพิ่ม Penalty Function เข้าไปในสมการถดถอยโลจิสติกด้วยวิธีดังนี้

3.1 วิธี L_0-norm penalty (Best Subset Selection)

3.2 วิธี L_1-norm penalty (Lasso)

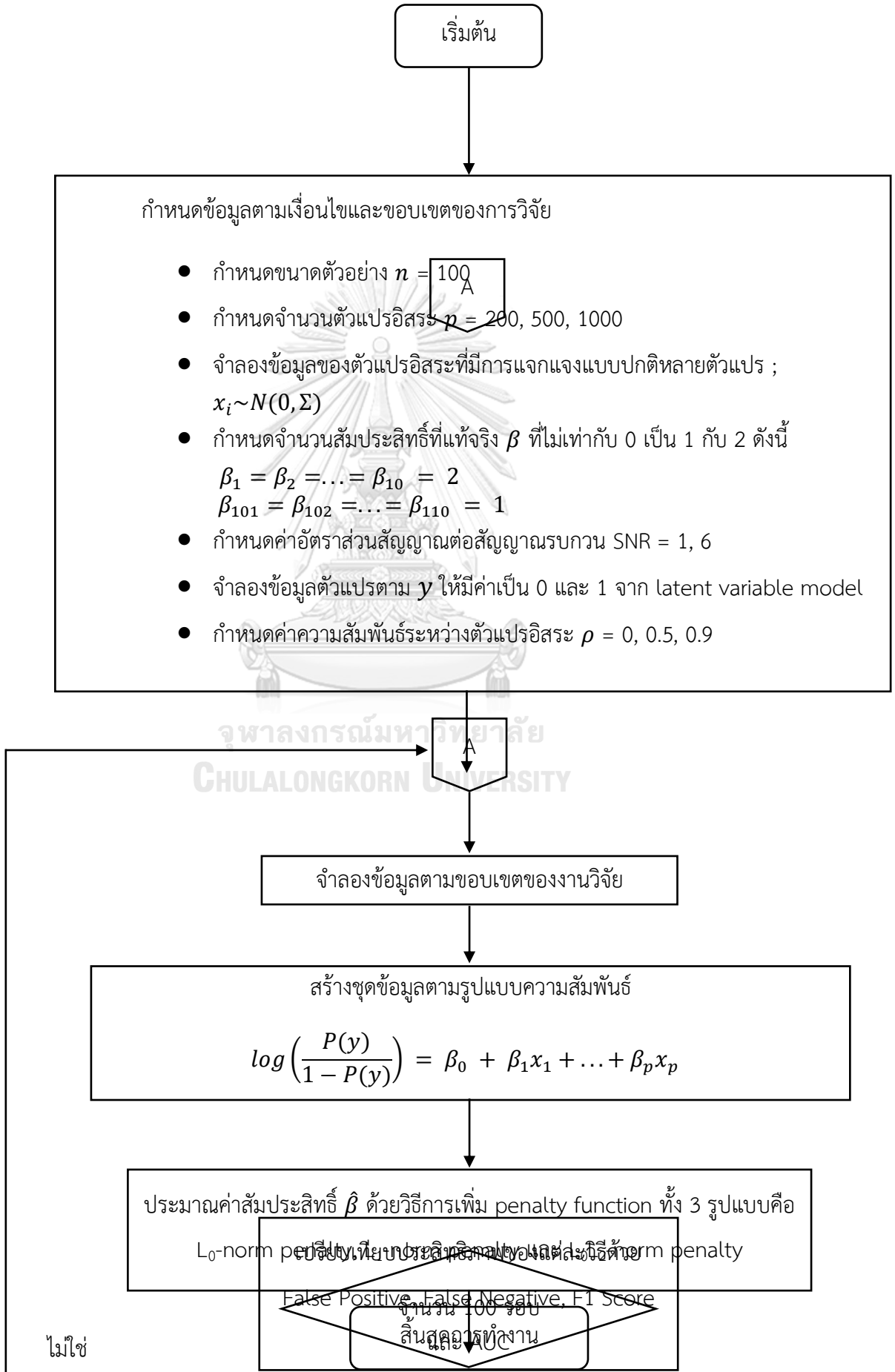
3.3 วิธี L_0L_2-norm penalty

4. นำข้อมูลที่ได้จากข้อ 3 มาหาความผิดพลาดในการตรวจจับเชิงบวก (False Positive: FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative: FN) ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score) และ พื้นที่ใต้โค้ง ROC (AUC) ทั้งสองเกณฑ์

5. เปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี

6. สรุปผลที่ได้จากการศึกษา

3.4 กรอบแนวคิดการวิจัย



บทที่ 4 ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการในการคัดกรองตัวแปร และประสิทธิภาพในการพยากรณ์ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม จากวิธีการใช้ Penalty function 3 รูปแบบได้แก่ L_0-norm penalty (Best Subset Selection), L_1-norm penalty (LASSO) และ L_0L_2-norm penalty โดยทำการศึกษากับข้อมูลที่จำลองขึ้นมาและข้อมูลมีลักษณะเป็นข้อมูลที่มีมิติสูงและพิจารณาแยกตามขนาดตัวแปรอิสระ p = 200, 500 และ 1000 รวมถึงค่าความสัมพันธ์ภายในตัวแปรอิสระและค่าอัตราสัญญาณต่อสัญญาณรบกวน (SNR) โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพในการพยากรณ์และค่าความถูกต้องในการคัดกรองตัวแปร ของแต่ละวิธีจาก ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ค่าความผิดพลาดนการตรวจจับเชิงลบ (False Negative : FN) ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score) พื้นที่

ใต้เส้นโค้ง ROC (AUC) ซึ่งใช้บ่งชี้ความสามารถในการพยากรณ์ของตัวแบบ และ Variable Selection AUC ที่ใช้วัดความถูกต้องในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ

โดยผลการวิจัยแบ่งออกเป็น 5 ส่วน ดังนี้

ส่วนที่ 1 ผลการเปรียบเทียบค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับ 1 และ 6 จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

ส่วนที่ 2 ผลการเปรียบเทียบค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับ 1 และ 6 จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

ส่วนที่ 3 ผลการเปรียบเทียบค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) ด้วยการเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้กับค่าสัมประสิทธิ์ที่แท้จริง ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับ 1 และ 6 จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

ส่วนที่ 4 ผลการเปรียบเทียบค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm

penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับ 1 และ 6 จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

ส่วนที่ 5 ผลการเปรียบเทียบค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับ 1 และ 6 จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

4.1 ผลการเปรียบเทียบค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวกเพื่อวัดความถูกต้องในการคัดเลือกตัวแปร ของวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ระหว่างวิธี L_0, วิธี L_1 และวิธี L_0L_2 ภายใต้ปัจจัยดังต่อไปนี้

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 100 : 200, 100 : 500 และ 100 : 1000

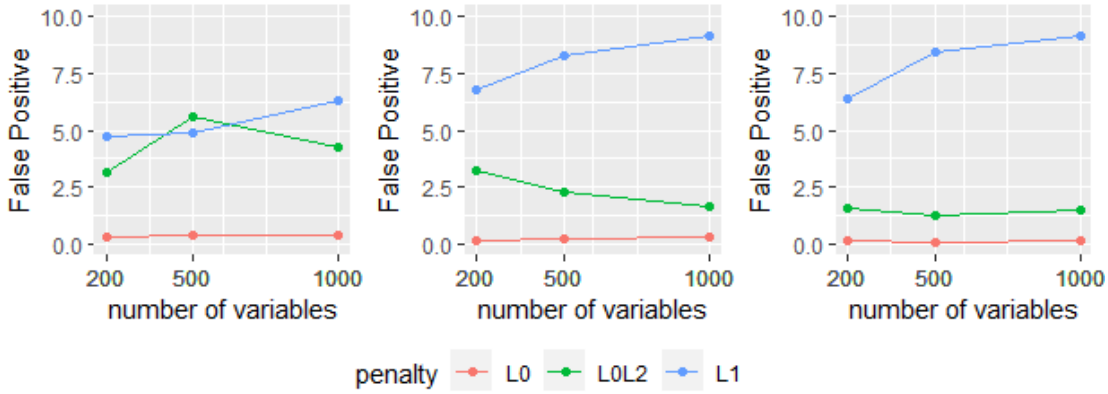
2. ศึกษาภายใต้ความสัมพันธ์ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0, 0.5, 0.9$

