# การเปรียบเทียบวิธีบูตสแทร็ปในการประมาณช่วงความเชื่อมั่นของค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่มีมิติสูงด้วยวิธีลาสโซ่แบบปรับปรุงและพาร์เชียลริดจ์

# A COMPARISON OF BOOTSTRAP METHODS FOR ADAPTIVE LASSO + PARTIAL RIDGE TO CONSTRUCT CONFIDENCE INTERVALS FOR PARAMETERS IN HIGH-DIMENSIONAL SPARSE LINEAR MODELS

พริษฐ์ ชาญเชิงพานิช* และวิฐรา พึ่งพาพงศ์

*ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย*

Parit Chancherngpanich* and Vitara Pungpapong

*Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University*

**Received: March 10, 2022 ; Accepted: June 27, 2022**

## บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอวิธีบูตสแทร็ปลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์ในการสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยสำหรับข้อมูลที่มีมิติสูงและเปรียบเทียบประสิทธิภาพกับวิธีบูตสแทร็ปลาสโซ่ + พาร์เชียลริดจ์ วิธีบูตสแทร็ปลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์เป็นตัวประมาณแบบ 2 ขั้นตอน คือ ใช้วิธีลาสโซ่แบบปรับปรุงในการคัดเลือกตัวแปรอิสระจากนั้นใช้วิธีพาร์เชียลริดจ์ในการประมาณค่าสัมประสิทธิ์การถดถอยอีกครั้ง การศึกษานี้ได้ทดลองบูตสแทร็ป 2 วิธีได้แก่ วิธีสุ่มส่วนเหลือและวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ อีกทั้งยังศึกษาสัมประสิทธิ์การถดถอยใน 2 ลักษณะได้แก่บางเบาอย่างอ่อนและบางเบาอย่างรุนแรง โดยลักษณะบางเบาอย่างอ่อนและบางเบาอย่างรุนแรง หมายถึง กรณีที่สัมประสิทธิ์การถดถอยส่วนใหญ่มีค่าใกล้เคียงศูนย์และเท่ากับศูนย์ ตามลำดับ การวิจัยครั้งนี้ใช้ข้อมูลจำลองที่ มีมิติสูง ซึ่งตัวแปรอิสระสร้างจากการแจกแจงแบบปกติหลายตัวแปรโดยใช้เมทริกซ์ความแปรปรวนร่วมที่แตกต่างกันทั้งหมด 8 กรณี เกณฑ์ที่ใช้วัดประสิทธิภาพที่ใช้ คือ ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นและความน่าจะเป็นคุ้มรวม ผลการศึกษาจากข้อมูลจำลอง พบว่าวิธีบูตสแทร็ปแบบสุ่มส่วนเหลือลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์ให้ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นต่ำที่สุดในเกือบทุกกรณี อย่างไรก็ตาม เมื่อพิจารณาความน่าจะเป็นคุ้มรวม พบว่าไม่ปรากฏวิธีการบูตสแทร็ปแบบใดแบบหนึ่งที่มีประสิทธิภาพสูงสุดสำหรับทุกกรณี อนึ่ง เมื่อนำวิธีบูตสแทร็ปแบบต่างๆไปปรับใช้กับข้อมูลจริง กล่าวคือ ข้อมูลไมโครอาร์เรย์ในโรคมะเร็งลำไส้ใหญ่ พบว่าวิธีบูตสแทร็ปแบบสุ่มส่วนเหลือลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์ยังคงให้ค่าเฉลี่ยความ

*Corresponding author: chancherngpanich@gmail.com

DOI: 10.14456/tjst.2021.40

กว้างของช่วงความเชื่อมั่นต่ำที่สุด และวิธีบูตสแทร็ปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์ให้ความน่าจะเป็นคุ้มรวมสูงสุด

**คำสำคัญ :** การถดถอยเชิงเส้นที่มีมิติสูง; การถดถอยแบบลาสโซ่; การถดถอยแบบลาซโซ่ปรับปรุง; การถดถอยแบบริดจ์; บูตสแทร็ปสแทร็ป; ช่วงความเชื่อมั่น

## Abstract

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 and paired bootstrap. We also consider two cases of coefficients which are weak sparsity and hard sparsity where weak sparsity and hard sparsity refer to the case that majority of coefficients have value closed to zero and equal to zero respectively. Simulation studies in 8 cases of high – dimensional covariates that are generated from multivariate normal distribution with different types of covariance matrix. Mean interval lengths and coverage probabilities are used to measure and compare performance of bootstrap methods. Our simulation studies show that the residual bootstrap adaptive lasso + partial ridge provides lowest mean interval lengths for most cases. However, it is not obvious that which bootstrap method is the best in terms of providing highest coverage probabilities. We also apply each bootstrap method with the real data, colon cancer microarray data set. The results show that the residual bootstrap adaptive lasso + partial ridge and the paired bootstrap adaptive lasso + partial ridge are the best method in terms of mean interval lengths and coverage probabilities, respectively.

**Keywords:** high – dimensional regression; lasso regression; adaptive lasso regression; ridge regression; bootstrap; confidence intervals

## 1. บทนำ

การวิเคราะห์การถดถอยเชิงเส้นเป็นวิธีทางสถิติที่นิยมใช้กันอย่างแพร่หลายในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยการวิเคราะห์การถดถอยเชิงเส้นจะใช้วิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Squares Method, OLS) ในการประมาณค่าสัมประสิทธิ์การถดถอย อย่างไรก็ตาม วิธี OLS มีข้อจำกัดคือ เมื่อข้อมูลมีมิติสูงหรือจำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่างของข้อมูล *(p>n)* วิธี OLS จะไม่สามารถหาค่าของตัวประมาณสัมประสิทธิ์การถดถอยได้ (James et al. 2013) นอกจากนี้อาจเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์กันเองสูงซึ่งส่งผลให้ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธี OLS มีความไม่เสถียร (Pungpapong, 2015) การวิเคราะห์ข้อมูลที่มีมิติสูงจึงนิยมใช้วิธีการ

ประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Regression) โดยการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษมีหลากหลายวิธีแต่ที่นิยมใช้กันอย่างแพร่หลายได้แก่ การถดถอยลาสโซ่ (Lasso Regression) การถดถอยลาสโซ่แบบปรับปรุง (Adaptive Lasso Regression) และการถดถอยแบบริดจ์ (Ridge Regression)

การวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษทั้งสามวิธีที่กล่าวมาสามารถหาค่าประมาณสัมประสิทธิ์การถดถอยในกรณีที่ข้อมูลมีมิติสูงได้ แต่ในการจะตอบคำถามว่าตัวแปรอิสระตัวใดบ้างที่มีความสัมพันธ์กับตัวแปรตามแบบมีนัยสำคัญทางสถิตินั้น ยังคงเป็นประเด็นที่ท้าทาย วิธีที่นิยมใช้ในการตอบคำถามดังกล่าวเมื่อข้อมูลมีมิติสูง คือ การสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยด้วยวิธีบูตสแทร็ป (Bootstrap) หากช่วงความเชื่อมั่นที่ $(1-\alpha)\%$ ไม่ครอบคลุมค่าศูนย์ จะสามารถระบุได้ว่าตัวแปรอิสระตัวดังกล่าวมีความสัมพันธ์กับตัวแปรตามที่ระดับนัยสำคัญ $\alpha$

Liu & Yu (2013) นำเสนอวิธีบูตสแทร็ปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + OLS ซึ่งเป็นตัวประมาณแบบสองขั้นตอน โดยขั้นตอนที่หนึ่งใช้วิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระ และขั้นตอนที่สองใช้วิธี OLS ในการประมาณค่าสัมประสิทธิ์การถดถอย ทว่าวิธีบูตสแทร็ป Lasso + OLS มักจะประสบปัญหาช่วงความเชื่อมั่นที่สร้างขึ้นมักจะไม่ครอบคลุมค่าของสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่มีค่าน้อยมากแต่ไม่เท่ากับศูนย์ (น้อยกว่า $\frac{1}{\sqrt{n}}$) เนื่องจากการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่ในขั้นตอนที่หนึ่งมักจะให้ค่าประมาณสัมประสิทธิ์การถดถอยของตัวแปรอิสระดังกล่าวเป็นศูนย์ ส่งผลให้ตัวแปรอิสระดังกล่าวไม่ได้รับคัดเลือกให้นำไปหาค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธี OLS เรียกปัญหาลักษณะนี้ว่าความน่าจะเป็นคุ้มรวมต่ำ (Liu et al. 2020)

Liu et al. (2020) ได้นำเสนอวิธีบูตสแทร็ปตัวประมาณสัมประสิทธิ์การถดถอยแบบ Lasso + Partial Ridge (LPR) ซึ่งเป็นตัวประมาณแบบสองขั้นตอนเช่นกัน โดยขั้นตอนที่หนึ่งใช้วิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระและขั้นตอนที่สองใช้ฟังก์ชันการลงโทษแบบ L2 – Norm แก่ตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่เท่านั้น ซึ่งเปรียบเสมือนเป็นการนำตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่ไปหาค่าประมาณสัมประสิทธิ์การถดถอยใหม่อีกครั้ง เนื่องจากวิธีริดจ์มักจะให้ค่าประมาณสัมประสิทธิ์การถดถอยเข้าใกล้ศูนย์แต่ไม่เท่ากับศูนย์ ผลการศึกษาพบว่าในกรณีที่ปรากฏสัมประสิทธิ์การถดถอยบางตัวที่มีค่าน้อยมากแต่ไม่เท่ากับศูนย์ วิธีบูตสแทร็ปแบบ Lasso + Partial Ridge ให้ความน่าจะเป็นคุ้มรวมสูงกว่าวิธี Lasso + OLS

ในการศึกษาครั้งนี้ ผู้วิจัยมีจุดประสงค์ที่จะนำเสนอวิธีบูตสแทร็ปตัวประมาณแบบ Adaptive Lasso + Partial Ridge โดยการเปลี่ยนจากวิธี Lasso เป็น Adaptive Lasso เนื่องจากวิธี Adaptive Lasso มีคุณสมบัติที่โดดเด่นประการหนึ่งคือ คุณสมบัติออราเคิล (Oracle Property) ซึ่งเป็นคุณสมบัติที่สามารถคัดเลือกตัวแปรเข้าตัวแบบเสมือนทราบตัวแบบที่แท้จริง (Zou, 2006) และจากการทบทวนวรรณกรรมที่ผ่านมายังไม่พบว่ามีการศึกษาวิธีบูตสแทร็ปตัวประมาณแบบดังกล่าว ดังนั้นผู้วิจัยจึงสนใจศึกษาเกี่ยวกับประเด็นนี้ โดยจะทำการเปรียบเทียบกับวิธีบูตสแทร็ปตัวประมาณแบบ Lasso + Partial Ridge ทั้งนี้ผู้วิจัยจะทดลองบูตสแทร็ป 2 วิธีคือ วิธีสุ่มส่วนเหลือ (Residual Bootstrap) และวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ (Paired Bootstrap) จากนั้นจึงทำ

การเปรียบเทียบและวิเคราะห์ผลลัพธ์โดยใช้ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่น (Mean Interval Lengths) และความน่าจะเป็นคุ้มรวม (Coverage Probabilities) เป็นเกณฑ์การวัดประสิทธิภาพเพื่อหาวิธีที่เหมาะสมและมีประสิทธิภาพที่สุดในการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูง

## 2. วิธีการ

### 2.1 การวิเคราะห์การถดถอยเชิงเส้น (OLS)

การวิเคราะห์การถดถอยเชิงเส้นเมื่อข้อมูลมีตัวอย่างขนาด n และตัวแปรอิสระขนาด p สามารถเขียนเป็นสมการของตัวแบบได้ดังนี้

$$Y = X\beta + \varepsilon \quad (1)$$

เมื่อ $Y$ คือ เวกเตอร์ของตัวแปรตามขนาด n

$X$ คือ เมทริกซ์ของตัวแปรอิสระขนาด n x p

$\beta$ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด p

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

ในการหาค่าของตัวประมาณสัมประสิทธิ์การถดถอย $(\hat{\beta})$ จะหาได้จากวิธีกำลังสองน้อยที่สุดซึ่งเขียนได้ดังสมการ

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j\beta_j \right\|^2 = (X^TX)^{-1}X^Ty \quad (2)$$

ในการตรวจสอบว่าตัวแปรอิสระตัวที่ $j$ มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ $\alpha$ สามารถทำได้โดยการสร้างช่วงความเชื่อมั่นที่ $(1-\alpha)\%$ ของพารามิเตอร์ $\beta_j$ กล่าวคือ $[L_j, U_j]$ โดยที่

$$L_j = \hat{\beta}_j - t_{(1-\frac{\alpha}{2},\ n-p-1)} SE[\hat{\beta}_j]$$

และ

$$U_j = \hat{\beta}_j + t_{(1-\frac{\alpha}{2},\ n-p-1)} SE[\hat{\beta}_j]$$

เมื่อ $t_{(1-\frac{\alpha}{2},\ n-p-1)}$ คือ ควอนไทล์ที่ $1-\frac{\alpha}{2}$ ของการแจกแจงแบบที ที่องศาอิสระ $n-p-1$ และ $SE[\hat{\beta}_j]$ คือ ความคลาดเคลื่อนมาตรฐานของตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ทั้งนี้ หากช่วงความเชื่อมั่นที่สร้างขึ้นไม่ครอบคลุมค่าศูนย์ จะสามารถสรุปได้ว่าตัวแปรอิสระตัวที่ $j$ มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ $\alpha$

