การเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่

นางสาวจุฑาทิพย์ นันทสุวรรณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

ON TUNING PARAMETER SELECTION OF LASSO REGRESSION

Miss Jutatip Nuntasuwan

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 2016

หัวข้อวิทยานิพนธ์	การเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่
โดย	นางสาวจุฑาทิพย์ นันทสุวรรณ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

.....กรรมการ

(อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.ฐิตารีย์ รุ่งรัตน์เกษม)

จุฑาทิพย์ นันทสุวรรณ : การเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่ (ON TUNING PARAMETER SELECTION OF LASSO REGRESSION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.วิฐรา พึ่งพาพงศ์, 60 หน้า.

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการเลือกพารามิเตอร์ปรับสำหรับวิธีการถดถอยแบบลาสโซ่โดยใช้การตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอย และเปรียบเทียบผลที่ได้กับการเลือกพารามิเตอร์ปรับจากสองวิธีที่ใช้กันอย่างแพร่หลายสำหรับการถดถอยแบบลาสโซ่ ได้แก่ วิธีการตรวจสอบไขว้ และวิธีการใช้เกณฑ์ข้อสนเทศของเบส์ โดยทำการจำลองข้อมูลให้ครอบคลุมกับเหตุการณ์ที่อาจก่อให้เกิดปัญหาเกี่ยวกับข้อบังคับเบื้องต้นของการถดถอยทั้งหมด 6 กรณี เน้นไปที่การเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ สำหรับเกณฑ์ที่ใช้วัดประสิทธิ์ภาพของผลที่ได้จากการวิเคราะห์การถดถอยด้วยพารามิเตอร์ปรับจากวิธีต่าง ๆ ได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าคลาดเคลื่อนจากการพยากรณ์ และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย ผลการศึกษาจากการจำลองข้อมูลพบว่าวิธีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยให้อัตราความผิดพลาดในการตรวจจับเชิงบวกต่ำที่สุด วิธีการตรวจสอบไขว้ให้อัตราความผิดพลาดในการตรวจจับเชิงลบต่ำกว่าอีกสองวิธี นอกจากนี้ วิธีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยและวิธีการตรวจสอบไขว้ ไม่มีวิธีใดวิธีหนึ่งที่เหมาะสมกว่าอย่างเด่นชัดกว่ากันเมื่อพิจารณาจากค่าคลาดเคลื่อนของการพยากรณ์และสัมประสิทธิ์การถดถอย

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559

5881509626 : MAJOR STATISTICS
KEYWORDS: HIGH-DIMENSIONAL DATA / LASSO REGRESSION / TUNING PARAMETER / REGRESSION DIAGNOSTICS / CROSS-VALIDATION / BAYESIAN INFORMATION CRITERIA

JUTATIP NUNTASUWAN: ON TUNING PARAMETER SELECTION OF LASSO REGRESSION. ADVISOR: ASST. PROF.VITARA PUNGPAPONG, Ph.D., 60 pp.

This research is aimed to propose a method to select a tuning parameter for lasso regression by using regression diagnostics. Here we compare the results with the two popular approaches in lasso tuning parameter selection including cross-validation and Bayesian Information Criteria. Simulation studies in 6 cases emphasizing on violation of the linearity and homoscedasticity assumptions are carried out. The performance of three methods are compared in terms of false positive rate, false negative rate, prediction error, and estimation error. Our simulation studies show that regression diagnostics approach yields the lowest false positive rates and cross-validation method provides the lower false negative rates than the other two methods. In addition, regression diagnostics and cross-validation methods are comparable in terms of prediction error and estimation error.

Department: Statistics
Field of Study: Statistics
Academic Year: 2016

Student's Signature .....

Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ด้วยดี เพราะความกรุณาและการอนุเคราะห์เป็นอย่างดีจากคณาจารย์และผู้เกี่ยวข้องทุกท่านโดยเฉพาะอย่างยิ่ง ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช และ อาจารย์ ดร. ฐิตารีย์ รุ่งรัตน์เกษม กรรมการสอบวิทยานิพนธ์ ที่ได้ให้คำปรึกษา ความรู้ คำแนะนำ ตลอดจนการเอาใจใส่ในการปรับปรุงงาน ตรวจสอบ แก้ไขข้อบกพร่องต่าง ๆ และส่งเสริมให้กำลังใจเป็นอย่างดีเสมอมา ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณาจารย์ในภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ได้กรุณาประสิทธิ์ประสาทวิชาความรู้ทางคณิตศาสตร์และสถิติ ทำให้ผู้วิจัยสามารถนำความรู้ที่ได้รับไปประยุกต์ใช้ให้เป็นประโยชน์สูงสุด และขอกราบขอบพระคุณบุคลากรทุกท่านในภาควิชาสถิติที่ได้อำนวยความสะดวกในด้านเอกสารและการประสานงานต่าง ๆ

สุดท้ายนี้ขอกราบขอบพระคุณ บิดา มารดา ญาติพี่น้องของผู้วิจัย ที่คอยให้กำลังใจ และส่งเสริมสนับสนุนด้านการเรียนด้วยดีมาโดยตลอด รวมทั้งเพื่อนๆ และทุกคนที่มีส่วนเกี่ยวข้องกับการวิจัยครั้งนี้ที่ได้ให้คำปรึกษา และกาลังใจตลอดระยะเวลาในการทำวิจัยได้เป็นอย่างดี

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การถดถอยเชิงเส้นเป็นกระบวนการทางสถิติที่ใช้ในการวิเคราะห์ข้อมูล เพื่อหาความสัมพันธ์ของตัวแปร 2 ประเภท คือ ตัวแปรอิสระและตัวแปรตาม นอกจากนี้ยังสามารถใช้พยากรณ์ค่าของตัวแปรตามเมื่อทราบค่าความสัมพันธ์ของข้อมูลที่หาได้จากการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัวด้วยวิธีกำลังสองน้อยสุด (Ordinary Least Squares Method) การวิเคราะห์นี้มีข้อจำกัดอยู่หลายประการ และข้อจำกัดที่สำคัญประการหนึ่งคือสามารถใช้วิเคราะห์ได้กับข้อมูลที่มีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระเท่านั้น หากข้อมูลที่นำมาวิเคราะห์มีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ จะเรียกข้อมูลประเภทนี้ว่าข้อมูลที่มีมิติสูง และจะไม่สามารถวิเคราะห์ได้โดยวิธีการวิเคราะห์การถดถอยเชิงเส้นแบบทั่วไป เนื่องจากไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีที่กล่าวมาข้างต้น นอกจากนี้ยังอาจเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์กันเองสูงซึ่งจะส่งผลให้การพยากรณ์เกิดความผิดพลาดสูง และ ปัญหาในการแปรผลลัพธ์ของตัวแบบที่ได้อีกด้วย [1]

การวิเคราะห์ข้อมูลที่มีมิติสูงจะนิยมใช้วิธี Penalized Regression ซึ่งเป็นวิธีที่เพิ่ม Penalty Term เข้าไปในสมการที่ใช้ประมาณค่าสัมประสิทธิ์การถดถอย ซึ่งจะอยู่ในรูปของการดำเนินการทางคณิตศาสตร์ต่อค่าสัมประสิทธิ์การถดถอยและถูกถ่วงน้ำหนักโดยค่าพารามิเตอร์ที่เรียกว่าพารามิเตอร์การปรับ (Tuning Parameter) วิธี Penalized Regression ที่เป็นที่รู้จักและนิยมใช้กันอย่างแพร่หลาย คือวิธี Least Absolute Shrinkage and Selection Operator หรือลาสโซ่ ที่ถูกเสนอโดย Tibshirani ในปี ค.ศ. 1996 [2] เพื่อคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์การถดถอยในคราวเดียวกัน โดยการบีบค่าสัมประสิทธิ์บางตัวให้เป็นศูนย์

ในการวิเคราะห์การถดถอยด้วยวิธีลาสโซ่นั้นการหาค่าพารามิเตอร์การปรับ (Tuning Parameter) ที่เหมาะสมเป็นอีกหนึ่งประเด็นสำคัญที่ต้องคำนึงถึง เนื่องจากอาจส่งผลในเรื่องของการพยากรณ์ได้ โดยทั่วไปการหาพารามิเตอร์การปรับจะนิยมใช้วิธี Cross-Validation (CV) เพื่อลดความผิดพลาดจากการทำนาย [2] นอกจากนี้ยังมีการศึกษาที่พบว่าวิธี Bayesian Information Criterion (BIC) เป็นอีกหนึ่งวิธีที่ใช้ในการประมาณค่าพารามิเตอร์การปรับอย่างมีประสิทธิภาพเช่นกัน [3-5]

จากการทบทวนวรรณกรรมที่ผ่านมา ยังไม่พบว่ามีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยทั้ง 5 ข้อ ซึ่งได้แก่ (1) การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น (2) การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ (3) การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ และ (4) การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน หลังจากการ

วิเคราะห์การถดถอยด้วยวิธีลาสโซ่ ผู้วิจัยจึงสนใจศึกษาเกี่ยวกับประเด็นนี้ และพบว่ามีการละเมิดข้อบังคับเบื้องต้นในเรื่องของฟังก์ชันการถดถอยไม่เป็นเชิงเส้นตรง

วิธีการหนึ่งที่ผู้วิจัยเสนอวิธีหาค่าพารามิเตอร์การปรับเพื่อหลีกเลี่ยงการเกิดการละเมิดข้อบังคับเบื้องต้น คือ การพิจารณาจากการตรวจสอบข้อบังคับเบื้องต้นทั้ง 4 ข้อเป็นหลัก โดยเลือกช่วงของค่าพารามิเตอร์การปรับที่ก่อให้เกิดการละเมิดข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยน้อยที่สุด และทำการเปรียบเทียบวิธีการคัดเลือกพารามิเตอร์การปรับในแต่ละวิธี โดยการวัดอัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate) อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate) ค่าคลาดเคลื่อนจากการพยากรณ์ (Prediction Error) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (Estimation Error) เพื่อหาวิธีการเลือกค่าพารามิเตอร์การปรับที่มีประสิทธิภาพและเหมาะสมที่สุด

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบวิธีการหาค่าพารามิเตอร์การปรับด้วยวิธี Cross-Validation (CV), Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (Regression Diagnosis (RD)) ที่นำเสนอ

1.3 ข้อตกลงเบื้องต้น

ในการศึกษาครั้งนี้จะทำการเปรียบเทียบวิธีการหาค่าค่าพารามิเตอร์การปรับในแบบจำลองเชิงเส้นตรง โดยกำหนดให้ข้อมูลจำลองจากตัวแบบดังนี้

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \ldots (1.1)$$

โดยที่

$\boldsymbol{Y_{n\times 1}}$	เป็นเวกเตอร์ของตัวแปรตาม
$\boldsymbol{X_{n\times p}}$	เป็นเมทริกซ์ของตัวแปรอิสระ
$\boldsymbol{\beta_{p\times 1}}$	เป็นเวกเตอร์ของพารามิเตอร์ในตัวแบบ
$\boldsymbol{\varepsilon_{n\times 1}}$	เป็นเวกเตอร์ของค่าความคลาดเคลื่อน โดย $E(\varepsilon_i) = 0$ และ $Var(\varepsilon_i) = \sigma^2 I_n$

ซึ่งสามารถเขียนได้ในรูปของ

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

1.4 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาโดยการจำลองข้อมูลแบบตัดขวาง (Cross-Sectional Data) ให้มีสถานการณ์ที่จะก่อให้เกิดปัญหาเกี่ยวกับข้อบังคับเบื้องต้น โดยเน้นไปที่การเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ ดังนี้

1. กำหนดขนาดตัวอย่าง (n) เท่ากับ 100 และ จำนวนตัวแปรอิสระ (p) เท่ากับ 1,000

2. ทำการจำลอง (Simulate) ข้อมูลทั้งหมด 6 กรณี ดังต่อไปนี้

กรณีที่ 1 ข้อมูลมีการแจกแจงปกติ

1) กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปรที่มีค่าเฉลี่ยเท่ากับ 0 และค่าความแปรปรวนเท่ากับ 1 โดยกำหนดให้ไม่มีความสัมพันธ์ระหว่างตัวแปรอิสระ x_i และ x_j

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) กำหนดค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 (Sparse Coefficient) ให้มีค่าเท่ากับ 1.5 มีจำนวน 25 ตัว

4) ทำการจำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y = X\beta + \varepsilon$$

สำหรับกรณีที่ 2 – 4 เป็นการจำลองข้อมูลให้เกิดปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ โดยได้แนวคิดในการจำลองข้อมูลมาจาก Dezeure *et al.* [6] ดังนี้

กรณีที่ 2 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นลักษณะ Equal Correlation)

1) กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปรที่มีค่าเฉลี่ยเท่ากับ 0 และค่าความแปรปรวนเท่ากับเมทริกซ์ Σ โดยที่ Σ เป็นเมทริกซ์ที่มีค่าสหสัมพันธ์คงที่ มีค่าดังนี้

$$\Sigma_{j,k} = \begin{cases} 0.8 \ ; j \neq k \\ 1 \ ; \text{ อื่นๆ} \end{cases}$$

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) กำหนดค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 (Sparse Coefficient) ให้มีค่าเท่ากับ 1.5 มีจำนวน 25 ตัว

4) ทำการจำลองตัวแปรตามจากตัวแบบดังนี้

$$Y = X\beta + \varepsilon$$

กรณีที่ 3 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นเมทริกซ์ Toeplitz)

1) กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปรที่มีค่าเฉลี่ยเท่ากับ 0 และค่าความแปรปรวนเท่ากับเมทริกซ์ Σ โดยที่ Σ เป็นเมทริกซ์โทพลิทซ์ มีค่าดังนี้

$$\Sigma_{j,k} = 0.9^{|j-k|} \quad \ldots (1.2)$$

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) กำหนดค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 (Sparse Coefficient) ให้มีค่าเท่ากับ 1.5 มีจำนวน 25 ตัว

4) ทำการจำลองตัวแปรตามจากตัวแบบดังนี้

$$Y = X\beta + \varepsilon$$

กรณีที่ 4 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นลักษณะ Exponential decay)

1) กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปรที่มีค่าเฉลี่ยเท่ากับ 0 และค่าความแปรปรวนเท่ากับเมทริกซ์ Σ โดยที่ Σ เป็นเมทริกซ์ที่มีค่าลดแบบชี้กำลัง มีค่าดังนี้

$$(\Sigma^{-1})_{j,k} = 0.4^{|j-k|/5} \quad \ldots (1.3)$$

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) กำหนดค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 (Sparse Coefficient) ให้มีค่าเท่ากับ 1.5 มีจำนวน 25 ตัว

4) ทำการจำลองตัวแปรตามจากตัวแบบดังนี้

$$Y = X\beta + \varepsilon$$

กรณีที่ 5 ข้อมูลเกิดปัญหาความผิดพลาดในการวัด

1) ทำการจำลองตัวแปรอิสระจากตัวแบบดังนี้

$$X_j = Z_j + \xi_j \quad \ldots (1.4)$$

โดยที่ Z_j มีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระต่อกัน

ξ_j มีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระต่อกัน

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) กำหนดค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 (Sparse Coefficient) ให้มีค่าเท่ากับ 1.5 มีจำนวน 25 ตัว