3. ศึกษาภายใต้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 1 และ 6

ค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive :FP) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 1

ภาพที่ 4.1. 1 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก กรณี SNR = 1

$\rho = 0$ $\rho = 0.5$ $\rho = 0.9$



ภาพที่ 4.1.1 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย โดยค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าไม่เท่ากับศูนย์ในขณะที่ สัมประสิทธิ์ที่แท้จริงมีค่าเท่ากับศูนย์ ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหนึ่ง จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.1.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุด และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากที่สุด และ วิธี L_0L_2 รองลงมา เมื่อจำนวนของตัวแปรอิสระมีจำนวน 200 ตัวแปร ในขณะที่เมื่อจำนวนของตัวแปรอิสระมีจำนวน 500 ตัวแปร ค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกของวิธี L_0L_2 มีค่ามากกว่าวิธี L_1 และเมื่อจำนวนของตัวแปรอิสระมีจำนวน 1000 ตัวแปร วิธี L_1 มีค่ามากที่สุด โดยจะเห็นได้ว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกเพิ่มขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

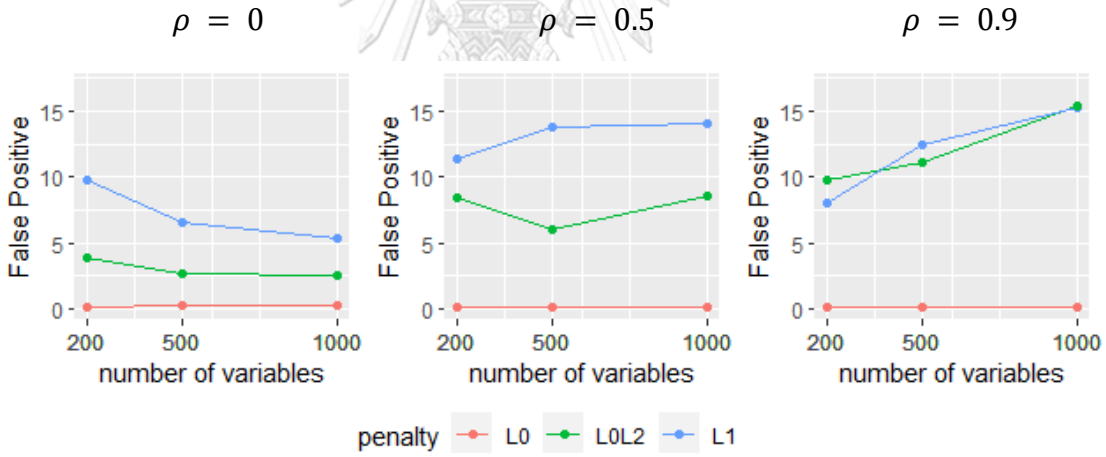
จากภาพที่ 4.1.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าเพิ่มขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ในขณะที่ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการ

ตรวจจับเชิงบวกมากกว่าวิธี L_0 และน้อยกว่าวิธี L_1 ในทุกกรณีของจำนวนของตัวแปรอิสระ แต่มีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.1.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าเพิ่มขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ในขณะที่ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากกว่าวิธี L_0 และน้อยกว่าวิธี L_1 ในทุกกรณีของจำนวนของตัวแปรอิสระ แต่มีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

ค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive :FP) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 6

ภาพที่ 4.1. 2 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก กรณี SNR = 6



ภาพที่ 4.1.2 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย โดยค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าไม่เท่ากับศูนย์ในขณะที่ สัมประสิทธิ์ที่แท้จริงมีค่าเท่ากับศูนย์ ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหก

จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.1.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ในขณะที่ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากกว่าวิธี L_0 และน้อยกว่าวิธี L_1 ในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.1.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าเพิ่มขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ในขณะที่ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากกว่าวิธี L_0 และน้อยกว่าวิธี L_1 ในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าใกล้เคียงกันเมื่อจำนวนของตัวแปรอิสระมีจำนวน 200 และ 1000 ตัวแปร แต่มีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวน 500 ตัวแปร

จากภาพที่ 4.1.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้น ในกรณีที่จำนวนของตัวแปรอิสระมีจำนวน 200 ตัวแปร วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกมากกว่าวิธี L_1 แต่เมื่อจำนวนของตัวแปรอิสระมีจำนวน 500 ตัวแปร วิธี L_0L_2 มีค่าเฉลี่ยที่น้อยกว่าวิธี L_1 และเมื่อจำนวนของตัวแปรอิสระมีจำนวน 1000 ตัวแปร ทั้งสองวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกใกล้เคียงกัน โดยทั้งสองวิธี มีค่าเฉลี่ยที่เพิ่มขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

4.2 ผลการเปรียบเทียบค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงลบเพื่อวัดความถูกต้องในการคัดเลือกตัวแปร ของวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ระหว่างวิธี L_0, วิธี L_1 และวิธี L_0L_2 ภายใต้ปัจจัยดังต่อไปนี้

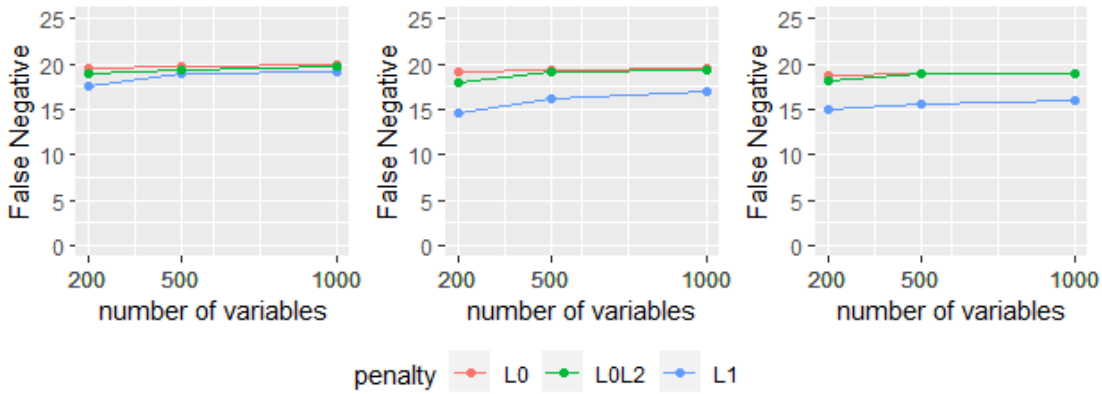
1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 100 : 200, 100 : 500 และ 100 : 1000

2. ศึกษาภายใต้ความสัมพันธ์ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0, 0.5, 0.9$

3. ศึกษาภายใต้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 1 และ 6

ค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 1

ภาพที่ 4.2. 1 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ กรณี SNR = 1



ภาพที่ 4.2.1 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย โดยค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าเท่ากับศูนย์ในขณะที่ สัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์ ระหว่าง

วิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหนึ่ง จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.2.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบน้อยที่สุด และ วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากที่สุด ในทุกกรณีของจำนวนของตัวแปรอิสระ และทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบสูงขึ้นเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

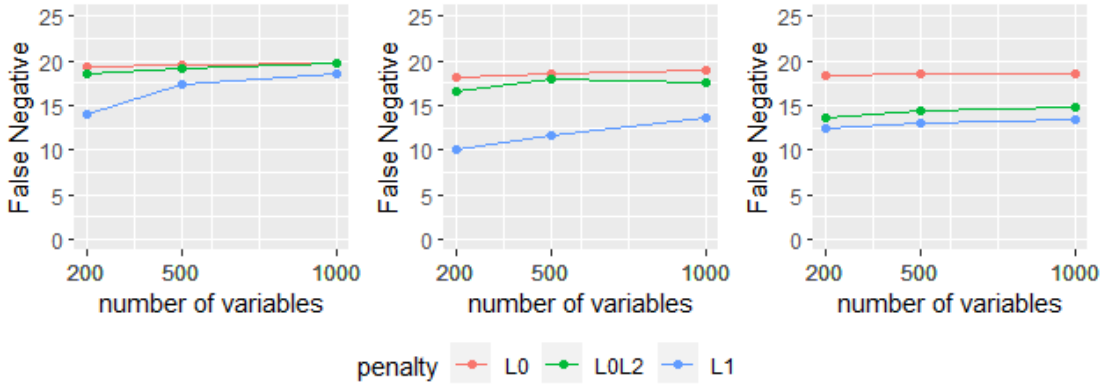
จากภาพที่ 4.2.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบน้อยที่สุด และ วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากที่สุด ในทุกกรณีของจำนวนของตัวแปรอิสระ และทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบสูงขึ้นเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น และทั้งสามวิธีมีค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบน้อยลงเมื่อเทียบกับในกรณีที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.2.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบน้อยที่สุด และ วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากที่สุด ในทุกกรณีของจำนวนของตัวแปรอิสระ และในกรณีที่จำนวนของตัวแปรอิสระมีจำนวน 500 และ 1000 ตัวแปร วิธี L_0 และ L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบใกล้เคียงกัน โดยทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบสูงขึ้นเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น และทั้งสามวิธีมีค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบน้อยลงเมื่อเทียบกับในกรณีที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5