### 2.2 การประมาณค่าสัมประสิทธิ์ด้วยการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ

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

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

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j\beta_j \right\|^2 + P_\lambda(\beta) \quad (4)$$

สำหรับฟังก์ชันการลงโทษ $P_\lambda(\beta)$ มีหลายรูปแบบแต่ที่นิยมใช้กันอย่างแพร่หลายคือฟังก์ชันการลงโทษแบบแอลวันนอร์ม (L1 – Norm) และแอลทูนอร์ม (L2 – Norm) เขียนแสดงได้ดังสมการที่ 5 และ 6 ตามลำดับ

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \quad (5)$$

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j{}^2 \quad (6)$$

**2.2.1 วิธี Lasso Regression**

วิธีลาสโซ่ (Lasso) นำเสนอโดย Tibshirani (1996) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L1 – Norm โดยสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 7

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

การใช้ฟังก์ชันลงโทษแบบแอลวันนอร์ม (L1 – Norm) จะทำให้เวกเตอร์ตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่ ซึ่งเขียนแทนด้วย $\hat{\beta}_{Lasso}$ ประกอบด้วยค่าศูนย์จำนวนมาก (Sparse Vector) ทั้งนี้ขึ้นอยู่กับพารามิเตอร์การปรับ ($\lambda$) หากพารามิเตอร์การปรับมีค่ามากจะส่งผลให้เวกเตอร์ $\hat{\beta}_{Lasso}$ มีจำนวนค่าศูนย์มาก หากพารามิเตอร์การปรับมีค่าน้อยจะส่งผลให้เวกเตอร์ $\hat{\beta}_{Lasso}$ มีจำนวนค่าศูนย์น้อย และในกรณีที่พารามิเตอร์การปรับเท่ากับ 0 การถดถอยลาสโซ่จะกลับมาเป็นการถดถอยแบบดั้งเดิม (Tibshirani, 1996)

การเลือกพารามิเตอร์การปรับ ($\lambda$) ที่เหมาะสมเป็นสิ่งที่จำเป็นสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยลาสโซ่ โดยทั่วไปนิยมใช้วิธี Cross Validation (CV) ซึ่งเป็นวิธีที่ใช้วัดประสิทธิภาพของตัวแบบเมื่อต้องการทดสอบว่าค่าพารามิเตอร์ปรับที่เท่าใดส่งผลให้ผลรวมความคลาดเคลื่อนกำลังสองน้อยที่สุด โดยหลักการของวิธี Cross - Validation มีดังนี้

2.2.1.1 กำหนด $\lambda$ เป็นเซ็ตของพารามิเตอร์ปรับที่ต้องการทดสอบ ซึ่งประกอบด้วยสมาชิกจำนวน m ตัว ดังนี้ $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$

2.2.1.2 แบ่งข้อมูลออกเป็นชุดย่อยๆ จำนวน k ชุด โดยใช้สัญลักษณ์ $d_1, d_2, \ldots, d_k$ แทนข้อมูลย่อยชุดที่ $1,2,\ldots,k$ ตามลำดับ ทั้งนี้จำนวนข้อมูลย่อยในแต่ละชุดต้องมีจำนวนเท่ากัน

2.2.1.3 สำหรับครั้งที่ $i$ เมื่อ $i = 1,2,\ldots,m$

2.2.1.3.1 ให้ใช้ $\lambda = \lambda_i$ และภายใต้ครั้งที่ $i$ ให้ทำการแบ่งข้อมูล $j$ ครั้ง เมื่อ $j = 1,2,\ldots,k$

2.2.1.3.2 สำหรับครั้งที่ $j$ เมื่อ $j = 1,2,\ldots,k$ ให้ใช้ข้อมูลย่อยทุกชุด ยกเว้นชุด $d_j$ ในการสร้างตัวแบบด้วยวิธีการถดถอยแบบลาสโซ่ จากนั้นใช้ข้อมูลย่อยชุด $d_j$ เป็นข้อมูลชุดทดสอบ เพื่อคำนวณผลรวมความคลาดเคลื่อนกำลังสอง โดยจะใช้สัญลักษณ์ $RSS_{\lambda_{i,j}}$

2.2.1.3.3 จากนั้นคำนวณค่าเฉลี่ยของ RSS ที่ได้จากการใช้ $\lambda = \lambda_i$ ซึ่งเขียนได้ดังสมการที่ 8

$$CV(\lambda_i) = \frac{1}{k} \sum_{j=1}^{k} RSS_{\lambda_{i,j}} \quad (8)$$

2.2.1.4 เลือก $\lambda_i$ ที่ทำให้ $CV(\lambda_i)$ มีค่าน้อยที่สุด ซึ่ง $\lambda_i$ ดังกล่าวจะเป็นพารามิเตอร์ปรับที่เหมาะสมที่สุดสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่

ด้วยเหตุที่วิธีลาสโซ่มีคุณสมบัติทำให้ได้เวกเตอร์ตัวประมาณสัมประสิทธิ์การถดถอยที่ประกอบด้วยค่าศูนย์จำนวนมาก ส่งผลให้วิธีลาสโซ่สามารถหาค่าประมาณสัมประสิทธิ์การถดถอยพร้อมทั้งคัดเลือกตัวแปรเข้าตัวแบบได้ในเวลาเดียวกันทำให้ตัวแบบที่ได้ง่ายต่อการแปรผลลัพธ์ อีกทั้งยังแก้ปัญหาที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงได้ (Multicollinearity) (Pungpapong 2015) อย่างไรก็ตาม ค่าประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธีลาสโซ่มักไม่มีความคงเส้นคงวา ส่งผลให้การคัดเลือกตัวแปรเข้าตัวแบบมักไม่มีความคงเส้นคงวา (Knight and Fu 2007)

**2.2.2 วิธี Adaptive Lasso Regression**

วิธีลาสโซ่แบบปรับปรุง นำเสนอโดย Zou (2006) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L1 – Norm และเพิ่มเงื่อนไขการให้ค่าน้ำหนักแก่พารามิเตอร์ที่แตกต่างกัน โดยสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 9