4) ทำการจำลองตัวแปรตามจากตัวแบบดังนี้

$$Y = Z\beta + \varepsilon \quad \ldots(1.5)$$

กรณีที่ 6 ข้อมูลเกิดปัญหาตัวแปรแฝง

1) ทำการจำลองตัวแปรอิสระจากตัวแบบดังนี้

$$X_j = sign(5.5 - j)Z_1 1_{\{j \le 10\}} + sign(15.5 - j)Z_2 1_{\{11 < j \le 20\}} + Z_3 1_{\{21 < j \le 25\}} + \xi_j \quad \ldots(1.6)$$

โดยที่ Z_1, Z_2, Z_3 มีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระต่อกัน

ξ_j มีการแจกแจงแบบปกติมาตรฐานและเป็นอิสระต่อกัน

2) กำหนดให้ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน

3) ทำการจำลองตัวแปรตามจากตัวแบบดังนี้

$$Y = 1.5Z_1 + 1.5Z_2 + 1.5Z_3 + \varepsilon \quad \ldots(1.7)$$

1.5 คำจำกัดความที่ใช้ในงานวิจัย

1. **ข้อมูลที่มีมิติสูง (High-Dimensional Data)** คือ ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่าง $(p > n)$

2. **ความผิดพลาดในการตรวจจับเชิงบวก (False Positive)** คือ การวัดจำนวนของการเกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าไม่เท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่แท้จริงเท่ากับศูนย์

3. **ความผิดพลาดในการตรวจจับเชิงลบ (False Negative)** คือ การวัดจำนวนของการเกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าเท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่แท้จริงไม่เท่ากับศูนย์

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีที่ใช้หาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุด จะพิจารณาจากอัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate (FPR)) อัตราความผิดพลาดในการตรวจจับเป็นลบ (False Negative Rate (FNR)) ค่าคลาดเคลื่อนจากการพยากรณ์ (Prediction Error (PE)) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (Estimation Error (EE)) ของข้อมูลที่จำลองขึ้นมาทั้งหมด 6 กรณีโดยที่

1. อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate) คือ การวัดความน่าจะเป็นที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์ b_j มีค่าไม่เท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์จริง β_j เท่ากับศูนย์ ซึ่งสามารถคำนวณได้ดังนี้

$$FPR = \frac{\sum_{j=1}^{p} 1_{\{b_j \neq 0 \ and \ \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{b_j \neq 0\}}} \qquad \ldots (1.8)$$

เมื่อ p คือจำนวนตัวแปรอิสระ

2. อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate) คือ การวัดความน่าจะเป็นที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์ b_j มีค่าเท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์จริง β_j ไม่เท่ากับศูนย์ ซึ่งสามารถคำนวณได้ดังนี้

$$FPR = \frac{\sum_{j=1}^{p} 1_{\{b_j = 0 \ and \ \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{b_j = 0\}}} \qquad \ldots (1.9)$$

เมื่อ p คือจำนวนตัวแปรอิสระ

3. ค่าคลาดเคลื่อนจากการพยากรณ์ (Prediction Error) ใช้บ่งบอกว่าค่าพยากรณ์ (Fitted value) ที่ได้จากการวิเคราะห์มีค่าใกล้เคียงกับค่าสังเกตเพียงใด โดยคำนวณได้ดังนี้

$$PE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad \ldots (1.10)$$

เมื่อ p คือจำนวนตัวแปรอิสระ

4. ค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (Estimation Error) ใช้บ่งบอกว่าค่าประมาณสัมประสิทธิ์ b_j ที่ได้จากการวิเคราะห์มีค่าใกล้เคียงกับค่าสัมประสิทธิ์จริง β_j เพียงใด โดยคำนวณได้ดังนี้

$$EE = \sum_{j=1}^{p} |b_j - \beta_j| \qquad \ldots (1.11)$$

เมื่อ p คือจำนวนตัวแปรอิสระ

1.7 วิธีการศึกษา

1. ค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง

2. กำหนดค่าเริ่มต้นสำหรับการจำลองข้อมูลในแต่ละกรณีที่ศึกษา

    2.1 กำหนดขนาดตัวอย่าง n

    2.2 จำนวนตัวแปรอิสระ p

    2.3 กำหนดค่าสัมประสิทธิ์การถดถอยเริ่มต้น β_j สำหรับแต่ละกรณี

3. ทำการจำลองข้อมูลจากค่าเริ่มต้นที่กำหนดให้แตกต่างกันไปทั้งหมด 6 กรณี

4. นำข้อมูลที่ได้จากการจำลองมาทำการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีการหาค่าพารามิเตอร์การปรับต่างๆ ดังนี้

    4.1 Cross-Validation (CV)

    4.2 Bayesian Information Criterion (BIC)

    4.3 วิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

5. ทำการเปรียบเทียบผลการวิเคราะห์โดยใช้อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE) เป็นเกณฑ์การตัดสินใจ และสรุปผล

1.8 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้วิธีการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยด้วยวิธีลาสโซ่

บทที่ 2
ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การวิเคราะห์การถดถอยเชิงเส้นโดยทั่วไป จะนิยมใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method) ในการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบ ซึ่งสามารถวิเคราะห์ได้กับข้อมูลที่มีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระเท่านั้น แต่เนื่องจากปัจจุบันเทคโนโลยีทางด้านวิทยาศาสตร์มีความก้าวหน้า ส่งผลให้มีข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างเกิดขึ้น ข้อมูลประเภทนี้เรียกว่าข้อมูลที่มีมิติสูง ทำให้ไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีกำลังสองน้อยที่สุด (OLS) ดังนั้นในงานวิจัยนี้จะกล่าวถึงวิธีการวิเคราะห์การถดถอยในข้อมูลที่มีมิติสูง นั่นคือ การวิเคราะห์การถดถอยลาสโซ่ (Least Absolute Shrinkage and Selection Operator) และวิธีการหาค่าพารามิเตอร์การปรับในการวิเคราะห์การถดถอยลาสโซ่ ซึ่งประกอบไปด้วยวิธีที่ใช้กันอย่างแพร่หลาย 2 วิธี ได้แก่วิธี Cross-validation (CV) และวิธี Bayesian Information Criteria (BIC) และอีกหนึ่งที่ผู้วิจัยนำเสนอ นั่นคือ วิธีการตรวจสอบข้อบังคับเบื้องต้นของการถดถอย (Regression Diagnosis (RD)) รวมไปถึงเกณฑ์ที่ใช้วัดประสิทธิภาพของผลที่ได้จากการวิเคราะห์การถดถอยด้วยพารามิเตอร์ปรับจากวิธีต่าง ๆ ซึ่งได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าคลาดเคลื่อนจากการพยากรณ์ และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย

2.1 การวิเคราะห์การถดถอย (Regression Analysis)

การวิเคราะห์การถดถอย เป็นวิธีทางสถิติที่ใช้ในการศึกษาความสัมพันธ์ระหว่าง 2 ตัวแปร หรือมากกว่า 2 ตัวแปร โดยการวิเคราะห์จะพิจารณาการพยากรณ์ตัวแปรหนึ่งจากตัวแปรอีกตัวหนึ่ง หรือตัวแปรอีกกลุ่มหนึ่ง ซึ่งตัวแปรที่สนใจเรียกว่า ตัวแปรตาม (Dependent Variable) และตัวแปรอิสระ (Independent Variable) ในการวิเคราะห์การถดถอยหากประกอบไปด้วยตัวแปรอิสระเพียงตัวเดียว จะเรียกว่า การวิเคราะห์การถดถอยอย่างง่าย (Simple Regression Analysis) หากมีตัวแปรอิสระมากกว่า 1 ตัวขึ้นไป จะเรียกว่าการวิเคราะห์การถดถอยเชิงพหุ (Multiple Regression Analysis) โดยมีตัวแบบการถดถอยอยู่ในรูป [7, 8]

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad ; i = 1,2,\ldots,n \qquad \ldots (2.1)$$

โดยที่

Y_i เป็นค่าสังเกตของตัวแปรตาม เมื่อ $i = 1,2,\ldots,n$ และ Y_i เป็นค่าศูนย์กลาง (Centering)

$\beta_1, \beta_2, \ldots, \beta_p$ เป็นค่าสัมประสิทธิ์การถดถอยของตัวแบบ เมื่อ $j = 1,2, \ldots, p$

$X_{i1}, X_{i2}, \ldots, X_{ip}$ เป็นค่าสังเกตของตัวแปรอิสระ เมื่อ $i = 1,2, \ldots, n$ และ X_i เป็นค่ามาตรฐาน (Standardize)

ε_i เป็นค่าความคลาดเคลื่อน โดยมี $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2,$ และ $Cov(\varepsilon_i, \varepsilon_j) = 0$ (สำหรับทุก ๆ ค่าของ i, j เมื่อ $i \neq j$)

หรือเขียนอยู่ในรูปของเมทริกซ์ได้ดังนี้

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad \ldots (2.2)$$

โดยที่

$\boldsymbol{Y_{n\times 1}}$ เป็นเวกเตอร์ของตัวแปรตาม

$\boldsymbol{X_{n\times(p+1)}}$ เป็นเมทริกซ์ของตัวแปรอิสระ

$\boldsymbol{\beta_{(p+1)\times 1}}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยในตัวแบบ

$\boldsymbol{\varepsilon_{n\times 1}}$ เป็นเวกเตอร์ของค่าความคลาดเคลื่อน โดย $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 I_n$

ซึ่งสามารถเขียนได้ดังนี้

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

การประมาณค่าพารามิเตอร์จะใช้วิธีกำลังสองน้อยที่สุดแบบทั่วไป (Ordinary Least Squares Method: OLS) โดยที่ $b_0, b_1, \ldots, b_p$ เป็นตัวประมาณของ $\beta_0, \beta_1, \ldots, \beta_p$ ตามลำดับ

ให้ Q เป็นค่าผลรวมความแตกต่างกำลังสองระหว่าง Y_i และค่าเฉลี่ยของ Y_i

ดังนั้น

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}))^2 \qquad \ldots (2.3)$$

หาค่า $b_0, b_1, \ldots, b_p$ ซึ่งเป็นตัวประมาณของ $\beta_0, \beta_1, \ldots, \beta_p$ ตามลำดับ โดยการหาอนุพันธ์ Q เทียบกับพารามิเตอร์ของตัวแบบ จะได้

$$\frac{\partial Q}{\partial \beta_0|_{b_0,..,b_p}} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1X_{i1} - \ldots - b_pX_{ip}\right) = 0$$

$$\frac{\partial Q}{\partial \beta_1|_{b_0,..,b_p}} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1X_{i1} - \ldots - b_pX_{ip}\right)X_{i1} = 0$$

$$\vdots$$

$$\frac{\partial Q}{\partial \beta_p|_{b_0,..,b_p}} = -2\sum_{i=1}^{n}\left(Y_i - b_0 - b_1X_{i1} - \ldots - b_pX_{ip}\right)X_{ip} = 0$$

จะได้ว่า

$$nb_0 + b_1\sum_{i=1}^{n}X_{i1} + b_2\sum_{i=1}^{n}X_{i2} + \ldots + b_p\sum_{i=1}^{n}X_{ip} = \sum_{i=1}^{n}Y_i$$

$$b_0\sum_{i=1}^{n}X_{i1} + b_1\sum_{i=1}^{n}X_{i1}^2 + b_2\sum_{i=1}^{n}X_{i1}X_{i2} + \ldots + b_p\sum_{i=1}^{n}X_{i1}X_{ip} = \sum_{i=1}^{n}X_{i1}Y_i$$

$$\vdots$$

$$b_0\sum_{i=1}^{n}X_{ip} + b_1\sum_{i=1}^{n}X_{i1}X_{ip} + \ldots + b_p\sum_{i=1}^{n}X_{ip}^2 = \sum_{i=1}^{n}X_{ip}Y_i$$

เมื่อแก้สมการ จะได้ค่าสัมประสิทธิ์การถดถอย $b_0, b_1, \ldots, b_p$ ตามลำดับ หรือเขียนให้อยู่ในรูป

$$b = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left\|Y_i - \sum_{j=1}^{p}\beta_jX_{ij}\right\|^2\right\} \qquad \ldots (2.4)$$

หรือเขียนให้อยู่ในรูปเมทริกซ์ได้ว่า

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \qquad \ldots (2.5)$$

จะได้ $\boldsymbol{b}$ ที่สามารถประมาณค่าพยากรณ์ $\hat{\boldsymbol{Y}}$ (Fitted value) ซึ่งคำนวณได้จาก

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b} \qquad \ldots (2.6)$$

และค่าเศษเหลือ (Residual) $\boldsymbol{e}$ โดยที่

$$e_i = Y_i - \hat{Y}_i \qquad \ldots (2.7)$$

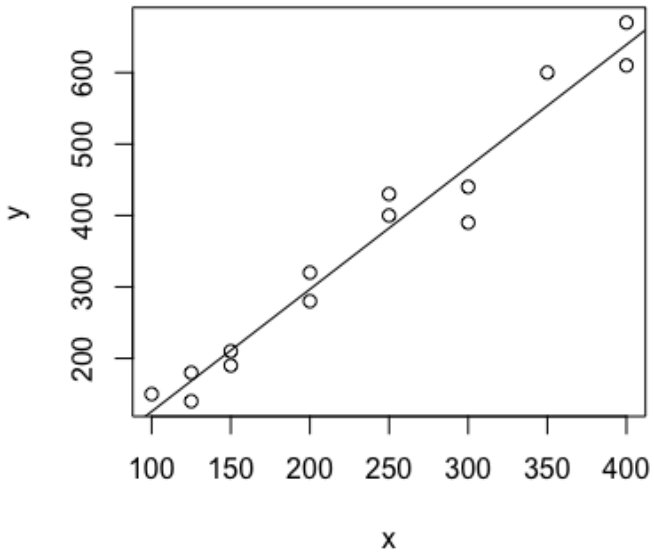
2.2 การตรวจสอบเงื่อนไขของการวิเคราะห์การถดถอย (Regression Diagnosis)

จากตัวแบบการถดถอยที่ได้จากการวิเคราะห์การถดถอย เราไม่สามารถสรุปได้อย่างแน่นอนว่าตัวแบบดังกล่าวมีความเหมาะสม การตรวจสอบตัวแบบการถดถอยว่ามีความเหมาะสมหรือไม่ อาจเริ่มต้นด้วยการเขียนแผนภาพการกระจายระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y) และพิจารณาว่าแนวโน้มของข้อมูลเป็นอย่างไร อย่างไรก็ตาม ในการวิเคราะห์การถดถอย ตัวแปรตาม (Y) เป็นฟังก์ชันของตัวแปรอิสระ (X) ส่งผลให้การเขียนแผนภาพดังกล่าวอาจไม่มีประโยชน์มากนัก ดังนั้นในการตรวจสอบตัวแบบการถดถอยจึงอาจทำได้โดยการพิจารณาค่าเศษเหลือ (Residuals) โดยการเขียนแผนภาพการกระจายหรือใช้วิธีการคำนวณและตรวจสอบความเหมาะสมของตัวแบบด้วยการทดสอบสมมติฐาน