ค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 6

ภาพที่ 4.2. 2 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ กรณี SNR = 6

$\rho = 0$ $\rho = 0.5$ $\rho = 0.9$



ภาพที่ 4.2.2 แสดงค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) จากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอย โดยค่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีค่าเท่ากับศูนย์ในขณะที่ สัมประสิทธิ์ที่แท้จริงมีค่าไม่เท่ากับศูนย์ ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหก จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.2.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าที่เปลี่ยนแปลงไปมากเมื่อจำนวนของตัวแปรอิสระเพิ่มขึ้นจาก 200 เป็น 500 ตัวแปร แต่มีค่าที่เปลี่ยนแปลงไปน้อยเมื่อจำนวนของตัวแปรอิสระเพิ่มสูงขึ้นจาก 500 เป็น 1000 ตัวแปร วิธี L_0 และ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบใกล้เคียงกันในทุกกรณีของจำนวนของตัวแปรอิสระโดย วิธี L_0L_2 มีค่าน้อยกว่าเพียงเล็กน้อย โดยทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบสูงขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.2.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ และทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบลดลงเมื่อเทียบกับกรณีที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.2.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_1 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงบวกน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0L_2 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากกว่าวิธี L_1 เล็กน้อยในทุกกรณีของจำนวนของตัวแปรอิสระ และวิธี L_0 มีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบมากที่สุด โดยทั้งสามวิธีมีค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบสูงขึ้นเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น และในกรณีของวิธี L_1 เมื่อเปรียบเทียบค่าเฉลี่ยของความผิดพลาดในการตรวจจับเชิงลบระหว่างกรณีความสัมพันธ์ของตัวแปรอิสระเป็น 0.5 และ 0.9 พบว่า มีค่าสูงขึ้น แต่ในกรณีของวิธี L_0L_2 กลับมีค่าลดลง

4.3 ผลการเปรียบเทียบค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) ด้วยการเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้กับค่าสัมประสิทธิ์ที่แท้จริง ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบค่าเฉลี่ยของ F1 Score เพื่อวัดความถูกต้องในการคัดเลือกตัวแปร ของวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ระหว่างวิธี L_0, วิธี L_1 และวิธี L_0L_2 ภายใต้ปัจจัยดังต่อไปนี้

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 100 : 200, 100 : 500 และ 100 : 1000

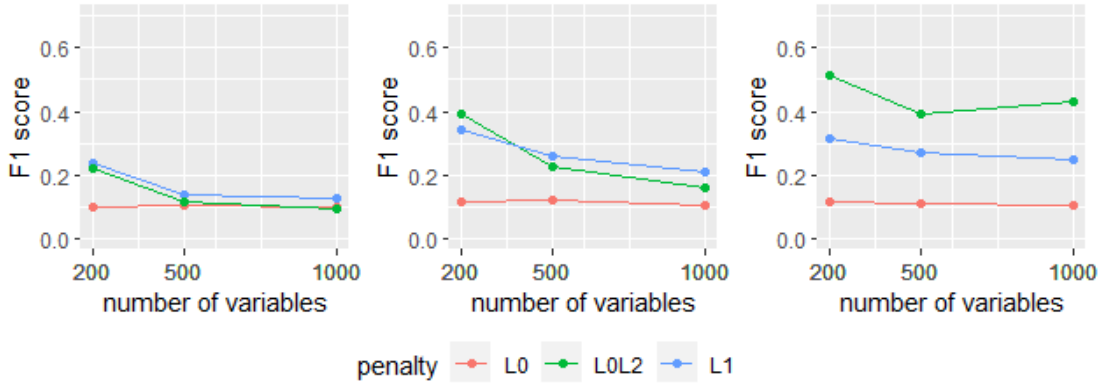
2. ศึกษาภายใต้ความสัมพันธ์ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0, 0.5, 0.9$

3. ศึกษาภายใต้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 1 และ 6

ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 1

ภาพที่ 4.3. 1 แสดงค่าเฉลี่ย F1 Score กรณี SNR = 1

$\rho = 0$ $\rho = 0.5$ $\rho = 0.9$



ภาพที่ 4.3.1 แสดงค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) เป็นค่าที่ใช้วัดความสามารถของโมเดล จากการเอาค่าความแม่นยำ (Precision) และ ค่าความไว (Sensitivity) มาคำนวณรวมกัน โดยแทนค่า Confusion Matrix ด้วยการเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้กับค่าสัมประสิทธิ์ที่แท้จริง ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหนึ่ง จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.3.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุดอย่างชัดเจน และวิธี L_1 กับวิธี L_0L_2 มีค่าใกล้เคียงกัน ในกรณีที่จำนวนของตัวแปรอิสระมีจำนวน 200 ตัวแปร วิธี L_0, L_1, L_0L_2 มีค่าเฉลี่ยของ F1 Score ใกล้เคียงกัน ในกรณีที่จำนวนของตัวแปรอิสระมีจำนวน 500 และ 1000 ตัวแปร โดยวิธี L_1 และ วิธี L_0L_2 มีค่าเฉลี่ยของ F1 Score ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

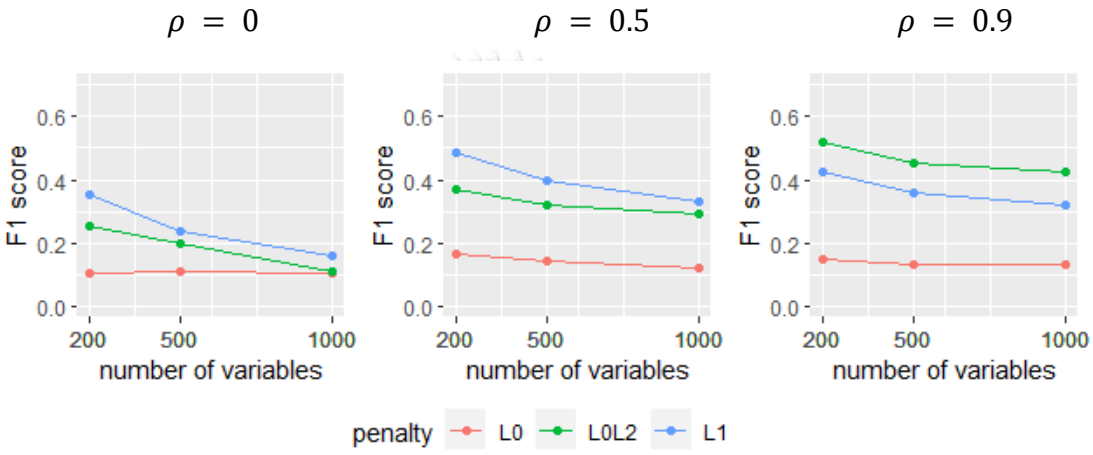
จากภาพที่ 4.3.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 และ L_0L_2 มีค่าใกล้เคียงกัน และมีค่าน้อยลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ทั้งวิธี L_1 และ L_0L_2 มีค่า F1 Score สูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.3.1 เมื่อพิจารณา ที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีค่าเฉลี่ยมากกว่า

วิธี L_0 และวิธี L_0L_2 มีค่าเฉลี่ยของ F1 Score มากที่สุด ทั้งวิธี L_1 และ L_0L_2 มีค่า F1 Score สูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5

ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 6

ภาพที่ 4.3. 2 แสดงค่าเฉลี่ย F1 Score กรณี SNR = 6



ภาพที่ 4.3.2 แสดงค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (Weighted average of Precision and Sensitivity : F1 Score) เป็นค่าที่ใช้วัดความสามารถของโมเดล จากการเอาค่าความแม่นยำ (Precision) และ ค่าความไว (Sensitivity) มาคำนวณรวมกัน โดยแทนค่า Confusion Matrix ด้วยการเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้กับค่าสัมประสิทธิ์ที่แท้จริง ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหก จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.3.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุด ในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0L_2 มีค่าเฉลี่ยมากกว่าวิธี L_0 ในขณะที่วิธี L_1 มีค่าเฉลี่ยสูงที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ โดยทั้งวิธี L_0L_2 และวิธี L_1 มีค่าเฉลี่ยของ F1 Score ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.3.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0L_2 มีค่าเฉลี่ยมากกว่าวิธี L_0 ในขณะที่วิธี L_1 มีค่าเฉลี่ยสูงที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ โดยทั้งสามวิธีมีค่าเฉลี่ยของ F1 Score ที่ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.3.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า L_0 มีค่าเฉลี่ยของ F1 Score น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีค่าเฉลี่ยมากกว่าวิธี L_0 ในขณะที่วิธี L_0L_2 มีค่าเฉลี่ย สูงที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ โดยทั้งสามวิธีมีค่าเฉลี่ยของ F1 Score ที่ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

4.4 ผลการเปรียบเทียบค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง ROC (AUC) เพื่อวัดความสามารถในการพยากรณ์ของตัวแบบ ของวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ระหว่างวิธี L_0, วิธี L_1 และวิธี L_0L_2 ภายใต้ปัจจัยดังต่อไปนี้

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 100 : 200, 100 : 500 และ 100 : 1000

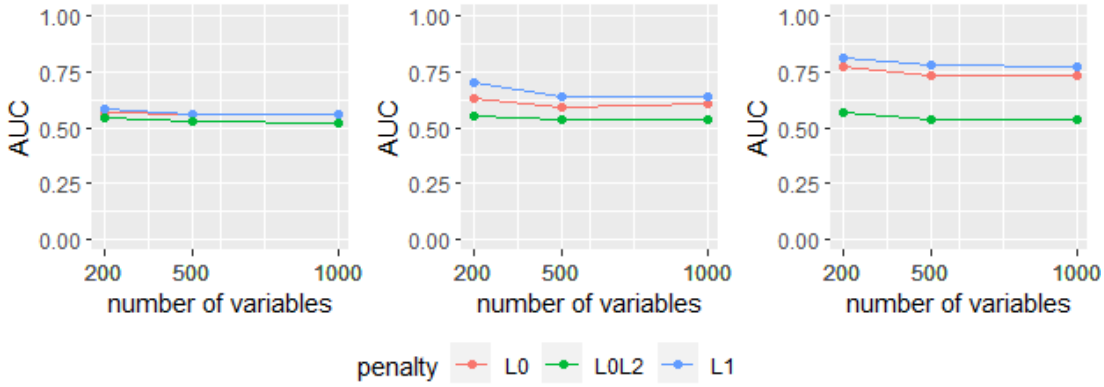
2. ศึกษาภายใต้ความสัมพันธ์ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0, 0.5, 0.9$

3. ศึกษาภายใต้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 1 และ 6

ค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 1

ภาพที่ 4.4. 1 แสดงพื้นที่ใต้เส้นโค้ง ROC (AUC) กรณี SNR = 1

$\rho = 0$ $\rho = 0.5$ $\rho = 0.9$



ภาพที่ 4.4.1 แสดงค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ หาได้จากพื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) ที่สร้างจาก กราฟของ Sensitivity และ 1-Specificity ของ Confusion Matrix ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหนึ่ง จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.4.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระเท่ากับ 0 พบว่า ทั้งสามวิธีมีค่าเฉลี่ยของ AUC ที่ใกล้เคียงกันในทุกกรณีของจำนวนของตัวแปรอิสระ และมีค่าน้อยลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

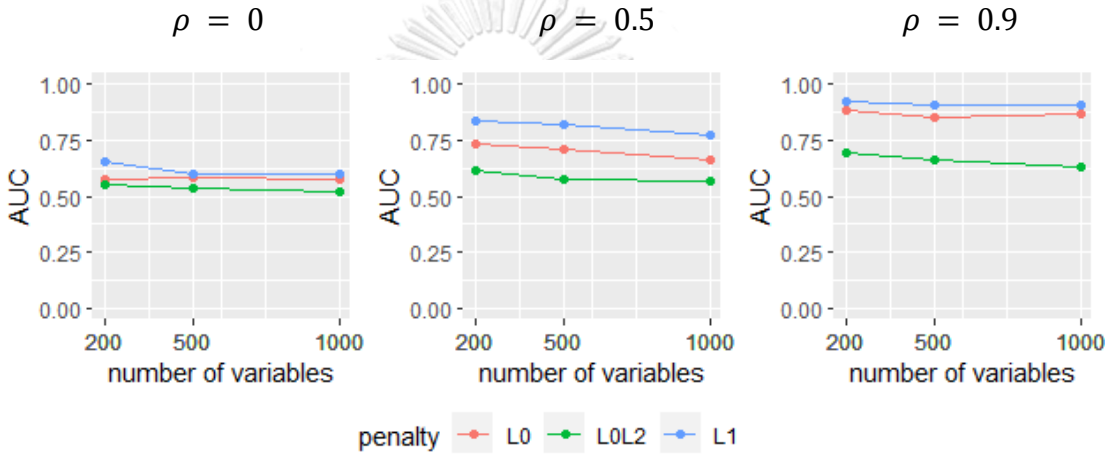
จากภาพที่ 4.4.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระเท่ากับ 0.5 พบว่า วิธี L_0L_2 มีค่าเฉลี่ยของ AUC ที่น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยที่มากกว่าวิธี L_0L_2 ในขณะที่ วิธี L_1 มีค่าเฉลี่ยของ AUC มากที่สุด และทั้งสามวิธีมีค่าเฉลี่ยของ AUC สูงเมื่อจำนวนของตัวแปรอิสระมีจำนวน 200 ตัวแปร และลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยค่าเฉลี่ยของ AUC ของวิธี L_0 และ วิธี L_1 มีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระเท่ากับ 0 แต่วิธี L_0L_2 มีค่าใกล้เคียงเดิม

จากภาพที่ 4.4.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระเท่ากับ 0.9 พบว่า วิธี L_0L_2 มีค่าเฉลี่ยของ AUC ที่น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยที่มากกว่าวิธี L_0L_2 ในขณะที่ วิธี L_1 มีค่าเฉลี่ยของ AUC มากที่สุด และทั้งสามวิธีมีค่าเฉลี่ยของ AUC สูงเมื่อจำนวน

ของตัวแปรอิสระมีจำนวน 200 ตัวแปร และลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยค่าเฉลี่ยของ AUC ของวิธี L_0 และ วิธี L_1 มีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระเท่ากับ 0.5 แต่วิธี L_0L_2 มีค่าใกล้เคียงเดิม

ค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 6

ภาพที่ 4.4. 2 แสดงพื้นที่ใต้เส้นโค้ง ROC (AUC) กรณี SNR = 6



ภาพที่ 4.4.2 แสดงค่าเฉลี่ยของพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ใช้บ่งชี้ความน่าเชื่อถือและความสามารถในการพยากรณ์ของตัวแบบ หาได้จากพื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) ที่สร้างจาก กราฟของ Sensitivity และ 1-Specificity ของ Confusion Matrix ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหก จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.4.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0L_2 มีค่าเฉลี่ยของ AUC น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยมากกว่าวิธี L_0L_2 และวิธี L_1 มีค่าเฉลี่ยของ AUC มากที่สุด โดยมีค่าลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.4.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0L_2 มีค่าเฉลี่ยของ AUC น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยมากกว่าวิธี L_0L_2 และวิธี L_1 มีค่าเฉลี่ยของ AUC มากที่สุด และ ทั้งสามวิธีมีค่าเฉลี่ยของ AUC ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยค่าเฉลี่ยของ AUC ของทั้งสามวิธีมีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.4.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0L_2 มีค่าเฉลี่ยของ AUC น้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_0 มีค่าเฉลี่ยมากกว่าวิธี L_0L_2 และวิธี L_1 มีค่าเฉลี่ยของ AUC มากที่สุด และ ทั้งสามวิธีมีค่าเฉลี่ยของ AUC ลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยค่าเฉลี่ยของ AUC ของทั้งสามวิธีมีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5

