การเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso + MLE และวิธี Lasso + Partial Ridge

น.ส.ณิชากร ไทยวงษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Efficiency Comparison on Method to Construct Confidence Intervals for Parameters in High-dimensional Logistic Regression Models between A Bootstrap Lasso + MLE and A Bootstrap Lasso + Partial Ridge

Miss Nichagorn Thaiwong

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	การเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso + MLE และวิธี Lasso + Partial Ridge
โดย	น.ส.ณิชากร ไทยวงษ์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.นัจชลี ศรีมณีกาญจน์)

ณิชากร ไทยวงษ์ : การเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso + MLE และวิธี Lasso + Partial Ridge. (Efficiency Comparison on Method to Construct Confidence Intervalsfor Parameters in High-dimensional Logistic Regression Models betweenA Bootstrap Lasso + MLE and A Bootstrap Lasso + Partial Ridge) อ.ที่ปรึกษาหลัก : ผศ. ดร.วิฐรา พึ่งพาพงศ์

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ซึ่งในการศึกษานี้จะจำลองข้อมูลทั้งหมด 8 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว

จากการศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่า วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นมากที่สุด รองลงมาคือ วิธี Paired Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+MLE ตามลำดับ และวิธีที่มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นน้อยที่สุด ก็คือ วิธี Parametric Bootstrap Lasso+MLE ดังนั้นจึงสรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+Partial Ridge มีประสิทธิภาพมากกว่าวิธี Lasso+MLE

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380116126 : MAJOR STATISTICS

KEYWORD: Binary Logistic Regression, Ridge Regression, Lasso Regression, Bootstrap Sampling, Confidence Intervals Estimation

Nichagorn Thaiwong : Efficiency Comparison on Method to Construct Confidence Intervalsfor Parameters in High-dimensional Logistic Regression Models betweenA Bootstrap Lasso + MLE and A Bootstrap Lasso + Partial Ridge. Advisor: Asst. Prof. VITARA PUNGPAPONG, Ph.D.

This research is aimed to compare the efficiency of methods to construct confidence intervals for parameters in high-dimensional logistic regression models between a bootstrap Lasso + MLE and a bootstrap Lasso + Partial Ridge. In this study, there are 8 simulation data sets. Also, the confidence intervals are constructed by 4 methods: (i) Parametric Bootstrap Lasso+MLE (ii) Parametric Bootstrap Lasso+Partial Ridge (iii) Paired Bootstrap Lasso+MLE, and (iv) Paired Bootstrap Lasso+Partial Ridge. The performance of all 4 methods is compared in terms of average width value, coverage probability value, precision value, and recall value.

From our simulation studies, they show that a Parametric Bootstrap Lasso+Partial Ridge is the best performance method to construct confidence intervals for parameters in high-dimensional logistic regression models, followed by a Paired Bootstrap Lasso+Partial Ridge method and a Paired Bootstrap Lasso+MLE method respectively, and the worse performance method is a Parametric Bootstrap Lasso+MLE method. So, we can conclude that a bootstrap Lasso + Partial Ridge method has the most effective more than that a bootstrap Lasso + MLE method.

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

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

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความเชื่อเหลือและเอาใจใส่จากผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูง ที่กรุณาให้คำปรึกษา อบรบสั่งสอน และให้ข้อคิดเห็นต่าง ๆ ตลอดจนความช่วยเหลือให้คำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน จนกระทั่งวิทยานิพนธ์เสร็ตสมบูรณ์ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณท่านผู้ช่วยศาสตราจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี ประธานกรรมการสอบวิทยานิพนธ์ และอาจารย์ผู้ช่วยศาสตราจารย์ ดร.ณัตติฤดี เจริญรักษ์ กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสองท่านได้สละเวลาเพื่อเป็นกรรมการสอบครั้งนี้ ตลอดจนช่วยตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ได้ให้โอกาสทางการศึกษา และอบรมสั่งสอนให้ความรู้ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่ให้กำลังใจและความห่วงใย ส่งเสริมและสนับสนุนมาโดยตลอด และขอขอบคุณเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือและให้คำแนะนำแก่ผู้วิจัยตลอดมา

ณิชากร ไทยวงษ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นการวิเคราะห์ที่ถูกนำมาใช้อย่างแพร่หลายและมีวัตถุประสงค์เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งการวิเคราะห์การถดถอยลอจิสติกที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่า การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) สำหรับการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกจะใช้วิธีภาวะน่าจะเป็นสูงสุดหรือ Maximum Likelihood Estimator (MLE) อันเป็นการคำนวณทวนซ้ำ (Iterative Algorithm) แต่มีข้อจำกัดว่าตัวแปรอิสระต้องไม่มีความสัมพันธ์กันเองสูง หรือไม่มีปัญหาเรื่องความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) และจะคำนวณได้ในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ (วิฐรา พึ่งพาพงศ์, 2015) กล่าวว่า ข้อมูลในปัจจุบันมีขนาดใหญ่และซับซ้อนมากขึ้น เนื่องจากความสามารถในการจัดเก็บข้อมูลที่มีความทันสมัย ทำให้เกิดข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง เรียกว่า ข้อมูลที่มีมิติสูง (High Dimensional Data) ซึ่งพบได้มากในข้อมูลด้านการแพทย์ วิทยาศาสตร์อวกาศและเทคโนโลยี โดยในการวิเคราะห์ข้อมูลที่มีมิติสูงจะเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ ทำให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดไม่มีประสิทธิภาพ ซึ่งมีอีกวิธีการหนึ่งที่นิยมใช้ในการวิเคราะห์ข้อมูลที่มีมิติสูง คือ วิธีการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Regression) อันเป็นการปรับตัวแบบภายใต้เงื่อนไขที่เรียกว่า ฟังก์ชันลงโทษ (Penalty Function)

การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษที่มีการใช้งานอย่างแพร่หลาย คือ การถดถอยแบบริดจ์ (Ridge Regression) และการถดถอยลาสโซ่ (Lasso regression) โดยในการคัดเลือกตัวแปรเข้าหรือออกจากตัวแบบสำหรับการถดถอยแบบริดจ์ สามารถใช้การทดสอบสมมติฐานสำหรับสัมประสิทธิ์การถดถอยแต่ละตัวได้ ในขณะที่วิธีการถดถอยแบบลาสโซ่เป็นวิธีที่ใช้อย่างแพร่หลายมากที่สุด เพราะสามารถประมาณค่าและคัดเลือกตัวแปรเข้าสู่ตัวแบบได้ในคราวเดียวกัน แต่จะทำให้ได้ค่าประมาณสัมประสิทธิ์การถดถอยส่วนใหญ่เท่ากับศูนย์ เรียกว่า ตัวประมาณมากเลขศูนย์ (Sparse Estimator) อย่างไรก็ตาม ตัวประมาณจากวิธีลาสโซ่ไม่สามารถหาการแจกแจงค่าตัวอย่าง (Sampling Distribution) ได้อย่างแน่ชัด ในการทดสอบสมมติฐานดังกล่าว จึงสามารถใช้การสุ่มตัวอย่างบูตสแตรป (Bootstrap Sampling) เพื่อสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอย ซึ่งหากช่วงความเชื่อมั่นไม่ครอบคลุมค่าศูนย์ จะสามารถแปลผลได้ว่าสัมประสิทธิ์การถดถอยนั้นมีค่าแตกต่างจากศูนย์แบบมีนัยสำคัญทางสถิติ

จากการศึกษางานวิจัย (Liu & Yu, 2013) ได้นำเสนอการวิเคราะห์ข้อมูลที่มีมิติสูงด้วยการสร้างความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยการสุ่มตัวอย่างบูตสแตรป และการประมาณค่าสัมประสิทธิ์การถดถอย 2 ขั้นตอน โดยเริ่มจากการสร้างตัวแบบการถดถอยเชิงเส้น (Linear Regression) ที่ปรับด้วยฟังก์ชันการลงโทษแบบลาสโซ่ และประมาณค่าสัมประสิทธิ์การถดถอยอีกครั้งด้วยวิธีกำลังสองน้อยที่สุด (Lasso+OLS) ซึ่งพบว่าสัมประสิทธิ์การถดถอยที่คำนวณได้จากวิธีนี้ประสบปัญหา Beta-min condition หรือก็คือ สัมประสิทธิ์การถดถอยมีค่าน้อยและเข้าใกล้ศูนย์จำนวนมาก ต่อมา (Dezeure, Bühlmann, Meier, & Meinshausen, 2015) ได้ศึกษาประสิทธิภาพของการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยวิธี Lasso+OLS กับข้อมูลจำลอง พบว่า วิธีนี้ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้น เนื่องจากสัมประสิทธิ์การถดถอยที่คำนวณได้มีค่าน้อยและเข้าใกล้ศูนย์จำนวนมาก ทำให้ความน่าจะเป็นครอบคลุม (Coverage Probability) ต่ำกว่า 50 เปอร์เซ็นต์

(Liu, Xu, & Li, 2020) ได้นำเสนอการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้น โดยใช้ตัวอย่างบูตสแตรป สำหรับการประมาณค่าสัมประสิทธ์การถดถอยที่ Liu et al. (2020) นำเสนอ คือ วิธี Lasso+Partial Ridge ซึ่งเป็นการประมาณ 2 ขั้นตอน เริ่มจากการสร้างตัวแบบการถดถอยเชิงเส้นที่ปรับด้วยฟังก์ชันการลงโทษแบบลาสโซ่ จากนั้นจึงใช้วิธีการถดถอยแบบริดจ์ในการหาสัมประสิทธิ์ในขั้นตอนที่ 2 ซึ่งในขั้นตอนนี้จะใช้ฟังก์ชันการลงโทษแบบริดจ์กับเฉพาะสัมประสิทธิ์ที่เท่ากับศูนย์จากวิธีลาสโซ่เท่านั้น สำหรับสัมประสิทธิ์ที่ไม่เท่ากับศูนย์จากวิธีลาสโซ่ ในขั้นตอนที่ 2 นี้จะไม่มีการปรับด้วยฟังก์ชันการลงโทษใด ๆ Liu et al. (2020) พบว่าช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้น ได้จากวิธี Lasso+Partial Ridge แคบกว่าของวิธี Lasso+OLS และยังได้ค่าความน่าจะเป็นครอบคลุมสูงกว่าวิธี Lasso+OLS อีกด้วย ดังนั้นการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยวิธี Lasso+Partial Ridge จึงเป็นเทคนิคที่น่าสนใจ เพราะสามารถใช้วิเคราะห์ข้อมูลที่มีมิติสูงและสามารถนำมาปรับใช้กับการวิเคราะห์การถดถอยเชิงเส้นได้ดี

แม้ว่าจะมีงานวิจัยได้นำเสนอการประมาณ 2 ขั้นตอนในการสร้างความเชื่อมั่นของสัมประสิทธิ์การถดถอยสำหรับข้อมูลที่มีมิติสูง อย่างไรก็ตาม งานวิจัยที่กล่าวมาทั้งหมดจะดำเนินการศึกษาเฉพาะกรณีของตัวแบบการถดถอยเชิงเส้นกรณีที่ตัวแปรตามมีการแจกแจงแบบปกติเท่านั้น ในงานวิจัยนี้จึงสนใจที่จะศึกษาการประมาณ 2 ขั้นตอนสำหรับตัวแบบการถดถอยลอจิสติกทวิภาค โดยเปรียบเทียบประสิทธิภาพการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกทวิภาคสำหรับข้อมูลที่มีมิติสูง โดยใช้การประมาณ 2 ขั้นตอนได้แก่ วิธี Lasso+MLE และวิธี Lasso+Partial Ridge

1.2 วัตถุประสงค์การวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอน ด้วยวิธี Lasso+MLE และวิธี Lasso+Partial Ridge

1.3 สมมติฐานการวิจัย

การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอน ด้วยวิธี Lasso+Partial Ridge จะให้ความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุดและมีค่าความน่าจะเป็นครอบคลุมสูง ค่าความแม่นยำและค่าความไวต่ำกว่าวิธี Lasso+MLE

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การวิเคราะห์การถดถอยลอจิสติกส์