การตรวจสอบเงื่อนไขเกี่ยวกับการวิเคราะห์การถดถอยที่สำคัญมีทั้งหมด 4 ข้อ ดังนี้

2.2.1 การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น

พิจารณาจากแผนภาพการกระจายระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y) เพื่อดูแนวโน้มของข้อมูลเป็นเส้นตรงหรือไม่ [7] ดังตัวอย่างในรูปที่ 2.1

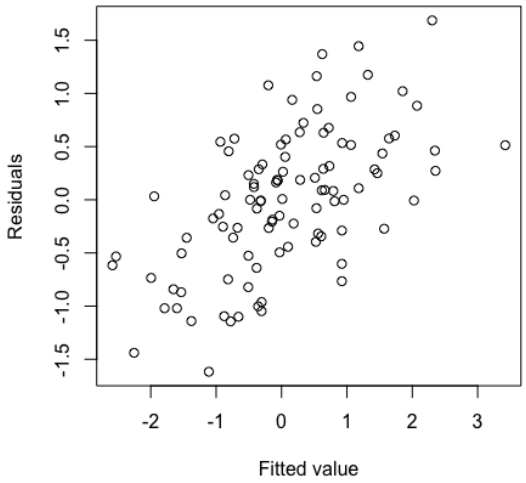


รูปที่ 2. 1 แสดงแผนภาพการกระจายระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y)

จากรูปที่ 2.1 จะเห็นได้ว่าแผนภาพการกระจายแสดงให้เห็นว่าตัวแปรตามมีแนวโน้มเพิ่มขึ้นเมื่อตัวแปรต้นเพิ่มขึ้นในลักษณะเชิงเส้นตรง ดังนั้น จึงสรุปได้ว่าฟังก์ชันการถดถอยมีลักษณะเป็นเชิงเส้น

อย่างไรก็ตามการพิจารณาแผนภาพการกระจายระหว่างตัวแปรอิสระ (X) และตัวแปรตาม (Y) เพียงค่าเดียวไม่อาจตัดสินใจได้ว่าตัวแบบมีลักษณะเป็นเชิงเส้นหรือไม่ ซึ่งอาจพิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าพยากรณ์ $(\hat{Y})$ ว่าเป็นไปอย่างสุ่มหรือไม่ หาก

มีการเพิ่มขึ้นหรือลดลงอย่างมีรูปแบบ แสดงว่าฟังก์ชันการถดถอยไม่เป็นเชิงเส้น ดังตัวอย่างในรูปที่ 2.2

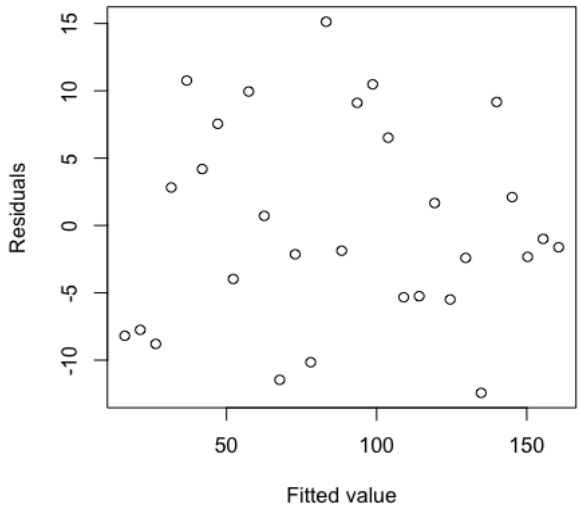


รูปที่ 2. 2 แสดงแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) *และค่าพยากรณ์* $(\hat{Y})$

จากรูปที่ 2.2 จะเห็นได้ว่าแผนภาพการกระจายไม่ได้เป็นไปอย่างสุ่ม กล่าวคือ ค่าเศษเหลือมีค่าเพิ่มขึ้นเมื่อค่าพยากรณ์มีค่าเพิ่มขึ้น ดังนั้น จึงสรุปได้ว่าฟังก์ชันการถดถอยไม่เป็นเชิงเส้น

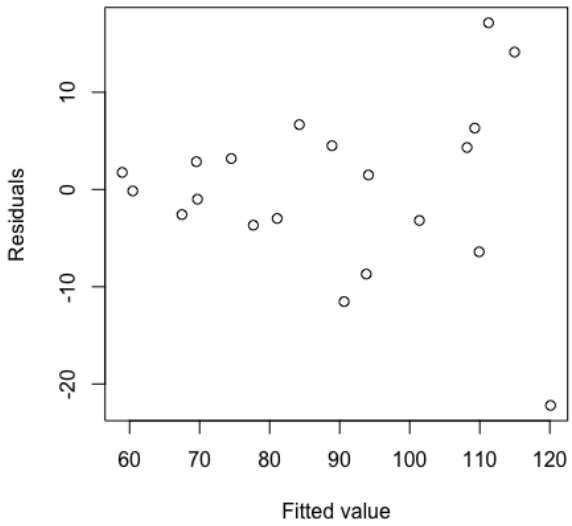
2.2.2 การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่

1) พิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าพยากรณ์ $(\hat{Y})$ ว่าเป็นไปอย่างสุ่มหรือไม่ หากมีค่าเศษเหลือมีแนวโน้มที่จะเพิ่มขึ้นหรือลดลงแปรผันตามการเพิ่มขึ้นหรือลดลงของค่าพยากรณ์ อยู่ในลักษณะการกระจายเป็นรูปคล้ายลำโพง แสดงว่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ ซึ่งจะเรียกว่า Heteroscedasticity



รูปที่ 2. 3 แสดงแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) *และค่าพยากรณ์* $(\hat{Y})$

จากรูปที่ 2.3 จะเห็นได้ว่าแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าพยากรณ์ $(\hat{Y})$ เป็นไปอย่างสุ่ม ดังนั้น จึงสรุปได้ว่าไม่เกิดปัญหา Heteroscedasticity หรือเรียกว่า Homoscedasticity



รูปที่ 2. 4 แสดงแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าพยากรณ์ $(\hat{Y})$

จากรูปที่ 2.4 จะเห็นได้ว่าแผนภาพมีรูปแบบเป็นรูปลำโพง ดังนั้น จึงสรุปได้ว่าเกิดปัญหา Heteroscedasticity ทั้งนี้การเกิดปัญหาค่าความแปรปรวนมีค่าไม่คงที่ มักจะเกิดในกรณีที่ใช้ข้อมูลแบบตัดขวาง (Cross-Sectional Data) มากกว่าข้อมูลที่เป็นอนุกรมเวลา (Time-Series Data)

2) ทำการตรวจสอบหานัยสำคัญทางสถิติ ซึ่งมีวิธีตรวจสอบได้หลากหลายวิธี เช่น Glejser's Test, Bartlett Test, Goldfeld-Quandt Test, Breusch-Pagan Test และ White's Test เป็นต้น ในที่นี้จะนำเสนอเฉพาะวิธีการทดสอบของ Breusch-Pagan Test [8]

โดยสามารถเขียนสมมติฐานของการทดสอบได้ดังนี้

$$H_0: Var(\varepsilon_i) = \sigma^2 \ ; \ \forall i = 1,2,\dots,n$$

$$H_1: not\ H_0$$

ขั้นตอนในการวิเคราะห์ Breusch-Pagan Test มีดังนี้

- จากค่าเศษเหลือ $e_i = Y_i - \hat{Y}_i \ ; \ i = 1,2,\dots,n$ ที่ได้จากตัวแบบการถดถอย นำมากำหนดตัวแบบการถดถอยใหม่ โดยกำหนดให้ค่าเศษเหลือกำลังสอง $(e_i{}^2)$ เป็นตัวแปรตาม และกำหนดให้ตัวแปรอิสระทั้งหมดจากตัวแบบการถดถอย เป็นตัวแปรอิสระของตัวแบบการถดถอยใหม่ ดังนี้

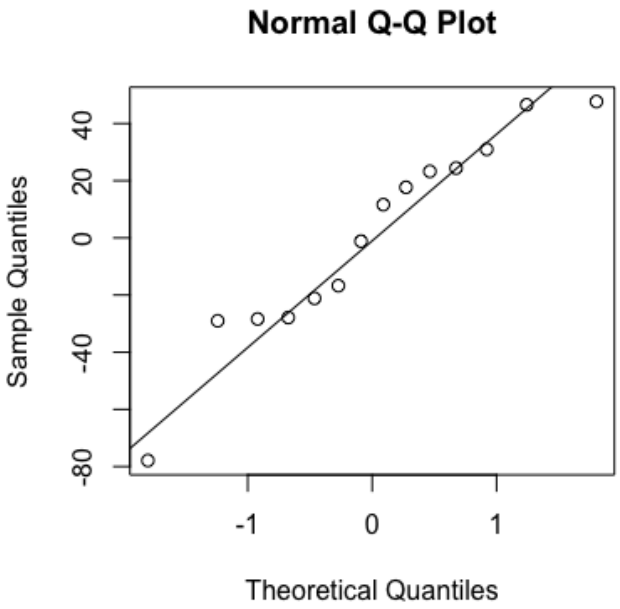
$$e_i{}^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_p X_{ip} + u_i \ ; i = 1,2,\dots,n \quad \dots (2.8)$$

โดยที่ u_i เป็นค่าความคลาดเคลื่อน โดยมี $E(u_i) = 0, Var(u_i) = \sigma^2$, และ $Cov(u_i, u_j) = 0$ (สำหรับทุก ๆ ค่าของ i, j เมื่อ $i \neq j$) ; $i = 1,2,\dots,n$

- ทำการวิเคราะห์การถดถอยจากตัวแบบใหม่ข้างต้น คำนวณค่าสัมประสิทธิ์การตัดสินใจ R^2
- ทำการทดสอบสมมติฐานเชิงสถิติ ดังนี้
  - สมมติฐานเชิงสถิติ $H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$
    $H_1: not\ H_0$
  - กำหนดระดับนัยสำคัญ α
  - ตัวสถิติทดสอบคือ $LM_c = nR^2 \sim {\chi_p}^2$
  - ทำการตัดสินใจ หาก $LM_c > {\chi_{\alpha,p}}^2$ จะปฏิเสธสมมติฐาน H_0

2.2.3 การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ

1) พิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าคาดหวังของค่าเศษเหลือ $E(e_i)$ โดยมีข้อสมมติว่า ถ้าการแจกแจงของค่าความคลาดเคลื่อนมีการแจกแจงปกติแล้ว กราฟจะเป็นลักษณะเส้นตรง ดังตัวอย่างในรูปที่ 2.5



รูปที่ 2.5 แสดงแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าคาดหวังของค่าเศษเหลือ $E(e_i)$

จากรูปที่ 2.5 จะเห็นได้ว่าแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าคาดหวังของค่าเศษเหลือ $E(e_i)$ มีแนวโน้มเป็นเส้นตรง ดังนั้น จึงสรุปได้ว่าค่าความคลาดเคลื่อนมีการแจกแจงปกติ [7, 8]

2) ทำการตรวจสอบหานัยสำคัญทางสถิติ ซึ่งมีวิธีตรวจสอบได้หลากหลายวิธี เช่น Shapiro-Wilk Test, Anderson-darling Test, Kolmogorov-Smirnov Test และ Goodness-of-fit Test เป็นต้น ในที่นี้จะนำเสนอเฉพาะวิธีการทดสอบของ Shapiro-Wilk Test [9]

โดยสามารถเขียนสมมติฐานของการทดสอบได้ดังนี้

$$H_0: \varepsilon_i \sim \text{Normal Distribution} \ ; \ i = 1,2,\dots,n$$

$$H_1: not\ H_0$$

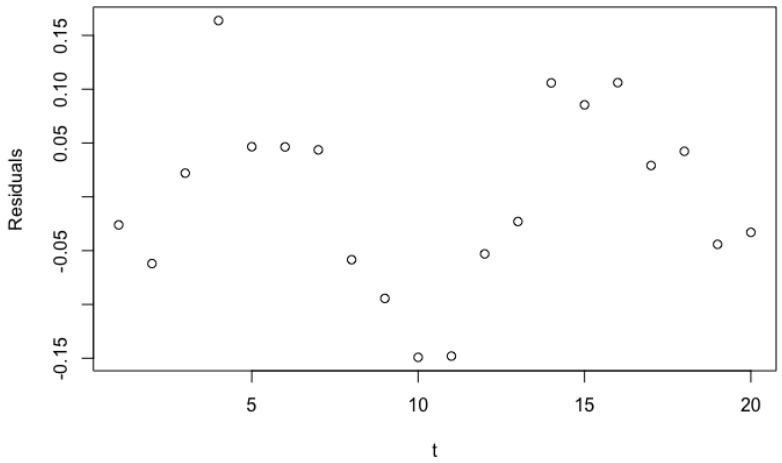
ขั้นตอนในการวิเคราะห์ Shapiro-Wilk Test มีดังนี้

- จากค่าเศษเหลือ $e_i = Y_i - \hat{Y}_i \ ; \ i = 1,2,\dots,n$ ที่ได้จากตัวแบบการถดถอย ทำการเรียงค่าเศษเหลือจากน้อยไปหามาก
- ตัวสถิติทดสอบคือ $W = \frac{\sum_{i=1}^{k} a_{n-i+1}(e_{n-i+1}-e_i)}{\sum_{i=1}^{n}(e_i-\bar{e})^2}$

  เมื่อ $k = \begin{cases} \frac{n}{2} & \text{เมื่อ } n \text{ เป็นเลขคู่} \\ \frac{n-1}{2} & \text{เมื่อ } n \text{ เป็นเลขคี่} \end{cases}$ และ a_{n-i+1} คือค่าสัมประสิทธิ์ของ W-Test
- ทำการตัดสินใจจากค่า $p-value$ ที่ได้จากตาราง W-Test หาก $p-value < \alpha$ จะปฏิเสธสมมติฐาน H_0

2.2.4 การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน

1) ในกรณีที่เก็บรวบรวมข้อมูลตามลำดับเวลา จะพิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และเวลา (t) ว่าเป็นไปอย่างสุ่มหรือมีการเพิ่มขึ้นหรือลดลงในลักษณะวัฏจักร ดังตัวอย่าง ในรูปที่ 2.6



รูปที่ 2. 6 แสดงแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และเวลา (t)

จากรูปที่ 2.6 จะเห็นได้ว่าความสัมพันธ์ระหว่างค่าเศษเหลือกับเวลามีลักษณะเพิ่มขึ้นและลดลงเป็นวัฏจักร ดังนั้น จึงสรุปได้ว่าเกิดปัญหาค่าความคลาดเคลื่อนไม่เป็นอิสระต่อกัน [7, 8]