4.5 ผลการเปรียบเทียบค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบค่าพื้นที่ใต้เส้นโค้ง ROC (AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) เพื่อวัดความถูกต้องในการคัดเลือกตัวแปรของวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ระหว่างวิธี L_0, วิธี L_1 และวิธี L_0L_2 ภายใต้ปัจจัยดังต่อไปนี้

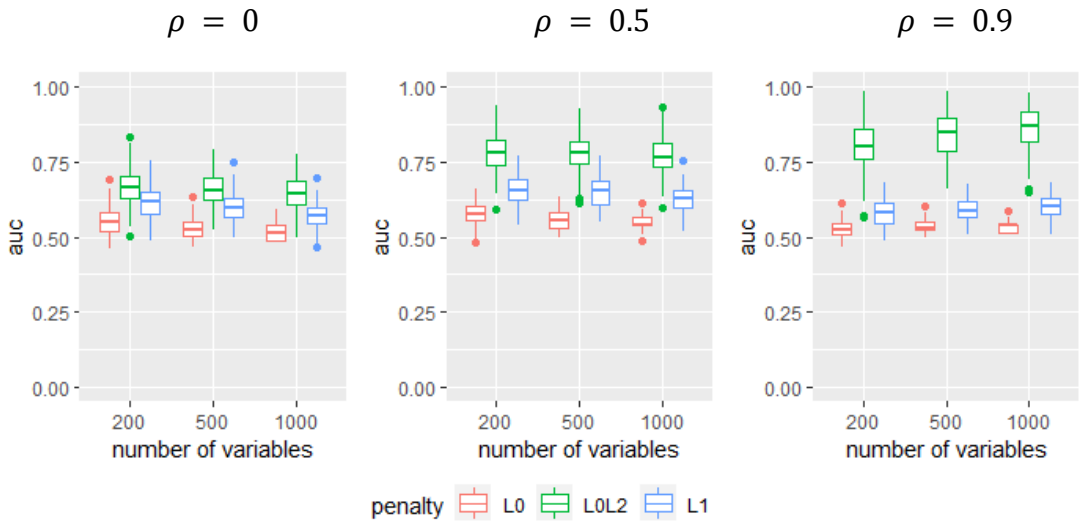
1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 100 : 200, 100 : 500 และ 100 : 1000

2. ศึกษาภายใต้ความสัมพันธ์ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0, 0.5, 0.9$

3. ศึกษาภายใต้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 1 และ 6

ค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 1

ภาพที่ 4.5. 1 แสดงพื้นที่ใต้เส้นโค้ง Variable Selection AUC กรณี SNR = 1



ภาพที่ 4.5.1 แสดงค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน เพื่อวัดความถูกต้องในการคัดเลือกตัวแปรของตัวแบบ หาได้จากพื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) ที่สร้างจาก กราฟของ Sensitivity และ 1-Specificity ของ Confusion Matrix ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหนึ่ง จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.5.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมาก

ที่สุด และทั้งสามวิธีมีความถูกต้องในการคัดเลือกตัวแปรลดลงเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

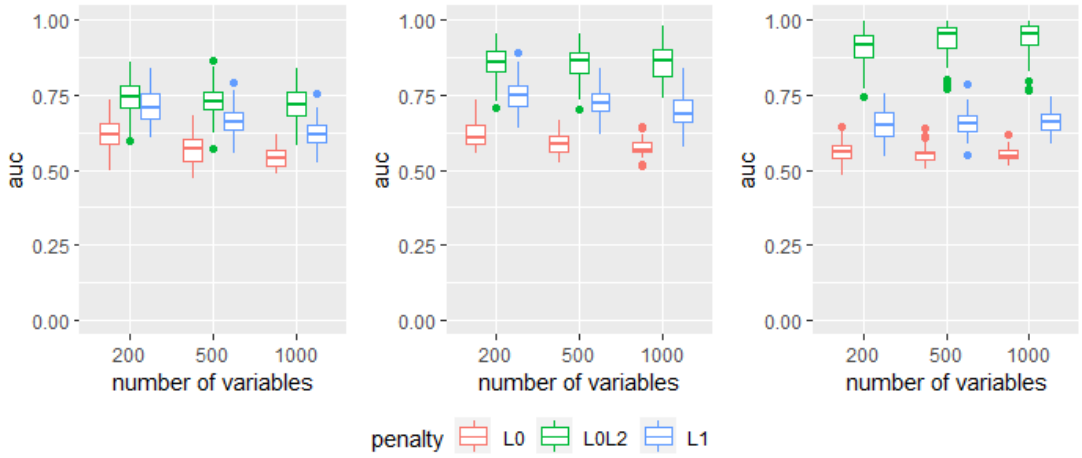
จากภาพที่ 4.5.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมากที่สุด และทั้งสามวิธีมีความถูกต้องในการคัดเลือกตัวแปรลดลงเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยความถูกต้องในการคัดเลือกตัวแปรของทั้งสามวิธีมีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.5.1 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมากที่สุด และทั้งสามวิธีมีความถูกต้องในการคัดเลือกตัวแปรเพิ่มขึ้นเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยความถูกต้องในการคัดเลือกตัวแปรของวิธี L_0 และ L_1 มีค่าลดลงเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 แต่วิธี L_0L_2 มีค่าความถูกต้องในการคัดเลือกตัวแปรสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5

ค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน กรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR = 6

ภาพที่ 4.5. 2 แสดงพื้นที่ใต้เส้นโค้ง Variable Selection AUC กรณี SNR = 6

$\rho = 0$ $\rho = 0.5$ $\rho = 0.9$



ภาพที่ 4.5.2 แสดงค่าพื้นที่ใต้เส้นโค้ง (Area under the curve : AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน เพื่อวัดความถูกต้องในการคัดเลือกตัวแปรของตัวแบบ หาได้จากพื้นที่ใต้เส้นโค้ง ROC (Receiver Operating Characteristic curve) ที่สร้างจาก กราฟของ Sensitivity และ 1-Specificity ของ Confusion Matrix ระหว่างวิธีการคัดกรองตัวแปรในการวิเคราะห์การถดถอยโลจิสติก ด้วยการเพิ่ม penalty function 3 รูปแบบ ได้แก่ L_0-norm penalty (Best Subset Selection) ,L_1-norm penalty (LASSO) และวิธี L_0L_2-norm penalty ในกรณีที่ข้อมูลมีค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) เท่ากับหก จำนวนของตัวแปรอิสระมีจำนวน 200, 500, 1000 และค่าความสัมพันธ์ของตัวแปรอิสระที่ 0, 0.5, 0.9

จากภาพที่ 4.5.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมากที่สุด และทั้งสามวิธีมีความถูกต้องในการคัดเลือกตัวแปรลดลงเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น

จากภาพที่ 4.5.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมากที่สุด และทั้งสามวิธีมีความถูกต้องในการคัดเลือกตัวแปรลดลงเล็กน้อยเมื่อจำนวนของตัว

แปรอิสระมีจำนวนมากขึ้น โดยความถูกต้องในการคัดเลือกตัวแปรของทั้งสามวิธีมีค่าสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0

จากภาพที่ 4.5.2 เมื่อพิจารณาที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 พบว่า วิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรน้อยที่สุดในทุกกรณีของจำนวนของตัวแปรอิสระ วิธี L_1 มีความถูกต้องในการคัดเลือกตัวแปรมากกว่าวิธี L_0 ในขณะที่ วิธี L_0L_2 มีความถูกต้องในการคัดเลือกตัวแปรมากที่สุด และวิธี L_0 มีความถูกต้องในการคัดเลือกตัวแปรใกล้เคียงเดิมในทุกกรณีของจำนวนของตัวแปรอิสระ ส่วนวิธี L_1 และ วิธี L_0L_2 มีค่าเพิ่มขึ้นเล็กน้อยเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น โดยความถูกต้องในการคัดเลือกตัวแปรของวิธี L_0 และ L_1 มีค่าลดลงเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 แต่วิธี L_0L_2 มีค่าความถูกต้องในการคัดเลือกตัวแปรสูงขึ้นเมื่อเทียบกับกรณีความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบวิธีการในการคัดกรองตัวแปร และประสิทธิภาพในการพยากรณ์ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม จากวิธีการใช้ Penalty function 3 รูปแบบได้แก่ L_0-norm penalty (Best Subset Selection), L_1-norm penalty (LASSO) และ L_0L_2-norm penalty โดยทำการศึกษากับข้อมูลที่จำลองขึ้นมาและข้อมูลมีลักษณะเป็นข้อมูลที่มีมิติสูงและพิจารณาแยกตามขนาดตัวแปรอิสระ p = 200, 500 และ 1000 รวมถึงค่าความสัมพันธ์ภายในตัว