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|$$

; โดยที่ $\widehat{w}_j \begin{cases} \frac{1}{|\hat{\beta}_{OLS}|} ; n > p \\ \frac{1}{|\hat{\beta}_{Ridge}|} ; n < p \end{cases}$ (9)

วิธีลาสโซ่แบบปรับปรุงมีคุณสมบัติที่ทำให้ได้ตัวประมาณสัมประสิทธิ์การถดถอยที่ประกอบด้วยค่าศูนย์จำนวนมาก ดังนั้นจึงมีคุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบเช่นเดียวกับวิธีลาสโซ่ และการเพิ่มเงื่อนไขการให้ค่าน้ำหนักแก่พารามิเตอร์ที่แตกต่างกันยังช่วยแก้ปัญหาความไม่คงเส้นคงวาที่ประสบในวิธีลาสโซ่ได้ส่งผลให้วิธีลาสโซ่แบบปรับปรุงมีคุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบเสมือนทราบตัวแบบที่แท้จริงหรือเรียกว่าคุณสมบัติออราเคิล (Oracle Property)

**2.2.3 วิธี Ridge Regression**

วิธีริดจ์ (Ridge Regression) เป็นการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษแบบ L2 – Norm สามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 10

$$\hat{\beta} = \underset{\beta}{argmin} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \quad (10)$$

โดยตัวประมาณสัมประสิทธิ์การถดถอยบางตัวที่ได้จากวิธีริดจ์จะถูกบีบค่าให้เข้าใกล้ศูนย์แต่ไม่เท่ากับศูนย์ ดังนั้นวิธีริดจ์จึงเหมาะแก่การวิเคราะห์ข้อมูลที่มีสัมประสิทธิ์ขนาดเล็กแต่ไม่เท่ากับศูนย์จำนวนมาก นอกจากนี้วิธีริดจ์ยังช่วยแก้ปัญหาที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงได้ อย่างไรก็ตาม คุณสมบัติการคัดเลือกตัวแปรเข้าตัวแบบไม่ปรากฏในวิธีริดจ์ (Hoerl and Kennard, 1970)

**2.3 ตัวประมาณ Lasso + Partial Ridge (LPR)**

ตัวประมาณ $\hat{\beta}_{LPR}$ นำเสนอโดย Liu et al. (2020) เป็นตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากสองขั้นตอน โดยขั้นตอนที่หนึ่งใช้วิธีการประมาณสัมประสิทธิ์การถดถอยด้วยวิธีลาสโซ่เพื่อคัดเลือกตัวแปรอิสระ และขั้นตอนที่สองใช้ฟังก์ชันการลงโทษ L2 – Norm เพื่อกอบกู้ตัวแปรอิสระที่ไม่ได้ถูกเลือกจากวิธีลาสโซ่ เนื่องจากการใช้ฟังก์ชัน L2 – Norm จะทำให้ได้สัมประสิทธิ์ขนาดเล็กแต่ไม่เท่ากับศูนย์ซึ่งสามารถเขียนเป็นฟังก์ชันได้ดังสมการที่ 11

$$\hat{\beta}_{LPR} = \underset{\beta}{argmin} \left\{ \frac{1}{2n} \| Y - X\beta \|_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}} \beta_j^2 \right\} \quad (11)$$

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 วิธี Residual Bootstrap Lasso + Partial Ridge (rBLPR) และ วิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)**

การทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยในกรณีที่ข้อมูลมีมิติสูงเป็นประเด็นที่ท้าทาย เนื่องจากการลู่เข้าเชิงการแจกแจง (Asymptotic Distribution) ของตัวประมาณสัมประสิทธิ์การถดถอยที่นิยมใช้ เช่นตัวประมาณลาสโซ่มีความซับซ้อน (Chatterjee and Lahiri, 2011) อีกทั้งตัวสถิติทดสอบทีหรือเอฟที่ใช้ในวิธีกำลังสองน้อยที่สุดไม่สามารถนำมาใช้ได้ ดังนั้นวิธีที่นิยมคือวิธีบูตสแทร็ป (Bootstrap)

**2.4.1 วิธี Residual Bootstrap Lasso + Partial Ridge (rBLPR)**

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

กำหนดให้

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

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

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

2.4.1.1 คำนวณค่าของสัมประสิทธิ์ $\hat{\beta}_{Lasso+OLS}$

$$\hat{\beta}_{Lasso+OLS} = \underset{\beta:\, \beta_{\hat{S}^c}=0}{argmin}\left\{\frac{1}{2n}\|Y - X\beta\|_2^2\right\};$$

เมื่อ $\beta_{\hat{S}^c} = \{\beta_j : j \notin \hat{S}\}$

2.4.1.2 คำนวณเวกเตอร์ของค่าพยากรณ์ $\hat{y} = X\hat{\beta}_{Lasso+OLS}$

2.4.1.3 คำนวณเวกเตอร์ส่วนเหลือ $\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta}_{Lasso+OLS}$

2.4.1.4 คำนวณเซตของ Centered residual $\{\hat{\epsilon}_i - \bar{\hat{\epsilon}}, i = 1,\dots,n\}$ เมื่อ $\bar{\hat{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i$

2.4.1.5 ทำการสุ่ม Centered residual แบบใส่คืน $\epsilon^* = (\epsilon_1^*,\dots,\epsilon_n^*)^T$

2.4.1.6 ทำการสร้างตัวอย่างบูตสแทร็ป

$$y^*_{rboot} = X\hat{\beta}_{Lasso+OLS} + \epsilon^*$$

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

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

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

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

2.4.1.9 ทำซ้ำในขั้นตอนที่ 5 – 8 ไป B รอบจนได้ $\hat{\beta}^*_{rBLPR}{}^{(1)}, \hat{\beta}^*_{rBLPR}{}^{(2)}, \dots, \hat{\beta}^*_{rBLPR}{}^{(B)}$ โดยที่ $\hat{\beta}^*_{rBLPR}{}^{(B)}$ คือสัมประสิทธิ์การถดถอยที่ได้จากการบูตสแทร็ปแบบสุ่มส่วนเหลือตัวประมาณลาสโซ่ + พาร์เชียลริดจ์ในครั้งที่ B

2.4.1.10 สร้างช่วงความเชื่อมั่นที่ $(1-\alpha)\%$ สำหรับสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ดังนี้ $[L_j, U_j]$; เมื่อ $L_j = (\hat{\beta}_{LPR})_j + (\hat{\beta}_{Lasso+OLS})_j - (\hat{\beta}^*_{rBLPR})_{j,1-\alpha/2}$ และ $U_j = (\hat{\beta}_{LPR})_j + (\hat{\beta}_{Lasso+OLS})_j - (\hat{\beta}^*_{rBLPR})_{j,\alpha/2}$ โดยที่ $(\hat{\beta}^*_{rBLPR})_{j,1-\alpha/2}$ และ $(\hat{\beta}^*_{rBLPR})_{j,\alpha/2}$ คือเปอร์เซ็นไทล์ที่ $1-\alpha/2$ และ $\alpha/2$ ของ $\hat{\beta}^*_{rBLPR}{}^{(1)},\dots,\hat{\beta}^*_{rBLPR}{}^{(B)}$ ตามลำดับ

