การเปรียบเทียบวิธีบูตแสตรปในการประมาณช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่มีมิติสูงด้วยวิธีลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์

นายพริษฐ์ ชาญเชิงพานิช

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

A COMPARISON OF RESIDUAL BOOTSTRAP AND PAIR BOOTSTRAP FOR ADAPTIVE LASSO + PARTIAL RIDGE METHOD TO CONSTRUCT CONFIDENCE INTEVALS FOR PARAMETERS IN HIGH-DIMENSIONAL SPARSE LINEAR MODELS

Mr. Parit Chancherngpanich

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 2021

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีบูตแสตรปในการประมาณช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่มีมิติสูงด้วยวิธีลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์
โดย	นายพริษฐ์ ชาญเชิงพานิช
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.นัจชลี ศรีมณีกาญจน์)

พริษฐ์ ชาญเชิงพานิช : การเปรียบเทียบวิธีบูตแสตรปในการประมาณช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่มีมิติสูงด้วยวิธีลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์. (A COMPARISON OF RESIDUAL BOOTSTRAP AND PAIR BOOTSTRAP FOR ADAPTIVE LASSO + PARTIAL RIDGE METHOD TO CONSTRUCT CONFIDENCE INTEVALS FOR PARAMETERS IN HIGH-DIMENSIONAL SPARSE LINEAR MODELS) อ.ที่ปรึกษาหลัก : ผศ. ดร.วิฐรา พึ่งพาพงศ์

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์ ซึ่งเป็นตัวประมาณแบบ 2 ขั้นตอน คือใช้วิธีลาสโซ่แบบปรับปรุงในการคัดเลือกตัวแปรอิสระจากนั้นใช้วิธีริดจ์ในการประมาณค่าสัมประสิทธิ์การถดถอย และเปรียบเทียบกับวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่และพาร์เชียลริดจ์ โดยทดลองบูตแสตรป 2 วิธีคือ วิธีสุ่มส่วนเหลือและวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ ซึ่งเกณฑ์ที่ใช้วัดประสิทธิภาพคือ ความกว้างของช่วงความเชื่อมั่น ความน่าจะเป็นครอบคลุม อัตราผลบวกเทียม และอัตราผลลบเทียม งานวิจัยนี้ศึกษาสัมประสิทธิ์การถดถอยใน 2 ลักษณะได้แก่ บางเบาอย่างอ่อนและบางเบาอย่างรุนแรง และจำลองข้อมูลจากการแจกแจงแบบปกติหลายตัวแปรโดยใช้เมทริกซ์ความแปรปรวนร่วมของค่าคลาดเคลื่อนที่แตกต่างกัน ทั้งหมด 8 กรณี ผลการศึกษาพบว่าวิธีบูตแสตรปแบบสุ่มส่วนเหลือตัวประมาณลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์มีประสิทธิภาพสูงสุดในแง่การให้ความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ยสั้นที่สุดในเกือบทุกกรณี และวิธีบูตแสตรปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระตัวประมาณลาสโซ่และพาร์เชียลริดจ์มีประสิทธิภาพสูงสุดเมื่อพิจารณาด้วยเกณฑ์อัตราผลบวกเทียม อย่างไรก็ตาม เมื่อพิจารณาด้วยเกณฑ์ความน่าจะเป็นครอบคลุมและอัตราผลลบเทียมพบว่าไม่ปรากฏวิธีการบูตแสตรปแบบใดแบบหนึ่งที่มีประสิทธิภาพสูงสุดอย่างชัดเจน

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380223626 : MAJOR STATISTICS

KEYWORD: high – dimensional regression, lasso regression, adaptive lasso regression, ridge regression, bootstrap, confidence intervals

Parit Chancherngpanich : A COMPARISON OF RESIDUAL BOOTSTRAP AND PAIR BOOTSTRAP FOR ADAPTIVE LASSO + PARTIAL RIDGE METHOD TO CONSTRUCT CONFIDENCE INTEVALS FOR PARAMETERS IN HIGH-DIMENSIONAL SPARSE LINEAR MODELS. Advisor: Asst. Prof. VITARA PUNGPAPONG, Ph.D.

This research is aimed to propose a method, called bootstrap adaptive lasso + partial ridge (ALPR), to construct confidence intervals of regression coefficients in high – dimensional data and compare its performance with bootstrap lasso + partial ridge (LPR). The ALPR is a two-stage estimator. The adaptive lasso is used to select variables and the partial ridge is used to refit the coefficients. Here we perform two techniques of bootstrap which are residual bootstrap (rB) and paired bootstrap (pB). Hence, there are four bootstrap methods to be studied which are rBALPR, rBLPR, pBALPR and pBLPR while mean intervals width, coverage probabilities, false positive rate and false negative rate are used to measure and compare their performance. Simulation studies in 8 cases of high – dimensional data and all of them are generated independently from multivariate normal distribution with different types of covariance matrix. We also consider two cases of coefficients which are weak sparsity and hard sparsity. Our simulation studies show that the residual bootstrap adaptive lasso + partial ridge (rBALPR) produces shortest width of confidence intervals of regression coefficients on average for most cases and the paired bootstrap lasso + partial ridge (pBLPR) is the most effective method in terms of providing lowest false positive rate. However, it is not obvious that which bootstrap method is the best in terms of providing highest coverage probabilities and lowest false negative rate.

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

Academic Year: 2021 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงไปได้ด้วยความเมตตากรุณาและเอาใจใส่อย่างดียิ่งจากผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้ความกรุณารับเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ ช่วยให้คำแนะนำ คำปรึกษา รวมไปถึงชี้แนะแนวทางในการเขียนวิทยานิพนธ์และแนวคิดข้อคิดเห็นต่างๆที่เป็นประโยชน์ เพื่อปรับปรุงแก้ไขวิทยานิพนธ์ ตลอดจนให้ความช่วยเหลือและอบรมสั่งสอน ผู้วิจัยขอขอบพระคุณท่านอาจารย์เป็นอย่างสูงด้วยความเคารพอย่างยิ่ง

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี และผู้ช่วยศาสตราจารย์ ดร.ทพญ.นัจชลี ศรีมณีกาญจน์ กรรมการสอบวิทยานิพนธ์ ท่านอาจารย์ทั้งสามท่านได้กรุณาสละเวลาเป็นกรรมการสอบครั้งนี้ ตลอดจนช่วยให้ความรู้ คำแนะนำที่มีประโยชน์ยิ่งในการเขียนวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ถ่ายทอดวิชาความรู้เสมอมาจนสำเร็จการศึกษาในครั้งนี้

ผู้วิจัยขอขอบพระคุณ คุณวิศิลป์ ชาญเชิงพานิช และคุณทิพยดา ชาญเชิงพานิช ผู้เป็นบิดาและมารดาของผู้วิจัยซึ่งให้โอกาสในการศึกษาที่ดีและเป็นผู้อยู่เบื้องหลังความสำเร็จของผู้วิจัยเสมอมา รวมถึงครอบครัวที่เป็นกำลังใจให้ผู้วิจัยสามารถจัดทำวิทยานิพนธ์เล่มนี้จนสำเร็จลุล่วง และขอขอบคุณเพื่อน ๆ ทุกคนที่ช่วยเหลือ ให้คำแนะนำ และเป็นกำลังใจตลอดมา

ท้ายสุดผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์เล่มนี้จักก่อเกิดประโยชน์แก่ผู้สนใจศึกษาค้นคว้าในเรื่องดังกล่าว คุณความดีใดที่เกิดขึ้นจากวิทยานิพนธ์เล่มนี้ผู้วิจัยขอมอบให้แก่บุคคลทุกท่านที่ได้กล่าวมาทั้งหมดนี้ ตลอดจนท่านผู้เขียนตำราที่ผู้วิจัยนำมาอ้างอิงและเรียบเรียงเป็นวิทยานิพนธ์เล่มนี้ หากวิทยานิพนธ์เล่มนี้มีข้อผิดพลาดประการใด ผู้วิจัยขอน้อมรับไว้แต่เพียงผู้เดียวและขออภัยไว้ ณ โอกาสนี้

พริษฐ์ ชาญเชิงพานิช

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การถดถอยเชิงเส้นเป็นวิธีทางสถิติที่นิยมใช้กันอย่างแพร่หลายในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยการวิเคราะห์การถดถอยเชิงเส้นจะใช้วิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Squares Method) ในการประมาณค่าสัมประสิทธิ์การถดถอย อย่างไรก็ตาม วิธี OLS มีข้อจำกัดคือ เมื่อข้อมูลมีมิติสูงหรือจำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่างของข้อมูล (p>n) วิธี OLS จะไม่สามารถหาค่าของตัวประมาณสัมประสิทธิ์การถดถอยได้ [1] นอกจากนี้อาจเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์กันเองสูงซึ่งส่งผลให้ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธี OLS มีความไม่เสถียร [2] การวิเคราะห์ข้อมูลที่มีมิติสูงจึงนิยมใช้วิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Regression) โดยการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษมีหลากหลายวิธีแต่ที่นิยมใช้กันอย่างแพร่หลายได้แก่ การถดถอยลาสโซ่ (Lasso Regression) การถดถอยลาสโซ่แบบปรับปรุง (Adaptive Lasso Regression) และการถดถอยแบบริดจ์ (Ridge Regression)

การวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษทั้งสามวิธีที่กล่าวมาสามารถหาค่าประมาณสัมประสิทธิ์การถดถอยในกรณีที่ข้อมูลมีมิติสูงได้แต่การทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยที่ว่า $H_0: \beta_j = 0$ และ $H_1: \beta_j \neq 0$ ยังคงเป็นประเด็นที่ท้าทายเนื่องจากตัวสถิติทดสอบทีหรือเอฟที่ใช้ในวิธี OLS ไม่สามารถนำมาใช้ได้ ดังนั้นวิธีที่นิยมใช้ในการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูงคือวิธีบูตสแตรป (Bootstrap)

Liu & Yu (2013) นำเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + OLS [3] ซึ่งเป็นตัวประมาณแบบสองขั้นตอน โดยขั้นตอนที่หนึ่งใช้วิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระและขั้นตอนที่สองใช้วิธี OLS ในการประมาณค่าสัมประสิทธิ์การถดถอย ทว่าวิธีบูตแสตรป Lasso + OLS มักจะประสบปัญหาช่วงความเชื่อมั่นที่สร้างขึ้นมักจะไม่ครอบคลุมค่าของสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่มีค่าน้อยมากแต่ไม่เท่ากับศูนย์ (น้อยกว่า $\frac{1}{\sqrt{n}}$) เนื่องจากการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่ในขั้นตอนที่หนึ่งมักจะให้ค่าประมาณสัมประสิทธิ์การถดถอยของตัวแปรอิสระดังกล่าวเป็นศูนย์ ส่งผลให้ตัวแปรอิสระดังกล่าวไม่ได้รับคัดเลือกให้นำไปหาค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธี OLS เรียกปัญหาลักษณะนี้ว่าความน่าจะเป็นครอบคลุมต่ำ (Low Coverage Probabilities) [4]

Liu et al. (2020) ได้นำเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge (LPR) ซึ่งเป็นตัวประมาณแบบสองขั้นตอนเช่นกัน โดยขั้นตอนที่หนึ่งใช้วิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระและขั้นตอนที่สองใช้ฟังก์ชันการลงโทษแบบ L2 – Norm แก่ตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่เท่านั้น ซึ่งเปรียบเสมือนเป็นการนำตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่ไปหาค่าประมาณสัมประสิทธิ์การถดถอยใหม่อีกครั้ง เนื่องจากวิธีริดจ์มักจะให้ค่าประมาณสัมประสิทธิ์การถดถอยเข้าใกล้ศูนย์แต่ไม่เท่ากับศูนย์ ผลการศึกษาพบว่าในกรณีที่ปรากฏสัมประสิทธิ์การถดถอยบางตัวที่มีค่าน้อยมากแต่ไม่เท่ากับศูนย์ วิธีบูตแสตรปแบบ Lasso + Partial Ridge ให้ความน่าจะเป็นครอบคลุมสูงกว่าวิธี Lasso + OLS [4]

ในการศึกษาครั้งนี้ ผู้วิจัยมีจุดประสงค์ที่จะนำเสนอวิธีบูตแสตรปตัวประมาณแบบ Adaptive Lasso + Partial Ridge โดยการเปลี่ยนจากวิธี Lasso เป็น Adaptive Lasso เนื่องจากวิธี Adaptive Lasso มีคุณสมบัติที่โดดเด่นประการหนึ่งคือคุณสมบัติออราเคิลซึ่งเป็นคุณสมบัติที่สามารถคัดเลือกตัวแปรเข้าตัวแบบเสมือนทราบตัวแบบที่แท้จริง [5] และจากการทบทวนวรรณกรรมที่ผ่านมายังไม่พบว่ามีการศึกษาวิธีบูตแสตรปตัวประมาณแบบดังกล่าว ดังนั้นผู้วิจัยจึงสนใจศึกษาเกี่ยวกับประเด็นนี้ โดยจะทำการเปรียบเทียบกับวิธีบูตแสตรปตัวประมาณแบบ Lasso + Partial Ridge ทั้งนี้ผู้วิจัยจะทดลองบูตแสตรป 2 วิธีคือ วิธีสุ่มส่วนเหลือ (Residual Bootstrap) และวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ (Paired Bootstrap) จากนั้นจึงทำการเปรียบเทียบและวิเคราะห์ผลลัพธ์โดยใช้ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals) ความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าของพารามิเตอร์หรือค่าจริง (Coverage Probabilities) อัตราผลบวกเทียม (False Positive Rate) และอัตราผลลบเทียม (False Negative Rate) เป็นเกณฑ์การวัดประสิทธิภาพเพื่อหาวิธีที่เหมาะสมและมีประสิทธิภาพที่สุดในการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูง

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบวิธี Paired Bootstrap และ Residual Bootstrap ของตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge และ Lasso + Partial Ridge โดยใช้ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals) ความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าของพารามิเตอร์ (Coverage Probability) อัตราผลบวกเทียม (False Positive Rate) และอัตราผลลบเทียม (False Negative Rate) เป็นเกณฑ์ในการวัดประสิทธิภาพ โดยทำการศึกษาทั้งในกรณีที่สัมประสิทธิ์การถดถอยมีลักษณะ Hard Sparsity และ Weak Sparsity อีกทั้งตัวแปรอิสระมีความสัมพันธ์กันเองในหลากหลายรูปแบบ

1.3 สมมติฐานการวิจัย

การใช้วิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge อาจมีประสิทธิภาพที่สูงกว่าวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge

1.4 ขอบเขตของการวิจัย

การศึกษาครั้งนี้ทำการจำลองข้อมูลแบบตัดขวาง (Cross - Sectional Data) และลักษณะของข้อมูลจัดอยู่ในประเภทข้อมูลที่มีมิติสูง (High - Dimensional Data) โดยลักษณะข้อมูลที่จำลองขึ้นเพื่อการศึกษาครั้งนี้จำลองข้อมูลในลักษณะเดียวกับการศึกษาของ Liu et al. (2020) โดยมาจากตัวแบบดังนี้

$$y = X\beta + \varepsilon \qquad \text{...(1.1)}$$

โดยที่

$\boldsymbol{y}$ คือ เวกเตอร์ของตัวแปรตามขนาด n

X คือ เมทริกซ์ของตัวแปรอิสระขนาด n x p

β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด p

ε คือ เวกเตอร์ของค่าความเคลื่อนขนาด n โดยที่ $E(\varepsilon_i) = 0$ และ $Var(\varepsilon_i) = \sigma^2 I_n$

ซึ่งสามารถเขียนได้ในรูปของ

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

โดยทำการศึกษาข้อมูลในลักษณะดังต่อไปนี้

1. กำหนดขนาดตัวอย่างข้อมูล (n) เท่ากับ 200 และจำนวนตัวแปรสระ (p) เท่ากับ 500
2. กำหนดให้เวกเตอร์ของตัวแปรอิสระ $x_1, x_2, \ldots x_n$ เป็นเวกเตอร์ที่มีความเป็นอิสระต่อกันและตัวแปรอิสระภายในเวกเตอร์มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) $N(0, \sum)$ โดยการศึกษาครั้งนี้จะพิจารณา $\sum$ ใน 2 กรณีหลักดังนี้

1) $\sum$ เป็นเมทริกซ์ที่มีค่าสหสัมพันธ์แบบ Toeplitz : $\sum_{ij} = \rho^{|i-j|}$, โดยพิจารณาเพิ่มเติมอีกสองกรณีย่อยคือ เมื่อ $\rho = 0.5$ และ $\rho = 0.9$
2) $\sum$ เป็นเมทริกซ์ที่มีค่าสหสัมพันธ์แบบ Equal Correlation : $\sum_{ij} = \rho$, โดยพิจารณาเพิ่มเติมอีกสองกรณีย่อยคือ เมื่อ $\rho = 0.5$ และ $\rho = 0.9$