แปรอิสระและค่าอัตราสัญญาณต่อสัญญาณรบกวน (SNR) โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพในการพยากรณ์และค่าความถูกต้องในการคัดกรองตัวแปร ของแต่ละวิธีจาก ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score) พื้นที่ใต้เส้นโค้ง ROC (AUC) ซึ่งใช้บ่งชี้ความสามารถในการพยากรณ์ของตัวแบบ และ Variable Selection AUC ที่ใช้วัดความถูกต้องในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ โดยสรุปผลการวิจัยได้ดังนี้

แบ่งผลการวิจัยออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร และประสิทธิภาพในการพยากรณ์ จากการคัดกรองตัวแปรทั้ง 3 วิธี โดยพิจารณาจากเกณฑ์ที่ใช้ในการตัดสินใจ คือ ค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) ค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score) พื้นที่ใต้เส้นโค้ง ROC (AUC) และพื้นที่ใต้เส้นโค้งจากการคัดเลือกตัวแปร (Variable Selection AUC)

ส่วนที่ 2 ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร, ประสิทธิภาพในการพยากรณ์จากการคัดกรองตัวแปรทั้ง 3 วิธี และ ระยะเวลาที่ใช้ในการประมวลผลของโปรแกรม โดยพิจารณาจากกรณีของการจำลองข้อมูล

5.1 ผลการวิจัยส่วนที่ 1

5.1.1 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP)

ตารางที่ 5.1. 1 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก

n : p	SNR = 1			SNR = 6		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
100:200	L_0	L_0	L_0	L_0	L_0	L_0
100:500	L_0	L_0	L_0	L_0	L_0	L_0
100:1000	L_0	L_0	L_0	L_0	L_0	L_0

จากตารางที่ 5.1.1 แสดงวิธีการคัดกรองตัวแปรที่ดีที่สุดในแต่ละกรณีของการจำลองข้อมูล เมื่อพิจารณาจากค่าความผิดพลาดในการตรวจจับเชิงบวก (FP)

จากตารางที่ 5.1.1 สรุปผลได้ว่า ในทุกกรณีของการจำลองข้อมูล ค่าความผิดพลาดในการตรวจจับเชิงบวกจากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์เทียบกับค่าสัมประสิทธิ์ที่แท้จริงของวิธี L_0 มีค่าน้อยที่สุด คือมีการคัดเลือกตัวแปรที่ไม่เกี่ยวข้องเข้าสู่ตัวแบบน้อยที่สุด

5.1.2 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN)

ตารางที่ 5.1. 2 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงลบ

n : p	SNR = 1			SNR = 6		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
100:200	L_1	L_1	L_1	L_1	L_1	L_1
100:500	L_1	L_1	L_1	L_1	L_1	L_1
100:1000	L_1	L_1	L_1	L_1	L_1	L_1

จากตารางที่ 5.1.2 แสดงวิธีการคัดกรองตัวแปรที่ดีที่สุดในแต่ละกรณีของการจำลองข้อมูล เมื่อพิจารณาจากค่าความผิดพลาดในการตรวจจับเชิงลบ (FN)

จากตารางที่ 5.1.2 สรุปผลได้ว่า ในทุกกรณีของการจำลองข้อมูล ค่าความผิดพลาดในการตรวจจับเชิงลบจากการวัดจำนวนความผิดพลาดในการประมาณค่าสัมประสิทธิ์เทียบกับค่าสัมประสิทธิ์ที่แท้จริงของวิธี L_1 มีค่าน้อยที่สุด คือมีการตัดตัวแปรที่แท้จริงออกจากตัวแบบน้อยที่สุด

5.1.3 ผลการเปรียบเทียบค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score)

ตารางที่ 5.1. 3 ผลการเปรียบเทียบค่าเฉลี่ย F1 Score

n : p	SNR = 1			SNR = 6		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
100:200	L_1	L_0L_2	L_0L_2	L_1	L_1	L_0L_2
100:500	L_1	L_1	L_0L_2	L_1	L_1	L_0L_2
100:1000	L_1	L_1	L_0L_2	L_1	L_1	L_0L_2

จากตารางที่ 5.1.3 แสดงวิธีการคัดกรองตัวแปรที่ดีที่สุดในแต่ละกรณีของการจำลองข้อมูล เมื่อพิจารณาจากค่าเฉลี่ยแบบฮาร์โมนิคของค่าความแม่นยำและค่าความไว (F1 Score)

จากตารางที่ 5.1.3 สรุปผลได้ว่า เมื่อพิจารณากรณีของอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) ทั้ง 1 และ 6 พบว่าที่ความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 และ 0.5 วิธี L_1 ให้ค่าเฉลี่ยของ F1 Score สูงที่สุด คือมีความถูกต้องในการคัดกรองตัวแปรมากที่สุด แต่ในกรณีของความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 วิธี L_0L_2 ให้ค่าความถูกต้องในการคัดกรองตัวแปรสูงที่สุด

5.1.4 ผลการเปรียบเทียบพื้นที่ใต้เส้นโค้ง ROC (AUC) ที่ใช้บ่งชี้ความสามารถในการพยากรณ์

ตารางที่ 5.1. 4 ผลการเปรียบเทียบพื้นที่ใต้เส้นโค้ง ROC (AUC)

n : p	SNR = 1			SNR = 6		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
100:200	L_1	L_1	L_1	L_1	L_1	L_1
100:500	L_1	L_1	L_1	L_1	L_1	L_1
100:1000	L_1	L_1	L_1	L_1	L_1	L_1

จากตารางที่ 5.1.4 แสดงวิธีการที่ดีที่สุดในแต่ละกรณีของการจำลองข้อมูล เมื่อพิจารณาจากพื้นที่ใต้เส้นโค้ง ROC (AUC) ที่ใช้บ่งชี้ความสามารถในการพยากรณ์

จากตารางที่ 5.1.4 สรุปผลได้ว่า ในทุกกรณีของการจำลองข้อมูล ค่าพื้นที่ใต้เส้นโค้ง (AUC) ของความสามารถในการพยากรณ์ของตัวแบบ วิธี L_1 ให้ค่าสูงที่สุด คือ ตัวแบบมีความน่าเชื่อถือและมีสามารถในการพยากรณ์สูงที่สุด

5.1.5 ผลการเปรียบเทียบพื้นที่ใต้เส้นโค้ง ROC (AUC) ในกรณีของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ (Variable Selection AUC) โดยเปรียบเทียบค่าสัมประสิทธิ์ที่ประมาณได้ กับค่าสัมประสิทธิ์ที่แท้จริง ในตัวแบบที่ใช้ พารามิเตอร์ Lambda (λ) ที่แตกต่างกัน

ตารางที่ 5.1. 5 ผลการเปรียบเทียบพื้นที่ใต้เส้นโค้ง Variable Selection AUC

n : p	SNR = 1			SNR = 6		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
100:200	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2
100:500	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2

100:1000	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2	L_0L_2

จากตารางที่ 5.1.5 แสดงวิธีการที่ดีที่สุดในแต่ละกรณีของการจำลองข้อมูล เมื่อพิจารณาจาก Variable Selection AUC

จากตารางที่ 5.1.5 สรุปผลได้ว่า ในทุกกรณีของการจำลองข้อมูล ค่าความถูกต้องในการคัดกรองตัวแปร (Variable Selection AUC) ของวิธี L_0L_2 ให้ค่าสูงที่สุด คือ มีความสามารถในการคัดกรองตัวแปรที่แท้จริงเข้าสู่ตัวแบบมากที่สุด