**2.4.2 วิธี Paired Bootstrap Lasso + Partial Ridge (pBLPR)**

วิธีบูตสแทร็ป pBLPR เป็นการสุ่มข้อมูลตัวอย่างแบบใส่คืนโดยจะทำการสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระเพื่อสร้างข้อมูลตัวอย่างบูตส

แทร็ปจากนั้นนำชุดข้อมูลตัวอย่างบูตสแทร็ปไปใช้คำนวณหาค่าสัมประสิทธิ์การถดถอย เพื่อสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอย $\beta_j$ เมื่อ $j = 1,2,\dots,p$ โดยใช้สัญลักษณ์ $\hat{\beta}^*_{pBLPR}$ แทนตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธีบูตสแทร็ป pBLPR ซึ่งวิธี pBLPR มีขั้นตอนดังนี้

กำหนดให้

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

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

2.4.2.1 สุ่มข้อมูลตัวอย่างแบบใส่คืนเพื่อสร้างข้อมูลตัวอย่างบูตสแทร็ป $\left(x^*_{pboot}, y^*_{pboot}\right) = \{(x_i^*, y_i^*), i = 1,2,\dots,n\}$ เมื่อ $y^*_{pboot} = (y_1^*, y_2^*, \dots, y_n^*)^T$ และ $x^*_{pboot} = (x_1^*, x_2^*, \dots, x_n^*)^T$

2.4.2.2 คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยใช้วิธีลาสโซ่จากชุดข้อมูล $\left(x^*_{pboot}, y^*_{pboot}\right)$

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

2.4.2.3 คำนวณค่าของตัวประมาณสัมประสิทธิ์โดยมีการเพิ่มฟังก์ชันการลงโทษแบบ L2 – Penaltyสำหรับตัวแปรอิสระที่ไม่ได้ถูกคัดเลือกและใช้ชุดข้อมูล $\left(x^*_{pboot}, y^*_{pboot}\right)$

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

2.4.2.4 ทำซ้ำในขั้นตอนที่ 1 – 3 ไป B รอบจนได้ $\hat{\beta}^*_{pBLPR}{}^{(1)}, \hat{\beta}^*_{pBLPR}{}^{(2)}, \dots, \hat{\beta}^*_{pBLPR}{}^{(B)}$ โดยที่ $\hat{\beta}^*_{pBLPR}{}^{(B)}$ คือสัมประสิทธิ์การถดถอยที่ได้จากการบูตสแทร็ปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระตัวประมาณลาสโซ่ + พาร์เชียลริดจ์ในครั้งที่ B

2.4.2.5 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 $1-\alpha/2$ และ $\alpha/2$ ของ $\hat{\beta}^*_{pBLPR}{}^{(1)}, \dots, \hat{\beta}^*_{pBLPR}{}^{(B)}$ ตามลำดับ

**2.5 วิธี Residual Bootstrap Adaptive Lasso + Partial Ridge (rBALPR) และ วิธี Paired Bootstrap Adaptive Lasso + Partial Ridge (pBALPR)**

จากคุณสมบัติออราเคิลของวิธี Adaptive Lasso ที่เหนือกว่าวิธี Lasso ในงานวิจัยนี้ จึงนำเสนอวิธี rBALPR และ pBALPR ซึ่งจะมีขั้นตอนเหมือนกับวิธี rBLPR และ pBLPR ตามลำดับ แต่ในขั้นตอนต่าง ๆ ที่มีการใช้วิธี Lasso จะถูกแทนที่ด้วยวิธี Adaptive Lasso ทั้งหมด

## 3. ผลการวิจัยและวิจารณ์ผล

จำลองข้อมูลแบบตัดขวาง โดยกำหนดให้ขนาดตัวอย่าง (n) เท่ากับ 200 และจำนวนตัวแปรอิสระ (p) เท่ากับ 500 และทำการศึกษาทั้งหมด 8 กรณี โดยทุกกรณีจำลองข้อมูลจากตัวแบบเชิงเส้นดังสมการ $y_i = x_i^T\beta + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$, $i = 1,2\dots,n$ และศึกษาภายใต้ค่าความแปรปรวนของค่าความคลาดเคลื่อน $\sigma^2$ โดยที่กำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) เท่ากับ 10 ทั้งนี้การจำลองข้อมูลอ้างอิงมาจาก Liu et al. (2020)

**กรณีที่ 1 และ 2 :** ใช้เมทริกซ์ความแปรปรวนร่วมแบบโทพลิทซ์ (Toeplitz) และสัมประสิทธิ์การถดถอยแบบบางเบาอย่างรุนแรง (Hard Sparsity) ซึ่งเขียนได้ดังนี้ $X_i \sim N(0, \sum)$ เมื่อ $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.5$ และ $\rho = 0.9$ ตามลำดับ และ $\beta_j = \begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots,10 \\ 0; \text{อื่นๆ} \end{cases}$

**กรณีที่ 3 และ 4 :** ใช้เมทริกซ์ความแปรปรวนร่วมแบบโทพลิทซ์ (Toeplitz) และสัมประสิทธิ์การถดถอยแบบบางเบาอย่างอ่อน (Weak Sparsity) ซึ่งเขียนได้ดังนี้ $X_i \sim N(0, \sum)$ เมื่อ $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.5$ และ $\rho = 0.9$ ตามลำดับ และ $\beta_j = \begin{cases} N(1, 0.001); j = 1, 2, \ldots, 10 \\ \beta_j = \frac{1}{(j+3)^2}; \ j = 1,2, \ldots, 490 \end{cases}$

**กรณีที่ 5 และ 6 :** ใช้เมทริกซ์ความแปรปรวนร่วมแบบเท่าเทียม (Equal Correlation) และสัมประสิทธิ์การถดถอยแบบบางเบาอย่างรุนแรง (Hard Sparsity) ซึ่งเขียนได้ดังนี้ $X_i \sim N(0, \sum)$ เมื่อ $\sum_{ij} = \begin{cases} \rho; i \neq j \\ 1 \ ; \text{อื่นๆ} \end{cases}$ ที่ $\rho = 0.5$ และ $\rho = 0.9$ ตามลำดับ และ $\beta_j = \begin{cases} U\left[\frac{1}{3}, 1\right]; j = 1,2 \ldots,10 \\ 0; \text{อื่นๆ} \end{cases}$