3. ศึกษาเวกเตอร์สัมประสิทธิ์การถดถอย (β) ใน 2 กรณี ดังนี้
   1) กรณี Hard Sparse Coefficients :

$$\beta_j = \begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots,10 \\ 0; \text{ อื่นๆ} \end{cases} \quad \ldots(1.2)$$

   2) กรณี Weak Sparse Coefficients

$$\beta_j = \begin{cases} N(1,0.001); j = 1,2,\ldots,10 \\ \beta_j = \frac{1}{(j+3)^2}; \ j = 1,2,\ldots,490 \end{cases} \quad \ldots(1.3)$$

4. ศึกษาภายใต้ค่าความแปรปรวนของค่าความคลาดเคลื่อน σ^2 โดยที่กำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$SNR = \frac{\|X\beta\|_2^2}{(n\sigma^2)} = 10 \quad \ldots(1.4)$$

5. ศึกษาภายใต้ตัวแบบเชิงเส้น

$$y_i = x_i^T \beta + \epsilon_i \, , \epsilon_i \sim N\,(0, \sigma^2), \ i = 1,\ldots,n \quad \ldots(1.5)$$

6. การจำลองข้อมูล x_i และเวกเตอร์สัมประสิทธิ์การถดถอย β ในแต่ละกรณีจะกระทำเพียงครั้งเดียวเท่านั้น แต่จะทำการจำลอง $Y = (y_1, y_2, .., y_n)^T$ จากตัวแบบเชิงเส้นในขั้นตอนที่ 5 โดยสุ่ม error terms จำนวน 50 replications

โดยสรุปจะทำการศึกษาทั้งหมด 8 กรณีดังนี้

กรณีที่ 1 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.5$)

กรณีที่ 2 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.9$)

กรณีที่ 3 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.5$)

กรณีที่ 4 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.9$)

กรณีที่ 5 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.5$)

กรณีที่ 6 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.9$)

กรณีที่ 7 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.5$)

กรณีที่ 8 : สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.9$)

1.5 วิธีดำเนินการวิจัย

1.5.1 ค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง

1.5.2 กำหนดค่าเริ่มต้นสำหรับการจำลองข้อมูลในแต่ละกรณีที่ทำการศึกษา

1) กำหนดขนาดตัวอย่าง (n) เท่ากับ 200

2) กำหนดจำนวนตัวแปรอิสระ (p) เท่ากับ 500

1.5.3 ทำการจำลองข้อมูลทั้งหมด 8 กรณีตามขอบเขตการวิจัย

1.5.4 ในแต่ละกรณีที่ทำการศึกษานั้นจะใช้วิธีบูตสแตรปแบบ Residual Bootstrap และ Paired Bootstrap สำหรับตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge และ Lasso + Partial Ridge เพื่อสร้างช่วงความเชื่อมั่น (Confidence Interval) ที่ 95% ทำให้ได้ (L_j, U_j) โดยที่ L_j และ U_j คือขอบเขตล่างและขอบเขตบนของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอย β_j

1.5.5 นำผลที่ได้จากข้อ 1.5.4 มาคำนวณหาค่าดังนี้

1) ความน่าจะเป็นครอบคลุม (Coverage Probability)

2) ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals)

3) อัตราผลบวกเทียม (False Positive Rate)

4) อัตราผลลบเทียม (False Negative Rate)

1.5.6 เปรียบเทียบผลการวิเคราะห์ในข้อที่ 1.5.5

1.5.7 สรุปผลการศึกษา

1.6 แนวทางการวิเคราะห์ข้อมูลและสถิติที่ใช้ในการวิเคราะห์

เกณฑ์ที่ใช้วัดประสิทธิภาพสำหรับวิธีบูตแสตรปตัวประมาณแบบต่างๆ มีดังต่อไปนี้

1.6.1 ความน่าจะเป็นครอบคลุม (Coverage Probability) คือ ความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าของพารามิเตอร์หรือค่าจริง ซึ่งสามารถคำนวณได้ดังนี้

$$CP = \frac{\sum_{i=1}^{B} I_{[L_i,U_i]}(\beta_j)}{B} \quad \text{...(1.6)}$$

โดยที่ $I_{[L_i,U_i]}(\beta_j)$ จะมีค่าเท่ากับ 1 เมื่อ β_j อยู่ในช่วง $[L_i, U_i]$ และเท่ากับ 0 เมื่ออยู่นอกช่วง L_i และ U_i เป็นขอบเขตล่างและขอบเขตบนของช่วงความเชื่อมั่นในรอบที่ i ตามลำดับ และ B เป็นจำนวนครั้งที่ทำการสร้างช่วงความเชื่อมั่น

1.6.2 ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals) คือ ขอบเขตบนของช่วงความเชื่อมั่นลบขอบเขตล่างของช่วงความเชื่อมั่น ซึ่งสามารถคำนวณได้ดังนี้

$$CI_j = U_j - L_j \quad \text{...(1.7)}$$

เมื่อ U_j และ L_j คือขอบเขตบนและขอบเขตล่างของช่วงความเชื่อมั่นสำหรับแต่ละ β_j ตามลำดับ

1.6.3 อัตราผลบวกเทียม (False Positive Rate) คือ การวัดความน่าจะเป็นที่เกิดจากความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ มีค่าไม่เท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์การถดถอยจริง β_j เท่ากับศูนย์ โดยในการวิจัยครั้งนี้จะถือว่า $\hat{\beta}_j$ ไม่เท่ากับศูนย์ก็ต่อเมื่อช่วงความเชื่อมั่นที่ 95% ของสัมประสิทธิ์การถดถอยไม่ครอบคลุมค่าศูนย์และจะเท่ากับศูนย์ก็ต่อเมื่อช่วงความเชื่อมั่นที่ 95% ครอบคลุมค่าศูนย์ ซึ่งสามารถคำนวณอัตราผลบวกเทียมได้ดังนี้

$$FPR = \frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0 \text{ and } \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{0 \notin (L_j, U_j)\}}} \quad \text{...(1.8)}$$

เมื่อ p คือจำนวนตัวแปรอิสระ

1.6.4 อัตราผลลบเทียม (False Negative Rate) คือ การวัดความน่าจะเป็นที่เกิดจากความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์ $\hat{\beta}_j$ มีค่าเท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์จริง β_j ไม่เท่ากับศูนย์ โดยในการวิจัยครั้งนี้จะถือว่า $\hat{\beta}_j$ ไม่เท่ากับศูนย์ก็ต่อเมื่อช่วงความเชื่อมั่นที่ 95% ของสัมประสิทธิ์การถดถอยไม่ครอบคลุมค่าศูนย์

และจะเท่ากับศูนย์ก็ต่อเมื่อช่วงความเชื่อมั่นที่ 95% ครอบคลุมค่าศูนย์ ซึ่งสามารถคำนวณอัตราผลลบเทียมได้ดังนี้

$$FNR = \frac{\sum_{j=1}^{p} 1_{\{\hat{\beta}_j = 0 \; and \; \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{0 \in (L_j, U_j\}}} \quad ...(1.9)$$

เมื่อ p คือจำนวนตัวแปรอิสระ

1.7 การนำเสนอ

นำเสนอข้อมูลในรูปตารางเพื่อเปรียบเทียบว่าวิธีการบูตแสตรปตัวประมาณแบบใดมีประสิทธิภาพสูงสุด

1.8 ประโยชน์ที่คาดว่าจะได้รับ

ผลการศึกษาจะแสดงให้เห็นถึงประสิทธิภาพและความแตกต่างระหว่างวิธี Residual Bootstrap และ Paired Bootstrap สำหรับตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge และ Lasso + Partial Ridge ซึ่งเป็นประโยชน์ในด้านการอนุมานเชิงสถิติสำหรับการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยในกรณีที่ข้อมูลมีมิติสูง

บทที่ 2
ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

2.1 การวิเคราะห์การถดถอยเชิงเส้น (OLS)

การวิเคราะห์การถดถอยเชิงเส้นเมื่อข้อมูลมีตัวอย่างขนาด n และตัวแปรอิสระขนาด p สามารถเขียนเป็นสมการของตัวแบบได้ดังนี้

$$Y = X\beta + \varepsilon \quad \text{...(2.1)}$$

เมื่อ Y คือ เวกเตอร์ของตัวแปรตามขนาด n

X คือ เมทริกซ์ของตัวแปรอิสระขนาด n x p

β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด p

ε คือ เวกเตอร์ของค่าความเคลื่อนขนาด n โดยที่ $E(\varepsilon_i) = 0$ และ $Var(\varepsilon_i) = \sigma^2 I_n$

ในการหาค่าของตัวประมาณสัมประสิทธิ์การถดถอย $(\hat{\beta})$ จะหาได้จากวิธีกำลังสองน้อยที่สุด ซึ่งเขียนได้ดังสมการ

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 = (X^T X)^{-1} X^T y \quad \text{...(2.2)}$$

สำหรับการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอย $\beta_1, \beta_2, \ldots, \beta_P$ ที่ได้จากตัวประมาณ $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ ถือเป็นประเด็นสำคัญเนื่องจากจะทำให้เข้าใจความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามได้และมีประโยชน์ในการคัดเลือกตัวแปรเข้าตัวแบบ โดยทั่วไปจะใช้ค่าสถิติ $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ เพื่อทดสอบว่าพารามิเตอร์ $\beta_1, \beta_2, \ldots, \beta_P$ มีค่าเท่ากับ 0 หรือไม่

$$\begin{matrix} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{matrix} ; j = 1,2,\ldots,p$$

ตัวสถิติทดสอบที่ใช้ในการทดสอบสมมติฐานคือตัวสถิติทดสอบทีที่องศาอิสระ $n - p - 1$

$$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{n-p-1} \quad \text{...(2.3)}$$

โดยจะปฏิเสธสมมติฐานว่าง (H_0) ที่ระดับนัยสำคัญ α ก็ต่อเมื่อค่า $|t|$ มีค่ามากกว่า $t_{(1-\frac{\alpha}{2}, n-p-1)}$ ทั้งนี้ ตัวสถิติทดสอบทีมีการแจกแจงแบบสติวเด้นส์ที

2.2 การประมาณค่าสัมประสิทธิ์ด้วยการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ

การประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ หรือ Penalized Regression เป็นวิธีที่พัฒนามาจากวิธีกำลังสองน้อยที่สุด เนื่องจากในกรณีที่ข้อมูลมีมิติสูง วิธีกำลังสองน้อยที่สุดจะประสบปัญหาเมทริกซ์ X^TX ในสมการที่ 2 จะเป็นเมทริกซ์เอกฐาน (Singular Matrix) ดังนั้นเมทริกซ์ X^TX จึงไม่มี เมทริกซ์ผกผัน ส่งผลให้ไม่สามารถแก้สมการได้ หรือกล่าวอีกนัยหนึ่งคือ มีตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุดมากกว่าหนึ่งชุดซึ่งทำให้ผลรวมความคลาดเคลื่อนกำลังสองมีค่าน้อยที่สุด

การประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษเป็นการเพิ่มฟังก์ชันการลงโทษ (Penalty function) $P_\lambda(\beta)$ เข้าไปในฟังก์ชันเป้าหมายของวิธีกำลังสองน้อยที่สุด

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j\beta_j \right\|^2 + P_\lambda(\beta) \quad ...(2.4)$$

สำหรับฟังก์ชันการลงโทษ $P_\lambda(\beta)$ มีหลายรูปแบบแต่ที่นิยมใช้กันอย่างแพร่หลายคือฟังก์ชันการลงโทษแอลวันนอร์ม (L1 – Norm) และแอลทูนอร์ม (L2 – Norm) เขียนแสดงได้ดังสมการที่ 2.5 และ 2.6 ตามลำดับ

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \quad ...(2.5)$$

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j{}^2 \quad ...(2.6)$$

2.2.1 วิธี Lasso Regression

วิธีลาสโซ่ (Lasso) นำเสนอโดย Tipshirani (1996) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L1 – Norm โดยสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 2.7

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j\beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad ...(2.7)$$

การใช้ฟังก์ชันลงโทษแบบแอลวันนอร์ม (L1 – Norm) จะทำให้เวกเตอร์ตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่ ซึ่งเขียนแทนด้วย $\hat{\beta}_{Lasso}$ ประกอบด้วยค่าศูนย์จำนวนมาก (Sparse Vector) ทั้งนี้ขึ้นอยู่กับพารามิเตอร์การปรับ (λ) หากพารามิเตอร์การปรับมีค่ามากจะส่งผลให้เวกเตอร์ $\hat{\beta}_{Lasso}$ มีจำนวนค่าศูนย์มาก หากพารามิเตอร์การปรับมีค่าน้อยจะส่งผลให้เวกเตอร์ $\hat{\beta}_{Lasso}$ มีจำนวนค่าศูนย์น้อย และในกรณีที่พารามิเตอร์การปรับเท่ากับ 0 การถดถอยลาสโซ่จะกลับมาเป็นการถดถอยแบบดั้งเดิม [6]

การเลือกพารามิเตอร์การปรับ (λ) ที่เหมาะสมเป็นสิ่งที่จำเป็นสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยลาสโซ่ โดยทั่วไปนิยมใช้วิธี Cross Validation (CV) ซึ่งเป็นวิธีที่ใช้วัดประสิทธิภาพของตัวแบบเมื่อต้องการทดสอบว่าค่าพารามิเตอร์ปรับที่เท่าใดส่งผลให้ผลรวมความคลาดเคลื่อนกำลังสองน้อยที่สุด โดยหลักการของวิธี Cross - Validation มีดังนี้

1. กำหนด λ เป็นเซ็ตของพารามิเตอร์ปรับที่ต้องการทดสอบ ซึ่งประกอบด้วยสมาชิกจำนวน m ตัว ดังนี้ $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$
2. แบ่งข้อมูลออกเป็นชุดย่อยๆ จำนวน k ชุด โดยใช้สัญลักษณ์ $d_1, d_2, \ldots, d_k$ แทนข้อมูลย่อยชุดที่ $1,2,\ldots,k$ ตามลำดับ ทั้งนี้จำนวนข้อมูลย่อยในแต่ละชุดต้องมีจำนวนเท่ากัน
3. สำหรับครั้งที่ i เมื่อ $i = 1,2,\ldots,m$

   3.1 ให้ใช้ $\lambda = \lambda_i$ และภายใต้ครั้งที่ $\boldsymbol{i}$ ให้ทำการแบ่งข้อมูล j ครั้ง เมื่อ $j = 1,2,\ldots,k$

   3.2 สำหรับครั้งที่ j เมื่อ $j = 1,2,\ldots,k$ ให้ใช้ข้อมูลย่อยทุกชุด ยกเว้นชุด d_j ในการสร้างตัวแบบด้วยวิธีการถดถอยแบบลาสโซ่ จากนั้นใช้ข้อมูลย่อยชุด d_j เป็นข้อมูลชุดทดสอบเพื่อคำนวณผลรวมความคลาดเคลื่อนกำลังสอง โดยจะใช้สัญลักษณ์ $RSS_{\lambda_i,j}$

   3.3 จากนั้นคำนวณค่าเฉลี่ยของ RSS ที่ได้จากการใช้ $\lambda = \lambda_i$ ซึ่งเขียนได้ดังสมการที่ 2.8

$$CV(\lambda_i) = \frac{1}{k}\sum_{j=1}^{k} RSS_{\lambda_i,j} \qquad \ldots(2.8)$$

4. เลือก λ_i ที่ทำให้ $CV(\lambda_i)$ มีค่าน้อยที่สุด ซึ่ง λ_i ดังกล่าวจะเป็นพารามิเตอร์ปรับที่เหมาะสมที่สุดสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่

ด้วยเหตุที่วิธีลาสโซ่มีคุณสมบัติทำให้ได้เวกเตอร์ตัวประมาณสัมประสิทธิ์การถดถอยที่ประกอบด้วยค่าศูนย์จำนวนมาก ส่งผลให้วิธีลาสโซ่สามารถหาค่าประมาณสัมประสิทธิ์การถดถอยพร้อมทั้งคัดเลือกตัวแปรเข้าตัวแบบได้ในเวลาเดียวกันทำให้ตัวแบบที่ได้ง่ายต่อการแปรผลลัพธ์อีกทั้งยังแก้ปัญหาที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงได้ (Multicollinearity) [2] อย่างไรก็ตามค่าประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธีลาสโซ่มักไม่มีความคงเส้นคงวา ส่งผลให้การคัดเลือกตัวแปรเข้าตัวแบบมักไม่มีความคงเส้นคงวา [7]

2.2.2 วิธี Adaptive Lasso Regression

วิธีลาสโซ่แบบปรับปรุง นำเสนอโดย Zou (2006) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L1 – Norm และเพิ่มเงื่อนไขการให้ค่าน้ำหนักแก่พารามิเตอร์ที่แตกต่างกัน โดยสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 2.9

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j| \text{ ; โดยที่ } \widehat{w}_j \begin{cases} \frac{1}{|\hat{\beta}_{OLS}|} ; n > p \\ \frac{1}{|\hat{\beta}_{Ridge}|} ; n < p \end{cases} \quad ...(2.9)$$