2) ทำการตรวจสอบหานัยสำคัญทางสถิติ ซึ่งมีวิธีตรวจสอบได้หลากหลายวิธี เช่น Frequency Test, Serial Test, Gap Test, Permutation Test, Runs Test และ Serial Correlation Test เป็นต้น ในที่นี้จะนำเสนอเฉพาะวิธีการทดสอบของ Runs Test [9]

โดยสามารถเขียนสมมติฐานของการทดสอบได้ดังนี้

$$H_0: \varepsilon_i \sim independent \ ; \ i = 1,2,\dots,n$$

$$H_1: not\ H_0$$

ขั้นตอนในการวิเคราะห์ Runs Test มีดังนี้

- จากค่าเศษเหลือ $e_i = Y_i - \hat{Y}_i \ ; \ i = 1,2,\dots,n$ ที่ได้จากตัวแบบการถดถอย พิจารณาลำดับการรันส์ของค่าเศษเหลือ ในที่นี้จะพิจารณาเปรียบเทียบจากค่าเฉลี่ยของค่าเศษเหลือ โดยข้อมูลแบ่งออกเป็น 2 กลุ่มคือ กลุ่มที่มีค่าเป็นบวก และกลุ่มที่มีค่าเป็นลบ
- ตัวสถิติทดสอบคือ T จำนวนรันส์ของค่าเศษเหลือ

$$Z = \frac{T - \mu_T}{\sigma_T} \ ; \ \mu_T = \frac{2n_1n_2}{n} + 1 \ , \ \sigma_T = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$$

- ทำการตัดสินใจ หาก $Z < Z_{\alpha/2}$ หรือ $Z > Z_{1-\alpha/2}$ จะปฏิเสธสมมติฐาน H_0

2.3 การถดถอยแบบลาสโซ่ (Least Absolute Shrinkage and Selection Operator (Lasso))

เนื่องจากการประมาณค่าพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุดแบบทั่วไป (OLS) มีปัญหาในการวิเคราะห์ข้อมูลที่มีมิติสูง $(p > n)$ อยู่ 2 ประการ นั่นคือปัญหาการแปลผลลัพธ์ของตัวแบบ (Interpretation) ในกรณีของข้อมูลที่ตัวแปรอิสระมีจำนวนมาก ส่งผลให้การแปรผลเกิดความยุ่งยากและซับซ้อน และปัญหาด้านความถูกต้องในการทำนาย (Prediction Accuracy) วิธีกำลังสองน้อยที่สุดเป็นตัวประมาณที่มีความเอนเอียง (bias) ต่ำ แต่มีค่าความแปรปรวน (Variance) สูง [2] ดังนั้นการถดถอยลาสโซ่จึงถูกพัฒนาเพื่อใช้ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง โดยการบีบให้ค่าสัมประสิทธิ์ b_j ส่วนใหญ่เป็นศูนย์ และ b_j บางส่วนไม่เท่ากับศูนย์ (Sparse Estimator) ดังนั้นวิธีนี้จะสามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์ b_j ได้ในคราวเดียวกัน ซึ่งจะสามารถหาค่าประมาณ b_j ได้ดังนี้

$$b = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left\| Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad \dots (2.8)$$

โดยที่

Y_i เป็นค่าสังเกตของตัวแปรตาม เมื่อ $i = 1,2,\dots,n$ และ Y_i เป็นค่าศูนย์กลาง (Centering)

X_{ij} เป็นค่าสังเกตของตัวแปรอิสระ เมื่อ $i = 1,2,\dots,n$ และ $j = 1,2,..,p$ และ X_{ij} เป็นค่ามาตรฐาน (Standardize)

β_j เป็นค่าสัมประสิทธิ์การถดถอยของตัวแบบ เมื่อ $j = 1,2,..,p$

λ เป็นพารามิเตอร์การปรับ (Tuning Parameter) โดยที่ $\lambda \geq 0$

อย่างไรก็ตามการวิเคราะห์การถดถอยด้วยวิธีนี้ยังมีข้อจำกัดอีกเล็กน้อย นั่นคือวิธีลาสโซ่สามารถเลือกตัวแปรเข้าสู่ตัวแบบได้มากที่สุดเท่ากับจำนวนของค่าสังเกต n ตัว $(p = n)$ และหากตัวแปรอิสระมีความสัมพันธ์กันสูง วิธีลาสโซ่มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวเดียวจากกลุ่มตัวแปรนั้นๆ ดังนั้นหากข้อมูลที่นำมาวิเคราะห์มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมากๆ หรือตัวแประอิสระมีความสัมพันธ์กันเองสูง ตัวแบบที่ได้จากการวิเคราะห์ด้วยวิธีลาสโซ่ก็อาจไม่มีความเหมาะสม [1]

2.4 วิธีเลือกพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่

การหาค่าพารามิเตอร์การปรับที่เหมาะสมเป็นสิ่งที่สำคัญในการวิเคราะห์การถดถอยลาสโซ่และการประมาณค่าสัมประสิทธิ์ b_j สำหรับวิธีการเลือกพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่ที่ใช้กันแพร่หลายมี 2 วิธีด้วยกัน คือวิธี Cross-validation (CV) และวิธี Bayesian Information Criterion (BIC) ดังนี้

2.4.1 วิธี Cross-Validation (CV)

เป็นวิธีการวัดประสิทธิภาพของตัวแบบ โดยแบ่งข้อมูลออกเป็น 2 ส่วน คือข้อมูลชุดการเรียนรู้ (Training Data) ใช้สร้างตัวแบบการถดถอย และข้อมูลชุดตรวจสอบ (Testing Data) ใช้ตรวจสอบตัวแบบการถดถอยที่ได้จากชุดการเรียนรู้ โดยพิจารณาจากค่าความผิดพลาดในการทำนาย

วิธี Cross-Validation เป็นวิธีที่นิยมใช้กันโดยทั่วไปในการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ โดยการแบ่งกลุ่มของชุดข้อมูลออกเป็นกลุ่มย่อย k กลุ่ม จากนั้นนำข้อมูลเพียง $k-1$ กลุ่ม (Training Data) มาทำนายกลุ่มที่เหลืออีก 1 กลุ่ม (Testing Data) ทำเช่นนี้ไป k ครั้ง แล้วพิจารณาค่าความผิดพลาดในการทำนาย โดยทั่วไปแล้วจะนิยมใช้ $k = 5$ หรือ 10 มีขั้นตอนในการคำนวณดังนี้ [10, 11]

1) กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1, \dots, \lambda_m\}$

2) แบ่งกลุ่มของข้อมูลทั้งหมด n ชุด ออกเป็น k กลุ่มๆละเท่าๆกัน

3) สำหรับ $k = 1,2,\dots,K$ ทำการทำนายดังนี้

ครั้งที่ k ใช้ข้อมูลจำนวน $K-1$ ชุด คือชุดที่ 1 ถึงชุดที่ K ใดๆ ที่ไม่ใช่ชุดที่ k

a) ในแต่ละ λ_l ; $l = 1,2,\dots,m$ ทำการประมาณค่าพารามิเตอร์ $\{b_j\}_l$ ในข้อมูลจำนวน $k-1$ ชุดที่เลือกมา เมื่อได้ค่าประมาณ $\{b_j\}_l$ แล้ว นำไปหาค่าพยากรณ์ $\{\hat{Y}_i\}_l$ ของข้อมูลชุดที่ k

b) ทำการคำนวณหาค่าความผิดพลาดจาก

$$e_k(\lambda_l) = \sum_{i=1}^{n_k} (Y_i - \{\hat{Y}_i\}_l)^2 \qquad \dots (2.9)$$

4) ในแต่ละ λ_l ; $l = 1,2,\dots,m$ คำนวณค่าความผิดพลาดเฉลี่ยของ λ_l จากทั้งหมด k ครั้ง ดังนี้

$$CV(\lambda_l) = \frac{1}{n}\sum_{k=1}^{K} e_k(\lambda_l) \qquad \dots (2.10)$$

5) ตัว λ_l ที่มีค่า $CV(\lambda_l)$ น้อยที่สุด จะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

2.4.2 วิธี Bayesian Information Criterion (BIC)

วิธี Bayesian Information Criterion (BIC) หรือวิธี Schwarz Information Criterion เป็นวิธีที่ถูกคิดค้นโดย Gideon Schwarz ในปี ค.ศ. 1978 [12] เพื่อใช้ในการคัดเลือกตัวแบบของการวิเคราะห์การถดถอยเชิงเส้นทั่วไป ซึ่งเป็นการดัดแปลงจากเกณฑ์ Akaike Information Criterion (AIC) โดยพิจารณาจากค่าสูงสุดของความน่าจะเป็นภายหลัง (Posterior Probability) ของตัวแบบการถดถอยนั้นๆ

ต่อมาถูกพัฒนามาใช้กับข้อมูลที่มีมิติสูงโดย Hansheng Wang และคณะ [4] ในการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยในข้อมูลมิติสูง เพื่อให้ได้ตัวแบบการถดถอยที่มีความคงเส้นคงวา (Consistency) มีขั้นตอนในการคำนวณดังนี้

1) กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1, \dots, \lambda_m\}$

2) ในแต่ละ λ_l ; $l = 1,2,\dots,m$ ทำการประมาณค่าพารามิเตอร์ $\{b_j\}_l$ และประมาณค่า BIC โดยคำนวณจาก

$$BIC_l = \log(\hat{\sigma}_l^2) + |S_l| \times \frac{\log n}{n} \times C_n \qquad \ldots (2.11)$$

เมื่อ $\hat{\sigma}_l^2 = SSE_l/n$

$C_n = \log\log p$

S_l คือเซต $\{b_j\}_l$ ที่ไม่เท่ากับ 0

3) ตัว λ_l ที่มีค่า BIC_l น้อยที่สุด จะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

2.5 วิธีการหาพารามิเตอร์การปรับโดยพิจารณาจากการตรวจสอบข้อบังคับเบื้องต้น (Regression Diagnostics (RD))

งานวิจัยนี้นำเสนอการใช้การตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยในการเลือกพารามิเตอร์ปรับ ซึ่งมีเกณฑ์ในการเลือกพารามิเตอร์การปรับโดยอ้างอิงจากการตรวจสอบข้อบังคับเบื้องต้นของการถดถอยทั้ง 4 ข้อ มีรายละเอียดขั้นตอนดังนี้

1) กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$

2) ในแต่ละ λ_l ; $l = 1,2, \ldots, m$ ทำการวิเคราะห์การถดถอยลาสโซ่

3) ในแต่ละ λ_l ; $l = 1,2, \ldots, m$ ทำการตรวจสอบข้อบังคับเบื้องต้นของการถดถอยทั้ง 4 ข้อ หลังจากทำการวิเคราะห์การถดถอยลาสโซ่ โดยมีเกณฑ์ในการตัดสินใจดังนี้

(1) การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น พิจารณาจากแนวโน้มของแผนภาพการกระจายระหว่างค่าเศษเหลือ (e_i) และค่าพยากรณ์ $(\hat{Y})$ หรือ Residuals plot และค่าสหสัมพันธ์ระหว่างค่าเศษเหลือและค่าพยากรณ์ (r)

(2) การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ ใช้ตัวสถิติทดสอบ Breusch-Pagan Test โดยมีสมมติฐานของการทดสอบดังนี้

$H_0: Var(\varepsilon_i) = \sigma^2$; $\forall i = 1,2, \ldots, n$

$H_1: not\ H_0$

ที่ระดับนัยสำคัญ 0.05

(3) การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน ใช้ตัวสถิติทดสอบ Runs Test โดยมีสมมติฐานของการทดสอบดังนี้

$$H_0: \varepsilon_i \sim independent \ ; \ i = 1,2, \dots, n$$

$$H_1: not \ H_0$$

ที่ระดับนัยสำคัญ 0.05

(4) การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ โดยใช้ตัวสถิติทดสอบ Shapiro-wilk Test โดยมีสมมติฐานของการทดสอบดังนี้

$$H_0: \varepsilon_i \sim \text{Normal Distribution} \ ; \ i = 1,2, \dots, n$$

$$H_1: not \ H_0$$

ที่ระดับนัยสำคัญ 0.05

4) ตัว λ_l ที่ไม่ปฏิเสธสมมติฐานหลักของการตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่, การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกันและการตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ ที่ระดับนัยสำคัญ 0.05 และมีค่าสหสัมพันธ์ระหว่างค่าเศษเหลือและค่าพยากรณ์ (r) น้อยที่สุดจะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

2.6 เกณฑ์การตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการหาค่าพารามิเตอร์การปรับวิธีใดมีความเหมาะสมในการหาค่าพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่ที่สุด จะพิจารณาจากอัตราความผิดพลาดในการตรวจจับเชิงบวก อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าคลาดเคลื่อนจากการพยากรณ์ และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย ดังนี้

2.6.1 อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate)

เป็นการวัดความน่าจะเป็นที่เกิดความผิดพลาดจากการประมาณค่าสัมประสิทธิ์ b_j ไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์จริง β_j เท่ากับศูนย์ ซึ่งถือเป็นความน่าจะเป็นของการเกิดความผิดพลาดประเภทที่ 1 (Type I error) โดยคำนวณได้ดังนี้

$$FPR = \frac{\sum_{j=1}^{p} 1_{\{b_j \neq 0 \ and \ \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{b_j \neq 0\}}}$$

2.6.2 อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate)

เป็นการวัดความน่าจะเป็นที่เกิดความผิดพลาดจากการประมาณค่าสัมประสิทธิ์ b_j เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์จริง β_j ไม่เท่ากับศูนย์ ซึ่งถือเป็นความน่าจะเป็นของการเกิดความผิดพลาดประเภทที่ 2 (Type II error) โดยคำนวณได้ดังนี้

$$FNR = \frac{\sum_{j=1}^{p} 1_{\{b_j=0 \; and \; \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{b_j=0\}}}$$

2.6.3 ค่าความผิดพลาดในการพยากรณ์ (Prediction Error)

ใช้บ่งบอกว่าค่าพยากรณ์ (Fitted value) ที่ได้จากการวิเคราะห์มีค่าใกล้เคียงกับค่าสังเกตเพียงใด หากค่าความผิดพลาดในการพยากรณ์เข้าใกล้ศูนย์ก็ยิ่งแสดงถึงความถูกต้องแม่นยำ โดยคำนวณได้ดังนี้

$$PE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

2.6.4 ค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอย (Estimation Error)

ใช้บ่งบอกว่าค่าประมาณสัมประสิทธิ์ b_j ที่ได้จากการวิเคราะห์มีค่าใกล้เคียงกับค่าสัมประสิทธิ์จริง β_j เพียงใด หากค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอยเข้าใกล้ศูนย์ก็ยิ่งแสดงถึงความถูกต้องแม่นยำ โดยคำนวณได้ดังนี้

$$EE = \sum_{i=1}^{n} |b_j - \beta_j|$$

บทที่ 3
วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการหาค่าพารามิเตอร์การปรับทั้ง 3 วิธี ซึ่งได้แก่วิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) ซึ่งทำการศึกษาในกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนของตัวแปรอิสระ โดยมีการจำลองข้อมูลแบบตัดขวาง (Cross-Sectional Data) ที่แตกต่างกันใน 6 กรณี ซึ่งในการเปรียบเทียบว่าวิธีการใดทำการประมาณค่าค่าพารามิเตอร์การปรับได้เหมาะสมกว่ากัน จะพิจารณาจาก 4 เกณฑ์ คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE)