(Menard, 2002) กล่าวว่า การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นการวิเคราะห์ที่มีวัตถุประสงค์เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยที่ตัวแปรตามจะเป็นตัวแปรจำแนกประเภท (Categorical Variable) และอาจแบ่งออกได้เป็นข้อมูลทวิภาค (Dichotomous Data) ซึ่งก็คือข้อมูล 2 กลุ่ม หรือมากกว่า 2 กลุ่ม สำหรับการวิเคราะห์การถดถอยลอจิสติกกรณีที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่าการวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression)

ในการศึกษาครั้งนี้เราสนใจการวิเคราะห์การถดถอยลอจิสติกทวิภาค ซึ่งจะสร้างตัวแบบตามสมการ

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{2.1}$$

โดยในการประมาณค่าสัมประสิทธิ์การถดถอยจะใช้วิธีภาวะน่าจะเป็นสูงสุด หรือ Maximum Likelihood Estimator (MLE) ซึ่งหาได้จากการทำให้ Log Likelihood function มีค่ามากที่สุด ดังสมการ

$$argmax(l(\beta)) = argmax(\sum_{i=1}^{n} y_i \log\frac{\pi_i}{1-\pi_i} + (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)) \tag{2.2}$$

เมื่อกำหนดให้ y_i คือ จำนวนผลสำเร็จของข้อมูลสังเกต i และ $i = 1,2,\ldots,n$

นอกจากการประมาณค่าสัมประสิทธิ์การถดถอยแล้ว เรามักสนใจที่จะตรวจสอบว่าตัวแปรอิสระใดบ้างที่มีความสัมพันธ์กับโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งสามารถเขียนได้อยู่ในรูปสมมติฐานดังนี้

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0 \tag{2.3}$$

เมื่อ $j = 1,2,\ldots,p$ ในการใช้ตัวประมาณ MLE ในการประมาณค่าสัมประสิทธิ์การถดถอย ทำให้ได้ว่าค่าประมาณสัมประสิทธิ์ของการถดถอยลอจิสติกแต่ละตัว มีการกระจายตัวของค่าตัวอย่างโดยประมาณแบบปกติ โดยที่ $\hat{\beta}_j \sim N(\beta_j, \hat{\sigma}_j^2)$ ดังนั้นจะได้ว่าตัวสถิติทดสอบสำหรับทดสอบของ Wald (Wald's Test Statistics) สมมติฐานตามสมการที่ (2.3) คือ

$$z = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim N(0,1) \tag{2.4}$$

และสามารถหาช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ $\hat{\beta}_j$ ได้ โดยมีขีดจำกัดล่างของช่วงความเชื่อมั่น คือ $\left(\hat{\beta}_j - Z_{1-\frac{\alpha}{2}}\hat{\sigma}_j\right)$ และขีดจำกัดบนของช่วงความเชื่อมั่น คือ $\left(\hat{\beta}_j + Z_{1-\frac{\alpha}{2}}\hat{\sigma}_j\right)$

2.2 การประมาณค่าสัมประสิทธิ์ด้วยการวิเคราะห์การถดถอยลอจิสติกส์ที่ปรับด้วยฟังก์ชันการลงโทษ

วิธีการหนึ่งที่ใช้กันอย่างแพร่หลายในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง คือ วิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Regression) ซึ่งจะช่วยลดค่าสัมประสิทธิ์ของตัวแปรที่ส่งผลต่อตัวแบบน้อยให้เป็นศูนย์ และคัดออกมา โดยปรับให้สมการ (2.5) มีค่าน้อยที่สุด

$$l_\lambda(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda P_\lambda(\beta) \quad (2.5)$$

เมื่อ $P_\lambda(\beta)$ คือ ฟังก์ชันการลงโทษ และ λ คือ พารามิเตอร์ที่มีการปรับค่าแล้ว (Tuning Parameter) ซึ่ง $\lambda \geq 0$

ต่อไปจะเสนอฟังก์ชันการลงโทษ 2 วิธีที่มีการใช้งานอย่างแพร่หลายได้แก่ การถดถอยแบบริดจ์ (Ridge Regression) และการถดถอยลาสโซ่ (Lasso Regression)

2.2.1 การถดถอยแบบริดจ์ (Ridge Regression)

(Hoerl & Kennard, 1970) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ซึ่งเป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง (Multicollinearity) หรือเกิดภาวะร่วมเชิงเส้น โดยวิธีการถดถอยแบบริดจ์จะประมาณค่าสัมประสิทธิ์การถดถอยจากการทำให้สมการที่ (2.6) มีค่าน้อยที่สุด

$$l_\lambda^R(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda_2 \sum_{j=1}^{p}\beta_j^2 \quad (2.6)$$

2.2.2 การถดถอยลาสโซ่ (Lasso Regression)

(Tibshirani, 1996) ได้เสนอวิธีการวิเคราะห์การถดถอยแบบลาสโซ่ (Least Absolute Shrinkage and Selection Operator Regression: Lasso Regression) โดยจะประมาณค่าสัมประสิทธิ์การถดถอยจากการทำให้สมการที่ (2.7) มีค่าน้อยที่สุด

$$l_\lambda^L(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda_1 \sum_{j=1}^{p}|\beta_j| \;\; ; \lambda > 0 \quad (2.7)$$

2.3 วิธีบูตสแตรปสำหรับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์

2.3.1 วิธี Parametric Bootstrap

(Efron & Tibshirani, 1994) กล่าวว่า วิธีบูตสแตรปเป็นวิธีการประมาณค่าโดยใช้การสุ่มตัวอย่างจากประชากรแบบใส่คืน (replacement) นั่นคือ มีโอกาสที่ตัวอย่างจะสุ่มได้ซ้ำกัน โดยที่แต่ละหน่วยตัวอย่างมีโอกาสในการถูกสุ่มเท่ากัน จากนั้นทำการสุ่มตัวอย่างด้วยจำนวนครั้งที่มากพอ เพื่อ

สร้างการแจกแจงของตัวสถิติตัวอย่างแล้วนำมาใช้ในการประมาณค่าพารามิเตอร์ที่สนใจ ทั้งนี้ การบูตสแตรปมีอยู่หลากหลายวิธี ซึ่งหนึ่งในวิธีที่น่าสนใจ คือ วิธี Parametric Bootstrap อันเป็นวิธีที่เหมาะสมสำหรับการประมาณค่าพารามิเตอร์ที่สนใจจากข้อมูลที่ทราบการแจกแจง โดยมีขั้นตอนดังต่อไปนี้