**กรณีที่ 7 และ 8 :** ใช้เมทริกซ์ความแปรปรวนร่วมแบบเท่าเทียม (Equal Correlation) และสัมประสิทธิ์การถดถอยแบบบางเบาอย่างอ่อน (Weak Sparsity) ซึ่งเขียนได้ดังนี้ $X_i \sim N\ (0, \sum)$ เมื่อ $\sum_{ij} = \begin{cases} \rho; i \neq j \\ 1 \ ; \text{อื่นๆ} \end{cases}$ ที่ $\rho = 0.5$ และ $\rho = 0.9$ ตามลำดับ และ $\beta_j$

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

การจำลองข้อมูล $X$ และสัมประสิทธิ์การถดถอย $\beta$ ในแต่ละกรณีจะกระทำเพียงครั้งเดียวเท่านั้น แต่จะทำการจำลอง $Y = (y_1, y_2, .., y_n)^T$ ทั้งหมด 50 รอบ โดยมาจากการสุ่มส่วนเหลือ (error terms) และทำการเปรียบเทียบประสิทธิภาพของการบูตสแทร็ปทั้ง 4 วิธี ได้แก่ วิธีบูตสแทร็ปแบบสุ่มส่วนเหลือลาสโซ่แบบปรับปรุง + พาร์เซียลริดจ์ (rBALPR) วิธีบูตสแทร็ปแบบสุ่มส่วนเหลือลาสโซ่ + พาร์เซียลริดจ์ (rBLPR) วิธีบูตสแทร็ปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระลาสโซ่แบบปรับปรุง + พาร์เซียลริดจ์ (pBALPR) และวิธีบูตสแทร็ปแบบสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระลาสโซ่ + พาร์เซียลริดจ์ (pBLPR) โดยใช้เกณฑ์วัดประสิทธิภาพดังต่อไปนี้

- ค่าเฉลี่ยของความกว้างของช่วงความเชื่อมั่น (Mean Interval lengths)

$$CI_j = \frac{\sum_{i=1}^{B} (U_j - L_j)}{B}$$

เมื่อ $U_j$ และ $L_j$ คือขอบเขตบนและขอบเขตล่างของช่วงความเชื่อมั่นสำหรับแต่ละ $\beta_j$ ตามลำดับและ $B$ เป็นจำนวนครั้งที่ทำการสร้างช่วงความเชื่อมั่น

- ความน่าจะเป็นคุ้มรวม (Coverage Probabilities)

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

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

**Table 1** Mean and standard deviation of confidence interval lengths obtained from each bootstrap method for 8 cases of simulation studies

| Case | rBALPR | | rBLPR | | pBALPR | | pBLPR | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation |
| Case 1 | **0.004** | 0.001 | 0.020 | 0.002 | 0.014 | 0.001 | 0.037 | 0.003 |
| Case 2 | **0.009** | 0.001 | 0.012 | 0.002 | 0.017 | 0.003 | 0.046 | 0.005 |
| Case 3 | **0.006** | 0.002 | 0.031 | 0.003 | 0.021 | 0.002 | 0.053 | 0.004 |
| Case 4 | **0.014** | 0.001 | 0.018 | 0.003 | 0.024 | 0.003 | 0.068 | 0.008 |
| Case 5 | **0.018** | 0.003 | 0.036 | 0.003 | 0.024 | 0.002 | 0.057 | 0.006 |
| Case 6 | 0.091 | 0.013 | 0.132 | 0.020 | **0.074** | 0.007 | 0.098 | 0.011 |
| Case 7 | **0.024** | 0.006 | 0.053 | 0.004 | 0.038 | 0.005 | 0.084 | 0.011 |
| Case 8 | 0.145 | 0.014 | 0.211 | 0.023 | **0.117** | 0.009 | 0.159 | 0.017 |

**Table 2** Mean and standard deviation of coverage probabilities obtained from each bootstrap method for 8 cases of simulation studies.

| Case | rBALPR | | rBLPR | | pBALPR | | pBLPR | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation |
| Case 1 | 0.933 | 0.03 | 0.940 | 0.03 | 0.942 | 0.03 | **0.999** | 0.01 |
| Case 2 | 0.939 | 0.04 | 0.939 | 0.03 | 0.954 | 0.03 | **0.999** | 0.01 |
| Case 3 | 0.808 | 0.22 | 0.830 | 0.28 | **0.940** | 0.08 | 0.881 | 0.21 |
| Case 4 | 0.813 | 0.31 | 0.772 | 0.35 | **0.912** | 0.11 | 0.864 | 0.26 |
| Case 5 | 0.919 | 0.07 | 0.934 | 0.05 | 0.789 | 0.11 | **0.973** | 0.05 |
| Case 6 | 0.933 | 0.06 | **0.940** | 0.05 | 0.647 | 0.19 | 0.558 | 0.21 |
| Case 7 | 0.737 | 0.35 | **0.745** | 0.32 | 0.591 | 0.26 | 0.710 | 0.25 |
| Case 8 | 0.623 | 0.37 | **0.806** | 0.18 | 0.398 | 0.22 | 0.434 | 0.21 |

**3.1 ผลจากการจำลองข้อมูล**

จากTable 1 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยที่ได้จากการบูตสแทร็ปจำนวน 50 รอบซึ่งพบว่าเมื่อจำลองข้อมูลในรูปแบบกรณีที่ 1, 2, 3, 4, 5, และ 7 วิธี rBALPR มีประสิทธิภาพสูงที่สุดในการให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุด นอกจากนี้ส่วนเบี่ยงเบนมาตรฐานของความกว้างของช่วงความเชื่อมั่นที่ได้จากวิธี

rBALPR มีค่าต่ำที่สุดในกรณีที่ 1 – 4 ในขณะที่เมื่อจำลองข้อมูลในรูปแบบกรณีที่ 6 และ 8 พบว่าวิธี pBALPR ให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุดและมีค่าส่วนเบี่ยงเบนมาตรฐานต่ำที่สุดเช่นกัน นอกจากนี้สังเกตได้ว่าการบูตสแทร็ปตัวประมาณ LPR ไม่ว่าจะด้วยวิธีบูตสแทร็ปแบบสุ่มส่วนเหลือหรือบูตสแทร็ปแบบสุ่มตัวแปรตามและตัวแปรอิสระนั้นให้ความกว้างของช่วงความเชื่อมั่นค่อนข้างสูงในเกือบทุกกรณีของข้อมูลจำลอง