สำหรับการจำลองข้อมูลและการวิเคราะห์ข้อมูลทั้งหมดจะทำงานด้วยโปรแกรม RStudio เวอร์ชัน 0.99.903 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาในส่วนของข้อมูลจำลองทั้งหมด 6 กรณี โดยกำหนดให้มีสถานการณ์ที่ก่อให้เกิดปัญหาเกี่ยวกับข้อบังคับเบื้องต้น โดยเน้นที่การเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ ภายใต้ขอบเขตการวิจัยดังต่อไปนี้

1. กำหนดขนาดตัวอย่าง (n) เท่ากับ 100 และ จำนวนตัวแปรอิสระ (p) เท่ากับ 1,000

2. ทำการจำลองข้อมูล (Simulate) ดังต่อไปนี้

กรณีที่ 1 ข้อมูลมีการแจกแจงปกติ

1) กำหนดตัวแปรอิสระมีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution)

$$X_i \overset{iid}{\sim} N_p(0, I_p)$$

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) กำหนดค่าสัมประสิทธิ์ (β_j) ของตัวแปรอิสระมีค่าเป็น 1.5 สำหรับตัวแปรอิสระ 25 ตัวแรก $(\beta_1, \beta_2, \dots, \beta_{25})$ และของตัวแปรที่เหลือมีค่าเป็น 0 $(\beta_{26}, \beta_{27}, \dots, \beta_{1,000})$ ดังนี้

4) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = X_i\beta_j + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,1000} \\ \vdots & \ddots & \vdots \\ X_{100,1} & \cdots & X_{100,1000} \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} 1.5 \\ \vdots \\ 1.5 \end{matrix}\right\} \text{25 ตัว} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{975 ตัว} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

กรณีที่ 2 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นลักษณะ Equal Correlation)

1) กำหนดตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ และเมทริกซ์ความแปรปรวนร่วม $\Sigma_{p\times p}$

$$X_i \sim N_p(0, \Sigma)$$

โดยที่ $\Sigma_{p\times p}$ เป็นเมทริกซ์ที่มีค่าความสัมพันธ์คงที่

$$\Sigma = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$$

(โดยกำหนดระดับความสัมพันธ์ $\rho = 0.8$)

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) กำหนดค่าสัมประสิทธิ์ (β_j) ของตัวแปรอิสระมีค่าเป็น 1.5 สำหรับตัวแปรอิสระ 25 ตัวแรก $(\beta_1, \beta_2, \dots, \beta_{25})$ และของตัวแปรที่เหลือมีค่าเป็น 0 $(\beta_{26}, \beta_{27}, \dots, \beta_{1,000})$

4) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = X_i\beta_j + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,1000} \\ \vdots & \ddots & \vdots \\ X_{100,1} & \cdots & X_{100,1000} \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} 1.5 \\ \vdots \\ 1.5 \end{matrix}\right\} \text{25 ตัว} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{975 ตัว} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

กรณีที่ 3 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นเมทริกซ์ Toeplitz)

1) กำหนดตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ และเมทริกซ์ความแปรปรวนร่วม $\Sigma_{p\times p}$

$$X_i \sim N_p(0, \Sigma)$$

โดยที่ $\Sigma_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์

$$\Sigma = \begin{bmatrix} 1 & \Sigma_{1,2} & \cdots & \Sigma_{1,p} \\ \Sigma_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \Sigma_{j,k} \\ \Sigma_{p,1} & \cdots & \Sigma_{j,k} & 1 \end{bmatrix}$$

(โดยกำหนด $\Sigma_{j,k} = 0.9^{|j-k|}$)

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) กำหนดค่าสัมประสิทธิ์ (β_j) ของตัวแปรอิสระมีค่าเป็น 1.5 สำหรับตัวแปรอิสระ 25 ตัวแรก $(\beta_1, \beta_2, \ldots, \beta_{25})$ และของตัวแปรที่เหลือมีค่าเป็น 0 $(\beta_{26}, \beta_{27}, \ldots, \beta_{1,000})$

4) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = X_i\beta_j + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,1000} \\ \vdots & \ddots & \vdots \\ X_{100,1000} & \cdots & X_{100,1000} \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} 1.5 \\ \vdots \\ 1.5 \end{matrix}\right\} \text{25 ตัว} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{975 ตัว} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

กรณีที่ 4 ข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคาดเคลื่อนมีค่าไม่คงที่ (เมทริกซ์ที่มีค่าสหสัมพันธ์เป็นลักษณะ Exponential decay)

1) กำหนดตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นเวกเตอร์ศูนย์ และเมทริกซ์ความแปรปรวนร่วม $\Sigma_{p\times p}$

$$X_i \sim N_p(0, \Sigma)$$

โดยที่ $\Sigma_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์

$$\Sigma = \begin{bmatrix} 1 & \Sigma_{1,2} & \cdots & \Sigma_{1,p} \\ \Sigma_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \Sigma_{j,k} \\ \Sigma_{p,1} & \cdots & \Sigma_{j,k} & 1 \end{bmatrix}$$

(โดยกำหนด $(\Sigma^{-1})_{j,k} = 0.4^{|j-k|/5}$)

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) กำหนดค่าสัมประสิทธิ์ (β_j) ของตัวแปรอิสระมีค่าเป็น 1.5 สำหรับตัวแปรอิสระ 25 ตัวแรก $(\beta_1, \beta_2, \ldots, \beta_{25})$ และของตัวแปรที่เหลือมีค่าเป็น 0 $(\beta_{26}, \beta_{27}, \ldots, \beta_{1,000})$

4) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = X_i\beta_j + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,1000} \\ \vdots & \ddots & \vdots \\ X_{100,1} & \cdots & X_{100,1000} \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} 1.5 \\ \vdots \\ 1.5 \end{matrix}\right\} 25 \text{ ตัว} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} 975 \text{ ตัว} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

กรณีที่ 5 ข้อมูลเกิดปัญหาความผิดพลาดในการวัด

1) กำหนดตัวแปรอิสระจากตัวแบบดังนี้

$$X_i = Z_i + \xi_i$$

โดยที่ Z_i และ ξ_i มีการแจกแจงปกติมาตรฐานหลายตัวแปร (Multivariate Standard Normal Distribution) ดังนี้

$$Z_i \overset{iid}{\sim} N_p(0, I_{\mathrm{p}})$$

$$\xi_i \overset{iid}{\sim} N_p(0, I_{\mathrm{p}})$$

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) กำหนดค่าสัมประสิทธิ์ (β_j) ของตัวแปรอิสระมีค่าเป็น 1.5 สำหรับตัวแปรอิสระ 25 ตัวแรก $(\beta_1, \beta_2, \dots, \beta_{25})$ และของตัวแปรที่เหลือมีค่าเป็น 0 $(\beta_{26}, \beta_{27}, \dots, \beta_{1,000})$

4) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = Z_i\beta_j + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} X_{1,1} & \cdots & X_{1,1000} \\ \vdots & \ddots & \vdots \\ X_{100,1} & \cdots & X_{100,1000} \end{bmatrix} = \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,1000} \\ \vdots & \ddots & \vdots \\ Z_{100,1} & \cdots & Z_{100,1000} \end{bmatrix} + \begin{bmatrix} \xi_{1,1} & \cdots & \xi_{1,1000} \\ \vdots & \ddots & \vdots \\ \xi_{100,1} & \cdots & \xi_{100,1000} \end{bmatrix}$$

และ

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = \begin{bmatrix} Z_{1,1} & \cdots & Z_{1,1000} \\ \vdots & \ddots & \vdots \\ Z_{100,1} & \cdots & Z_{100,1000} \end{bmatrix} \begin{bmatrix} \left.\begin{matrix} 1.5 \\ \vdots \\ 1.5 \end{matrix}\right\} \text{25 ตัว} \\ \left.\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}\right\} \text{975 ตัว} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

กรณีที่ 6 ข้อมูลเกิดปัญหาตัวแปรแฝง

1) กำหนดตัวแปรอิสระจากตัวแบบดังนี้

$$X_j = sign(5.5 - j)Z_1 1_{\{j \le 10\}} + sign(15.5 - j)Z_2 1_{\{11 \le j \le 20\}} + Z_3 1_{\{21 \le j \le 25\}} + \xi_j$$

โดยที่ Z_1, Z_2, Z_3 และ ξ_j มีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution) ดังนี้

$$Z_1, Z_2, Z_3 \overset{iid}{\sim} N(0,1)$$
$$\xi_j \overset{iid}{\sim} N(0,1)$$

2) กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

3) จำลองตัวแปรตามจากตัวแบบดังต่อไปนี้

$$Y_i = 1.5Z_1 + 1.5Z_2 + 1.5Z_3 + \varepsilon_i$$

หรือเขียนเป็นเมทริกซ์ได้ดังนี้

- การจำลองตัวแปรอิสระในกรณีนี้จะทำการจำลองทีละหลัก ดังนี้

สำหรับ $j = 1,2,\dots,10$

$$\begin{bmatrix} X_{1,j} \\ X_{2,j} \\ \vdots \\ X_{100,j} \end{bmatrix} = sign(5.5 - j) \begin{bmatrix} Z_{1,1} \\ Z_{2,1} \\ \vdots \\ Z_{100,1} \end{bmatrix} + \begin{bmatrix} \xi_{1,j} \\ \xi_{2,j} \\ \vdots \\ \xi_{100,j} \end{bmatrix}$$

สำหรับ $j = 11,12,\dots,20$

$$\begin{bmatrix} X_{1,j} \\ X_{2,j} \\ \vdots \\ X_{100,j} \end{bmatrix} = sign(15.5 - j) \begin{bmatrix} Z_{1,2} \\ Z_{2,2} \\ \vdots \\ Z_{100,2} \end{bmatrix} + \begin{bmatrix} \xi_{1,j} \\ \xi_{2,j} \\ \vdots \\ \xi_{100,j} \end{bmatrix}$$

สำหรับ $j = 21,22,\dots,25$

$$\begin{bmatrix} X_{1,j} \\ X_{2,j} \\ \vdots \\ X_{100,j} \end{bmatrix} = \begin{bmatrix} Z_{1,3} \\ Z_{2,3} \\ \vdots \\ Z_{100,3} \end{bmatrix} + \begin{bmatrix} \xi_{1,j} \\ \xi_{2,j} \\ \vdots \\ \xi_{100,j} \end{bmatrix}$$

สำหรับ $j = 26,27,\dots,1000$

$$\begin{bmatrix} X_{1,j} \\ X_{2,j} \\ \vdots \\ X_{100,j} \end{bmatrix} = \begin{bmatrix} \xi_{1,j} \\ \xi_{2,j} \\ \vdots \\ \xi_{100,j} \end{bmatrix}$$

และ

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{100} \end{bmatrix} = 1.5 \begin{bmatrix} Z_{1,1} \\ Z_{2,1} \\ \vdots \\ Z_{100,1} \end{bmatrix} + 1.5 \begin{bmatrix} Z_{1,2} \\ Z_{2,2} \\ \vdots \\ Z_{100,2} \end{bmatrix} 1.5 \begin{bmatrix} Z_{1,3} \\ Z_{2,3} \\ \vdots \\ Z_{100,3} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{100} \end{bmatrix}$$

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง

2. กำหนดค่าเริ่มต้นสำหรับการจำลองข้อมูลในแต่ละกรณีที่ศึกษา

2.1 กำหนดขนาดตัวอย่าง (n) เท่ากับ 100

2.2 กำหนดจำนวนตัวแปรอิสระ (p) เท่ากับ 1,000

3. ทำการจำลองข้อมูลจากค่าเริ่มต้นที่กำหนดทั้งหมด 6 กรณี

4. ในแต่ละกรณีที่ทำการศึกษา นำข้อมูลที่ได้จากการจำลองมาทำการวิเคราะห์ดังต่อไปนี้

4.1 หาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธี CV โดยใช้ $k = 10$

4.2 หาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธี BIC

4.3 หาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธี RD

5. นำค่าพารามิเตอร์การปรับที่ได้จากข้อที่ 4. มาคำนวณหาค่าดังนี้

5.1 อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate)

5.2 อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate)

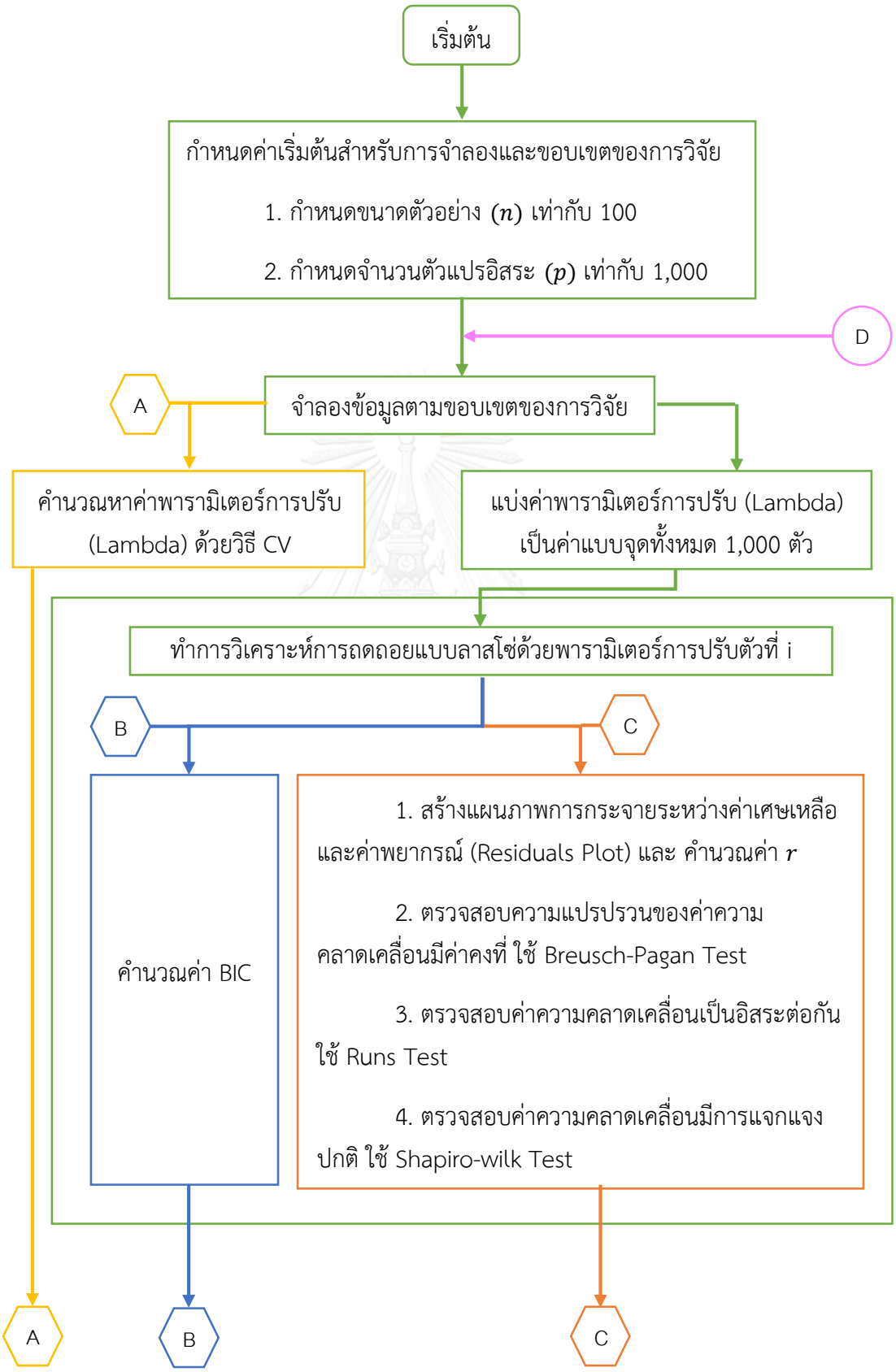
5.3 ค่าคลาดเคลื่อนจากการพยากรณ์ (Prediction Error)