1) นำข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคตามสมการที่ (2.1) เพื่อประมาณ ค่า $\hat{\pi}$
2) สร้างตัวอย่างบูตสแตรป $\boldsymbol{Y}^*$ จากการประมาณค่า $Y^* \sim Bin(1, \hat{\pi})$ ทั้งหมด n ตัว โดยที่เวกเตอร์ $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \dots, Y_n^*)^T$
3) คำนวณค่าสัมประสิทธิ์การถดถอยลอจิสติกทวิภาคตามสมการ (2.2) โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)^T$
4) ทำตามขั้นตอน 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \dots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$
5) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\dots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \dots, \hat{\beta}_j^{(B)}\right)$

2.3.2 วิธี Paired Bootstrap

เรียกอีกชื่อว่าวิธี vector bootstrap เป็นวิธีการประมาณค่าโดยใช้การสุ่มตัวอย่างจากประชากรแบบใส่คืนคล้ายกับวิธี Parametric Bootstrap ซึ่งจับสุ่มเป็นคู่ และมีขั้นตอนดังต่อไปนี้

1) กำหนดให้ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ แล้วนำมาสุ่มตัวอย่างแบบใส่คืน จำนวน n ตัว จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ โดยที่ $(\boldsymbol{X}^*, \boldsymbol{Y}^*) = (X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$
2) นำตัวอย่างที่สุ่มมาจากข้อ 1) มาประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกทวิภาคตามสมการ (2.2) โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\right)^T$
3) ทำตามขั้นตอน 1) และ 2) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \dots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$
4) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\dots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \dots, \hat{\beta}_j^{(B)}\right)$

2.4 วิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์โดยใช้การประมาณสองขั้นตอน

2.4.1 วิธี Parametric Bootstrap Lasso+MLE

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำของบูตสแตรปเท่ากับ B ครั้งเป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ เมื่อ $j = 1,2,\ldots,p$ ขั้นตอน :

1) นำข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคด้วยวิธีลาสโซ่
2) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ; \hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ; \hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j
3) นำค่า $\widehat{\boldsymbol{\beta}}$ ที่ได้ในข้อ 2) ประมาณค่า $\hat{\pi}$
4) สร้างตัวอย่างบูตสแตรป $\boldsymbol{Y}^*$ จากการประมาณค่า $Y^* \sim Bin(1, \hat{\pi})$ ทั้งหมด n ตัว โดยที่เวกเตอร์ $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^T$
5) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (2.7) มีค่าน้อยที่สุด
6) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ; \hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ; \hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j ที่ได้ในขั้นตอนที่ 5) และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j
7) ทำตามขั้นตอนที่ 4), 5) และ 6) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$
8) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\ldots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \ldots, \hat{\beta}_j^{(B)}\right)$

2.4.2 วิธี Parametric Bootstrap Lasso+Partial Ridge

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำของบูตสแตรปเท่ากับ B ครั้งเป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ เมื่อ $j = 1,2,\ldots,p$ ขั้นตอน :

1) นำข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคด้วยวิธีลาสโซ่
2) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ; \hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ; \hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j
3) นำค่า $\widehat{\boldsymbol{\beta}}$ ที่ได้ในข้อ 2) ประมาณค่า $\hat{\pi}$
4) สร้างตัวอย่างบูตสแตรป $\boldsymbol{Y}^*$ จากการประมาณค่า $Y^* \sim Bin(1, \hat{\pi})$ ทั้งหมด n ตัว โดยที่เวกเตอร์ $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^T$
5) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (2.7) มีค่าน้อยที่สุด
6) คำนวณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ โดยการใช้ฟังก์ชันการลงโทษแบบริดจ์สำหรับสัมประสิทธิ์ตัวที่เท่ากับศูนย์จากวิธีการลาสโซ่เท่านั้น กล่าวคือ เมื่อให้ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และให้ $\hat{S}_{Lasso} = \left\{j: \hat{\beta}_{Lasso,j} \neq 0\right\}$ แล้วสัมประสิทธิ์ที่ได้จากวิธีการริดจ์ในขั้นตอนที่ 5) จะเขียนแทนด้วยเวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ โดยหาได้จาก

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{argmin} \left\{ \frac{1}{2n} \|Y^* - X^*\beta\|_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}_{Lasso}} \beta_j^2 \right\} \qquad (2.8)$$

7) ทำตามขั้นตอนที่ 4), 5) และ 6) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$
8) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\ldots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \ldots, \hat{\beta}_j^{(B)}\right)$

2.4.3 วิธี Paired Bootstrap Lasso+MLE

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำของบูตสแตรปเท่ากับ B ครั้งเป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ เมื่อ $j = 1,2,\ldots,p$ ขั้นตอน :

1) สุ่มตัวอย่างบูตสแตรปจำนวน n ตัว จากข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ แบบใส่คืน จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ โดยที่ $(\boldsymbol{X}^*, \boldsymbol{Y}^*) = (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
2) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (2.7) มีค่าน้อยที่สุด
3) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ; \hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ; \hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j
4) ทำตามขั้นตอนที่ 1), 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$
5) คำนวณขีดจำกัดล่าง $\left(l_j\right)$ และขีดจำกัดบน $\left(u_j\right)$ ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\ldots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \ldots, \hat{\beta}_j^{(B)}\right)$

2.4.4 วิธี Paired Bootstrap Lasso+Partial Ridge

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำของบูตสแตรปเท่ากับ B ครั้งเป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ เมื่อ $j = 1,2,\ldots,p$ ขั้นตอน :

1) สุ่มตัวอย่างบูตสแตรปจำนวน n ตัว จากข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ แบบใส่คืน จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ โดยที่ $(\boldsymbol{X}^*, \boldsymbol{Y}^*) = (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$
2) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (2.7) มีค่าน้อยที่สุด
3) คำนวณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ โดยการใช้ฟังก์ชันการลงโทษแบบริดจ์สำหรับสัมประสิทธิ์ตัวที่เท่ากับศูนย์จากวิธีการลาสโซ่เท่านั้น กล่าวคือ เมื่อให้ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และให้ $\hat{S}_{Lasso} = \{j: \hat{\beta}_{Lasso,j} \neq 0\}$ แล้วสัมประสิทธิ์ที่ได้จากวิธีการริดจ์ในขั้นตอนที่ 3) จะเขียนแทนด้วยเวกเตอร์ $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ โดยหาได้จาก