วิธีลาสโซ่แบบปรับปรุงมีคุณสมบัติที่ทำให้ได้ตัวประมาณสัมประสิทธิ์การถดถอยที่ประกอบด้วยค่าศูนย์จำนวนมาก ดังนั้นจึงมีคุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบเช่นเดียวกับวิธีลาสโซ่ และการเพิ่มเงื่อนไขการให้ค่าน้ำหนักแก่พารามิเตอร์ที่แตกต่างกันยังช่วยแก้ปัญหาความไม่คงเส้นคงวาที่ประสบในวิธีลาสโซ่ได้ส่งผลให้วิธีลาสโซ่แบบปรับปรุงมีคุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบเสมือนทราบตัวแบบที่แท้จริงหรือเรียกว่าคุณสมบัติออราเคิล (Oracle Property) [5]

2.2.3 วิธี Ridge Regression

วิธีริดจ์ (Ridge Regression) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L2 – Norm สามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 2.10

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \quad ...(2.10)$$

โดยตัวประมาณสัมประสิทธิ์การถดถอยบางตัวที่ได้จากวิธีริดจ์จะถูกบีบค่าให้เข้าใกล้ศูนย์แต่ไม่เท่ากับศูนย์ ดังนั้นวิธีริดจ์จึงเหมาะแก่การวิเคราะห์ข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กแต่ไม่เท่ากับศูนย์จำนวนมาก นอกจากนี้วิธีริดจ์ยังช่วยแก้ปัญหาที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงได้ อย่างไรก็ตาม คุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบไม่ปรากฏในวิธีริดจ์ [8]

2.3 ตัวประมาณ Lasso + Partial Ridge (LPR)

ตัวประมาณ $\hat{\beta}_{LPR}$ นำเสนอโดย Liu et al. (2020) เป็นตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากสองขั้นตอน โดยขั้นตอนที่หนึ่งใช้วิธีการประมาณสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระ และขั้นตอนที่สองใช้ฟังก์ชันการลงโทษ L2 – Norm เพื่อหาค่าประมาณสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่ เนื่องจากการใช้ฟังก์ชัน L2 – Norm จะทำให้ได้สัมประสิทธิ์ขนาดเล็กแต่ไม่เท่ากับศูนย์ซึ่งสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 2.11

$$\hat{\beta}_{LPR} = \underset{\beta}{argmin}\left\{\frac{1}{2n}\|Y - X\beta\|_2^2 + \frac{\lambda_2}{2}\sum_{j \notin \hat{S}}\beta_j^2\right\} \quad ...(2.11)$$

โดยที่ $S = \{j \in \{1,\ldots,p\} : \beta_j \neq 0\}$ คือ ซัพพอร์ทเซตของตัวแปรอิสระ

$\hat{S} = \{j \in \{1,\ldots,p\} : (\hat{\beta}_{lasso}) \neq 0\}$ คือ เซตของตัวแปรอิสระที่ถูกคัดเลือกโดยวิธีลาสโซ่

2.4 วิธีบูตสแตรป rBLPR และ pBLPR

การทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยในกรณีที่ข้อมูลมีมิติสูงเป็นประเด็นที่ท้าทาย เนื่องจากการลู่เข้าเชิงการแจกแจง (Asymptotic Distribution) ของตัวประมาณสัมประสิทธิ์การถดถอยที่นิยมใช้ เช่นตัวประมาณลาสโซ่มีความซับซ้อน [9] อีกทั้งตัวสถิติทดสอบทีหรือเอฟที่ใช้ในวิธีกำลังสองน้อยที่สุดไม่สามารถนำมาใช้ได้ ดังนั้นวิธีที่นิยมคือวิธีบูตสแตรป (Bootstrap)

2.4.1 วิธี rBLPR (Residual Bootstrap Lasso + Partial Ridge)

วิธีบูตแสตรปส่วนเหลือ rBLPR เป็นการสุ่มวนส่วนเหลือเพื่อสร้างตัวอย่างบูตแสตรปซึ่งมีจุดประสงค์เพื่อใช้หาช่วงความเชื่อมั่น (Confidence Intervals) สำหรับสัมประสิทธิ์การถดถอย β_j เมื่อ $j = 1,2,\ldots,p$ ทั้งนี้ส่วนเหลือที่เลือกใช้ในวิธี rBLPR คำนวณได้จากส่วนต่างระหว่างค่าสังเกต (y_i) และค่าพยากรณ์ $(\hat{y}_i = X\hat{\beta}_{Lasso+OLS})$ โดยวิธี rBLPR มีขั้นตอนดังนี้

กำหนดให้

$S = \{j \in \{1,\ldots,p\} : \beta_j \neq 0\}$ คือซัพพอร์ทเซตของตัวแปรอิสระ

$\hat{S} = \{j \in \{1,\ldots,p\} : (\hat{\beta}_{lasso}) \neq 0\}$ คือเซตของตัวแปรอิสระที่ถูกคัดเลือกโดยวิธีลาสโซ่

$\hat{S}^*_{rBlasso} = \{j \in \{1,\ldots,p\} : (\hat{\beta}^*_{rBlasso}) \neq 0\}$ คือ เซตของตัวแปรอิสระที่ถูกคัดเลือกโดยวิธีลาสโซ่ ซึ่งใช้ข้อมูลชุด (X, y^*_{rboot})

1. คำนวณค่าของสัมประสิทธิ์ $\hat{\beta}_{Lasso+OLS}$

$$\hat{\beta}_{Lasso+OLS} = \underset{\beta:\, \beta_{\hat{S}^c}=0}{argmin}\left\{\frac{1}{2n}\|Y - X\beta\|_2^2\right\}; \text{เมื่อ } \beta_{\hat{S}^c} = \{\beta_j : j \notin \hat{S}\} \quad ...(2.12)$$

2. คำนวณเวกเตอร์ของค่าพยากรณ์

$$\hat{y} = X\hat{\beta}_{Lasso+OLS} \quad ...(2.13)$$

3. คำนวณเวกเตอร์ส่วนเหลือ

$$\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta}_{Lasso+OLS} \quad ...(2.14)$$

4. คำนวณเซตของ Centered residual

$$\{\hat{\epsilon}_i - \bar{\hat{\epsilon}}, i = 1,\dots,n\} \text{ เมื่อ } \bar{\hat{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i \quad \text{...(2.15)}$$

5. ทำการสุ่ม Centered residual แบบใส่คืน

$$\epsilon^* = (\epsilon_1^*,\dots,\epsilon_n^*)^T \quad \text{...(2.16)}$$

6. ทำการสร้างตัวอย่างบูตแสตรป

$$y^*{}_{rboot} = X\hat{\beta}_{Lasso+OLS} + \epsilon^* \quad \text{...(2.17)}$$

7. คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยใช้วิธีลาสโซ่จากชุดข้อมูล $(X, y^*{}_{rboot})$

$$\hat{\beta}^*_{rBLasso} = \underset{\beta}{argmin}\left\{\frac{1}{2n}\left\|y^*{}_{rboot} - X\beta\right\|_2^2 + \lambda_1\|\beta\|_1\right\} \quad \text{...(2.18)}$$

8. คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยมีการเพิ่มฟังก์ชันการลงโทษแบบ L2 – Penalty สำหรับตัวแปรอิสระที่ไม่ได้ถูกคัดเลือกและใช้ชุดข้อมูล $(X, y^*{}_{rboot})$

$$\hat{\beta}^*_{rBLPR} = \underset{\beta}{argmin}\left\{\frac{1}{2n}\left\|y^*{}_{rboot} - X\beta\right\|_2^2 + \frac{\lambda_2}{2}\sum_{j\notin \hat{S}_{*rBlasso}}\beta_j^2\right\} \quad \text{...(2.19)}$$

9. ทำซ้ำในขั้นตอนที่ 5 – 8 ไป B รอบจนได้ $\hat{\beta}^*_{rBLPR}{}^{(1)}, \hat{\beta}^*_{rBLPR}{}^{(2)}, \dots, \hat{\beta}^*_{rBLPR}{}^{(B)}$
10. สร้างช่วงความเชื่อมั่นที่ $(1-\alpha)\%$ สำหรับสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ดังนี้ $[L_j, U_j]$; เมื่อ

$$L_j = (\hat{\beta}_{LPR})_j + (\hat{\beta}_{Lasso+OLS})_j - (\hat{\beta}^*_{rBLPR})_{j,1-\alpha/2} \text{ และ} \quad \text{...(2.20)}$$

$$U_j = (\hat{\beta}_{LPR})_j + (\hat{\beta}_{Lasso+OLS})_j - (\hat{\beta}^*_{rBLPR})_{j,\alpha/2} \quad \text{...(2.21)}$$

โดยที่ $(\hat{\beta}^*_{rBLPR})_{j,1-\alpha/2}$ และ $(\hat{\beta}^*_{rBLPR})_{j,\alpha/2}$ คือเปอร์เซ็นไทล์ที่ $(1-\frac{\alpha}{2})\times 100$ และ $(\frac{\alpha}{2})\times 100$ ของ $\hat{\beta}^*_{rBLPR}{}^{(1)},\dots,\hat{\beta}^*_{rBLPR}{}^{(B)}$ ตามลำดับ

11. ในการทดสอบสมมติฐาน $H_0: \beta_j = 0$ และ $H_a: \beta_j \neq 0$ จะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α ก็ต่อเมื่อช่วงความเชื่อมั่นที่ $(1-\alpha)\times 100\%$ สำหรับสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ไม่ครอบคลุมค่า 0

2.4.2 วิธี pBLPR (Paired Bootstrap Lasso + Partial Ridge)

วิธีบูตสแตรป pBLPR เป็นการสุ่มข้อมูลตัวอย่างแบบใส่คืนเพื่อสร้างข้อมูลตัวอย่างบูตสแตรป จากนั้นนำชุดข้อมูลตัวอย่างบูตแสตรปไปใช้คำนวณหาสัมประสิทธิ์การถดถอย $\hat{\beta}^*_{pBLPR}$ เพื่อสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอย β_j เมื่อ $j = 1,2,\dots,p$ โดยวิธี pBLPR มีขั้นตอนดังนี้

กำหนดให้

$\{(x_i, y_i), i = 1,2,\ldots,n\}$ คือ เซตของชุดข้อมูลตัวอย่าง

$\{(x_i^*, y_i^*), i = 1,2,\ldots,n\}$ คือ เซตตัวอย่างบูตแสตรป

1. สุ่มข้อมูลตัวอย่างแบบใส่คืนเพื่อสร้างข้อมูลตัวอย่างบูตแสตรป

$$\left(x_{pboot}^*, y_{pboot}^*\right) = \{(x_i^*, y_i^*), i = 1,2,\ldots,n\} \quad \ldots(2.22)$$

เมื่อ $y_{pboot}^* = (y_1^*, y_2^*, \ldots, y_n^*)^T$ และ $x_{pboot}^* = (x_1^*, x_2^*, \ldots, x_n^*)^T$

2. คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยใช้วิธีลาสโซ่จากชุดข้อมูล $\left(x_{pboot}^*, y_{pboot}^*\right)$

$$\hat{\beta}_{pBLasso}^* = \underset{\beta}{argmin}\left\{\frac{1}{2n}\left\|y^*{}_{pboot} - X^*{}_{pboot}\beta\right\|_2^2 + \lambda_1\|\beta\|_1\right\} \quad \ldots(2.23)$$

3. คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยมีการเพิ่มฟังก์ชันการลงโทษแบบ L2 – Penalty สำหรับตัวแปรอิสระที่ไม่ได้ถูกคัดเลือกและใช้ชุดข้อมูล $\left(x_{pboot}^*, y_{pboot}^*\right)$

$$\hat{\beta}_{pBLPR}^* = \underset{\beta}{argmin}\left\{\frac{1}{2n}\left\|y^*{}_{pboot} - X^*{}_{pboot}\beta\right\|_2^2 + \frac{\lambda_2}{2}\sum_{j \notin \hat{S}_{pBlasso}^*}\beta_j^2\right\} \quad \ldots(2.24)$$

4. ทำซ้ำในขั้นตอนที่ 1 – 3 ไป B รอบจนได้ $\hat{\beta}_{pBLPR}^{*(1)}, \hat{\beta}_{pBLPR}^{*(2)}, \ldots, \hat{\beta}_{pBLPR}^{*(B)}$
5. สร้างช่วงความเชื่อมั่นที่ $(1-\alpha)\%$ สำหรับสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ดังนี้

$$(\hat{\beta}_{pBLPR_{j,\alpha/2}}^*, \hat{\beta}_{pBLPR_{j,1-\alpha/2}}^*)$$

เมื่อ $\hat{\beta}_{pBLPR_{j,1-\alpha/2}}^*$ และ $\hat{\beta}_{pBLPR_{j,\alpha/2}}^*$ คือ เปอร์เซ็นไทล์ที่ $(1-\frac{\alpha}{2}) \times 100$ และ $(\frac{\alpha}{2}) \times 100$ ของ $\hat{\beta}_{pBLPR}^{*(1)}, \ldots, \hat{\beta}_{pBLPR}^{*(B)}$ ตามลำดับ

ในการทดสอบสมมติฐาน $H_0: \beta_j = 0$ และ $H_a: \beta_j \neq 0$ จะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α ก็ต่อเมื่อช่วงความเชื่อมั่นที่ $(1-\alpha) \times 100\%$ สำหรับสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ไม่ครอบคลุมค่าศูนย์

บทที่ 3
วิธีการดำเนินการศึกษา

การวิจัยนี้เป็นการศึกษาเปรียบเทียบประสิทธิภาพวิธีการบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยทั้ง 2 วิธี ได้แก่ วิธีบูตแสตรปตัวประมาณ Adaptive Lasso + Partial Ridge และวิธีบูตแสตรปตัวประมาณ Lasso + Partial Ridge นอกจากนี้ผู้วิจัยได้ทดลองใช้วิธีบูตแสตรปแบบสุ่มส่วนเหลือและวิธีบูตแสตรปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระสำหรับการบูตแสตรปตัวประมาณทั้งสอง สำหรับข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลจำลองที่มีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระหรือเรียกว่าข้อมูลที่มีมิติสูง โดยข้อมูลจำลองจะแบ่งออกเป็น 8 กรณีย่อย ซึ่งในการเปรียบเทียบประสิทธิภาพและวิเคราะห์ผลลัพธ์ จะพิจารณาจาก 4 เกณฑ์ ได้แก่ ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals) ความน่าจะเป็นครอบคลุม (Coverage Probabilities) อัตราผลบวกเทียม (False Positive Rate) และอัตราผลลบเทียม (False Negative Rate) สำหรับการจำลองข้อมูลและวิเคราะห์ข้อมูลจะดำเนินการโดยใช้โปรแกรม R เวอร์ชัน 3.6.1 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการวิจัย

การศึกษาครั้งนี้ใช้การจำลองข้อมูลขึ้นแบบตัดขวาง (Cross - Sectional Data) ทั้งหมด 8 กรณีและลักษณะของข้อมูลจัดอยู่ในประเภทข้อมูลที่มีมิติสูง (High Dimensional Data) โดยลักษณะข้อมูลที่จำลองขึ้นเพื่อการศึกษาครั้งนี้จำลองข้อมูลในลักษณะเดียวกับการศึกษาของ Liu et al. (2020)

โดยทำการศึกษาข้อมูลในลักษณะดังต่อไปนี้

1. กำหนดขนาดตัวอย่างข้อมูล (n) เท่ากับ 200 และจำนวนตัวแปรสระ (p) เท่ากับ 500
2. ทำการจำลองข้อมูลดังต่อไปนี้

กรณีที่ 1 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.5$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N\,(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์ (Toeplitz)

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

โดยกำหนดให้ $\sum_{ij} = \rho^{|i-j|}$; $\rho = 0.5$

2. กำหนดค่าสัมประสิทธิ์การถดถอยแบบ Hard Sparsity คือ $\beta_j =$

$$\begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots, 10 \\ 0; \text{ อื่นๆ} \end{cases}$$

3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐานโดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N(0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{10} \end{bmatrix} U\left[\frac{1}{3}, 1\right] \\ \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{bmatrix} \text{มีค่าเท่ากับ } 0 \Bigg]_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

กรณีที่ 2 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.9$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์ (Toeplitz)

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

โดยกำหนดให้ $\sum_{ij} = \rho^{|i-j|}$; $\rho = 0.9$

2. กำหนดค่าสัมประสิทธิ์การถดถอยแบบ Hard Sparsity คือ $\beta_j =$

$$\begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots, 10 \\ 0; \text{ อื่นๆ} \end{cases}$$

3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N\,(0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} U\,[\frac{1}{3}, 1] \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \text{มีค่าเท่ากับ } 0 \end{bmatrix}_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

กรณีที่ 3 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.5$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร(Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N\,(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์ (Toeplitz)

$$\sum = \begin{bmatrix} 1 & \sum_{1,2} & \cdots & \sum_{1,p} \\ \sum_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \sum_{i,j} \\ \sum_{p,1} & \cdots & \sum_{i,j} & 1 \end{bmatrix}$$