จากTable 2 แสดงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของความน่าจะเป็นคุ้มรวมที่ได้จากการบูตสแทร็ป จำนวน 50 รอบ ซึ่งพบว่าไม่มีวิธีการบูตสแทร็ปตัวประมาณแบบใดที่มีความโดดเด่นในด้านการให้ความน่าจะเป็นสูงสุดสำหรับทุกกรณีของข้อมูลจำลอง โดยวิธีการบูตสแทร็ปแต่ละวิธีจะเหมาะสมกับข้อมูลจำลองลักษณะที่แตกต่างกันออกไป กล่าวคือเมื่อจำลองข้อมูลในรูปแบบกรณีที่ 1, 2 และ 5 วิธี pBLPR ให้ความน่าจะเป็นคุ้มรวมสูงที่สุด ในขณะที่ถ้าจำลองข้อมูลตามกรณีที่ 3 และ 4 นั้นวิธี pBALPR ให้ความน่าจะเป็นคุ้มรวมสูงที่สุด และเมื่อจำลองข้อมูลในรูปแบบ 6, 7 และ 8 วิธี rBLPR ให้ความน่าจะเป็นคุ้มรวมสูงที่สุด นอกจากนี้สังเกตได้ว่าเมื่อข้อมูลจำลองขึ้นในรูปแบบที่ 5 – 8 นั้นวิธี pBALPR ให้ความน่าจะเป็นคุ้มรวมที่ค่อนข้างต่ำกว่าวิธีอื่นๆ

**4.2 การปรับใช้กับข้อมูลจริง: การวิเคราะห์ข้อมูลไมโครอาร์เรย์ในโรคมะเร็งลำไส้ใหญ่**

ข้อมูลงานวิจัยของ Alon et al. (1999) ซึ่งเป็นข้อมูล Oligonucleotide Microarray ที่ได้จากการสกัดการแสดงออกของยีนส์ (Gene Expression) บริเวณเนื้อเยื่อลำไส้ใหญ่จำนวน 62 ตัวอย่าง ประกอบด้วยเนื้อเยื่อมะเร็ง 40 ตัวอย่าง และเนื้อเยื่อปกติ 22 ตัวอย่าง

ผู้วิจัยได้ใช้การแสดงออกของยีนส์จำนวน 2,000 ตำแหน่งเป็นตัวแปรอิสระ ซึ่งสามารถเขียนเป็นเมทริกซ์ได้ดังนี้

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,2000} \\ \vdots & \ddots & \vdots \\ x_{62,1} & \cdots & x_{62,2000} \end{bmatrix}_{62\times 2000}$$

และจำลองตัวแปรตาม $Y = (y_1, y_2, .., y_{62})^T$ จากตัวแบบเชิงเส้น $y_i = x_i^T\beta + \epsilon_i$ , $\epsilon_i \sim N(0, \sigma^2)$ โดยที่กำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวนเท่ากับ 10 สำหรับค่าความแปรปรวนของค่าความคลาดเคลื่อน $(\sigma^2)$ และใช้สัมประสิทธิ์การถดถอยแบบบางเบาอย่างอ่อน (Weak Sparsity) คือ

$$\beta_j = \begin{cases} N(1, 0.001); j = 1, 2, \ldots, 10 \\ \beta_j = \frac{1}{(j+3)^2};\ j = 1, 2, \ldots, 1990 \end{cases}$$

ดังนั้นข้อมูลที่นำมาวิเคราะห์จึงเป็นกึ่งข้อมูลจริง

ในการวิเคราะห์ข้อมูลชุดนี้ ผู้วิจัยทำการจำลองตัวแปรตามจำนวน 50 รอบ โดยมาจากการสุ่มส่วนเหลือ (error terms) แล้วจึงทำการบูตสแทร็ปทั้ง 4 วิธี ได้แก่ rBALPR, rBLPR, pBALPR, และ pBLPR เพื่อสร้างช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยและเปรียบเทียบประสิทธิภาพในแง่ของค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นและความน่าจะเป็นคุ้มรวม

**Table 3** Mean and standard deviation of confidence interval lengths obtained from each bootstrap method for colon cancer microarray data set.

| rBALPR | | rBLPR | | pBALPR | | pBLPR | |
|---|---|---|---|---|---|---|---|
| Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation |
| **0.022** | 0.006 | 0.042 | 0.006 | 0.050 | 0.003 | 0.067 | 0.004 |

**Table 4** Mean and standard deviation of coverage probabilities obtained from each bootstrap method for colon cancer microarray data set.

| rBALPR | | rBLPR | | pBALPR | | pBLPR | |
|---|---|---|---|---|---|---|---|
| Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation | Mean | Standard Deviation |
| 0.814 | 0.20 | 0.744 | 0.18 | **0.942** | 0.21 | 0.896 | 0.27 |

จากTable 3 แสดงค่าเฉลี่ยของความกว้างของช่วงความเชื่อมั่นที่ได้จากการบูตสแทร็ปจำนวน 50 รอบ สำหรับข้อมูลไมโครอาร์เรย์ในโรคมะเร็งลำไส้ใหญ่ พบว่าเมื่อนำวิธีบูตสแทร็ปแบบต่างๆมาปรับใช้กับข้อมูลดังกล่าว การบูตสแทร็ปแบบ rBALPR ให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุดซึ่งเท่ากับ 0.022

จากTable 4 แสดงค่าเฉลี่ยของความน่าจะเป็นที่ช่วงความเชื่อมั่นครอบคลุมค่าของสัมประสิทธิ์การถดถอยที่ได้จากการบูตสแทร็ปจำนวน 50 รอบ สำหรับข้อมูลไมโครอาร์เรย์ในโรคมะเร็งลำไส้ใหญ่ พบว่าเมื่อนำวิธีบูตสแทร็ปแบบต่างๆมาปรับใช้กับข้อมูลดังกล่าว การบูตสแทร็ปแบบ pBALPR ให้ความน่าจะเป็นคุ้มรวมสูงที่สุดซึ่งเท่ากับ 0.942 และมีส่วนเบี่ยงเบนมาตรฐานใกล้เคียงกับวิธีอื่นๆ