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{argmin}\left\{\frac{1}{2n}\|Y^* - X^*\beta\|_2^2 + \frac{\lambda_2}{2}\sum_{j \notin \hat{S}_{Lasso}}\beta_j^2\right\} \quad (2.9)$$

4) ทำตามขั้นตอนที่ 1), 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$

5) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น $(1-\alpha)100\%$ สำหรับβ_j เมื่อ $j = 1,2,\ldots,p$ โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของสัมประสิทธิ์ที่คำนวณจากตัวอย่างบูตสแตรป $\left(\hat{\beta}_j^{(1)}, \hat{\beta}_j^{(2)}, \ldots, \hat{\beta}_j^{(B)}\right)$

บทที่ 3
วิธีการดำเนินงานวิจัย

3.1 ขอบเขตของการวิจัย

การวิจัยนี้กระทำภายใต้การวิเคราะห์การถดถอยลอจิสติกทวิภาคสำหรับข้อมูลที่มีมิติสูง โดยผู้วิจัยได้กำหนดขอบเขตของการวิจัยดังต่อไปนี้

1) กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2) สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$

กำหนด $\sum$ จาก 2 วิธี ได้แก่ Toeplitz และ Equal Correlation ดังสมการต่อไปนี้

- วิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ด้วย $\rho = 0.5, 0.9$
- วิธี Equal Correlation: $\sum_{ij} = \rho$ ด้วย $\rho = 0.5, 0.9$

3) ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จาก 2 วิธี ดังนี้

- วิธีที่ 1 (Hard Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2,\ldots,p$
- วิธีที่ 2 (Weak Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1,2,\ldots,p$

4) สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

5) กำหนดจำนวนทำซ้ำของบูตสแตรป B = 1000 ครั้ง

6) กำหนดระดับนัยสำคัญทางสถิติของการทดสอบ เท่ากับ 0.05

7) กำหนดพารามิเตอร์ที่มีการปรับค่าแล้ว (λ_1, λ_2) ด้วยวิธี 5-fold cross validation และเก็บค่าไว้เพื่อตรวจสอบการแจกแจง โดยค่าพารามิเตอร์ที่มีการปรับค่าแล้วควรมีค่าใกล้เคียงกันทุกรอบการทำซ้ำ

<u>หมายเหตุ</u> : ในแต่ละกรณีจะสุ่มเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ และ $\boldsymbol{X}$ เพียงครั้งเดียวและจะทำซ้ำจำนวน 50 ครั้ง (50 Replicates)

3.2 วิธีดำเนินการวิจัย

1) สร้างข้อมูลจำลองตามขอบเขตที่ได้ศึกษา จำนวน 8 ชุด ดังต่อไปนี้

1.1) กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

1.2) สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ กำหนด $\sum$ จาก 2 วิธี ได้แก่ Toeplitz และ Equal Correlation ดังสมการต่อไปนี้

- วิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ด้วย $\rho = 0.5, 0.9$
- วิธี Equal Correlation: $\sum_{ij} = \rho$ ด้วย $\rho = 0.5, 0.9$

1.3) ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จาก 2 วิธี ดังนี้

- วิธีที่ 1 (Hard Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1, 2, \ldots, p$
- วิธีที่ 2 (Weak Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1, 2, \ldots, p$

1.4) สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1 + exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

2) สร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติก โดยใช้การประมาณ 2 ขั้นตอน ทั้งหมด 4 วิธี ดังนี้

- วิธี Parametric Bootstrap Lasso+MLE ซึ่งมีขั้นตอนตามหัวข้อที่ 2.4.1
- วิธี Parametric Bootstrap Lasso+ Partial Ridge ซึ่งมีขั้นตอนตามหัวข้อที่ 2.4.2
- วิธี Paired Bootstrap Lasso+MLE ซึ่งมีขั้นตอนตามหัวข้อที่ 2.4.3
- วิธี Paired Bootstrap Lasso+Partial Ridge ซึ่งมีขั้นตอนตามหัวข้อที่ 2.4.4

3) เปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากทั้ง 4 วิธี โดยใช้ความกว้างเฉลี่ยของช่วงความเชื่อมั่น (Average width: AW), ค่าความน่าจะเป็นครอบคลุม (Coverage probability: CP), ค่าความแม่นยำ (Precision) และค่าความไว (Recall) ดังสมการต่อไปนี้

$$AW = \frac{\sum_{j=1}^{p}(u_j - l_j)}{p} \quad (8)$$

$$CP = \frac{\text{จำนวนช่วงความเชื่อมั่นที่ครอบคลุมค่าพารามิเตอร์จริง}}{p} \quad (9)$$