5.2 ผลการวิจัยส่วนที่ 2

5.2.1 ความแตกต่างจากจำนวนของตัวแปรอิสระ

จากผลการทดลองที่ได้ พบว่า ทั้งในกรณีของอัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าเท่ากับ 1 และ 6 เมื่อจำนวนของตัวแปรอิสระมีจำนวน 200 ทั้งค่าความถูกต้องในการคัดกรองตัวแปรและค่าความสามารถในการพยากรณ์ของตัวแบบทั้งสามวิธี ให้ค่าที่ดีที่สุด และลดลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้นเป็น 500 และน้อยที่สุดเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้นเป็น 1000 กลับกัน เมื่อพิจารณาที่ค่าความผิดพลาดในการตรวจจับเชิงบวก ในกรณี อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับหนึ่ง และความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.5 และ 0.9 วิธี L_0L_2 ให้ค่าความผิดพลาดในการตรวจจับเชิงบวกน้อยลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ในกรณีอัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับหก และความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 ทั้งสามวิธี ให้ค่าความผิดพลาดในการตรวจจับเชิงบวกน้อยลงเมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น เมื่อพิจารณาที่ค่าความผิดพลาดในการตรวจจับเชิงลบ เมื่อจำนวนของตัวแปรอิสระมีจำนวนมากขึ้น ค่าความผิดพลาดในการตรวจจับเชิงลบของทุกกรณีมีแนวโน้มสูงขึ้น

5.2.2 ความแตกต่างจากความสัมพันธ์ของตัวแปรอิสระ

จากผลการทดลองที่ได้ พบว่า ทั้งในกรณีของอัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าเท่ากับ 1 และ 6 เมื่อความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0 ทั้งค่าความถูกต้องในการคัดกรองตัวแปร และค่าความสามารถในการพยากรณ์ของตัวแบบทั้งสามวิธี ให้ค่าน้อยที่สุด และสูงขึ้นเมื่อความสัมพันธ์ของตัวแปรอิสระมีค่าเพิ่มขึ้นเป็น 0.5 และสูงที่สุดเมื่อความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.9 กลับกัน เมื่อพิจารณาที่ค่าความผิดพลาดในการตรวจจับเชิงบวก ในกรณี อัตราส่วน

สัญญาณต่อสัญญาณรบกวนเท่ากับหนึ่ง วิธี L_1 ให้ค่าความผิดพลาดในการจรวจจับเชิงบวกมากขึ้นเมื่อความสัมพันธ์ของตัวแปรอิสระมีค่ามากขึ้น ในกรณี อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับหก วิธี L_1 และ วิธี L_0L_2 ให้ค่าความผิดพลาดในการตรวจจับเชิงบวกมากขึ้นเมื่อความสัมพันธ์ของตัวแปรอิสระมีจำนวนมากขึ้น

5.2.3 ความแตกต่างจากอัตราส่วนสัญญาณต่อสัญญาณรบกวน

จากผลการทดลองที่ได้ พบว่า ในทุกกรณีของความสัมพันธ์ของตัวแปรอิสระ ทั้งค่าความถูกต้องในการคัดกรองตัวแปร และค่าความสามารถในการพยากรณ์ของตัวแบบทั้งสามวิธี ให้ค่าสูงขึ้นเมื่อ อัตราส่วนสัญญาณติสัญญาณรบกวนเพิ่มขึ้นจาก 1 เป็น 6 กลับกัน เมื่อพิจารณาที่ค่าความผิดพลาดในการตรวจจับเชิงบวก วิธี L_1 และ วิธี L_0L_2 ให้ค่าความผิดพลาดในการตรวจจับเชิงบวกสูงขึ้นเมื่อ อัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าสูงขึ้นจาก 1 เป็น 6 เมื่อพิจารณาที่ค่าความผิดพลาดในการตรวจจับเชิงลบ ทั้งสามวิธีให้ค่าความผิดพลาดในการตรวจจับเชิงลบน้อยลงเมื่ออัตราส่วนสัญญาณต่อสัญญาณรบกวนมีค่าสูงขึ้นจาก 1 เป็น 6

5.2.4 ระยะเวลาที่ใช้ในการประมวลผล (Computation Time)

จากการทดลอง ผู้วิจัยได้บันทึกระยะเวลาที่ใช้ในการประมวลผลของแต่ละวิธีด้วยการนำข้อมูล 3 กรณี กรณีละ 100 ชุดข้อมูล มาทำการวิเคราะห์การถดถอยและบันทึกระยะเวลาที่ใช้ในการประมวลผล โดยแบ่งออกเป็น 3 กรณี คือ กรณีที่จำนวนของตัวแปรอิสระเป็น 200 ตัวแปร, กรีณีที่จำนวนของตัวแปรอิสระเป็น 500 ตัวแปร และ กรณีที่จำนวนของตัวแปรอิสระเป็น 1000 ตัวแปร ซึ่งทำการประมวลผลโดยใช้โปรแกรม R version 3.6.1 ผ่านคอมพิวเตอร์ด้วยระบบปฏิบัติการ Windows 10 HOME 64-bit ชิปประมวลผล Intel Core I5-8265U CPU @ 1.6GHz 1.8GHz RAM 8.00 GB แสดงผลดังตาราง

ตารางที่ 5.2. 1 ตารางแสดงระยะเวลาในการประมวลผลของโปรแกรม หน่วยเป็นวินาที

วิธี	จำนวนของตัวแปรอิสระ		
	200	500	1000
L_0	522.78	524.61	538.80
L_1	22.09	23.86	26.19

L_0L_2	3869.97	9717.88	19445.75

จากตารางที่ 5.2.4.1 สรุปผลได้ว่า วิธี L_1 มีระยะเวลาในการประมวลผลน้อยที่สุด รองลงมาคือวิธี L_0 และวิธี L_0L_2 มีระยะเวลาในการประมวลผลมากที่สุด และทั้งสามวิธีใช้ระยเวลาในการประมวลผลมากขึ้น เมื่อจำนวนของข้อมูลมากขึ้น

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ ผู้ที่สนใจอาจนำไปศึกษาต่อในเรื่องของ

1. ขอบเขตการศึกษา ในเรื่องของ จำนวนตัวอย่าง, จำนวนตัวแปรอิสระ, อัตราส่วนสัญญาณต่อสัญญาณรบกวน, ความสัมพันธ์ของตัวแปรอิสระ และ ขนาดของสัมประสิทธิ์ β_j อาจมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้น

2. วิธีการคัดกรองตัวแปรในงานวิจัยนี้ เลือกมาศึกษาทั้งหมด 3 วิธี ซึ่งในความเป็นจริงยังมีวิธีคัดกรองตัวแปรวิธีอื่นๆอีก ผู้ที่สนใจอาจนำวิธีคัดกรองตัวแปรอื่นมาร่วมพิจารณาเพื่อเปรียบเทียบประสิทธิภาพได้

3. การจำลองตัวแปรอิสระในกรณีที่ไม่ใช่การแจกแจงแบบปกติ

บรรณานุกรม

วิฐรา พึ่งพาพงศ์. (2558). บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี, 212-223.

Agresti, A. (2003). *Categorical data analysis* (Vol. 482): John Wiley & Sons.

Beale, E., Kendall, M., & Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika, 54*(3-4), 357-366.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The annals of statistics, 44*(2), 813-852.

Duffy, D. E., & Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics-Theory and Methods, 18*(3), 959-980.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*: Spinger.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learnin. *Cited on*, 33.

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research, 68*(5), 1517-1537.

Hocking, R. R., & Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics, 9*(4), 531-540.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 41*(1), 191-201.

Mazumder, R., Radchenko, P., & Dedieu, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *arXiv preprint arXiv:1708.03288*.

Miller, A. (2002). *Subset selection in regression*: CRC Press.