## 4. สรุป

งานวิจัยชิ้นนี้นำเสนอวิธีบูตสแทร็ปตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่แบบปรับปรุง + พาร์เซียลริดจ์และเปรียบเทียบกับวิธีบูตสแทร็ปตัวประมาณลาสโซ่ + พาร์เซียลริดจ์ โดยผู้วิจัยได้ทำการทดลองบูตสแทร็ป 2 วิธีคือ วิธีสุ่มส่วนเหลือและวิธีสุ่มตัวแปรตามพร้อมกับตัวแปรอิสระ จากนั้นจึงทำการเปรียบเทียบและวิเคราะห์ผลลัพธ์โดยใช้ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นและความน่าจะเป็นคุ้มรวมเป็นเกณฑ์การวัดประสิทธิภาพเพื่อหาวิธีที่มีประสิทธิภาพสูงสุดในการทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูง โดยผู้วิจัยได้ทำการศึกษาข้อมูลจำลองทั้งหมด 8 กรณี และได้นำวิธีบูตสแทร็ปแบบต่างๆมาปรับใช้กับข้อมูลไมโครอาร์เรย์ของโรคมะเร็งลำไส้ใหญ่

การวิจัยพบว่าเมื่อใช้ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยเป็นเกณฑ์การวัดประสิทธิภาพ วิธีบูตสแทร็ปแบบ rBALPR มีประสิทธิภาพสูงสุด โดยวิธีดังกล่าวให้ค่าความกว้างของช่วงความเชื่อมั่นโดยเฉลี่ยน้อยที่สุดถึง 6 กรณีจากทั้งหมด 8 กรณี อีกทั้งวิธี rBALPR ให้ค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นน้อยกว่า

วิธีอื่นอย่างมีนัยสำคัญเมื่อข้อมูลเกิดปัญหาค่าความแปรปรวนของค่าคลาดเคลื่อนไม่คงที่โดยเป็นเมทริกซ์ที่มีค่าสหสัมพันธ์แบบโทพลิทซ์ (Toeplitz) ซึ่งได้แก่ กรณีที่ 1 – 4 และเมื่อใช้ความน่าจะเป็นคุ้มรวมเป็นเกณฑ์การวัดประสิทธิภาพ พบว่าไม่ปรากฏวิธีการบูตสแทร็ปวิธีใดวิธีหนึ่งที่มีประสิทธิภาพสูงสุดสำหรับทุกกรณีของข้อมูลจำลอง อนึ่ง เป็นที่น่าสังเกตว่าวิธี rBALPR เป็นวิธีที่โดดเด่นในแง่การให้ช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยสั้นที่สุด แต่วิธีดังกล่าวไม่ได้ให้ความน่าจะเป็นคุ้มรวมสูงสุดในข้อมูลจำลองกรณีใดเลย อย่างไรก็ตาม หากพิจารณาค่าเฉลี่ยความกว้างของช่วงความเชื่อมั่นร่วมกับความน่าจะเป็นคุ้มรวมแล้วนั้น วิธี rBALPR ถือว่าเป็นวิธีที่มีประสิทธิภาพเพราะสามารถให้ความน่าจะเป็นคุ้มรวมใกล้เคียงกับวิธีบูตสแทร็ปแบบอื่นๆ

ผลการนำวิธีบูตสแทร็ปแบบต่างๆไปปรับใช้กับข้อมูลไมโครอาร์เรย์ในโรคมะเร็งลำไส้ใหญ่ของ Alon และคณะ พบว่าได้ผลลัพธ์ค่อนข้างสอดคล้องกับการวิเคราะห์ข้อมูลจำลองคือวิธี rBALPR มีประสิทธิภาพสูงสุดในแง่การให้ความกว้างของช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยโดยเฉลี่ยน้อยที่สุด และวิธี pBALPR เป็นวิธีที่ให้ความน่าจะเป็นคุ้มรวมโดยเฉลี่ยสูงสุด

อนึ่ง ในปัจจุบันความก้าวหน้าทางเทคโนโลยีมีมากขึ้นทำให้ประสิทธิภาพการเก็บข้อมูลเพิ่มสูงขึ้น ดังนั้นข้อมูลที่มีมิติสูงจึงแพร่หลายมากขึ้น หากแต่การอนุมานเชิงสถิติในกรณีที่ข้อมูลมีมิติสูงเพื่อทำความเข้าใจความสัมพันธ์ระหว่างตัวแปรยังคงเป็นประเด็นที่ท้าทายและซับซ้อน ทั้งนี้วิธีที่นิยมใช้เพื่อทดสอบสมมติฐานทางสถิติของสัมประสิทธิ์การถดถอยคือวิธีบูตสแทร็ป อย่างไรก็ตาม การบูตสแทร็ปนั้นก็มีหลากหลายรูปแบบซึ่งหากพิจารณาด้วยเกณฑ์ความกว้างของช่วงความเชื่อมั่นและความน่าจะเป็นคุ้มรวมที่ได้จากการศึกษาในงานวิจัยชิ้นนี้แล้วนั้น วิธีบูตสแทร็ปส่วนเหลือตัวประมาณลาสโซ่แบบปรับปรุง + พาร์เชียลริดจ์ (rBALPR) ก็เป็นอีกวิธีการหนึ่งที่น่าสนใจในการทดสอบสมมติฐานทางสถิติในกรณีที่ข้อมูลมีมิติสูง

## 5. References


Alon, U., Barkai, N., Notterman, DA., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl Acad. Sci. USA. 96(12): 6745 – 6750.

Chatterjee, A. and Lahiri, S. N., 2011, Bootstrapping lasso estimators. J. Am. Stat. Assoc. 106: 608 – 625.

Hoerl, A.E. and Kennard, R.W., 1970, Ridge regression: Biased estimation for nonorthogonal problems. J. Am. Stat. Assoc. 12: 55 – 67.

James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013, An introduction to statistical learning: with applications in R, 2nd Ed., Springer, 607 p.

Knight, K. and Fu, W.J., 2000, Asymptotics for lasso-type estimators. Ann. Stat. 28: 1356 – 1378.

Liu, H. and Yu, B., 2013, Asymptotic properties of lasso + mLS and lasso + Ridge in sparse high- dimensional linear regression. Electron. J. Statist. 7: 3124 – 3169.

Liu, H., Xu, X. and Li, JJ., 2020, A Bootstrap Lasso + Partial Ridge Method to Construct Confidence Intervals for Parameters in High – dimensional Sparse Linear Models. Statistica Sinica 30(3): 1333 – 1355.

Pungpapong, V., 2015, A brief review on high - dimensional linear regression, Thai Sci. Technol. J. 23(2): 212-223. (in Thai)

Tibshirani, R., 1996, Regression shrinkage and selection via the lasso. J. Royal Stat. Soc. B. 58: 267 – 288.

Tibshirani, RJ., 2013, The lasso problem and uniqueness. Electron. J. Statist. 7: 1456 – 1490.

Wasserman, L. and Roeder, K., 2009, Weak Signal Identification and Inference in Penalized Model Selection. Ann. Stat. 45: 1214 – 1253.

Zou, H., 2006, The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101: 1418 – 1429.