$$Precision = \frac{|\hat{S} \cap S|}{|\hat{S}|} \quad (10)$$

$$Recall \quad = \frac{|\hat{S} \cap S|}{|S|} \tag{11}$$

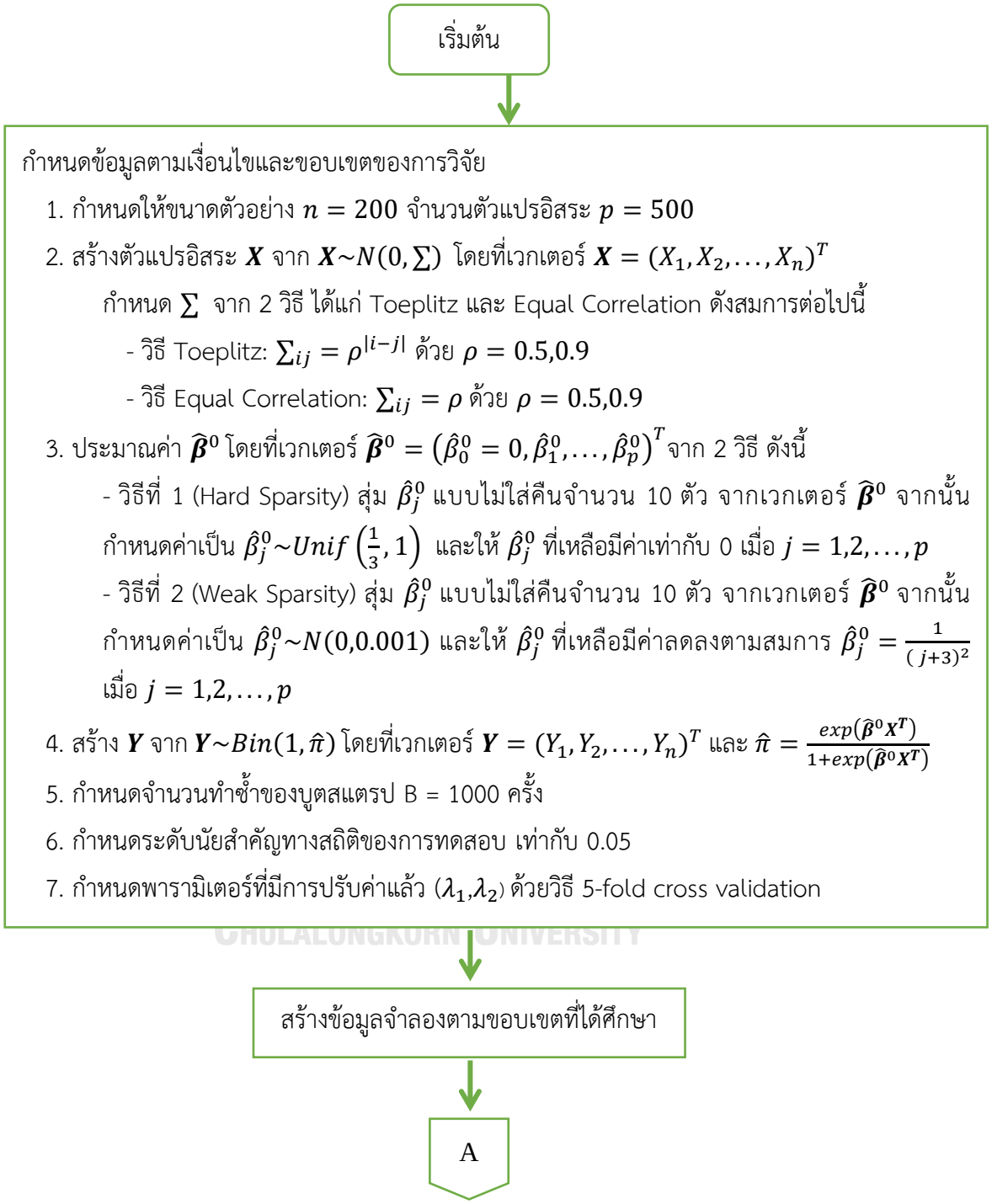
<u>กำหนดให้</u>

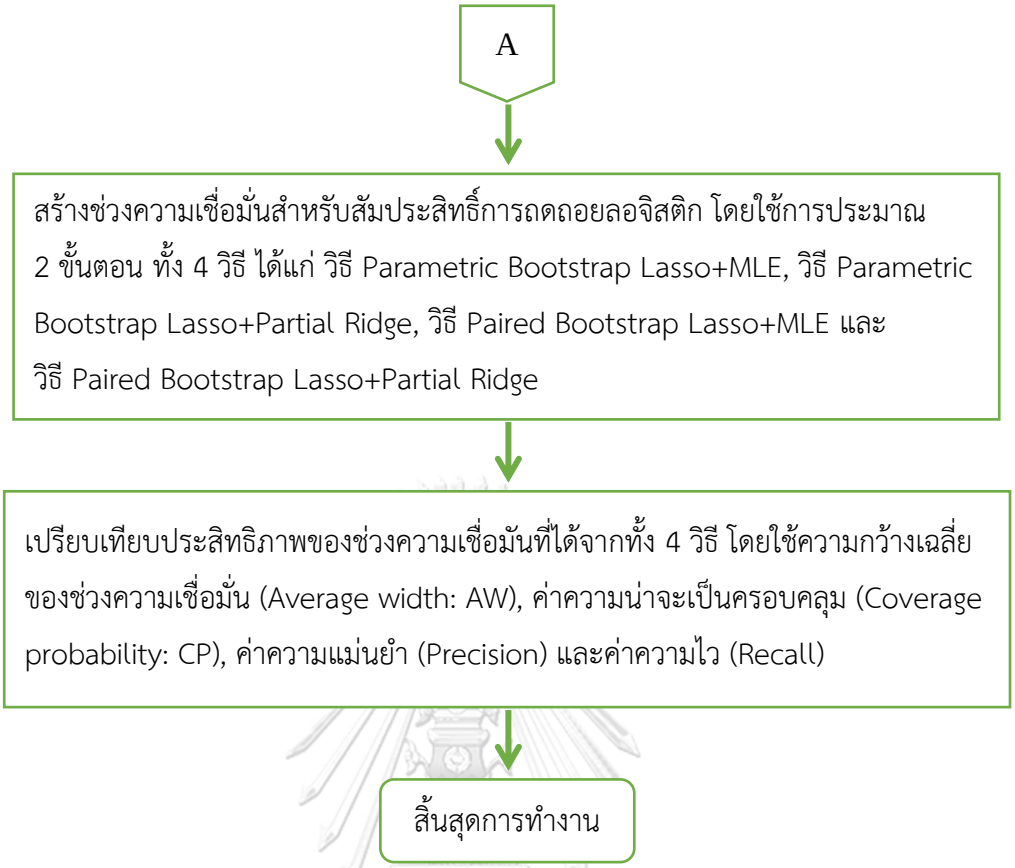
S คือ เซตของ j ที่ค่าสัมประสิทธิ์การถดถอยที่แท้จริงที่มีค่าไม่เท่ากับ 0 หรือ $S = \{j : \beta_j \neq 0\}$

$\hat{S}$ คือ เซตของ j ที่ค่าประมาณของสัมประสิทธิ์การถดถอยปฏิเสธสมมติฐานว่างว่าค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับ 0 หรือ $\hat{S} = \{j :$ ปฏิเสธ $H_0\}$ เมื่อ $j = 1,2,\ldots,p$