โดยกำหนดให้ $\sum_{ij} = \rho^{|i-j|}$; $\rho = 0.5$

2. กำหนดค่าสัมประสิทธิ์การถดถอย Weak Sparsity คือ $\beta_j =$

$$\begin{cases} N(1, 0.001); j = 1, 2, \ldots, 10 \\ \beta_j = \frac{1}{(j+3)^2};\ j = 1, 2, \ldots, 490 \end{cases}$$

3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N\,(0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} N(1,0.001) \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \beta_j = \dfrac{1}{(j+3)^2} \end{bmatrix}_{500\times1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times1}$$

กรณีที่ 4 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Toeplitz ($\rho = 0.9$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์ (Toeplitz)

$$\sum = \begin{bmatrix} 1 & \sum_{1,2} & \cdots & \sum_{1,p} \\ \sum_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \sum_{i,j} \\ \sum_{p,1} & \cdots & \sum_{i,j} & 1 \end{bmatrix}$$

โดยกำหนดให้ $\sum_{ij} = \rho^{|i-j|}$; $\rho = 0.5$

2. กำหนดค่าสัมประสิทธิ์การถดถอย Weak Sparsity คือ $\beta_j =$

$$\begin{cases} N(1,0.001); j = 1,2,\ldots,10 \\ \beta_j = \frac{1}{(j+3)^2}; \; j = 1,2,\ldots,490 \end{cases}$$

3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N(0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} N(1,0.001) \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \beta_j = \dfrac{1}{(j+3)^2} \end{bmatrix}_{500\times1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times1}$$

กรณีที่ 5 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.5$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N\,(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์ Equal Correlation

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

โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$

2. กำหนดค่าสัมประสิทธิ์การถดถอยแบบ Hard Sparsity คือ $\beta_j =$

$$\begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots, 10 \\ 0; \text{ อื่นๆ} \end{cases}$$

3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N\,(0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} U\left[\frac{1}{3}, 1\right] \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \text{มีค่าเท่ากับ } 0 \end{bmatrix}_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

กรณีที่ 6 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Hard Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.9$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N\,(0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์ Equal Correlation

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

โดยกำหนดระดับความสัมพันธ์ $\rho = 0.9$

2. กำหนดค่าสัมประสิทธิ์การถดถอยแบบ Hard Sparsity คือ $\beta_j =$
$\begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots, 10 \\ 0; \text{ อื่นๆ} \end{cases}$
3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10
$$\epsilon_i \sim N(0, \sigma^2)$$
4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)
$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} U\left[\frac{1}{3}, 1\right] \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \text{มีค่าเท่ากับ } 0 \end{bmatrix}_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

กรณีที่ 7 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.5$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$
$$x_i \sim N(0, \sum)$$
โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์ Equal Correlation

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

โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$

2. กำหนดค่าสัมประสิทธิ์การถดถอย Weak Sparsity คือ $\beta_j =$
$\begin{cases} N(1,0.001); j = 1,2,\ldots,10 \\ \beta_j = \frac{1}{(j+3)^2};\ j = 1,2,\ldots,490 \end{cases}$
3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N\ (0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} N(1,0.001) \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \beta_j = \frac{1}{(j+3)^2} \end{bmatrix}_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

กรณีที่ 8 : สัมประสิทธิ์การถดถอยเป็นลักษณะ Weak Sparsity และเมทริกซ์ความแปรปรวนร่วมแบบ Equal Correlation ($\rho = 0.9$)

1. กำหนดตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$

$$x_i \sim N\ (0, \sum)$$

โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์ Equal Correlation

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

โดยกำหนดระดับความสัมพันธ์ $\rho = 0.9$

2. กำหนดค่าสัมประสิทธิ์การถดถอย Weak Sparsity คือ $\beta_j =$
$\begin{cases} N(1,0.001); j = 1,2,\ldots,10 \\ \beta_j = \frac{1}{(j+3)^2};\ j = 1,2,\ldots,490 \end{cases}$
3. กำหนดค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐาน โดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10

$$\epsilon_i \sim N\ (0, \sigma^2)$$

4. จำลองตัวแปรตามจากตัวแบบเชิงเส้น (Linear Regression Model)

$$y_i = x_i^T \beta + \epsilon_i$$

สามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,500} \\ \vdots & \ddots & \vdots \\ x_{200,1} & \cdots & x_{200,500} \end{bmatrix}_{200\times 500} \begin{bmatrix} \left.\begin{matrix} \beta_1 \\ \vdots \\ \beta_{10} \end{matrix}\right\} N(1,0.001) \\ \left.\begin{matrix} \beta_{11} \\ \vdots \\ \beta_{500} \end{matrix}\right\} \beta_j = \frac{1}{(j+3)^2} \end{bmatrix}_{500\times 1} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{200} \end{bmatrix}_{200\times 1}$$

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ค้นคว้าเอกสาร ทฤษฎี และกรอบแนวคิดที่เกี่ยวข้อง

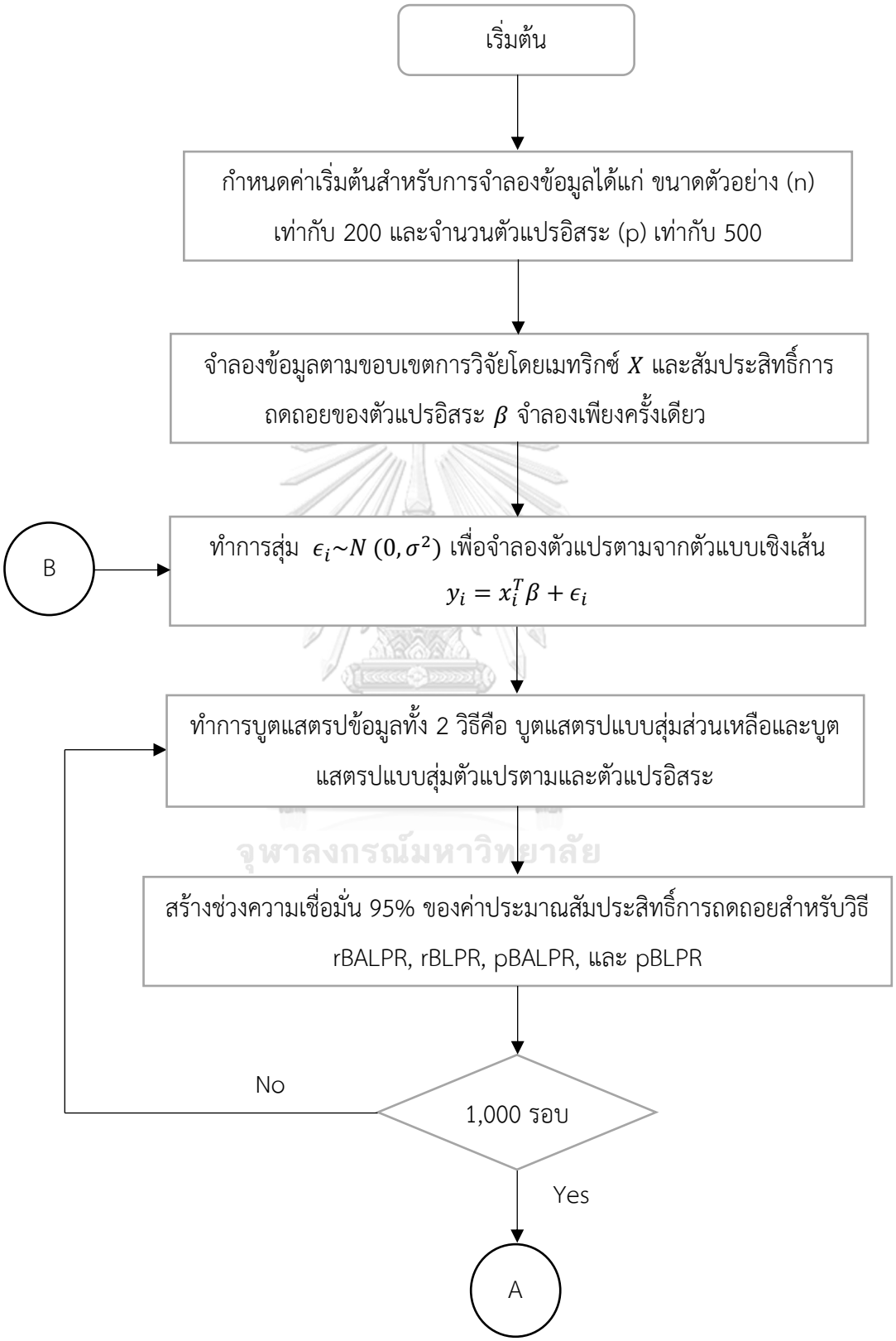
2. กำหนดค่าเริ่มต้นสำหรับการจำลองข้อมูลในแต่ละกรณีที่ทำการศึกษา

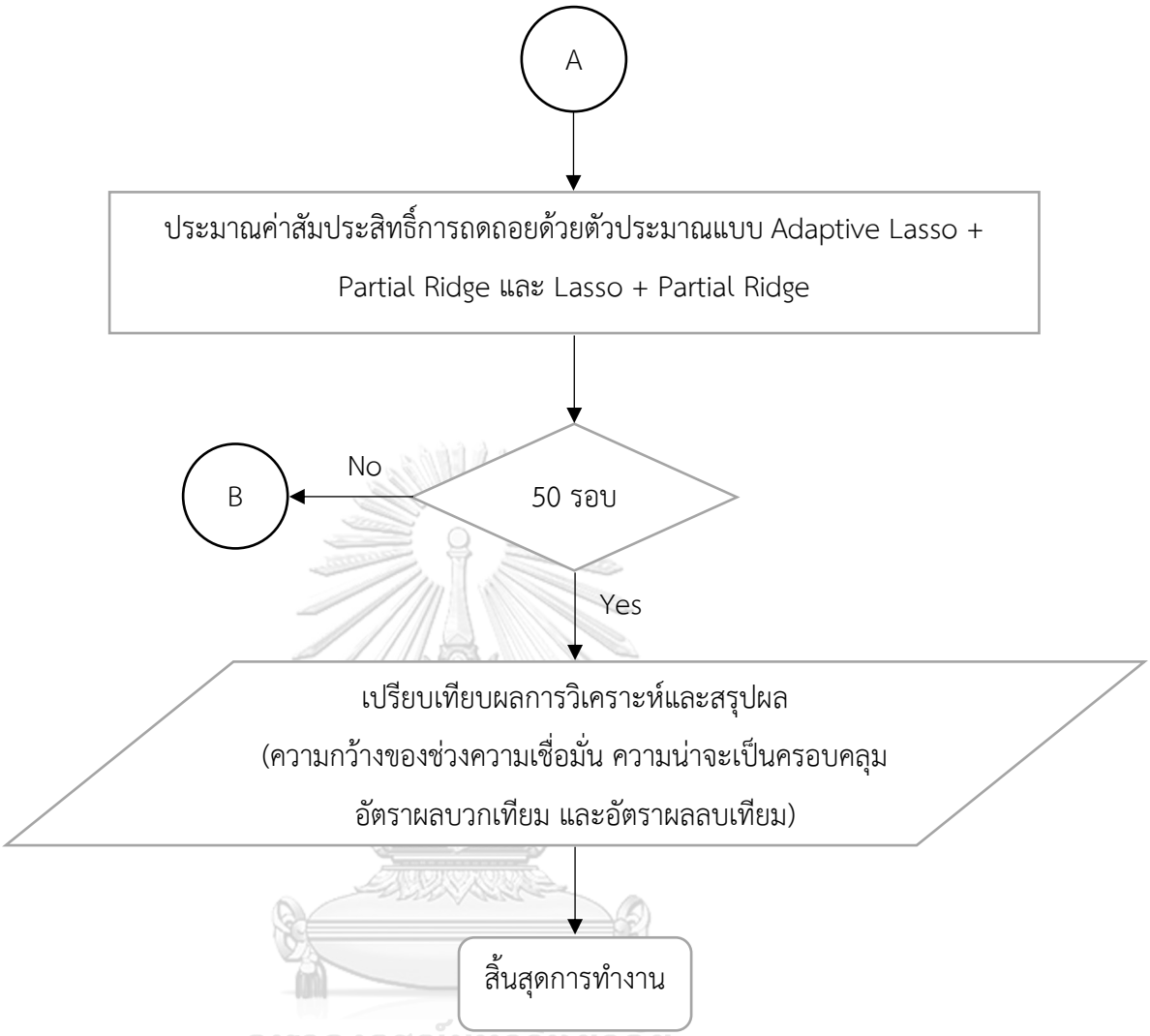
2.1 กำหนดขนาดตัวอย่าง (n) เท่ากับ 200

2.2 กำหนดจำนวนตัวแปรอิสระ (p) เท่ากับ 500

3. ทำการจำลองข้อมูลทั้งหมด 8 กรณีตามขอบเขตการวิจัย
4. ในแต่ละกรณีที่ทำการศึกษานั้นจะใช้วิธีบูตสแตรปแบบ Residual Bootstrap และ Paired Bootstrap สำหรับตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge และ Lasso + Partial Ridge เพื่อสร้างช่วงความเชื่อมั่น (Confidence Interval)
5. นำผลที่ได้จากข้อ 4 มาคำนวณหาค่าดังนี้
    5.1 ความน่าจะเป็นครอบคลุม (Coverage Probability)
    5.2 ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals)
    5.3 อัตราผลบวกเทียม (False Positive Rate)
    5.4 อัตราผลลบเทียม (False Negative Rate)
6. เปรียบเทียบผลการวิเคราะห์ในข้อที่ 5
7. สรุปผลการศึกษา

3.3 ขั้นตอนการทำงานของโปรแกรม R





บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์และเปรียบเทียบประสิทธิภาพกับวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่และพาร์เชียลริดจ์ โดยการศึกษาครั้งนี้ใช้ข้อมูลจำลองจากการแจกแจงแบบปกติหลายตัวแปรซึ่งใช้เมทริกซ์ความแปรปรวนร่วมของค่าคลาดเคลื่อนที่แตกต่างกันและศึกษาสัมประสิทธิ์การถดถอยใน 2 ลักษณะได้แก่ บางเบาอย่างอ่อน (Weak Sparsity) และบางเบาอย่างรุนแรง (Hard Sparsity) รวมทั้งหมด 8 กรณี โดยเกณฑ์ที่ใช้พิจารณาประสิทธิภาพของผลลัพธ์ที่ได้จากการบูตแสตรปแบบต่างๆ ได้แก่ ความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ย (Width of Confidence Intervals) ความน่าจะเป็นครอบคลุมโดยเฉลี่ย (Coverage Probabilities) อัตราผลบวกเทียม (FPR) และอัตราผลลบเทียม (FNR) โดยถ้าความกว้างของช่วงความเชื่อมั่น อัตราผลบวกเทียม และอัตราผลลบเทียม มีค่ายิ่งน้อยจะถือว่ายิ่งมีประสิทธิภาพสูง และถ้าความน่าจะเป็นครอบคลุมยิ่งมีค่ามากจะถือว่ายิ่งมีประสิทธิภาพสูงเช่นกัน

อักษรย่อและสัญลักษณ์ต่างๆที่ปรากฏในการนำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆแทนความหมายดังนี้

pBALPR แทน การหาค่าสัมประสิทธิ์การถดถอยด้วยวิธี Paired Bootstrap Adaptive Lasso + Partial Ridge

pBLPR แทน การหาค่าสัมประสิทธิ์การถดถอยด้วยวิธี Paired Bootstrap Lasso + Partial Ridge

rBALPR แทน การหาค่าสัมประสิทธิ์การถดถอยด้วยวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge

rBLPR แทน การหาค่าสัมประสิทธิ์การถดถอยด้วยวิธี Residual Bootstrap Lasso + Partial Ridge

LPR แทน ตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge

ALPR แทน ตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge

WCI แทน ความกว้างของช่วงความเชื่อมั่นหรือ Width of Confidence Intervals

CP แทน ความน่าจะเป็นครอบคลุมหรือ Coverage Probabilities

FPR แทน อัตราผลบวกเทียมหรือ False Positive Rate

FNR แทน อัตราผลลบเทียมหรือ False Negative Rate

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 4 ส่วน ได้แก่

ส่วนที่ 1 ผลการเปรียบเทียบค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยที่ได้จาก วิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ส่วนที่ 2 ผลการเปรียบเทียบค่าเฉลี่ยความน่าจะเป็นที่ช่วงความเชื่อมั่นที่ได้ครอบคลุมค่าของสัมประสิทธิ์การถดถอยซึ่งได้จาก วิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ส่วนที่ 3 ผลการเปรียบเทียบค่าเฉลี่ยอัตราผลบวกเทียมซึ่งได้จากวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ส่วนที่ 4 ผลการเปรียบเทียบค่าเฉลี่ยอัตราผลลบเทียมซึ่งได้จากวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