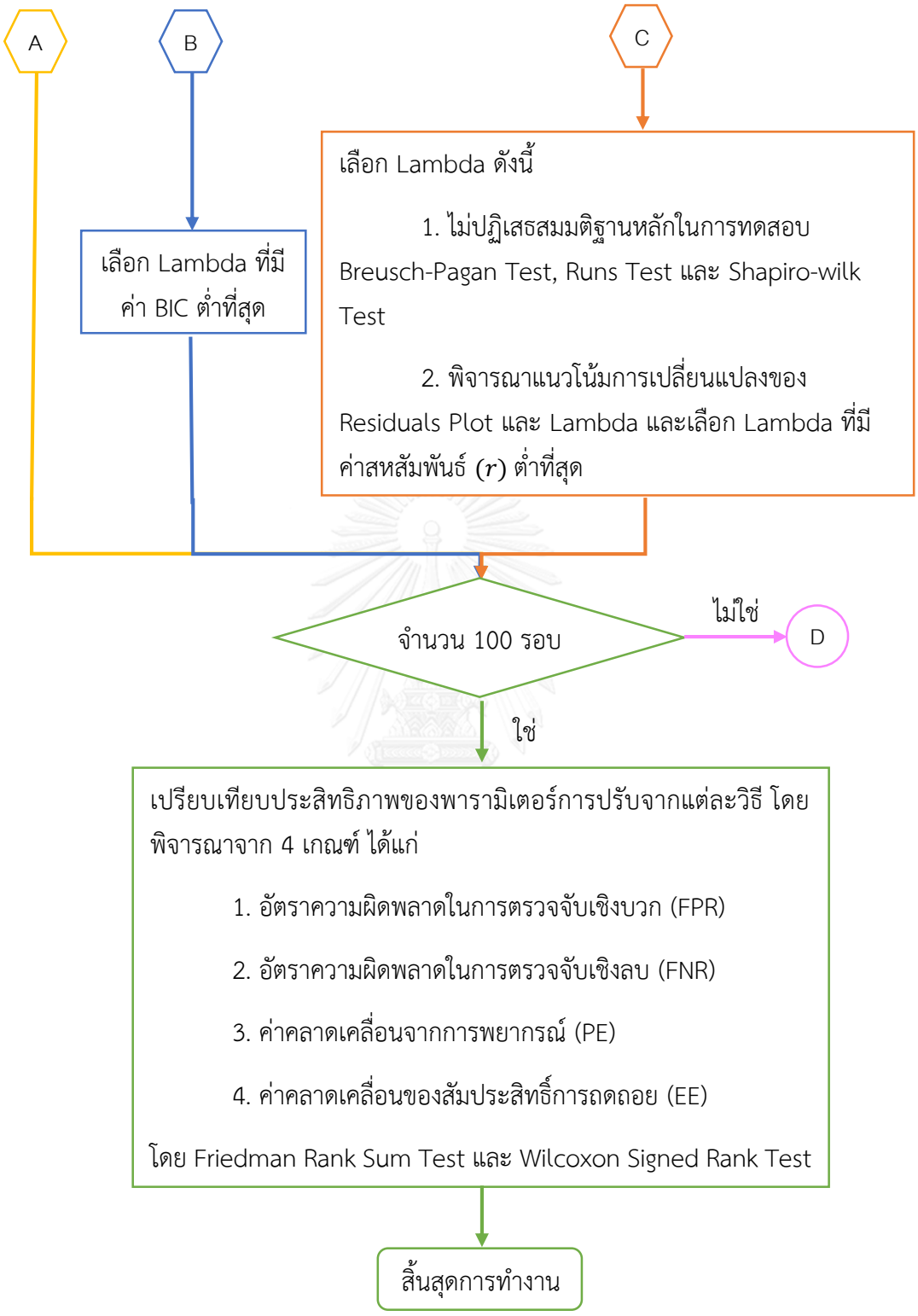
5.4 ค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (Estimation Error)

6. เปรียบเทียบผลการวิเคราะห์ในข้อที่ 5. โดย Friedman Rank Sum Test และทำการเปรียบเทียบเชิงคู่ (Pairwise Comparison) โดย Wilcoxon Signed Rank Test

7. สรุปผลการศึกษา

3.2 ขั้นตอนการทำงานของโปรแกรม R




A
B
C
เลือก Lambda ที่มีค่า BIC ต่ำที่สุด
เลือก Lambda ดังนี้
1. ไม่ปฏิเสธสมมติฐานหลักในการทดสอบ Breusch-Pagan Test, Runs Test และ Shapiro-wilk Test
2. พิจารณาแนวโน้มการเปลี่ยนแปลงของ Residuals Plot และ Lambda และเลือก Lambda ที่มีค่าสหสัมพันธ์ (r) ต่ำที่สุด
จำนวน 100 รอบ
ไม่ใช่
D
ใช่
เปรียบเทียบประสิทธิภาพของพารามิเตอร์การปรับจากแต่ละวิธี โดยพิจารณาจาก 4 เกณฑ์ ได้แก่
1. อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR)
2. อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR)
3. ค่าคลาดเคลื่อนจากการพยากรณ์ (PE)
4. ค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE)
โดย Friedman Rank Sum Test และ Wilcoxon Signed Rank Test
สิ้นสุดการทำงาน

บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการหาค่าพารามิเตอร์การปรับทั้ง 3 วิธี ซึ่งได้แก่วิธี วิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) โดยการจำลองข้อมูลที่มีขอบเขตแตกต่างกันจำนวน 6 กรณี โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของของผลที่ได้จากการวิเคราะห์การถดถอยด้วยพารามิเตอร์ปรับแต่ละวิธี ได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE) โดยถ้าเกณฑ์ทั้ง 4 มีค่าต่ำที่สุด จะถือว่าเป็นวิธีที่มีประสิทธิภาพและมีความเหมาะสมในการหาค่าพารามิเตอร์การปรับมากที่สุด

อักษรย่อและสัญลักษณ์ต่างๆที่ปรากฏในการนำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆแทนความหมายดังนี้

CV แทน การหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation

BIC แทน การหาค่าพารามิเตอร์การปรับโดยวิธี Bayesian Information Criterion

RD แทน การหาค่าพารามิเตอร์การปรับโดยวิธีวิธีการตรวจสอบข้อบังคับเบื้องต้น

FPR แทน อัตราความผิดพลาดในการตรวจจับเชิงบวก

FNR แทน อัตราความผิดพลาดในการตรวจจับเชิงลบ

PE แทน ค่าคลาดเคลื่อนจากการพยากรณ์

EE แทน ค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย

โดยผลการวิจัยแบ่งออกเป็น 5 ส่วน ดังนี้

ส่วนที่ 1 ค่าพารามิเตอร์การปรับจากข้อมูลจำลองที่ได้จากการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ส่วนที่ 2 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ส่วนที่ 3 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ส่วนที่ 4 ผลการเปรียบเทียบค่าความผิดพลาดในการพยากรณ์ (PE) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ส่วนที่ 5 ผลการเปรียบเทียบค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอย (BE) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

4.1 ค่าพารามิเตอร์การปรับจากข้อมูลจำลองที่ได้จากการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ในส่วนนี้ผู้วิจัยต้องการเปรียบเทียบค่าพารามิเตอร์การปรับที่ได้จากการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) เพื่อพิจารณาว่าค่าพารามิเตอร์ปรับที่ได้มีความแตกต่างกันมากหรือน้อยเพียงใด

ตารางที่ 4.1 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าพารามิเตอร์การปรับสำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1	170.621 (3.507)	1.000 (0.000)	244.756 (5.606)
กรณีที่ 2	862.954 (9.004)	1.000 (0.000)	7.004 (0.657)
กรณีที่ 3	25.718 (0.419)	27.973 (0.523)	30.470 (0.639)
กรณีที่ 4	80.339 (1.109)	107.893 (2.209)	75.675 (2.201)
กรณีที่ 5	194.063 (3.492)	1.000 (0.000)	218.282 (3.846)
กรณีที่ 6	40.715 (0.515)	112.638 (3.261)	116.135 (3.410)

จากตารางที่ 4.1 พบว่าการหาค่าพารามิเตอร์การปรับด้วยวิธี BIC จะสามารถแบ่งออกเป็น 3 กลุ่ม นั่นคือ (1) สามารถหาค่าพารามิเตอร์การปรับได้ใกล้เคียงหรือมากกว่าการหาค่าพารามิเตอร์ด้วยวิธี CV นั่นคือเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 3 และ 4 เมื่อพิจารณาค่าความคลาดเคลื่อนมาตรฐานของทั้ง 2 กรณี พบว่ามีค่าใกล้เคียงกับค่าความคลาดเคลื่อนมาตรฐานของค่าพารามิเตอร์การปรับที่ได้จากการหาค่าพารามิเตอร์ด้วยวิธี CV กลุ่มที่ (2) สามารถหาค่าพารามิเตอร์การปรับได้แต่ไม่ใกล้เคียงกับการหาค่าพารามิเตอร์ด้วยวิธี CV นั่นคือเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 6 และเมื่อพิจารณาส่วนเบี่ยงเบนยังพบว่ามีค่าสูงกว่าส่วนเบี่ยงเบนของค่าพารามิเตอร์การปรับที่ได้จากการหาค่าพารามิเตอร์ด้วยวิธี CV และกลุ่มที่ (3) สามารถหาค่าพารามิเตอร์การปรับได้เท่ากับ 1 เสมอ นั่นคือเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 1, 2 และ 5 ซึ่งจะส่งผลให้มีค่าสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์มีจำนวนเท่ากับจำนวนขนาดตัวอย่าง จึงจะไม่พิจารณาประสิทธิภาพของการหาค่าพารามิเตอร์การปรับด้วยวิธี BIC ในกรณีที่ที่ 1, 2 และ 5

ในขณะเดียวกันการหาค่าพารามิเตอร์การปรับด้วยวิธีการตรวจสอบข้อบังคับเบื้องต้นสามารถแบ่งออกเป็น 2 กลุ่ม นั่นคือ (1) สามารถหาค่าพารามิเตอร์การปรับได้ใกล้เคียงหรือมากกว่าการหาค่าพารามิเตอร์ด้วยวิธี CV เล็กน้อย นั่นคือเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 1, 3, 4 และ 5 และเมื่อพิจารณาค่าความคลาดเคลื่อนมาตรฐานของทั้ง 4 กรณี พบว่ายังมีค่าใกล้เคียงกับค่าความคลาดเคลื่อนมาตรฐานของค่าพารามิเตอร์การปรับที่ได้จากการหาค่าพารามิเตอร์ด้วยวิธี CV และกลุ่มที่ (2) สามารถหาค่าพารามิเตอร์การปรับได้ แต่ไม่ใกล้เคียงกับการหาค่าพารามิเตอร์ด้วยวิธี CV นั่นคือเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 2 และ 6 และเมื่อพิจารณาค่าความคลาดเคลื่อนมาตรฐานยังพบว่ามีความแตกต่างกันอย่างชัดเจนเมื่อเทียบกับค่าความคลาดเคลื่อนมาตรฐานของค่าพารามิเตอร์การปรับที่ได้จากการหาค่าพารามิเตอร์ด้วยวิธี CV

4.2 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) เพื่อพิจารณาว่าวิธีใดที่มีประสิทธิภาพมากที่สุด โดยใช้เกณฑ์อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR)

ตารางที่ 4.2 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	0.444 (0.020)	-	**0.333 (0.034)**
กรณีที่ 2‡	0.850 (0.008)	-	**0.780 (0.004)**
กรณีที่ 3*	0.074 (0.007)	0.038 (0.005)	**0.000 (0.018)**
กรณีที่ 4*	0.839 (0.013)	**0.500 (0.044)**	0.889 (0.017)
กรณีที่ 5‡	0.720 (0.024)	-	**0.683 (0.030)**
กรณีที่ 6*	0.250 (0.012)	**0.000 (0.013)**	**0.000 (0.019)**

หมายเหตุ ตัวหนา คือวิธีที่เหมาะสมที่สุด, * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman Rank Test), ‡ มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Wilcoxon Rank Sum Test)

จากตารางที่ 4.2 พบว่าเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 1, 2, 3 และ 5 การหาค่าพารามิเตอร์การปรับด้วยวิธีการตรวจสอบข้อบังคับเบื้องต้น จะมีความเหมาะสมมากที่สุด เนื่องจากมีค่ามัธยฐานของอัตราความผิดพลาดในการตรวจจับเชิงบวกต่ำที่สุดเมื่อเปรียบเทียบกับวิธี CV ในขณะที่เมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 4 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี BIC และเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 6 ผลการตัดสินไม่เด่นชัดระหว่างการหาค่าพารามิเตอร์การปรับด้วยวิธีการตรวจสอบข้อบังคับเบื้องต้นและวิธี BIC ว่าวิธีไดมีความเหมาะสมมากที่สุด เนื่องจากได้มัธยฐานของค่าอัตราความผิดพลาดในการตรวจจับเชิงบวกเท่ากัน อีกทั้งค่าความคลาดเคลื่อนมาตรฐานของค่าอัตราความผิดพลาดในการตรวจจับเชิงบวกมีความใกล้เคียงกันด้วย

4.3 ผลการเปรียบเทียบอัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) เพื่อพิจารณาว่าวิธีใดที่มีประสิทธิภาพมากที่สุด โดยใช้เกณฑ์อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR)

ตารางที่ 4.3 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**0.020 (<0.01)**	**-**	0.023 (<0.01)
กรณีที่ 2‡	0.020 (<0.01)	-	**0.005 (<0.01)**
กรณีที่ 3	0	0	0
กรณีที่ 4	0.024 (<0.01)	0.024 (<0.001)	0.024 (0.001)
กรณีที่ 5‡	**0.023 (<0.01)**	**-**	0.024 (<0.01)
กรณีที่ 6*	**0.011 (<0.01)**	0.024 (<0.001)	0.023 (<0.01)