Pungpapong, V., Zhang, M., & Zhang, D. (2015). Selecting massive variables using an iterated conditional modes/medians algorithm. *Electronic Journal of Statistics, 9*(1), 1243-1266.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

การจำลองข้อมูล

```
set.seed(5)

### correlation ###

corX0 <- matrix(,nrow=200,ncol=200)

corX05 <- matrix(,nrow=200,ncol=200)

corX09 <- matrix(,nrow=200,ncol=200)

rho0 <- 0

rho05 <- 0.5

rho09 <- 0.9


for (i in 1:200){

  for (j in 1:200){

    if(i!=j){

      corX0[i,j]=rho0^abs(i-j)

    }else{

      corX0[i,j]=1}

}}


for (i in 1:200){
```

```
  for (j in 1:200){
    if(i!=j){
      corX05[i,j]=rho05^abs(i-j)
    }else{
      corX05[i,j]=1}
}}


for (i in 1:200){
  for (j in 1:200){
    if(i!=j){
      corX09[i,j]=rho09^abs(i-j)
    }else{
      corX09[i,j]=1}
}}

### x and y ###
#mvX <- rmvnorm(100,rep(0,200),corX)
b <- c(rep(2,10),rep(0,90),rep(1,10),rep(0,90))
x0 <- replicate(n=100,rmvnorm(100,rep(0,200),corX0))
x05 <- replicate(n=100,rmvnorm(100,rep(0,200),corX05))
x09 <- replicate(n=100,rmvnorm(100,rep(0,200),corX09))
```

```
XB0 <- matrix(,nrow=100,ncol=100)

XB05 <- matrix(,nrow=100,ncol=100)

XB09 <- matrix(,nrow=100,ncol=100)

s1r0 <- matrix(,nrow=1,ncol=100)

s1r05 <- matrix(,nrow=1,ncol=100)

s1r09 <- matrix(,nrow=1,ncol=100)

s6r0 <- matrix(,nrow=1,ncol=100)

s6r05 <- matrix(,nrow=1,ncol=100)

s6r09 <- matrix(,nrow=1,ncol=100)

eL1r0 <- matrix(,nrow=100,ncol=100)

eL1r05 <- matrix(,nrow=100,ncol=100)

eL1r09 <- matrix(,nrow=100,ncol=100)

eL6r0 <- matrix(,nrow=100,ncol=100)

eL6r05 <- matrix(,nrow=100,ncol=100)

eL6r09 <- matrix(,nrow=100,ncol=100)

Y1r0 <- matrix(,nrow=100,ncol=100)

Y1r05 <- matrix(,nrow=100,ncol=100)

Y1r09 <- matrix(,nrow=100,ncol=100)

Y6r0 <- matrix(,nrow=100,ncol=100)

Y6r05 <- matrix(,nrow=100,ncol=100)

Y6r09 <- matrix(,nrow=100,ncol=100)
```

```
for (i in 1:100){
  XB0[,i] <- x0[,,i]%*%b
  s1r0[,i] <- sqrt((3*var(XB0[,i]))/pi^2)
  s6r0[,i] <- sqrt((var(XB0[,i]))/(2*pi^2))
  eL1r0[,i] <- rlogis(100,0,scale = s1r0[,i])
  eL6r0[,i] <- rlogis(100,0,scale = s6r0[,i])
  Y1r0[,i] <- as.numeric(ifelse((XB0[,i]+eL1r0[,i])>0,"1","0"))
  Y6r0[,i] <- as.numeric(ifelse((XB0[,i]+eL6r0[,i])>0,"1","0"))
}
for (i in 1:100){
  XB05[,i] <- x05[,,i]%*%b
  s1r05[,i] <- sqrt((3*var(XB05[,i]))/pi^2)
  s6r05[,i] <- sqrt((var(XB05[,i]))/(2*pi^2))
  eL1r05[,i] <- rlogis(100,0,scale = s1r05[,i])
  eL6r05[,i] <- rlogis(100,0,scale = s6r05[,i])
  Y1r05[,i] <- as.numeric(ifelse((XB05[,i]+eL1r05[,i])>0,"1","0"))
  Y6r05[,i] <- as.numeric(ifelse((XB05[,i]+eL6r05[,i])>0,"1","0"))
}
for (i in 1:100){
  XB09[,i] <- x09[,,i]%*%b
```

```
  s1r09[,i] <- sqrt((3*var(XB09[,i]))/pi^2)
  s6r09[,i] <- sqrt((var(XB09[,i]))/(2*pi^2))
  eL1r09[,i] <- rlogis(100,0,scale = s1r09[,i])
  eL6r09[,i] <- rlogis(100,0,scale = s6r09[,i])
  Y1r09[,i] <- as.numeric(ifelse((XB09[,i]+eL1r09[,i])>0,"1","0"))
  Y6r09[,i] <- as.numeric(ifelse((XB09[,i]+eL6r09[,i])>0,"1","0"))
}
```

การแบ่งข้อมูลเพื่อทำ Cross Validation

```
n = 100
trainIndex = sample(1:n, size = round(0.8*n), replace=FALSE)
Y1r0.tr <- Y1r0[trainIndex,]
Y1r05.tr <- Y1r05[trainIndex,]
Y1r09.tr <- Y1r09[trainIndex,]
Y6r0.tr <- Y6r0[trainIndex,]
Y6r05.tr <- Y6r05[trainIndex,]
Y6r09.tr <- Y6r09[trainIndex,]
x0.tr <- replicate(100,matrix(nrow=80,ncol=200))
x05.tr <- replicate(100,matrix(nrow=80,ncol=200))
x09.tr <- replicate(100,matrix(nrow=80,ncol=200))
for (i in 1:100){
  x0.tr[,,i] <- x0[trainIndex,,i]
```

```
  x05.tr[,,i] <- x05[trainIndex,,i]
  x09.tr[,,i] <- x09[trainIndex,,i]
}

Y1r0.ts <- Y1r0[-trainIndex,]
Y1r05.ts <- Y1r05[-trainIndex,]
Y1r09.ts <- Y1r09[-trainIndex,]
Y6r0.ts <- Y6r0[-trainIndex,]
Y6r05.ts <- Y6r05[-trainIndex,]
Y6r09.ts <- Y6r09[-trainIndex,]
x0.ts <- replicate(100,matrix(nrow=20,ncol=200))
x05.ts <- replicate(100,matrix(nrow=20,ncol=200))
x09.ts <- replicate(100,matrix(nrow=20,ncol=200))
for (i in 1:100){
  x0.ts[,,i] <- x0[-trainIndex,,i]
  x05.ts[,,i] <- x05[-trainIndex,,i]
  x09.ts[,,i] <- x09[-trainIndex,,i]
}
การ fit model
b <- txtProgressBar(min = 0,max = 100,style = 3)
```

```
cvfitL0 <- list()

cvfitL1 <- list()

cvfit <- list()

for (i in 1:100){

  cvfitL0[[i]] <- L0Learn.cvfit(x09.tr[,,i],Y6r09.tr[,i],nFolds=5,seed=5,penalty = "L0",loss =
"Logistic")

  cvfitL1[[i]] <- cv.glmnet(x09.tr[,,i],Y6r09.tr[,i],nfolds = 5,family = "binomial",alpha=1)

  Sys.sleep(0.1)

  setTxtProgressBar(pb,i)

}

close(pb)



ldg <- replicate(n=100,list())

for (j in 1:100){

  for (i in 1:100){

    ldg[[j]][[i]] <- cvfitL0[[j]]$fit$lambda[[1]][i]

}}



for (i in 1:100){

  cvfit[[i]] <- L0Learn.cvfit(x09.tr[,,i],Y6r09.tr[,i],lambdaGrid = ldg[[i]],nFolds = 5,seed =
5,penalty = "L0L2",loss = "Logistic")
```

```
  Sys.sleep(0.1)
  setTxtProgressBar(pb,i)
}
close(pb)
```

การเก็บค่าที่ใช้วัดผลการทดลอง

```
bMatL0 <- replicate(n=100,matrix(nrow=1000,ncol=100))
bMatL1 <- replicate(n=100,matrix(nrow=1000,ncol=100))
bMatL0L2 <- replicate(n=100,matrix(nrow=1000,ncol=100))
for (j in 1:100){
  for (i in 1:100){
    bMatL0[,i,j] <- coef(cvfitL0[[j]],lambda = cvfitL0[[j]]$fit$lambda[[1]][i])[2:1001,1]
    bMatL1[,i,j] <- coef(cvfitL1[[j]],s=cvfitL1[[j]]$lambda[i])[2:1001,1]
    bMatL0L2[,i,j] <- coef(cvfit[[j]],lambda = cvfit[[j]]$fit$lambda[[i]],gamma =
cvfit[[j]]$fit$gamma[i])[2:1001,1]
  }}
```

ประวัติผู้เขียน

ชื่อ-สกุล	รัชพงศ์ ปรัชญาเศรษฐ
วัน เดือน ปี เกิด	11 กุมภาพันธ์ 2538
สถานที่เกิด	จังหวัดกรุงเทพฯ
วุฒิการศึกษา	B.Econ.
ที่อยู่ปัจจุบัน	144/80 หมู่ 1 ตำบลปากเกร็ด อำเภอปากเกร็ด จังหวัดนนทบุรี 11120