4.1 ผลการเปรียบเทียบค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยที่ได้จากวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิภาพในการบูตแสตรปเพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยจากวิธี rBALPR, วิธี rBLPR, วิธี pBALPR และวิธี pBLPR โดยใช้เกณฑ์วัดประสิทธิภาพเป็นความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ย (Width of Confidence Intervals)

ตารางที่ 1 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยซึ่งได้จากวิธีการบูตแสตรปแบบต่างๆ

กรณีศึกษา	วิธีการบูตแสตรป			
	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1: Hard Sparse + Toeplitz ρ = 0.5	**0.004** (0.001)	0.020 (0.002)	0.014 (0.001)	0.037 (0.003)
กรณีที่ 2: Hard Sparse + Toeplitz ρ = 0.9	**0.009** (0.001)	0.012 (0.002)	0.017 (0.003)	0.046 (0.005)
กรณีที่ 3: Weak Sparse + Toeplitz ρ = 0.5	**0.006** (0.002)	0.031 (0.003)	0.021 (0.002)	0.053 (0.004)
กรณีที่ 4: Weak Sparse + Toeplitz ρ = 0.9	**0.014** (0.001)	0.018 (0.003)	0.024 (0.004)	0.068 (0.008)
กรณีที่ 5: Hard Sparse + Equal Correlation ρ = 0.5	**0.018** (0.003)	0.036 (0.003)	0.024 (0.002)	0.057 (0.006)
กรณีที่ 6: Hard Sparse + Equal Correlation ρ = 0.9	0.091 (0.013)	0.132 (0.020)	**0.074** (0.007)	0.098 (0.011)

กรณีที่ 7: Weak Sparse + Equal Correlation ρ = 0.5	**0.024** (0.006)	0.053 (0.004)	0.038 (0.005)	0.084 (0.011)
กรณีที่ 8: Weak Sparse + Equal Correlation ρ = 0.9	0.145 (0.014)	0.211 (0.023)	**0.117** (0.009)	0.159 (0.017)

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 1 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยที่ได้จากการบูตแสตรป จำนวน 50 replications ซึ่งพบว่าเมื่อจำลองข้อมูลในรูปแบบกรณีที่ 1, 2, 3, 4, 5, และ 7 วิธี rBALPR มีประสิทธิภาพสูงที่สุดในการให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุด นอกจากนี้ส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงความเชื่อมั่นที่ได้จากวิธี rBALPR มีค่าต่ำที่สุดในกรณีที่ 1 – 4 ในขณะที่เมื่อจำลองข้อมูลในรูปแบบกรณีที่ 6 และ 8 พบว่าวิธี pBALPR ให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุดและมีค่าส่วนเบี่ยงเบนมาตรฐานต่ำที่สุดเช่นกัน นอกจากนี้สังเกตได้ว่าการบูตแสตรปตัวประมาณ LPR ไม่ว่าจะด้วยวิธีบูตแสตรปแบบสุ่มส่วนเหลือหรือบูตแสตรปแบบสุ่มตัวแปรตามและตัวแปรอิสระนั้นให้ความกว้างของช่วงความเชื่อมั่นค่อนข้างสูงในเกือบทุกกรณีของข้อมูลจำลอง

4.2 ผลการเปรียบเทียบค่าเฉลี่ยความน่าจะเป็นที่ช่วงความเชื่อมั่นที่ได้ครอบคลุมค่าของสัมประสิทธิ์การถดถอยซึ่งได้จาก วิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิภาพในการบูตแสตรปเพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยจากวิธี rBALPR, วิธี rBLPR, วิธี pBALPR และวิธี pBLPR โดยใช้เกณฑ์วัดประสิทธิภาพเป็นความน่าจะเป็นที่ช่วงความเชื่อมั่นที่สร้างขึ้นครอบคลุมค่าจริงของสัมประสิทธิ์การถดถอย (Coverage Probabilities)

ตารางที่ 2 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าจริงของสัมประสิทธิ์การถดถอยซึ่งได้จากวิธีการบูตแสตรปแบบต่างๆ

กรณีศึกษา	วิธีการบูตแสตรป			
	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1: Hard Sparse + Toeplitz ρ = 0.5	0.933 (0.03)	0.940 (0.03)	0.942 (0.03)	**0.999** (0.01)
กรณีที่ 2: Hard Sparse + Toeplitz ρ = 0.9	0.939 (0.04)	0.939 (0.03)	0.954 (0.03)	**0.999** (0.01)
กรณีที่ 3: Weak Sparse + Toeplitz ρ = 0.5	0.808 (0.22)	0.830 (0.28)	**0.940** (0.08)	0.881 (0.21)
กรณีที่ 4: Weak Sparse + Toeplitz ρ = 0.9	0.813 (0.31)	0.772 (0.35)	**0.912** (0.11)	0.864 (0.26)
กรณีที่ 5: Hard Sparse + Equal Correlation ρ = 0.5	0.919 (0.07)	0.934 (0.05)	0.789 (0.11)	**0.973** (0.05)
กรณีที่ 6: Hard Sparse + Equal Correlation ρ = 0.9	0.933 (0.06)	**0.940** (0.05)	0.647 (0.19)	0.558 (0.21)

กรณีที่ 7: Weak Sparse + Equal Correlation ρ = 0.5	0.737 (0.35)	**0.745** (0.32)	0.591 (0.26)	0.710 (0.25)
กรณีที่ 8: Weak Sparse + Equal Correlation ρ = 0.9	0.623 (0.37)	**0.806** (0.18)	0.398 (0.22)	0.434 (0.21)

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 2 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าจริงของสัมประสิทธิ์การถดถอยที่ได้จากการบูตแสตรป จำนวน 50 replications ซึ่งพบว่าไม่มีวิธีการบูตแสตรปตัวประมาณแบบใดที่มีความโดดเด่นในด้านการให้ความน่าจะเป็นสูงสุดสำหรับทุกกรณีของข้อมูลจำลอง โดยวิธีการบูตแสตรปแต่ละวิธีจะเหมาะสมกับข้อมูลจำลองลักษณะที่แตกต่างกันออกไป กล่าวคือเมื่อจำลองข้อมูลในรูปแบบกรณีที่ 1, 2 และ 5 วิธี pBLPR ให้ความน่าจะเป็นครอบคลุมสูงที่สุด ในขณะที่ถ้าจำลองข้อมูลตามกรณีที่ 3 และ 4 นั้นวิธี pBALPR ให้ความน่าจะเป็นครอบคลุมสูงที่สุด และเมื่อจำลองข้อมูลในรูปแบบ 6, 7 และ 8 วิธี rBLPR ให้ความน่าจะเป็นครอบคลุมสูงที่สุด นอกจากนี้ สังเกตได้ว่าเมื่อข้อมูลจำลองขึ้นในรูปแบบที่ 5 – 8 นั้นวิธี pBALPR ให้ความน่าจะเป็นครอบคลุมที่ค่อนข้างต่ำกว่าวิธีอื่นๆ

4.3 ผลการเปรียบเทียบค่าเฉลี่ยอัตราผลบวกเทียมซึ่งได้จากวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิภาพในการบูตแสตรปเพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยจากวิธี rBALPR, วิธี rBLPR, วิธี pBALPR และวิธี pBLPR โดยใช้เกณฑ์วัดประสิทธิภาพเป็นอัตราผลบวกเทียม (False Positive Rate)

ตารางที่ 3 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราผลบวกเทียมซึ่งได้จากวิธีการบูตแสตรปแบบต่างๆ

กรณีศึกษา	วิธีการบูตแสตรป			
	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1: Hard Sparse + Toeplitz ρ = 0.5	0.388 (0.375)	0.611 (0.195)	0.735 (0.037)	**0.004** (0.018)
กรณีที่ 2: Hard Sparse + Toeplitz ρ = 0.9	0.468 (0.403)	0.451 (0.260)	0.666 (0.114)	**0.017** (0.044)
กรณีที่ 3: Weak Sparse + Toeplitz ρ = 0.5	-	-	-	-
กรณีที่ 4: Weak Sparse + Toeplitz ρ = 0.9	-	-	-	-
กรณีที่ 5: Hard Sparse + Equal Correlation ρ = 0.5	0.757 (0.149)	0.753 (0.046)	0.906 (0.022)	**0.430** (0.207)
กรณีที่ 6: Hard Sparse + Equal Correlation ρ = 0.9	**0.724** (0.165)	0.733 (0.099)	0.979 (0.008)	0.988 (0.006)
กรณีที่ 7: Weak Sparse + Equal Correlation ρ = 0.5	-	-	-	-

กรณีที่ 8: Weak Sparse + Equal Correlation ρ = 0.9	-	-	-	-

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 3 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราผลบวกเทียมที่ได้จากการบูตแสตรป จำนวน 50 replications โดยที่การจำลองข้อมูลในรูปแบบกรณีที่ 3, 4, 7, และ 8 ซึ่งเป็นกรณีที่สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะบางเบาอย่างอ่อน (Weak Sparsity) หรือเป็นกรณีที่มีสัมประสิทธิ์การถดถอยทุกตัวมีค่าไม่เท่ากับศูนย์ ส่งผลให้การใช้ FPR เป็นเกณฑ์การวัดประสิทธิภาพไม่สามารถกระทำได้ เนื่องจากทุกกรณีจะให้ค่า FPR เท่ากับ 0 เพราะไม่ปรากฏค่า False Positive (FP) อย่างไรก็ตาม เมื่อพิจารณาเฉพาะข้อมูลจำลองในรูปแบบกรณีที่ 1, 2, 5 และ 6 ซึ่งเป็นกรณีที่สัมประสิทธิ์การถดถอยมีลักษณะบางเบาอย่างรุนแรง (Hard Sparsity) หรือมีสัมประสิทธิ์การถดถอยมีค่าไม่เท่ากับศูนย์เพียง 10 ตัว พบว่าวิธี pBLPR มีประสิทธิภาพสูงสุดในด้านการให้อัตราผลบวกเทียมต่ำสุดสำหรับกรณีข้อมูลจำลองรูปแบบที่ 1, 2 และ 5 และวิธี rBALPR ให้อัตราผลบวกเทียมต่ำสุดสำหรับกรณีข้อมูลจำลองรูปแบบที่ 6

4.4 ผลการเปรียบเทียบค่าเฉลี่ยอัตราผลลบเทียมซึ่งได้จากวิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Parital Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิภาพในการบูตแสตรปเพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยจากวิธี rBALPR, วิธี rBLPR, วิธี pBALPR และวิธี pBLPR โดยใช้เกณฑ์วัดประสิทธิภาพเป็นอัตราผลลบเทียม (False Negative Rate)

ตารางที่ 4 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราผลลบเทียมซึ่งได้จากวิธีการบูตแสตรปแบบต่างๆ

กรณีศึกษา	วิธีการบูตแสตรป			
	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1: Hard Sparse + Toeplitz ρ = 0.5	**0.0000** (0.0000)	**0.0000** (0.0000)	**0.0000** (0.0000)	**0.0000** (0.0000)
กรณีที่ 2: Hard Sparse + Toeplitz ρ = 0.9	**0.0003** (0.0008)	0.0005 (0.0010)	0.0003 (0.0008)	0.0007 (0.0011)
กรณีที่ 3: Weak Sparse + Toeplitz ρ = 0.5	-	-	-	-
กรณีที่ 4: Weak Sparse + Toeplitz ρ = 0.9	-	-	-	-
กรณีที่ 5: Hard Sparse + Equal Correlation ρ = 0.5	**0.0000** (0.0000)	**0.0000** (0.0000)	0.00005 (0.0003)	0.00012 (0.0005)
กรณีที่ 6: Hard Sparse + Equal Correlation ρ = 0.9	0.0038 (0.0028)	**0.0033** (0.0029)	0.0195 (0.0043)	0.0262 (0.0045)
กรณีที่ 7: Weak Sparse + Equal Correlation ρ = 0.5	-	-	-	-

กรณีที่ 8: Weak Sparse + Equal Correlation ρ = 0.9	-	-	-	-

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บคือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 4 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของอัตราผลลบเทียมที่ได้จากการบูตแสตรป จำนวน 50 replications โดยที่การจำลองข้อมูลในรูปแบบกรณีที่ 3, 4, 7, และ 8 ซึ่งเป็นกรณีที่สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะบางเบาอย่างอ่อน (Weak Sparsity) หรือเป็นกรณีที่มีสัมประสิทธิ์การถดถอยทุกตัวมีค่าไม่เท่ากับศูนย์ ส่งผลให้การใช้ FNR เป็นเกณฑ์การวัดประสิทธิภาพไม่สามารถกระทำได้ เนื่องจากทุกกรณีจะให้ค่า FNR เท่ากับ 1 เพราะไม่ปรากฏค่า True Negative (TN) อย่างไรก็ตาม เมื่อพิจารณาเฉพาะข้อมูลจำลองในรูปแบบกรณีที่ 1, 2, 5 และ 6 ซึ่งเป็นกรณีที่สัมประสิทธิ์การถดถอยมีลักษณะบางเบาอย่างรุนแรง (Hard Sparsity) หรือมีสัมประสิทธิ์การถดถอยมีค่าไม่เท่ากับศูนย์เพียง 10 ตัว พบว่าเมื่อจำลองข้อมูลในรูปแบบกรณีที่ 1, 2 และ 5 ทุกวิธีการบูตแสตรปมีประสิทธิภาพใกล้เคียงกันคือ ให้ค่า FNR สำหรับกรณีดังกล่าวใกล้เคียงกัน ในขณะที่การจำลองข้อมูลในรูปแบบกรณีที่ 6 วิธีการบูตแสตรป rBLPR มีประสิทธิภาพสูงสุด เนื่องจากให้ค่า FNR ต่ำที่สุด

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยชิ้นนี้นำเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive Lasso + Partial Ridge และเปรียบเทียบกับวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge โดยทำการทดลองบูตแสตรป 2 วิธีคือ วิธีสุ่มส่วนเหลือ (Residual Bootstrap) และวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ (Paired Bootstrap) ดังนั้นจึงมีการศึกษาเปรียบเทียบประสิทธิภาพของวิธีการบูตแสตรปทั้งหมด 4 วิธี ได้แก่ วิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR), วิธี Residual Bootstrap Lasso + Partial Ridge (rBLPR), วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR) และวิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR) โดยการจำลองข้อมูลที่มีขอบเขตแตกต่างกันทั้งหมด 8 กรณีและมีเกณฑ์การพิจารณาประสิทธิภาพของผลลัพธ์ที่ได้จากการบูตแสตรป ได้แก่ ความกว้างของช่วงความเชื่อมั่น (Width of Confidence Intervals) ความน่าจะเป็นครอบคลุม (Coverage Probabilities) อัตราผลบวกเทียม (False Positive Rate) และ อัตราผลลบเทียม (False Negative Rate) โดยมีการสรุปผลการวิจัยดังนี้

5.1 สรุปผลการวิจัย

การเปรียบเทียบประสิทธิภาพของวิธีบูตแสตรปโดยใช้เกณฑ์ความกว้างของช่วงความเชื่อมั่นและความน่าจะเป็นครอบคลุมปรากฏดังตารางต่อไปนี้

ตารางที่ 5 แสดงวิธีบูตแสตรปที่เหมาะสมที่สุดเมื่อพิจารณาความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ยและความน่าจะเป็นครอบคลุมโดยเฉลี่ย โดยตัวเลขในวงเล็บแสดงถึงส่วนเบี่ยงเบนมาตรฐาน

กรณีศึกษา	เกณฑ์การตัดสินใจ							
	WCI				CP			
	rBALPR	rBLPR	pBALPR	pBLPR	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1	**0.004** (0.001)	0.020 (0.002)	0.014 (0.001)	0.037 (0.003)	0.933 (0.03)	0.940 (0.03)	0.942 (0.03)	**0.999** (0.01)
กรณีที่ 2	**0.009** (0.001)	0.012 (0.002)	0.017 (0.003)	0.046 (0.005)	0.939 (0.04)	0.939 (0.03)	0.954 (0.03)	**0.999** (0.01)
กรณีที่ 3	**0.006** (0.002)	0.031 (0.003)	0.021 (0.002)	0.053 (0.004)	0.808 (0.22)	0.830 (0.28)	**0.940** (0.08)	0.881 (0.21)

กรณีที่ 4	**0.014** (0.001)	0.018 (0.003)	0.024 (0.004)	0.068 (0.008)	0.813 (0.31)	0.772 (0.35)	**0.912** (0.11)	0.864 (0.26)
กรณีที่ 5	**0.018** (0.003)	0.036 (0.003)	0.024 (0.002)	0.057 (0.006)	0.919 (0.07)	0.934 (0.05)	0.789 (0.11)	**0.973** (0.05)
กรณีที่ 6	0.091 (0.013)	0.132 (0.020)	**0.074** (0.007)	0.098 (0.011)	0.933 (0.06)	**0.940** (0.05)	0.647 (0.19)	0.558 (0.21)
กรณีที่ 7	**0.024** (0.006)	0.053 (0.004)	0.038 (0.005)	0.084 (0.011)	0.737 (0.35)	**0.745** (0.32)	0.591 (0.26)	0.710 (0.25)
กรณีที่ 8	0.145 (0.014)	0.211 (0.023)	**0.117** (0.009)	0.159 (0.017)	0.623 (0.37)	**0.806** (0.18)	0.398 (0.22)	0.434 (0.21)