หมายเหตุ ตัวหนา คือวิธีที่เหมาะสมที่สุด, * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman Rank Test) , ‡ มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Wilcoxon Rank Sum Test)

จากตารางที่ 4.3 พบว่าเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 1, 5 และ 6 การหาค่าพารามิเตอร์การปรับด้วยวิธี CV จะมีความเหมาะสมมากที่สุด เนื่องจากมีค่ามัธยฐานของอัตราการเกิดความผิดพลาดในการตรวจจับข้อมูลเชิงลบต่ำที่สุด ในขณะที่การจำลองข้อมูลในรูปแบบของกรณีที่ 2 การหาค่าพารามิเตอร์การปรับด้วยวิธีการตรวจสอบข้อบังคับเบื้องต้นจะมีความเหมาะสมมากที่สุด แต่เมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 3 และ 4 การหาค่าพารามิเตอร์การปรับทั้ง 3 วิธีไม่มีวิธีใดที่เหมาะสมที่สุดอย่างเด่นชัด เนื่องจากอัตราการเกิดความผิดพลาดในการตรวจจับข้อมูลเชิงลบไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

4.4 ผลการเปรียบเทียบค่าความผิดพลาดในการพยากรณ์ (PE) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) เพื่อพิจารณาว่าวิธีใดที่มีประสิทธิภาพมากที่สุด โดยใช้เกณฑ์ค่าความผิดพลาดในการพยากรณ์ (PE)

ตารางที่ 4.4 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าความผิดพลาดในการพยากรณ์ (PE) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**5190 (103.1)**	-	5503 (96.31)
กรณีที่ 2‡	9155 (173.8)	-	**4.661 (0.880)**
กรณีที่ 3*	**159.5 (4.347)**	165.7 (4.855)	178.5 (5.391)
กรณีที่ 4*	**998.7 (20.29)**	1031 (23.64)	1011 (21.02)
กรณีที่ 5‡	5871 (103.7)	-	**5860 (102.5)**
กรณีที่ 6*	**340.9 (9.297)**	722.6 (22.88)	762.8 (20.54)

หมายเหตุ ตัวหนา คือวิธีที่เหมาะสมที่สุด, * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman Rank Test) , ‡ มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Wilcoxon Rank Sum Test)

จากตารางที่ 4.4 พบว่าเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 2 และ 4 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธีการตรวจสอบข้อบังคับเบื้องต้น เนื่องจากค่ามัธยฐานของค่าความผิดพลาดในการพยากรณ์ต่ำที่สุด โดยเฉพาะในข้อมูลจำลองกรณีที่ 2 ได้ค่ามัธยฐานเท่ากับ 4.661 ซึ่งมีค่าต่ำมากเมื่อเทียบกับวิธี CV ที่ได้ค่ามัธยฐานสูงถึง 9155 แต่ในขณะเดียวกันการจำลองข้อมูลในรูปแบบของกรณีที่ 1, 3, 5 และ 6 การหาค่าพารามิเตอร์การปรับด้วยวิธี CV มีความเหมาะสมมากที่สุด

4.5 ผลการเปรียบเทียบค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอย (BE) ระหว่างการหาค่าพารามิเตอร์การปรับโดยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการหาค่าพารามิเตอร์การปรับของการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธีวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) เพื่อพิจารณาว่าวิธีใดที่มีประสิทธิภาพมากที่สุด โดยใช้เกณฑ์ค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอย (EE)

ตารางที่ 4.5 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าความผิดพลาดค่าสัมประสิทธิ์การถดถอย (EE) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**36.23 (0.195)**	-	37.19 (0.076)
กรณีที่ 2‡	54.36 (0.513)	-	**44.41 (0.811)**
กรณีที่ 3*	**7.598 (0.195)**	7.659 (0.197)	7.800 (0.205)
กรณีที่ 4*	37.67 (0.036)	**37.44 (0.027)**	37.75 (0.047)
กรณีที่ 5‡	37.87 (0.109)	-	**37.60 (0.067)**
กรณีที่ 6*	**37.07 (0.065)**	37.43 (0.037)	37.46 (0.029)

หมายเหตุ ตัวหนา คือวิธีที่เหมาะสมที่สุด, * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman Rank Test) , ‡ มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Wilcoxon Rank Sum Test)

จากตารางที่ 4.5 พบว่าเมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 1, 3 และ 6 การหาค่าพารามิเตอร์การปรับด้วยวิธี CV จะมีความเหมาะสมมากที่สุด เนื่องจากมีค่ามัธยฐานของค่าความผิดพลาดค่าสัมประสิทธิ์การถดถอยต่ำที่สุด ในขณะที่เมื่อจำลองข้อมูลในรูปแบบของกรณีที่ 2 และ 5 การหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือ RD และสำหรับข้อมูลจำลองกรณีที่ 4 การหาค่าพารามิเตอร์การปรับด้วยวิธี BIC จะมีความเหมาะสมมากที่สุด

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาการเปรียบเทียบวิธีการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ทั้ง 3 วิธี ซึ่งได้แก่วิธี วิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) โดยการจำลองข้อมูลที่มีขอบเขตแตกต่างกันจำนวน 6 กรณี และมีเกณฑ์ในการพิจารณาประสิทธิภาพของของผลที่ได้จากการวิเคราะห์การถดถอยด้วยพารามิเตอร์ปรับแต่ละวิธี ได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE) โดยมีการสรุปผลการวิจัยดังนี้

5.1 สรุปผลการวิจัย

จากการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของวิธีการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ด้วยวิธี Cross-validation (CV), วิธี Bayesian Information Criterion (BIC) และวิธีการตรวจสอบข้อบังคับเบื้องต้น (RD) โดยพิจารณาวิธีที่ให้ค่ามัธยฐานของอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) ต่ำที่สุด จะถือว่าวิธีนั้นมีความเหมาะสมในการใช้หาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ เช่นเดียวกันกับอัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE) ผลการวิเคราะห์ข้อมูลจำลองสามารถสรุปผลการวิจัยได้ดังนี้

ตารางที่ 5. 1 แสดงวิธีการหาค่าพารามิเตอร์ปรับสำหรับการถดถอยลาสโซ่เหมาะสมที่สุด เมื่อพิจารณาอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) ค่าคลาดเคลื่อนจากการพยากรณ์ (PE) และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย (EE) จำแนกตามกรณีศึกษา

กรณีศึกษา	เกณฑ์การตัดสินใจ											
	FPR			FNR			PE			EE		
	CV	BIC	RD	CV	BIC	RD	CV	BIC	RD	CV	BIC	RD
กรณีที่ 1			✓	✓			✓			✓		
กรณีที่ 2			✓			✓			✓			✓
กรณีที่ 3			✓	✓	✓	✓	✓			✓		
กรณีที่ 4		✓		✓	✓	✓	✓				✓	
กรณีที่ 5			✓	✓					✓			✓
กรณีที่ 6		✓	✓	✓			✓			✓		

หมายเหตุ CV หมายถึงวิธี Cross-validation

BIC หมายถึงวิธี Bayesian Information Criterion

RD หมายถึงวิธีวิธีการตรวจสอบข้อบังคับเบื้องต้น

จากตารางที่ 5.1 สามารถแบ่งผลการวิจัยได้ 2 ส่วน ดังนี้

ส่วนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถดถอยลาสโซ่ โดยพิจารณาที่ความต่างของข้อมูลจำลอง

สำหรับข้อมูลจำลองกรณีที่ 1 และ 6 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี CV โดยวิธี CV มีค่ามัธยฐานของเกณฑ์ FNR, PE และ EE ต่ำที่สุด ในขณะเดียวกันสำหรับเกณฑ์ FPR วิธี RD มีค่ามัธยฐานต่ำที่สุด

สำหรับข้อมูลจำลองกรณีที่ 2 และ 5 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี RD โดยวิธี RD มีค่ามัธยฐานของเกณฑ์ FPR, PE และ EE ต่ำที่สุด

สำหรับข้อมูลจำลองกรณีที่ 3 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี CV และ RD ซึ่งไม่มีวิธีใดวิธีหนึ่งเด่นชัดกว่ากัน

และสำหรับข้อมูลจำลองกรณีที่ 4 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี CV และ BIC ซึ่งไม่มีวิธีใดวิธีหนึ่งเด่นชัดกว่ากัน

ส่วนที่ 2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ โดยพิจารณาที่เกณฑ์การตัดสินใจทั้ง 4 เกณฑ์

จากตารางที่ 5.1 พบว่าเมื่อพิจารณาเกณฑ์ FPR ด้วยค่ามัธยฐานของ FPR จะได้ว่าวิธี RD มีความเหมาะสมในการหาค่าพารามิเตอร์การปรับมากที่สุดสำหรับทุกกรณี ยกเว้นกรณีที่ 4 โดยวิธี RD จะสามารถทำงานได้ดีเมื่อจำลองข้อมูลแบบกรณีที่ 3 และ 6

สำหรับเกณฑ์ FNR จะได้ว่าวิธี CV มีความเหมาะสมในการหาค่าพารามิเตอร์การปรับมากที่สุดสำหรับทุกกรณี ยกเว้นกรณีที่ 2 วิธี อย่างไรก็ตามในกรณีที่ 3 และ 4 วิธีการหาค่าพารามิเตอร์การปรับทั้ง 3 วิธีมีความเหมาะสมในการหาค่าพารามิเตอร์การปรับ จึงไม่สามารถสรุปได้ว่า CV เหมาะสมอย่างเด่นชัด

และเมื่อพิจารณาด้วยเกณฑ์ PE และ EE ด้วยค่ามัธยฐานของทั้งสองเกณฑ์ จะได้ว่าไม่มีวิธีใดวิธีหนึ่งที่เหมาะสมอย่างเด่นชัด

อนึ่ง สามารถสรุปโดยรวมได้ว่าวิธี BIC เป็นวิธีที่ไม่เหมาะสมในการหาค่าพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่ เนื่องจากได้ค่าพารามิเตอร์การปรับเท่ากับ 1 ในหลายกรณีทำให้เลือกพารามิเตอร์เข้าตัวแบบเท่ากับจำนวนตัวอย่างตามข้อจำกัดของการถดถอยแบบลาสโซ่

ในขณะที่วิธี RD มีประสิทธิภาพในการหาค่าพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่มากกว่าอีก 2 วิธี ในกรณีที่พิจารณาอัตราความผิดพลาดในการตรวจจับเชิงบวก แต่ในแง่ของอัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าคลาดเคลื่อนจากการพยากรณ์ และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย พบว่าวิธี CV และวิธี RD ไม่มีวิธีใดวิธีหนึ่งที่เหมาะสมกว่าอย่างเด่นชัด

หากจะพิจารณาว่าเกณฑ์ที่ใช้ในการตัดสินใจเกณฑ์ใดสำคัญที่สุด ในที่นี้คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก ซึ่งถือเป็นความน่าจะเป็นของการเกิดความผิดพลาดประเภทที่ 1 (Type I error) โดยส่วนใหญ่ผู้วิจัยทางด้านวิทยาศาสตร์ชีวการแพทย์ได้ให้ความสำคัญกับอัตราความผิดพลาดในการตรวจจับเชิงบวกเป็นอย่างมาก ตัวอย่างเช่น ถ้าต้องการทดสอบ

H_0 : ยีน XX ไม่มีความสัมพันธ์กับการเกิดโรค Y

H_1 : ยีน XX มีความสัมพันธ์กับการเกิดโรค Y

ดังนั้นความผิดพลาดในการตรวจจับเชิงบวก หรือความน่าจะเป็นของการเกิดความผิดพลาดประเภทที่ 1 คือ ยีน XX ที่ไม่เกี่ยวข้องกับโรค Y จริงนั้นถูกระบุว่ามีความสัมพันธ์กับโรคนั้นๆ ซึ่งผู้ป่วยที่มียีนดังกล่าวอาจจะได้รับการรักษาที่ไม่เหมาะสมและส่งผลกระทบที่ร้ายแรงต่อผู้ป่วยก็เป็นได้ ดังนั้นอัตราความผิดพลาดเชิงบวกจึงส่งผลกระทบที่ร้ายแรงกว่าอัตราความผิดพลาดเชิงลบ

5.2 ข้อจำกัดของงานวิจัย

งานวิจัยนี้ผู้วิจัยทำการศึกษาภายใต้ขอบเขตการศึกษาที่ตายตัว ซึ่งกำหนดขนาดตัวอย่าง (n) เท่ากับ 100 และจำนวนตัวแปรอิสระ (p) เท่ากับ 1,000 นั่นคือมีอัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระ $(n:p)$ เท่ากับ 1:10 นอกจากนี้ยังกำหนดจำนวนตัวแปรอิสระที่ไม่เท่ากับศูนย์ (Sparsity level) จำนวน 25 ตัว หรือเท่ากับ 2.5% ของตัวแปรอิสระ ซึ่งข้อจำกัดที่กล่าวมาข้างต้นทั้ง 2 ประการนี้อาจส่งผลต่อการหาค่าพารามิเตอร์การปรับที่เหมาะสมในแต่ละกรณีศึกษาได้

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ ผู้ที่สนใจอาจนำไปศึกษาต่อในเรื่องของ

1. ขอบเขตการศึกษา ในเรื่องของขนาดตัวอย่าง, จำนวนตัวแปรอิสระ, จำนวนและค่าสัมประสิทธิ์ β_j ของตัวแปรอิสระที่ไม่เท่ากับ 0 อาจเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้นได้

2. การจำลองข้อมูลในการศึกษาครั้งนี้เน้นเพียงการเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่เท่านั้น โดยผู้ที่สนใจอาจทำการจำลองข้อมูลที่อาจก่อให้เกิดปัญหาความคลาดเคลื่อนมีการแจกแจงไม่ปกติ และปัญหาค่าความคลาดเคลื่อนไม่เป็นอิสระต่อกัน

3. การจำลองข้อมูลตัวแปรอิสระให้มีการแจกแจงอื่นๆ ที่ไม่เป็นการแจกแจงปกติ

4. วิธีการหาค่าพารามิเตอร์การปรับในงานวิจัยนี้ศึกษาเพียง 3 วิธีเท่านั้น ในความเป็นจริงแล้วยังมีอีกหลายวิธีที่น่าสนใจ โดยผู้ที่สนใจอาจนำวิธีอื่นๆมาร่วมพิจารณาเพื่อเปรียบเทียบประสิทธิภาพได้