4) สรุปผล โดยนำเสนอข้อมูลในรูปตาราง เพื่อตรวจสอบดูว่าวิธีการแบบใดให้ผลดีกว่ากัน

3.3 ขั้นตอนการทำงานของโปรแกรม





บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ซึ่งในการศึกษานี้จะจำลองข้อมูลทั้งหมด 8 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว โดยถ้าวิธีใดให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมสูง และให้ค่าความแม่นยำ และค่าความไวต่ำที่สุดจะถือเป็นวิธีที่มีประสิทธิภาพและมีความเหมาะสมในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูงมากที่สุด

อักษรย่อและสัญลักษณ์ต่าง ๆ ที่ปรากฏในการผลการวิจัยทั้งในตารางและจ้อความต่าง ๆ แทนความหมายดังนี้

AW	แทน ความกว้างเฉลี่ยของช่วงความเชื่อมั่น (Average width)
CP	แทน ค่าความน่าจะเป็นครอบคลุม (Coverage probability)
Precision	แทน ค่าความแม่นยำ (Precision)
Recall	แทน ค่าความไว (Recall)
Med	แทน ค่ามัธยฐาน (Median)
SD	แทน ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation)

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกเป็น 8 ส่วน ตามจำนวนชุดข้อมูล ซึ่งแต่ละส่วนจะเป็นการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว ภายใต้ขอบเขตของข้อมูลชุดที่ 1 ถึง 8 ตามลำดับ โดยมีผลลัพธ์ดังต่อไปนี้

4.1 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าแม่นยำ และค่าความไว โดยใช้ข้อมูลชุดที่ 1

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 1 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ และให้ $\rho = 0.5$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Hard Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น$\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 1 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 1

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	11.73	2.98e +12	0.07	0.42	0.05	0.03	0.65	0.02
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.07**	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	10.58	0.80	**0.99**	0.01	0.30	0.14	0.12	0.21
วิธี Paired Bootstrap Lasso+Partial Ridge	0.09	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 1 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 1 มากที่สุด และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 1 ด้วยเช่นกัน เนื่องจากมีค่าความกว้าง

เฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงที่สุด แต่วิธีนี้ถือว่าไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 1 เนื่องจากมีค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 และยังให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ และแม้ว่าวิธี Parametric Bootstrap Lasso+MLE จะให้ค่าความไวค่อนข้างสูง แต่เมื่อพิจารณาว่าให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างที่สุด มีค่าความน่าจะเป็นครอบคลุมและค่าความแม่นยำต่ำกว่าวิธีอื่น จึงสรุปได้ว่าวิธี Parametric Bootstrap Lasso+MLE ไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 1 ด้วยเช่นกัน

4.2 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 2

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 2 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ และให้ $\rho = 0.5$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Weak Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1, 2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1 + exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 2 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 2

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	6.44	1.82e +13	0.94	0.03	0.02	0.19	0.07	0.03
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.07**	0.02	0.98	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	15.78	7.17e +12	**1.00**	0.01	0.98	0.02	0.86	0.01
วิธี Paired Bootstrap Lasso+Partial Ridge	0.10	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 2 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 2 มากที่สุด และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 2 ด้วยเช่นกัน เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงที่สุด มีค่าความแม่นยำและค่าความไวใกล้เคียงกับทั้งวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge แต่เนื่องจากวิธีนี้ให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ ดังนั้นจึงกล่าวได้ว่า วิธี Paired Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 2 และวิธี Parametric Bootstrap Lasso+MLE จะถือได้ว่าเป็นวิธีที่มี

ประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 2 ต่ำที่สุด เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก มีค่าความน่าจะเป็นครอบคลุมต่ำกว่าวิธีอื่น และให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50

4.3 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 3

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 3 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ และให้ $\rho = 0.9$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Hard Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น$\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 3 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 3

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	3.00e +12	1.55e +13	0.81	0.29	0.03	0.04	0.11	0.03
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.05**	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	24.73	2.02e +12	**0.99**	0.01	0.02	0.14	0.01	0.23
วิธี Paired Bootstrap Lasso+Partial Ridge	0.05	0.01	0.98	0.02	0.05	0.05	0.21	0.01

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาใน Table 3 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 3 มากที่สุด และพบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+Partial Ridge จะให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด และให้ค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.81 แต่วิธีนี้กลับให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 จึงถือได้ว่าวิธี Paired Bootstrap Lasso+Partial Ridge ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 3

นอกจากนี้การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE และวิธี Parametric Bootstrap Lasso+MLE ก็ถือว่าไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 3 ด้วยเช่นกัน เนื่องจากมีค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 และมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ทำให้ได้ค่าความน่าจะเป็นครอบคลุมสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ

4.4 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 4

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 4 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ และให้ $\rho = 0.9$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Weak Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1, 2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y}\sim Bin(1,\hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y}=(Y_1,Y_2,\dots,Y_n)^T$ และ $\hat{\pi}=\frac{exp(\hat{\boldsymbol{\beta}}^0\boldsymbol{X}^T)}{1+exp(\hat{\boldsymbol{\beta}}^0\boldsymbol{X}^T)}$

Table 4 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 4

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	4.24	1.04e+13	0.92	0.03	0.02	0.26	0.08	0.01
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.01**	0.02	0.98	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	9.22e+10	2.83e+12	**1.00**	0.01	0.93	0.04	0.64	0.03
วิธี Paired Bootstrap Lasso+Partial Ridge	0.14	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 4 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 4 มากที่สุด และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 4 ด้วยเช่นกัน เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงที่สุด มีค่าความแม่นยำใกล้เคียงกับทั้งวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge และให้ค่าความไวสูงกว่า 0.50 แต่เนื่องจากวิธีนี้ให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็น

ครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ ดังนั้นจึงกล่าวได้ว่า วิธี Paired Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 4 และวิธี Parametric Bootstrap Lasso+MLE จะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 4 ต่ำที่สุด เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก มีค่าความน่าจะเป็นครอบคลุมต่ำกว่าวิธีอื่น และให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50

4.5 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ โดยใช้ข้อมูลชุดที่ 5

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 5 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ และให้ $\rho = 0.5$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Hard Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น$\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 5 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 5

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	3.36	4.24e +12	0.08	0.39	0.02	0.01	0.53	0.06
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.04**	0.01	**0.98**	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap	15.24	1.67e	0.83	0.04	0.01	0.02	0.06	0.33