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 5 แสดงช่วงความกว้างของความเชื่อมั่นและความน่าจะเป็นครอบคลุมที่ได้จากวิธีการบูตแสตรปแบบต่างๆ โดยตัวหนาในตารางแสดงถึงวิธีการบูตแสตรปที่ดีที่สุด ผลการวิเคราะห์ตารางที่ 5 ได้ดังนี้

1. การวิจัยพบว่าเมื่อใช้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยเป็นเกณฑ์การวัดประสิทธิภาพ วิธีบูตแสตรปแบบ rBALPR มีประสิทธิภาพสูงสุด โดยวิธีดังกล่าวให้ค่าความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ยน้อยที่สุดถึง 6 กรณีจากทั้งหมด 8 กรณี อีกทั้งวิธี rBALPR ให้ค่าความกว้างของช่วงความเชื่อมั่นน้อยกว่าวิธีอื่นอย่างมีนัยสำคัญเมื่อข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าคลาดเคลื่อนไม่คงที่โดยเป็นเมทริกซ์ที่มีค่าสหสัมพันธ์แบบโทพลิทซ์ (Toeplitz) ซึ่งได้แก่ กรณีที่ 1 – 4 สำหรับวิธีที่มีประสิทธิภาพต่ำที่สุดในด้านความกว้างของช่วงความเชื่อมั่นคือวิธี pBLPR เนื่องจากให้ความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ยมากที่สุดถึง 6 กรณีจากทั้งหมด 8 กรณี

2. เมื่อใช้ความน่าจะเป็นครอบคลุมเป็นเกณฑ์การวัดประสิทธิภาพ พบว่าไม่ปรากฏวิธีการบูตแสตรปวิธีใดวิธีหนึ่งที่มีประสิทธิภาพสูงสุดสำหรับทุกกรณีของข้อมูลจำลอง โดยแต่ละวิธีบูตแสตรปจะทำงานได้ดีในลักษณะข้อมูลจำลองที่แตกต่างกันออกไป ดังนั้นจึงไม่สามารถสรุปได้ว่าวิธีการบูตแสตรปใดมีประสิทธิภาพสูงสุดในแง่การให้ความน่าจะเป็นครอบคลุมสูงสุด อย่างไรก็ตาม วิธี rBLPR เป็นเพียงวิธีบูตแสตรปรูปแบบเดียวเท่านั้นที่สามารถให้ความน่าจะเป็นครอบคลุมในระดับที่ไม่น้อยกว่า 75% สำหรับทุกรูปแบบของข้อมูลจำลอง ดังนั้นอาจกล่าวได้ว่าวิธี rBLPR เป็นวิธีที่สามารถทำงานได้กับข้อมูลหลากหลายลักษณะ

การเปรียบเทียบประสิทธิภาพของวิธีบูตแสตรปโดยใช้เกณฑ์อัตราผลบวกเทียมและอัตราผลลบเทียมได้ดังตารางต่อไปนี้

ตารางที่ 6 แสดงวิธีบูตแสตรปที่เหมาะสมที่สุดเมื่อพิจารณาอัตราผลบวกเทียมโดยเฉลี่ยและอัตราผลลบเทียมโดยเฉลี่ย โดยตัวเลขในวงเล็บแสดงถึงส่วนเบี่ยงเบนมาตรฐาน

กรณีศึกษา	เกณฑ์การตัดสินใจ							
	FPR				FNR			
	rBALPR	rBLPR	pBALPR	pBLPR	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1	0.388 (0.375)	0.611 (0.195)	0.735 (0.037)	**0.004** (0.018)	**0.0000** (0.0000)	**0.0000** (0.0000)	**0.0000** (0.0000)	**0.0000** (0.0000)
กรณีที่ 2	0.468 (0.403)	0.451 (0.260)	0.666 (0.114)	**0.017** (0.044)	**0.0003** (0.0008)	0.0005 (0.0010)	**0.0003** (0.0008)	0.0007 (0.0011)
กรณีที่ 3	-	-	-	-	-	-	-	-
กรณีที่ 4	-	-	-	-	-	-	-	-
กรณีที่ 5	0.757 (0.149)	0.753 (0.046)	0.906 (0.022)	**0.430** (0.207)	**0.0000** (0.0000)	**0.0000** (0.0000)	0.00005 (0.0003)	0.00012 (0.0005)
กรณีที่ 6	**0.724** (0.165)	0.733 (0.099)	0.979 (0.008)	0.988 (0.006)	0.0038 (0.0028)	**0.0033** (0.0029)	0.0195 (0.0043)	0.0262 (0.0045)
กรณีที่ 7	-	-	-	-	-	-	-	-
กรณีที่ 8	-	-	-	-	-	-	-	-

หมายเหตุ ตัวหนา คือวิธีที่มีประสิทธิภาพสูงที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน

จากตารางที่ 6 แสดงอัตราผลบวกเทียมและอัตราผลลบเทียมที่ได้จากวิธีการบูตแสตรปแบบต่างๆ โดยตัวหนาในตารางแสดงถึงวิธีการบูตแสตรปที่ดีที่สุด สำหรับข้อมูลจำลองกรณีที่ 3, 4, 7 และ 8 ค่า FPR และ FNR จะเท่ากับ 0 และ 1 ตามลำดับ สำหรับทุกวิธีการบูตแสตรป เนื่องจากเป็นกรณี

ที่สัมประสิทธิ์การถดถอยของตัวแปรอิสระทุกตัวมีค่าไม่เท่ากับศูนย์ (Weak Sparsity) ดังนั้นผู้วิจัยจึงไม่นำมาวิเคราะห์ผลลัพธ์ในครั้งนี้ และจะวิเคราะห์ผลลัพธ์เฉพาะกรณีที่สัมประสิทธิ์การถดถอยของตัวแปรอิสระมีลักษณะบางเบาอย่างรุนแรง (Hard Sparsity) ได้แก่กรณีที่ 1, 2, 5 และ 6 โดยผลการวิเคราะห์ตารางที่ 6 ได้ดังนี้

1. เมื่อพิจารณาโดยใช้ FPR เป็นเกณฑ์วัดประสิทธิภาพพบว่า วิธี pBLPR เป็นวิธีที่มีประสิทธิภาพสูงสุดเนื่องจากให้ค่า FPR ต่ำที่สุดถึง 3 กรณีจากทั้งหมด 4 กรณีและต่ำกว่าวิธีอื่นอย่างมีนัยสำคัญ อย่างไรก็ตาม หากพิจารณา FPR ร่วมกับ WCI ในตารางที่ 5 จะพบว่าสาเหตุที่วิธี pBLPR ให้ FPR ต่ำกว่าวิธีการบูตแสตรปแบบอื่นๆนั้นเกิดจากวิธี pBLPR ให้ความกว้างของช่วงความเชื่อมั่นกว้างที่สุดและช่วงความเชื่อมั่นดังกล่าวครอบคลุมค่าศูนย์ซึ่งสอดคล้องกับลักษณะของข้อมูลจำลองที่สัมประสิทธิ์การถดถอยของตัวแปรอิสระมากเลขศูนย์ ส่งผลให้ค่า False Positive (FP) ต่ำ

2. เมื่อสัมประสิทธิ์การถดถอยส่วนใหญ่เท่ากับศูนย์และมีสัมประสิทธิ์การถดถอยบางตัวเท่านั้นที่มีค่ามากและไม่ใกล้เคียง 0 การใช้ FNR เป็นเกณฑ์การวัดประสิทธิภาพของวิธีบูตแสตรปแบบต่างๆอาจจะได้ผลลัพธ์ไม่ชัดเจนนัก เนื่องจากทุกวิธีบูตแสตรปให้ค่า FNR ใกล้เคียงกับ 0 เกือบทุกกรณี ซึ่งหมายความว่าทุกวิธีบูตแสตรปมีประสิทธิภาพในการระบุได้ว่าตัวแปรอิสระใดมีความสัมพันธ์กับตัวแปรตามหากสัมประสิทธิ์การถดถอยของตัวแปรอิสระนั้นมีค่ามากเพียงพอ

5.2 สรุปและอภิปรายผล

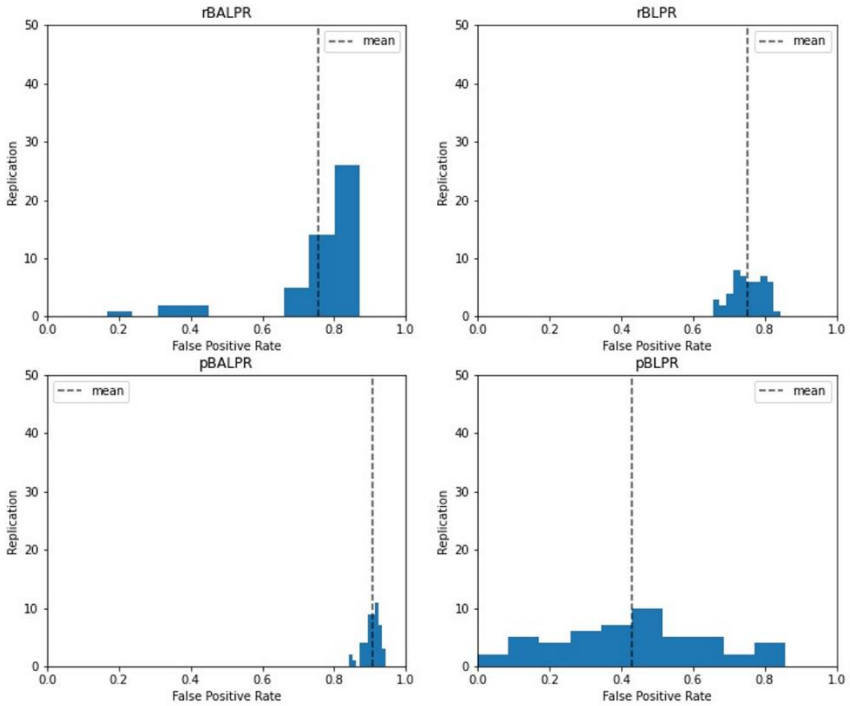
การอนุมานเชิงสถิติในกรณีที่ข้อมูลมีมิติสูงเพื่อทำความเข้าใจความสัมพันธ์ระหว่างตัวแปรยังคงเป็นประเด็นที่ท้าทายและซับซ้อน ทั้งนี้วิธีที่นิยมใช้เพื่อทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยคือวิธีบูตแสตรป อย่างไรก็ตาม การบูตแสตรปนั้นก็มีหลากหลายรูปแบบ ในงานวิจัยชิ้นนี้ได้นำเสนอวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบสองขั้นตอนคือ Adaptive Lasso + Partial Ridge โดยได้รับแนวคิดมาจากวิธีบูตแสตรปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge ซึ่งนำเสนอโดย Liu et.al (2020) การวิจัยครั้งนี้ ผู้วิจัยตั้งสมมติฐานว่าการปรับใช้ตัวประมาณ Adaptive Lasso แทนตัวประมาณ Lasso จะทำให้การบูตแสตรปเพื่อทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยมีประสิทธิภาพที่สูงขึ้น

เมื่อพิจารณาในภาพรวม พบว่าการปรับใช้ตัวประมาณ Adaptive Lasso แทนตัวประมาณ Lasso ส่งผลให้การบูตแสตรปตัวประมาณแบบสองขั้นตอนมีประสิทธิภาพสูงขึ้นในด้านการให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยที่สั้นลงเท่านั้น ในขณะที่การใช้ตัวประมาณ

Lasso นั้นมีประสิทธิภาพในด้านการให้ความน่าจะเป็นครอบคลุมที่สูงกว่า ดังพิจารณาได้จากตารางที่ 5 พบว่าการบูตสแตรปตัวประมาณ Lasso + Partial Ridge ให้ความน่าจะเป็นครอบคลุมสูงที่สุดถึง 6 กรณีจากทั้งหมด 8 กรณี ซึ่งสอดคล้องกับการศึกษาของ Liu et al. (2020) ที่พบว่าการใช้วิธีบูตแสตรปตัวประมาณ Lasso + Partial Ridge สามารถให้ความน่าจะเป็นครอบคลุมที่สูงถึงประมาณ 80 - 90% เมื่อสัมประสิทธิ์การถดถอยมีลักษณะบางเบาอย่างรุนแรง (Hard Sparsity) และให้ความน่าจะเป็นครอบคลุมที่ระดับใกล้เคียงกันเมื่อสัมประสิทธิ์การถดถอยมีลักษณะบางเบาอย่างอ่อน (Weak Sparsity)

สำหรับการเปรียบเทียบประสิทธิภาพในแง่ของอัตราผลบวกเทียมนั้นค่อนข้างมีความซับซ้อน เนื่องจากการใช้ค่าเฉลี่ยอัตราผลบวกเทียมเพียงอย่างเดียวในการเปรียบเทียบประสิทธิภาพอาจจะไม่เหมาะสม เพราะผลลัพธ์ที่ได้จากวิธีบูตแสตรปรูปแบบต่างๆมีค่าส่วนเบี่ยงเบนมาตรฐานที่ค่อนข้างสูง ดังแสดงได้ในภาพที่ 1 ซึ่งเป็นฮิสโตแกรมของอัตราผลบวกเทียมที่ได้จากวิธีบูตแสตรปทุกรูปแบบ จำนวน 50 รอบ (replications) โดยใช้ผลลัพธ์จากการทดลองข้อมูลจำลองกรณีที่ 5 เป็นตัวอย่างในการอภิปรายผล ซึ่งสังเกตได้ว่าวิธี pBLPR เป็นวิธีที่ให้ค่าเฉลี่ยอัตราผลบวกเทียมต่ำที่สุดแต่วิธีดังกล่าวให้ส่วนเบี่ยงเบนมาตรฐานของอัตราผลบวกเทียมสูงที่สุดเช่นกัน ซึ่งแปลผลได้ว่าในการบูตแสตรปแต่ละรอบนั้น วิธี pBLPR อาจจะให้อัตราผลบวกเทียมที่ต่ำมากหรือสูงมากก็ได้ ส่งผลให้ผู้ใช้งานที่จะนำวิธีการบูตแสตรปที่ได้มีการศึกษาในครั้งนี้ไปปรับใช้จริงอาจจะได้ผลลัพธ์ที่แตกต่างออกไปจากการวิจัยในครั้งนี้ กล่าวคือวิธี pBLPR อาจจะไม่ใช่วิธีที่ให้อัตราผลบวกเทียมต่ำที่สุดก็เป็นได้ ดังนั้นผู้วิจัยจึงเห็นว่าการเปลี่ยนตัวประมาณจาก Lasso เป็น Adaptive นั้นยังไม่สามารถสรุปผลได้อย่างชัดเจนว่าจะส่งผลอย่างไรในแง่ของอัตราผลบวกเทียม

ภาพที่ 1 แสดงฮิสโตแกรมของอัตราผลบวกเทียมที่ได้จากการบูตแสตรปจำนวน 50 รอบ (Replications) จากทั้ง 4 วิธี ได้แก่ rBALPR, rBLPR, pBALPR และ pBALPR



อนึ่ง นอกเหนือไปจากเกณฑ์ความกว้างของช่วงความเชื่อมั่น ความน่าจะเป็นครอบคลุมอัตราผลบวกเทียม และอัตราผลลบเทียมแล้ว ผู้วิจัยเห็นว่าการเปรียบเทียบการใช้ทรัพยากรการคำนวณ (Computational Resources Utilization) สำหรับวิธีบูตแสตรปแต่ละวิธีก็เป็นสิ่งที่สำคัญและอาจเป็นประโยชน์แก่ผู้ที่สนใจจะนำวิธีบูตแสตรปแบบต่างๆไปใช้งาน เนื่องจากการบูตแสตรปเพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยจำเป็นต้องอาศัยการคำนวณซ้ำจำนวนหลายครั้ง ดังนั้นจึงใช้ทรัพยากรการคำนวณสูง

ตารางที่ 7 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของเวลาที่ใช้สำหรับวิธีบูตแสตรปแบบ rBALPR, rBLPR, pBALPR, และ pBALPR โดยใช้ Central Processing Unit (CPU) รุ่น Intel i9-12900H

กรณีศึกษา	Bootstrap Method			
	rBALPR	rBLPR	pBALPR	pBLPR
กรณีที่ 1: Toeplitz Matrix + Hard Sparsity (ρ = 0.5)	5.22 (0.27)	**1.72 (0.24)**	6.32 (0.23)	1.82 (0.18)
กรณีที่ 3: Toeplitz Matrix + Weak Sparsity (ρ = 0.5)	5.97 (0.57)	**2.02 (0.08)**	6.84 (0.47)	2.11 (0.10)
กรณีที่ 5: Equal Correlation + Hard Sparsity (ρ = 0.5)	4.85 (0.27)	**3.23 (0.24)**	5.06 (0.23)	3.35 (0.18)