รายการอ้างอิง

1. วิฐรา พึ่งพาพงศ์, บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี, 2558. 23(2): p. 12.
2. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of Royal Statistical Society, Series B., 1996. 58(1): p. 21.
3. Inoue, T. and Y. Nagata, *Study of model evaluation method and of selection method of tuning parameter in Lasso.* Journal of the Japanese Society for Quanlity Control, 2016. 2(1): p. 8.
4. Wang, H., B. Li, and C. Leng, *Shrikage Tuning Parameter Selection with a Diverging Number of Parameters.* Journal of Royal Statistical Society, Series B., 2009. 71(3): p. 13.
5. Chand, S. *On tuning parameter selection of lasso-type methods - a monte carlo study*. in 2012 9*th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. 2012. Islamabad, Pakistan: IEEE.
6. Dezeure, R., et al., *High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi.* Statistical Science, 2015. 30(4): p. 533-558.
7. พิษณุ เจียวคุณ, การวิเคราะห์การถดถอย, ed. 1. 2550, เชียงใหม่: สถานบริการวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเชียงใหม่.
8. สุพล ดุรงค์วัฒนา, *Regression Models: Analytics-based Approach*, ed. 1. 2558, กรุงเทพมหานคร: บริษัท แดเน็กซ์ อินเตอร์คอร์ปอเรชั่น จำกัด.
9. มานะชัย รอดชื่น, สถิตินอนพาราเมตริก. คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่: เชียงใหม่. p. 35-50.
10. Syed, A.R., *A Review of Cross Validation and Adaptive Model Selection*, in *Mathematics and Statistics*. 2011, Georgia State University.
11. กัลยา วานิชย์บัญชา, การวิเคราะห์ข้อมูลหลายตัวแปร, ed. 4. 2552, กรุงเทพมหานคร: สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.
12. Schwarz, G., *Estimating the Dimension of a Model.* The Annals of Statistics, 1978. 6(2): p. 4.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่างกรณีที่ 1 การจำลองข้อมูลมีการแจกแจงปกติ

```
library(mvtnorm)
library(penalized)
library(randtests)

n<-100
p<-1000
corr<-0
sparse<-25
criteria<-0.05
selectedlambda<-rep(0, 100)
optlambda<-selectedlambda
BIC<-selectedlambda
fnbic<-selectedlambda
fpbic<-selectedlambda
fndiag<-selectedlambda
fpdiag<-selectedlambda
fnopt<-selectedlambda
fpopt<-selectedlambda
predicterror.diag<-selectedlambda
predicterror.bic<-selectedlambda
predicterror.opt<-selectedlambda
betaerror.diag<-selectedlambda
```

```
betaerror.bic<-selectedlambda

betaerror.opt<-selectedlambda



## generate cov matrix

Ip<-diag(p)

for (run in 1:100)

{

## simulate x ~ N(0,Ip)

x<-rmvnorm(n,rep(0,p),Ip)


## simulate beta = (1.5,..,1.5,0,..0)
beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)


## simulate error ~ N(0,1)

error<-rnorm(n)


## calculate y = xb+e

y<-x%*%beta+error


##------------------------------------------------------------------

## ( 1 ) find optimal value of lambda

##------------------------------------------------------------------

opt<-optL1(y,penalized = x,standardize = TRUE, fold = 10)

optimallambda<-opt$lambda
```

```
optlambda[run]<-opt$lambda


##-----------------------------------------------

##   ( 2 ) diagnostic

##-----------------------------------------------

pdf(paste('plot', run, '.pdf',sep=""))
e<-matrix(rep(0, 1000*n), ncol=1000)
predY<-e
par(mfrow=c(2,2))
lambda<-seq(1,500, length.out = 1000)

col<-0


correlation<-rep(1,1000)

pshapiro<-rep(1,1000)

chibp<-rep(1,1000)

prun<-rep(1,1000)

rejectbp<-rep(1,1000)

rejectnormal<-rep(1,1000)

rejectind<-rep(1,1000)

rejectlinear<-rep(1,1000)

diagnostic<-rep(1,1000)

bic<-rep(100,1000)


c<-log(log(p))

logn<-log(n)/n
```

```
for (i in lambda)
{
  col<-col+1
  model<-penalized(response=y,penalized=x,lambda1=i,standardize = TRUE)
  e[,col]<-residuals(model)
  predY[,col]<-fitted(model)
  #plot(predY[,col], e[,col], main=paste("lambda = ",i, sep=""))

  if (sd(predY[,col])>0)
  {
   # choose the minimum lambda by BIC
    SSE<-sum(residuals(model)^2)
    s<-length(which((abs(coefficients(model, "penalized"))>0)))
    sigma2<-SSE/n
    bic[col]<-log(sigma2)+(s*c*logn)
   # choose the minimum lambda such that the slope is 0
    correlation[col]<-cor(predY[,col],e[,col])
     if (0.5>correlation[col])
      {
          rejectlinear[col]<-0
      }
   # test of constant variance
    newx<-cbind(x[,which(abs(coefficients(model, "penalized"))>0)])
    resmodel<-lm(e[,col]^2~newx)
    sanova<-anova(resmodel)
    ssrstar<-sanova$`Sum Sq`[1]
    chibp[col]<-n*ssrsta^2
```

```
    if (chibp[col]<qchisq(1-criteria,s))
     {
        rejectbp[col]<-0
     }
   # test of independent error
    runtest<-runs.test(e[,col])
    prun[col]<-runtest$p.value
     if (criteria<prun[col])
     {
         rejectind[col]<-0
     }
   # test of normality
    shapiro<-shapiro.test(e[,col])
    pshapiro[col]<-shapiro$p.value
     if (criteria<pshapiro[col])
     {
         rejectnormal[col]<-0
     }
   # diagnostic conclusion
    if (rejectlinear[col]==0 & rejectbp[col]==0 & rejectind[col]==0 &
rejectnormal[col]==0)
     {
         diagnostic[col]<-0
     }
```

```
  }
}
dev.off()



BIC[run]<-lambda[which.min(bic)]
selected<-which(diagnostic==0)

if (length(selected)>0)
{
selectedlambda[run]<-lambda[min(selected)]
}

falserate<-function(beta,Beta)
{
    FN<-0
    FP<-0
    betanonzero<-length(which(beta!=0))
    betazero<-length(which(beta==0))

    for (k in 1:length(beta))
     {
        if (beta[k]!=0 & Beta[k]==0)
         {
            FP<-FP+1
```

```
        }
       else if (beta[k]==0 & Beta[k]!=0)
        {
          FN<-FN+1
        }
     }

   FNR<-FN/betazero
   FPR<-FP/betanonzero

   return(c(FNR,FPR))
}

testx<-rmvnorm(n,rep(0,p),lp)
testerror<-rnorm(n)
testy<-testx%*%beta+testerror

modeloptimal<-
penalized(response=y,penalized=x,lambda1=optlambda[run],standardize = TRUE)
modelbic<-penalized(response=y,penalized=x,lambda1=BIC[run],standardize = TRUE)
modeldiag<-
penalized(response=y,penalized=x,lambda1 = selectedlambda[run],standardize =
TRUE)

beta.bic<-coefficients(modelbic, "penalized")
beta.opt<-coefficients(modeloptimal, "penalized")
```

```
beta.diag<-coefficients(modeldiag, "penalized")
all.beta<-cbind(beta,beta.opt,beta.diag,beta.bic)

predicterror.opt[run]<-sum((testy-testx%*%beta.opt)^2)
predicterror.bic[run]<-sum((testy-testx%*%beta.bic)^2)
predicterror.diag[run]<-sum((testy-testx%*%beta.diag)^2)


betaerror.bic[run]<-sum(abs(beta-beta.bic))
betaerror.opt[run]<-sum(abs(beta-beta.opt))
betaerror.diag[run]<-sum(abs(beta-beta.diag))

bicrate<-falserate(beta.bic,beta)
fnbic[run]<-bicrate[1]
fpbic[run]<-bicrate[2]

diagrate<-falserate(beta.diag,beta)
fndiag[run]<-diagrate[1]
fpdiag[run]<-diagrate[2]

optrate<-falserate(beta.opt,beta)
fnopt[run]<-optrate[1]
fpopt[run]<-optrate[2]
}
```

```
output_s<-
data.frame(optlambda,BIC,selectedlambda,fnopt,fpopt,fnbic,fpbic,fndiag,fpdiag,predict
error.diag,

predicterror.bic,predicterror.opt,betaerror.diag,betaerror.bic,betaerror.opt)

write.table(output_s,'output_s.csv',quote = F,row.names = F,col.names = T,sep = ",")


# median of lambda

median(optlambda)

sd(optlambda)*1.253/sqrt(n)

median(BIC)

sd(BIC)*1.253/sqrt(n)

median(selectedlambda)

sd(selectedlambda)*1.253/sqrt(n)
```



```
# compare FNR by Wilcoxon Rank Sum

median(fnopt)

sd(fnopt)*1.253/sqrt(n)

median(fndiag)

sd(fndiag)*1.253/sqrt(n)

wilcox.test(fnopt,fndiag,paired = TRUE)
```

```
# compare FPR by  Wilcoxon Rank Sum

median(fpopt)

sd(fpopt)*1.253/sqrt(n)

median(fpdiag)

sd(fpdiag)*1.253/sqrt(n)

wilcox.test(fpopt,fpdiag,paired = TRUE)



# compare Prediction error by  Wilcoxon Rank Sum


median(predicterror.opt)

sd(predicterror.opt)*1.253/sqrt(n)

median(predicterror.diag)

sd(predicterror.diag)*1.253/sqrt(n)

wilcox.test(predicterror.opt,predicterror.diag,paired = TRUE)



# compare beta error by  Wilcoxon Rank Sum


median(betaerror.opt)

sd(betaerror.opt)*1.253/sqrt(n)

median(betaerror.diag)

sd(betaerror.diag)*1.253/sqrt(n)

wilcox.test(betaerror.opt,betaerror.diag,paired = TRUE)
```

ตัวอย่างกรณีที่ 2 การจำลองข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ (Equal Correlation) ในที่นี้จะแสดงคำสั่งโปรแกรมเพียงส่วนที่เป็นการจำลองข้อมูลเริ่มต้นเท่านั้น เนื่องจากคำสั่งการวิเคราะห์เหมือนกับคำสั่งในกรณีที่ 1

```
## generate cov matrix

Ip<-diag(p)

Ip<-ifelse(Ip<=0,corr,Ip)


## simulate x ~ N(0,Ip)

x<-rmvnorm(n,rep(0,p),Ip)


## simulate beta = (1.5,..,1.5,0,..0)

beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)


## simulate error ~ N(0,1)

error<-matrix(rnorm(n,0,1),nrow=n)


## calculate y = xb+e

y<-x%*%beta+error
```

ตัวอย่างกรณีที่ 3 การจำลองข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ (Toeplitz) ในที่นี้จะแสดงคำสั่งโปรแกรมเพียงส่วนที่เป็นการจำลองข้อมูลเริ่มต้นเท่านั้น เนื่องจากคำสั่งการวิเคราะห์เหมือนกับคำสั่งในกรณีที่ 1

```
## generate cov matrix

Jk<-c()

for (i in 0:999)

{

        j<-0.9^i

        Jk<-c(Jk, j)

}

Ip<-toeplitz(Jk)


## simulate x ~ N(0,Ip)

x<-rmvnorm(n,rep(0,p),Ip)


## simulate beta = (1.5,..,1.5,0,..0)

beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)


## simulate error ~ N(0,1)

error<-matrix(rnorm(n,0,1),nrow=n)


## calculate y = xb+e

y<-x%*%beta+error
```

ตัวอย่างกรณีที่ 4 การจำลองข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ (Exponential decay) ในที่นี้จะแสดงคำสั่งโปรแกรมเพียงส่วนที่เป็นการจำลองข้อมูลเริ่มต้นเท่านั้น เนื่องจากคำสั่งการวิเคราะห์เหมือนกับคำสั่งในกรณีที่ 1

```
## generate cov matrix

Jk<-c()

for (i in 0:999)

{

        j<-0.4^(i/5)

        Jk<-c(Jk, j)

}

ip<-toeplitz(Jk)

Ip<-solve(ip)


## simulate x ~ N(0,Ip)

x<-rmvnorm(n,rep(0,p),Ip)


## simulate beta = (1.5,..,1.5,0,..0)

beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)


## simulate error ~ N(0,1)

error<-matrix(rnorm(n,0,1),nrow=n)


## calculate y = xb+e

y<-x%*%beta+error
```

ตัวอย่างกรณีที่ 5 การจำลองข้อมูลเกิดปัญหาความผิดพลาดในการทำนาย ในที่นี้จะแสดงคำสั่งโปรแกรมเพียงส่วนที่เป็นการจำลองข้อมูลเริ่มต้นเท่านั้น เนื่องจากคำสั่งการวิเคราะห์เหมือนกับคำสั่งในกรณีที่ 1

```
## generate cov matrix

Ip<-diag(p)


## simulate z ~ N(0,Ip)

z<-rmvnorm(n,rep(0,p),Ip)


## simulate Xi ~ N(0,Ip)

Xi<-rmvnorm(n,rep(0,p),Ip)


## simulate x = z+Xi

x<-z+Xi


## simulate beta = (1.5,..,1.5,0,..0)

beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)


## simulate error ~ N(0,1)

error<-matrix(rnorm(n,0,1),nrow=n)


## calculate y = zb+e

y<-z%*%beta+error
```

ตัวอย่างกรณีที่ 6 การจำลองข้อมูลเกิดปัญหาตัวแปรแฝงในที่นี้จะแสดงคำสั่งโปรแกรมเพียงส่วนที่เป็นการจำลองข้อมูลเริ่มต้นเท่านั้น เนื่องจากคำสั่งการวิเคราะห์เหมือนกับคำสั่งในกรณีที่ 1

```
## simulate z ~ N(0,1)
z1<-matrix(rnorm(n,0,1),nrow=n)
z2<-matrix(rnorm(n,0,1),nrow=n)
z3<-matrix(rnorm(n,0,1),nrow=n)

## simulate x
designm<-matrix(rep(1,n),nrow=n)
for (i in 1:p)
{
        err<-matrix(rnorm(n,0,1),nrow=n)
        if (i==1|i==2|i==3|i==4|i==5|i==6|i==7|i==8|i==9|i==10)
                {
                        X<-(sign(5.5-i)*z1)+err
                        designm<-cbind(designm,X)
                }
        else if (i==11|i==12|i==13|i==14|i==15|i==16|i==17|i==18|i==19|i==20)
                {
                        X<-(sign(15.5-i)*z2)+err
                        designm<-cbind(designm,X)
                }
        else if (i==21|i==22|i==23|i==24|i==25)
                {
```

```
                X<-z3+err
                designm<-cbind(designm,X)
        }
    else
        {
                X<-err
                designm<-cbind(designm,X)
        }

}

x<-designm[,-1]

## simulate error ~ N(0,1)
error<-matrix(rnorm(n,0,1),nrow=n)

## simulate beta = (1.5,..,1.5,0,..0)
beta<-matrix(c(rep(1.5,sparse),rep(0,p-sparse)),nrow=p)

## calculate y = 1.5z1+1.5z2+1.5z3+error
y<-(1.5*z1)+(1.5*z2)+(1.5*z3)+error
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวจุฑาทิพย์ นันทสุวรรณ เกิดวันอังคารที่ 29 กันยายน พ.ศ. 2535 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาสถิติ เกียรตินิยมอันดับสอง (วิชาเอก: สถิติ, วิชาโท: การสื่อสารมวลชน) ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ในปีการศึกษา 2557 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2558