หมายเหตุ ตัวหนา คือวิธีที่ใช้เวลาน้อยที่สุด, ตัวเลขในวงเล็บ คือค่าส่วนเบี่ยงเบนมาตรฐาน (หน่วย: นาที)

จากตารางที่ 7 เป็นตารางแสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของเวลาที่ใช้ในการทำบูตแสตรปของแต่ละวิธี โดยทำการบูตแสตรปจำนวน 30 รอบ (Replications) ซึ่งผู้วิจัยเลือกที่จะนำเสนอผลการทำบูตแสตรปจากข้อมูลจำลอง 3 กรณี โดยครอบคลุมการใช้สัมประสิทธิ์การถดถอยแบบ Hard Sparsity และ Weak Sparsity รวมถึงเมทริกซ์ที่มีค่าสหสัมพันธ์แบบ Toeplitz และ Equal Correlation จากตารางข้างต้น สังเกตได้ว่าเมื่อปรับใช้ Adaptive Lasso แทน Lasso ระยะเวลาที่ใช้ในการทำบูตแสตรปจะเพิ่มขึ้นอย่างชัดเจนทั้งในวิธี Residual Bootstrap และ Paired Bootstrap โดยสาเหตุของการใช้เวลามากขึ้นเนื่องจากการบูตแสตรปตัวประมาณ Adaptive Lasso + Partial Ridge จำเป็นต้องคำนวณพารามิเตอร์การปรับถึง 3 ครั้ง ในขณะที่การบูตแสตรปตัวประมาณ Lasso + Partial Ridge คำนวณพารามิเตอร์การปรับเพียง 2 ครั้ง นอกจากนี้ยังมีข้อสังเกตเพิ่มเติมคือการบูตแสตรปด้วยวิธี Residual Bootstrap จะใช้เวลาน้อยกว่าการบูตสแตรปด้วยวิธี Paired Bootstrap อยู่เล็กน้อยทั้งในตัวประมาณ Adaptive Lasso + Partial Ridge และ Lasso + Partial Ridge

อย่างไรก็ดี ผู้วิจัยเห็นว่าการใช้ตัวประมาณ Adaptive Lasso แทน Lasso ในการบูตแสตรปตัวประมาณแบบสองขั้นตอนก็เป็นอีกวิธีหนึ่งที่มีประสิทธิภาพและสามารถเป็นวิธีทางเลือกให้แก่ผู้ใช้งานได้แม้ว่าจะใช้ทรัพยากรการคำนวณที่สูงขึ้น โดยเฉพาะอย่างยิ่งเมื่อทำการบูตแสตรปตัว

ประมาณ Adaptive Lasso + Partial Ridge ด้วยวิธี Residual Bootstrap ซึ่งจากผลการศึกษาวิจัยจะเห็นได้ว่าวิธีดังกล่าวให้ความกว้างของช่วงความเชื่อมั่นสั้นที่สุดในหลายกรณีอีกทั้งยังสามารถให้ความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธีบูตแสตรปแบบอื่นๆ ดังนั้นการบูตแสตรปตัวประมาณ Adaptive Lasso + Partial Ridge ด้วยวิธี Residual Bootstrap จึงเป็นอีกวิธีการหนึ่งที่น่าสนใจในการทดสอบสมมติฐานทางสถิติในกรณีที่ข้อมูลมีมิติสูง

5.3 ข้อเสนอแนะ

จากงานวิจัยชิ้นนี้ผู้สนใจอาจจะนำไปศึกษาต่อได้อีกในกรณีดังนี้

1. วิธีการทดสอบสมมติฐานของตัวแปรอิสระในกรณีที่ข้อมูลมีมิติสูง ในงานวิจัยนี้เลือกมาศึกษาทั้งหมด 4 วิธี ซึ่งในความเป็นจริงยังมีวิธีอื่นๆ ที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการทดสอบสมมติฐานอื่นๆมาพิจารณาร่วมด้วย

2. ขอบเขตของการวิจัย ในเรื่องลักษณะของข้อมูลเช่น อาจจะใช้เมทริกซ์ความแปรปรวนร่วมของค่าคลาดเคลื่อนที่แตกต่างกันออกไป ขนาดตัวอย่างของข้อมูลและขนาดของตัวแปรอิสระอาจจะมีการเพิ่มลดได้ หรืออาจจะจำลองข้อมูลตัวแปรอิสระจากการแจกแจงรูปแบบอื่นๆที่นอกเหนือไปจากการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) เป็นต้น

3. Lasso และ Ridge เป็นตัวประมาณที่มีความเอนเอียง (Biased Estimators) ส่งผลให้เมื่อทำการหาค่าของตัวประมาณสัมประสิทธิ์การถดถอยของตัวแปรอิสระด้วยตัวประมาณดังกล่าว ค่าที่ได้จึงมีความเอนเอียงไปจากค่าจริงหรือพารามิเตอร์ ดังนั้นผู้ที่สนใจอาจจะศึกษาต่อในส่วนของการทำ Biased Correction ตัวประมาณ Lasso หรือ Ridge มาพิจารณาร่วมด้วย

บรรณานุกรม

1. James, G., et al., *An introduction to statistical learning*. Vol. 112. 2013: Springer.
2. Pungpapong, V., *A brief review on high-dimensional linear regression.* Thammasat Journal of Science and Technology, 2015. **23**(2): p. 212-223.
3. Liu, H. and B. Yu, *Asymptotic properties of Lasso+ mLS and Lasso+ Ridge in sparse high-dimensional linear regression.* Electronic Journal of Statistics, 2013. **7**: p. 3124-3169.
4. Liu, H., X. Xu, and J.J. Li, *A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models.* Statistica Sinica, 2020. **30**(3): p. 1333-1355.
5. Zou, H., *The adaptive lasso and its oracle properties.* Journal of the American statistical association, 2006. **101**(476): p. 1418-1429.
6. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), 1996. **58**(1): p. 267-288.
7. Fu, W. and K. Knight, *Asymptotics for lasso-type estimators.* The Annals of statistics, 2000. **28**(5): p. 1356-1378.
8. Hoerl, A.E. and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 1970. **12**(1): p. 55-67.
9. Chatterjee, A. and S.N. Lahiri, *Bootstrapping lasso estimators.* Journal of the American Statistical Association, 2011. **106**(494): p. 608-625.
10. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences, 1999. **96**(12): p. 6745-6750.
11. Shi, P., *Weak signal identification and inference in penalized model selection*. 2015: University of Illinois at Urbana-Champaign.
12. Tibshirani, R.J., *The lasso problem and uniqueness.* Electronic Journal of statistics, 2013. **7**: p. 1456-1490.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่างข้อมูลจำลองกรณีที่ 1: ขนาดตัวอย่างเท่ากับ 200 และจำนวนตัวแปรอิสระเท่ากับ 500

- สัมประสิทธิ์การถดถอยที่แท้จริงมีลักษณะบางเบาอย่างรุนแรง (Hard Sparsity)
- ตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$ โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์โทพลิทซ์ (Toeplitz) โดยกำหนดให้ $\sum_{ij} = \rho^{|i-j|}$; $\rho = 0.5$
- ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐานโดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10
- ทำการบูตแสตรปแบบ rBALPR, rBLPR, pBALPR, และ pBLPR

```
# จำลองข้อมูล
n <- 200
p <- 500
# Toeplizt - Covariance Matrix
Sigma_Toeplitz <- function(p,rho) {
    sigma <- c()
    for (i in 0:(p-1)){
        sig <- rho^i
        sigma <- c(sigma,sig)
    }
    cov_matrix <- toeplitz(sigma)
    return(cov_matrix)
}
cov_Toeplitz05 <- Sigma_Toeplitz(p,0.5)
set.seed(1)
```

```
mu <- rep(0,p)

X_Toeplitz05 <- rmvnorm(n, mean = mu, sigma = cov_Toeplitz05)

# จำลองสัมประสิทธิ์การถดถอย Hard Sparse Beta

Hard_Sparse_Signal <- runif(10,min = 1/3, max = 1)

Hard_Sparse_noSignal <- rep(0,490)

Hard_Sparse_Beta <- sort(c(Hard_Sparse_Signal,Hard_Sparse_noSignal),decreasing=T)

data1_HardSparse_Toeplitz05 <- data.frame(X_Toeplitz05)

write.csv(data1_HardSparse_Toeplitz05,'data1_HardSparse_Toeplitz05_R.csv',
row.names=FALSE)

write.csv(data.frame(Hard_Sparse_Beta),'Parameter_Hard_Sparse_Beta_R.csv',
row.names=FALSE)


##################### Residual Bootstrap ALPR and LPR #####################

# ตัวประมาณ ALPR

ALPR <- function(X,y) {
   # This function perform Adaptive Lasso + Partial Ridge

      # Fit ridge to obtain weight (w)

   cv.ridge <- cv.glmnet(X,y, type.measure='mse', nfold = 10, alpha=0) # finding
optimal lambda that provide lowest mse

   best_ridge_coef <- coef(cv.ridge, s = cv.ridge$lambda.min)[-1] # drop intercept

   w <- 1/abs(best_ridge_coef)

      # Fit Adaptive Lasso
```

```
    Alasso_cv <- cv.glmnet(X, y, type.measure='mse', nfold=10, alpha=1, penalty.factor
= w)

    best_alasso_coef <- coef(Alasso_cv, s = Alasso_cv$lambda.min)[-1] # drop intercept

        # Select variables to be penalized: Create Index Vector that can indicate the
position of beta whose value equal 0 so that they will be penalized during the process
of partial ridge

    selectvars <- ifelse(best_alasso_coef==0, 1, 0)

        # Fit Partial Ridge Regression

    Ridge_cv <- cv.glmnet(X,y,type.measure='mse',nfold=10,alpha=0,
penalty.factor=selectvars)

    Alasso_PartialRidge <- coef(Ridge_cv, s= Ridge_cv$lambda.min)[-1] # drop intercept


    return(Alasso_PartialRidge)
}
# ตัวประมาณ LPR
LPR <- function(X,y) {
    # This function performs Lasso + Partial Ridge

        # Fit Lasso

    lasso_cv <- cv.glmnet(X, y, type.measure='mse', nfold=10, alpha=1)

    best_lasso_coef <- coef(lasso_cv, s = lasso_cv$lambda.min)[-1] # drop intercept

        # Select variables to be penalized: Create Index Vector that can indicate the
position of beta whose value equal 0 so that they will be penalized during the process
of partial ridge

    selectvars <- ifelse(best_lasso_coef==0, 1, 0)
```

```
    # Fit Partial Ridge Regression
  Ridge_cv <- cv.glmnet(X,y,type.measure='mse',nfold=10,alpha=0,
penalty.factor=selectvars)
  Lasso_PartialRidge <- coef(Ridge_cv, s= Ridge_cv$lambda.min)[-1] # drop intercept

  return(Lasso_PartialRidge)
}

y_generator <- function(X,Beta,error) {
  # This functions generate y
  y <- X%*%Beta + error
  colnames(y) <- 'y'
  return(y)
}
# Bootstrap rBALPR และ rBLPR จำนวน 1 replication

rBALPR_LPR <- function(dataframe, B) {
  # This function performs residuals bootstrap Adaptive Lasso + Partial Ridge and
Lasso + PartialRidge for 1 replication with number of bootstrap B times

  X = data.matrix(dataframe[,1:500])
  y = dataframe[,501]
  n = 200
```

```
p = 500

startTime <- Sys.time()

# STEP 1: Finding Beta Lasso + OLS
fit_LassoOLS <- LassoOLS(X,y,fix.lambda= FALSE,cv.method='cv') # LassoOLS is a function from HDCI package.
Beta_LassoOLS <- fit_LassoOLS$beta

# STEP 2: Compute Residual
residuals <- y - X%*%Beta_LassoOLS
centered_residual <- residuals - mean(residuals)

mat_ALPR <- matrix(,nrow=B,ncol=p)
mat_LPR <- matrix(,nrow=B,ncol=p)

# STEP6: Repeat step 3-5: for loop B times
for (i in 1:B) {
    # STEP 3: Resample from the empirical distribution of the centered residual
    sample_residuals <- sample(centered_residual,size=n,replace=  TRUE)

    # STEP 4: Generate residual Bootstrap response Y*rboot
    y_rboot <- (X%*%Beta_LassoOLS) + sample_residuals
```

```
    # STEP 5: Compute Beta ALPR and Beta LPR based on (X,y*rboot)
    Alasso_PartialRidge <- ALPR(X,y_rboot)
    Lasso_PartialRidge <- LPR(X,y_rboot)

    # Collect estimated coefficient in matrix with row = B, columns = variables(p)
    mat_ALPR[i,1:p] <- Alasso_PartialRidge
    mat_LPR[i,1:p] <- Lasso_PartialRidge
    }

# STEP7: Compute quantile of 0.025 and 0.975
qt_ALPR <- t(apply(mat_ALPR,MARGIN=2,quantile,probs=c(0.025,0.975)))
qt_LPR <- t(apply(mat_LPR,MARGIN=2,quantile,probs=c(0.025,0.975)))

a_ALPR <- qt_ALPR[,1]
b_ALPR <- qt_ALPR[,2]

a_LPR <- qt_LPR[,1]
b_LPR <- qt_LPR[,2]

# STEP8: return 1−alpha confidence interval
Beta_AdaptiveLasso_PartialRidge <- ALPR(X,y)
Beta_Lasso_PartialRidge <- LPR(X,y)
```

```
    lower_ALPR <- Beta_AdaptiveLasso_PartialRidge + Beta_LassoOLS - b_ALPR
    upper_ALPR <- Beta_AdaptiveLasso_PartialRidge + Beta_LassoOLS - a_ALPR

    lower_LPR <- Beta_Lasso_PartialRidge + Beta_LassoOLS - b_LPR
    upper_LPR <- Beta_Lasso_PartialRidge + Beta_LassoOLS - a_LPR

    CI95_coef_residboot_ALPR <- matrix(c(lower_ALPR,upper_ALPR),ncol=2)
    CI95_coef_residboot_LPR <- matrix(c(lower_LPR,upper_LPR),ncol=2)

    colnames(CI95_coef_residboot_ALPR) <- c('2.5%','97.5%')
    colnames(CI95_coef_residboot_LPR) <- c('2.5%','97.5%')

    endTime <- Sys.time()
    totalTime <- endTime - startTime
    print(totalTime)

    object <- list()
    object$CI95_rBALPR <- CI95_coef_residboot_ALPR
    object$CI95_rBLPR <- CI95_coef_residboot_LPR

    return(object)
}
```

```
# Bootstrap rBALPR และ rBLPR จำนวน 50 replications
rboot_ALPR_LPR_Rep <-function(X,Beta,B,rep) {
    # This function perform paired bootstrap Adaptive Lasso + Partial Ridge B times with r replications. It returns confidence intervals in each replication.
        # X -> data - fixed
        # Beta -> parameter - fixed
        # B -> number of bootstrap
        # rep -> replication for generating new y

    startTime <- Sys.time()
    n = 200
    p = 500
    mat_rboot_ALPR <- matrix(,ncol=rep*2,nrow=p)
    mat_rboot_LPR <- matrix(,ncol=rep*2,nrow=p)

    col_index <- 1

    for (i in 1:rep) {
        # generate new y each replication
        error <- error_generator(X,Beta)
        y <- y_generator(X,Beta,error)
        dataframe <- cbind(X,y)
        print(paste('rep',i))
```

```
    rboot <- rBALPR_LPR(dataframe,B=B)

    CI95_rBALPR <- rboot$CI95_rBALPR
    CI95_rBLPR <- rboot$CI95_rBLPR

    # Collect data into matrix
    mat_rboot_ALPR[1:p, c(col_index,col_index+1)] <- CI95_rBALPR
    mat_rboot_LPR[1:p, c(col_index,col_index+1)] <- CI95_rBLPR

    col_index <- col_index + 2
}
# rename columns and rows in matrix
replication <- rep(paste('Rep',c(1:rep),sep=""), each=2)
interval <- rep(c("_2.5%","_97.5%"), times=2)
column_names  <- paste(replication,interval, sep="")
colnames(mat_rboot_ALPR) <- column_names
colnames(mat_rboot_LPR) <- column_names

row_names <- paste('Beta',c(1:p),sep="")
rownames(mat_rboot_ALPR) <- row_names
rownames(mat_rboot_LPR) <- row_names
```

```
    endTime <- Sys.time()
    print('TotalTime')
    print(endTime-startTime)

    object <- list()
    object$rBALPR <- mat_rboot_ALPR
    object$rBLPR <- mat_rboot_LPR

    return(object)
}
##################### Paired Bootstrap ALPR and LPR ####################
# Bootstrap pBALPR และ pBLPR จำนวน 1 replication
pBALPR_LPR<- function(dataframe,B) {
    # This function performs paired bootstrap Adaptive Lasso + Partial Ridge and Lasso
+ PartialRidge for 1 replication with number of bootstrap B times

    startTime <- Sys.time()

    n = 200
    p = 500
    mat_ALPR <- matrix(,nrow=B,ncol=p)
    mat_LPR <- matrix(,nrow=B,ncol=p)
    # STEP4: Performs paired bootstrap of Adaptive Lasso + Paritial Ridge for B times
```

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

    # STEP1 : Sampling data with replacement from (X,y) -> (X*,y*) with size = 200
    sampling_data <- dataframe[sample(nrow(dataframe),size=200,replace=T),]
    Xpboot <- data.matrix(sampling_data[,1:500])
    ypboot <- sampling_data[,501]

    # STEP2-3 : Adaptive Lasso + Partial Ridge and Lasso + Partial Ridge
    Alasso_PartialRidge <- ALPR(Xpboot,ypboot)
    Lasso_PartialRidge <- LPR(Xpboot,ypboot)

    # Collect estimated coefficient in matrix with row = B, columns = variables(p)
    mat_ALPR[i,1:p] <- Alasso_PartialRidge
    mat_LPR[i,1:p] <- Lasso_PartialRidge
}
# STEP5: Construct 95% Confidence Interval
CI95_coef_pairboot_ALPR <- t(apply(mat_ALPR,MARGIN=2,quantile,probs=c(0.025,0.975)))
CI95_coef_pairboot_LPR <- t(apply(mat_LPR,MARGIN=2,quantile,probs=c(0.025,0.975)))

endTime <- Sys.time()
print(endTime - startTime)
```

```
    object <- list()
    object$CI95_pBALPR <- CI95_coef_pairboot_ALPR
    object$CI95_pBLPR <- CI95_coef_pairboot_LPR

    return(object)
}
# Bootstrap pBALPR และ pBLPR จำนวน 50 replication
pboot_ALPR_LPR_Rep <- function(X,Beta,B,rep) {
    # This function perform paired bootstrap Adaptive Lasso + Partial Ridge B times with
r replications. It returns confidence intervals of each replication.
        # X -> data - fixed
        # Beta -> parameter - fixed
        # B -> number of bootstrap
        # rep -> replication for generating new y

    startTime <- Sys.time()
    n = 200
    p = 500
    mat_pboot_ALPR <- matrix(,ncol=rep*2,nrow=p)
    mat_pboot_LPR <- matrix(,ncol=rep*2,nrow=p)

    col_index <- 1
```

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

    # generate new y each replication
    error <- error_generator(X,Beta)
    y <- y_generator(X,Beta,error)
    dataframe <- cbind(X,y)

    pboot <- pBALPR_LPR(dataframe,B=B)

    CI95_pBALPR <- pboot$CI95_pBALPR
    CI95_pBLPR <- pboot$CI95_pBLPR

    # Collect data into matrix
    mat_pboot_ALPR[1:p, c(col_index,col_index+1)] <- CI95_pBALPR
    mat_pboot_LPR[1:p, c(col_index,col_index+1)] <- CI95_pBLPR

    col_index <- col_index + 2
}

# rename columns in matrix
replication <- rep(paste('Rep',c(1:rep),sep=""), each=2)
interval <- rep(c("_2.5%","_97.5%"), times=2)
```

```
    column_names  <- paste(replication,interval, sep="")
    colnames(mat_pboot_ALPR) <- column_names
    colnames(mat_pboot_LPR) <- column_names

    row_names <- paste('Beta',c(1:p),sep="")
    rownames(mat_pboot_ALPR) <- row_names
    rownames(mat_pboot_LPR) <- row_names

    endTime <- Sys.time()
    print('TotalTime')
    print(endTime-startTime)

    object <- list()
    object$pBALPR <- mat_pboot_ALPR
    object$pBLPR <- mat_pboot_LPR

    return(object)
}
########## Run rBALPR, rBLPR, pBALPR, pBLPR ##########
X <- data.matrix(read.csv('data1_HardSparse_Toeplitz05_R.csv'))
Beta <- data.matrix(read.csv('Parameter_Hard_Sparse_Beta_R.csv'))
rbootALPR_LPR <- rboot_ALPR_LPR_Rep(X, Beta, B=1000, rep=50)
pbootALPR_LPR <- pboot_ALPR_LPR_Rep(X, Beta, B=1000, rep=50)
```

ตัวอย่างข้อมูลจำลองกรณีที่ 7: ขนาดตัวอย่างเท่ากับ 200 และจำนวนตัวแปรอิสระเท่ากับ 500

- สัมประสิทธิ์การถดถอยที่แท้จริงมีลักษณะบางเบาอย่างอ่อน (Weak Sparsity)
- ตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์และเมทริกซ์ความแปรปรวนร่วม $\sum_{p\times p}$ โดยที่ $\sum_{p\times p}$ เป็นเมทริกซ์ Equal Correlation โดยกำหนดระดับความสัมพันธ์ $\rho = 0.5$
- ค่าความคลาดเคลื่อนมีการแจกแจงปกติมาตรฐานโดยกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) = 10
- ทำการบูตแสตรปแบบ rBALPR, rBLPR, pBALPR, และ pBLPR

```
# จำลองข้อมูล
n <- 200
p <- 500
# Equal Correlation - Covariance Matrix
Sigma_EqualCorr <- function(p,rho){
    cov_matrix <- diag(p)
    cov_matrix <- ifelse(cov_matrix<=0,rho,cov_matrix)
    return(cov_matrix)
}
cov_EqualCorr05 <- Sigma_EqualCorr(p,0.5)
mu <- rep(0,p)
X_EqualCorr05 <- rmvnorm(n, mean = mu, sigma = cov_EqualCorr05)
# Weak Sparse Beta
Weak_Sparse_large <- rnorm(10,mean = 1, sd = sqrt(0.001))
Weak_Sparse_small <-c()
```

```
for (j in 1:490) {
    beta_small <- 1/(((j)+3)^2)
    Weak_Sparse_small <- c(Weak_Sparse_small,beta_small)
}
Weak_Sparse_Beta <- sort(c(Weak_Sparse_large, Weak_Sparse_small), decreasing=T)
Data7_WeakSparse_EqualCorr05 <- data.frame(X_EqualCorr05)
write.csv(data7_WeakSparse_EqualCorr05,'data7_WeakSparse_EqualCorr05_R.csv',row.names=FALSE)
write.csv(data.frame(Weak_Sparse_Beta),'Parameter_Weak_Sparse_Beta_R.csv',row.names=FALSE)

# ใช้ Functions การ bootstrap rBALPR, rBLPR, pBALPR, และ pBLPR เช่นเดียวกับกรณีที่ 1
########## Run rBALPR, rBLPR, pBALPR, pBLPR ##########
X <- data.matrix(read.csv('data7_WeakSparse_EqualCorr05_R.csv))
Beta <- data.matrix(read.csv('Parameter_Weak_Sparse_Beta_R.csv'))
rbootALPR_LPR <- rboot_ALPR_LPR_Rep(X, Beta, B=1000, rep=50)
pbootALPR_LPR <- pboot_ALPR_LPR_Rep(X, Beta, B=1000, rep=50)

### การวิเคราะห์ผลลัพธ์จากการบูตแสตรปจะดำเนินการในภาษา python version 3.9.7 ###
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
def length_ci(dataframe,rep=50):
    ''' Calculate Length of Confidence Intervals for each Coefficient'''
    ci_length = []
    column_index = 0
    for r in range(rep):
        temp = dataframe.iloc[:,column_index:column_index+2]
        length = temp.iloc[:,1] - temp.iloc[:,0]
        ci_length.append(length)
        column_index += 2
    ci_length = np.array(ci_length).T
    ci_length = pd.DataFrame(ci_length,columns=['Length_CI_Rep{}'.format(i) for i in
range(1,51)])
    return ci_length
```



```
def cov_prob(dataframe,parameter,rep=50):

    ''' Calculate Coverage probability for each Coefficient'''

    parameter = parameter.values.ravel()
    In_confidence_interval = []
    column_index = 0
```

```
    for r in range(rep):
        temp = dataframe.iloc[:,column_index:column_index+2]
        upper = temp.iloc[:,1]
        lower = temp.iloc[:,0]
        in_ci = np.where((parameter >= lower) & (parameter <= upper),1,0)
        In_confidence_interval.append(in_ci)
        column_index += 2
    In_confidence_interval = np.array(In_confidence_interval).T
    In_confidence_interva =
pd.DataFrame(In_confidence_interval,columns=['In_confidence_interval_Rep{}'.format(
i) for i in range(1,51)])
    coverage_probability = In_confidence_interval.sum(axis=1)/rep
    return coverage_probability


def perf_measure(y_actual, y_hat): # From
https://stackoverflow.com/questions/31324218/scikit-learn-how-to-obtain-true-
positive-true-negative-false-positive-and-fal
    '''
Finding numbers of TP, FP, TN, FN '''
    TP = 0
    FP = 0
    TN = 0
    FN = 0
```

```
    for i in range(len(y_hat)):
        if y_actual[i]==y_hat[i]==1:
            TP += 1
        if y_hat[i]==1 and y_actual[i]!=y_hat[i]:
            FP += 1
        if y_actual[i]==y_hat[i]==0:
            TN += 1
        if y_hat[i]==0 and y_actual[i]!=y_hat[i]:
            FN += 1

    return(TP, FP, TN, FN)
def FPR_FNR(dataframe,parameter,rep=50):
    '''
    This function calculates false positive rate and false negative for each replication
from dataframe
    '''

    column_index = 0
    conclusion = []
    for r in range(rep):
        temp = dataframe.iloc[:,column_index:column_index+2]
        upper = temp.iloc[:,1]
        lower = temp.iloc[:,0]
```

```
        condition_check = np.where((lower <= 0) & (upper>=0),0,1)# if confidence interval contains 0 -> assign 0 AND if not -> assign 1
        column_index += 2
        conclusion.append(condition_check)
    conclusion = np.array(conclusion).T
    result = pd.DataFrame(conclusion,columns=['Not_contain0_Rep{}'.format(i) for i in range(1,51)])

    transform_parameter = np.where(parameter != 0,1,0).ravel()

    FPR_eachRep = []
    FNR_eachRep = []

    for r in range(rep):
        temp = pd.DataFrame()
        temp['parameter'] = transform_parameter # y_actual
        temp['estimator'] = result.iloc[:,r] # y-hat

        TP, FP, TN, FN = perf_measure(temp['parameter'],temp['estimator'])

        try:
            FPR = FP/(FP+TN)
```

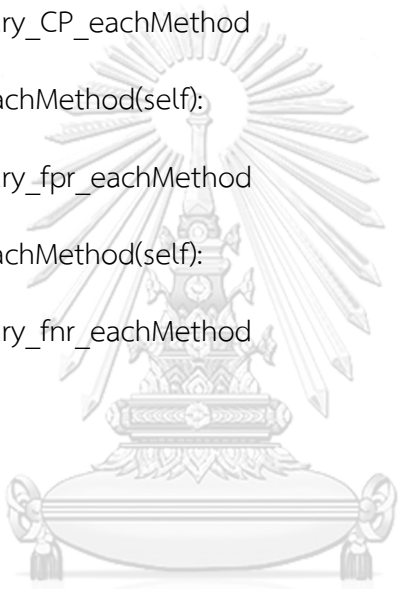
```
        except:
            FPR = np.nan # if denominator is 0 -> assign FPR = np.nan

        try:
            FNR = FN/(FN+TP)
        except:
            FNR = np.nan # if denominator is 0 -> assign FNR = np.nan

        FPR_eachRep.append(FPR)
        FNR_eachRep.append(FNR)
    return FPR_eachRep, FNR_eachRep

class data_result:
    def __init__(self, rboot_ALPR, rboot_LPR, pboot_ALPR, pboot_LPR, parameter,
title):
        self.rboot_ALPR = rboot_ALPR
        self.rboot_LPR = rboot_LPR
        self.pboot_ALPR = pboot_ALPR
        self.pboot_LPR = pboot_LPR
        self.parameter = parameter
        self.title = title

    ### Length of Confidence Interval
```

```
    LCI_rboot_ALPR = length_ci(self.rboot_ALPR).mean(axis=1)
    LCI_rboot_LPR = length_ci(self.rboot_LPR).mean(axis=1)
    LCI_pboot_ALPR = length_ci(self.pboot_ALPR).mean(axis=1)
    LCI_pboot_LPR = length_ci(self.pboot_LPR).mean(axis=1)
    LCI = pd.DataFrame({'LCI_rboot_ALPR':LCI_rboot_ALPR,
                        'LCI_rboot_LPR':LCI_rboot_LPR,
                        'LCI_pboot_ALPR':LCI_pboot_ALPR,
                        'LCI_pboot_LPR':LCI_pboot_LPR})
    self.LCI = LCI

### Coverage Probability
    CP_rboot_ALPR = cov_prob(self.rboot_ALPR, parameter) # sum By rows
    CP_rboot_LPR = cov_prob(self.rboot_LPR, parameter) # sum By rows
    CP_pboot_ALPR = cov_prob(self.pboot_ALPR, parameter) # sum By rows
    CP_pboot_LPR = cov_prob(self.pboot_LPR, parameter) # sum By rows
    CP = pd.DataFrame({'CP_rboot_ALPR':CP_rboot_ALPR,
                        'CP_rboot_LPR':CP_rboot_LPR,
                        'CP_pboot_ALPR':CP_pboot_ALPR,
                        'CP_pboot_LPR':CP_pboot_LPR})
    self.CP = CP

### False Positive Rate & False Negative Rate
    FPR_rboot_ALPR, FNR_rboot_ALPR = FPR_FNR(self.rboot_ALPR, parameter)
```

```
    FPR_rboot_LPR, FNR_rboot_LPR = FPR_FNR(self.rboot_LPR, parameter)
    FPR_pboot_ALPR, FNR_pboot_ALPR = FPR_FNR(self.pboot_ALPR, parameter)
    FPR_pboot_LPR, FNR_pboot_LPR = FPR_FNR(self.pboot_LPR, parameter)

    fpr = pd.DataFrame({'FPR_rboot_ALPR':FPR_rboot_ALPR,
                        'FPR_rboot_LPR':FPR_rboot_LPR,
                        'FPR_pboot_ALPR':FPR_pboot_ALPR,
                        'FPR_pboot_LPR':FPR_pboot_LPR})

    fnr = pd.DataFrame({'FNR_rboot_ALPR':FNR_rboot_ALPR,
                        'FNR_rboot_LPR':FNR_rboot_LPR,
                        'FNR_pboot_ALPR':FNR_pboot_ALPR,
                        'FNR_pboot_LPR':FNR_pboot_LPR})
    self.fpr = fpr
    self.fnr = fnr

### Summary of each method performance

    # Summary LCI
    mean_LCI = self.LCI.mean(axis=0)
    sd_LCI = self.LCI.std(axis=0)
    self.summary_LCI_eachMethod = pd.DataFrame({'Mean':mean_LCI,'SD':sd_LCI})
```

```
        # Summary CP
        mean_CP = self.CP.mean(axis=0)
        sd_CP = self.CP.std(axis=0)
        self.summary_CP_eachMethod =  pd.DataFrame({'Mean':mean_CP,'SD':sd_CP})

        # Summary fpr
        mean_fpr = self.fpr.mean(axis=0)
        sd_fpr = self.fpr.std(axis=0)
        self.summary_fpr_eachMethod =
pd.DataFrame({'Mean_fpr':mean_fpr,'SD_fpr':sd_fpr})

        # Summary fnr
        mean_fnr = self.fnr.mean(axis=0)
        sd_fnr = self.fnr.std(axis=0)
        self.summary_fnr_eachMethod =
pd.DataFrame({'Mean_fnr':mean_fnr,'SD_fnr':sd_fnr})
    def get_LCI(self):
        return self.LCI
    def get_CP(self):
        return self.CP
    def get_fpr(self):
        return self.fpr
    def get_fnr(self):
```

```
    return self.fnr
def get_title(self):
    return self.title
def get_result_LCI_eachMethod(self):
    return self.summary_CP_eachMethod
def get_result_CP_eachMethod(self):
    return self.summary_CP_eachMethod
def get_result_fpr_eachMethod(self):
    return self.summary_fpr_eachMethod
def get_result_fnr_eachMethod(self):
    return self.summary_fnr_eachMethod
```



ประวัติผู้เขียน

ชื่อ-สกุล	นายพริษฐ์ ชาญเชิงพานิช
วัน เดือน ปี เกิด	10 พฤศจิกายน 2537
สถานที่เกิด	จังหวัดสงขลา
วุฒิการศึกษา	ปริญญาตรีบริหารธุรกิจบัณฑิต (หลักสูตรนานาชาติ) สาขาการธนาคารและการเงิน คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ปัจจุบัน	848/549 ถนนประชาชื่น แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800
รางวัลที่ได้รับ	รองชนะเลิศเหรียญเงิน การแข่งขันเศรษฐศาสตร์เพชรยอดมงกุฎ ครั้งที่ 5 ระดับอุดมศึกษา ประจำปีการศึกษา 2560