Lasso+MLE		+12						
วิธี Paired Bootstrap Lasso+Partial Ridge	0.17	0.02	**0.98**	0.01	0.99	0.01	0.94	0.04

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 5 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมสูงที่สุด และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 5 มากที่สุด และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 5 ด้วยเช่นกัน เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge และให้ค่าความแม่นยำและค่าความไวใกล้เคียงกับหนึ่ง

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมค่อนข้างสูง แต่เมื่อพิจารณาเกณฑ์อื่น ๆ พบว่า วิธีนี้ให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 และมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ ดังนั้นจึงกล่าวได้ว่า วิธี Paired Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 5 และแม้ว่าวิธี Parametric Bootstrap Lasso+MLE จะให้ค่าความไวเท่ากับ 0.50 แต่เมื่อพิจารณาช่วงความเชื่อมั่นที่ได้จากวิธีนี้จะพบว่า ให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก มีค่าความน่าจะเป็นครอบคลุมต่ำกว่าวิธีอื่น และให้ค่าความแม่นยำต่ำกว่า 0.50 จึงกล่าวได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 5 ด้วยเช่นกัน

4.6 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 6

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 6 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X}\sim N(0,\sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ และให้ $\rho = 0.5$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Weak Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0\sim N(0,0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1,2,\ldots,p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y}\sim Bin(1,\hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0\boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0\boldsymbol{X}^T)}$

Table 6 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 6

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	5.33e +11	1.67e +13	0.92	0.05	0.02	0.26	0.10	0.01
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.01**	0.03	0.98	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	18.42	6.94e +12	**1.00**	0.01	0.93	0.03	0.63	0.05
วิธี Paired Bootstrap Lasso+Partial Ridge	0.15	0.02	0.98	0.01	0.99	0.01	0.94	0.01

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 6 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 6 มากที่สุด และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 6 ด้วยเช่นกัน เนื่องจากมีค่าความกว้าง

เฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge และให้ค่าความแม่นยำและค่าความไวใกล้เคียงกับหนึ่ง

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงที่สุด มีค่าความแม่นยำและค่าความไวใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge แต่เนื่องจากวิธีนี้ให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ ดังนั้นจึงกล่าวได้ว่า วิธี Paired Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 6 และวิธี Parametric Bootstrap Lasso+MLE จะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 6 ต่ำที่สุด เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก มีค่าความน่าจะเป็นครอบคลุมต่ำกว่าวิธีอื่น และให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50

4.7 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 7

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 7 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ และให้ $\rho = 0.9$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Hard Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น$\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 7 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 7

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	6.36e +12	1.48e +13	0.56	0.44	0.02	0.01	0.57	0.15
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.04**	0.01	**0.98**	0.00	**1.00**	0.00	**1.00**	0.00
วิธี Paired Bootstrap Lasso+MLE	51.41	1.034e +12	0.62	0.08	0.01	0.01	0.27	0.43
วิธี Paired Bootstrap Lasso+Partial Ridge	0.10	0.01	0.97	0.01	0.12	0.01	0.12	0.19

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 7 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบที่สุด มีค่าความน่าจะเป็นครอบคลุมสูงที่สุด และให้ค่าความแม่นยำและค่าความไวเท่ากับหนึ่ง กล่าวคือ วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 7 มากที่สุด และพบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+Partial Ridge จะให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นและค่าความน่าจะเป็นครอบคลุมใกล้เคียงกับวิธี Parametric Bootstrap Lasso+Partial Ridge แต่วิธีนี้กลับให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 จึงถือได้ว่าวิธี Paired Bootstrap Lasso+Partial Ridge ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับข้อมูลชุดที่ 7

นอกจากนี้การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE และวิธี Parametric Bootstrap Lasso+MLE ก็ถือว่าไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 7 ด้วยเช่นกัน เนื่องจากมีค่าความแม่นยำและค่าความไวต่ำกว่าวิธีอื่น และมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ทำให้ได้ค่าความน่าจะเป็นครอบคลุมสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ

4.8 ผลการเปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยใช้ข้อมูลชุดที่ 8

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 6 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ และให้ $\rho = 0.9$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Weak Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1, 2, \ldots, p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1 + exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

Table 8 แสดงค่ามัธยฐานและส่วนเบี่ยงเบนมาตรฐานของความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว ซึ่งคำนวณได้จากการทำซ้ำทั้งหมด 50 ครั้ง โดยใช้ข้อมูลชุดที่ 8

เกณฑ์ / วิธีการ	AW		CP		Precision		Recall	
	Med	SD	Med	SD	Med	SD	Med	SD
วิธี Parametric Bootstrap Lasso+MLE	7.61e+12	4.86e+13	0.95	0.13	0.03	0.27	0.11	0.01
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.01**	0.08	0.98	0.00	0.99	0.01	0.98	0.01
วิธี Paired Bootstrap Lasso+MLE	39.34	2.01e+12	**1.00**	0.01	0.93	0.06	0.62	0.13
วิธี Paired Bootstrap Lasso+Partial Ridge	0.33	0.01	0.98	0.00	**1.00**	0.00	**1.00**	0.00

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละเกณฑ์

จากการศึกษาในตาราง Table 8 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired

Bootstrap Lasso+Partial Ridge จะให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบ มีค่าความน่าจะเป็นครอบคลุมเท่ากับ 0.98 และให้ค่าความแม่นยำและค่าความไวใกล้เคียงกับศูนย์ กล่าวคือ ทั้งวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 8

นอกจากนี้จากการสังเกต พบว่า วิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงที่สุด มีค่าความแม่นยำใกล้เคียงกับทั้งวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge และให้ค่าความไวเท่ากับ 0.62 แต่เนื่องจากวิธีนี้ให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ ดังนั้นจึงกล่าวได้ว่า วิธี Paired Bootstrap Lasso+MLE ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 8 และวิธี Parametric Bootstrap Lasso+MLE จะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 8 ต่ำที่สุด เนื่องจากมีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก มีค่าความน่าจะเป็นครอบคลุมต่ำกว่าวิธีอื่น และให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเพื่อเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+ Partial Ridge ซึ่งในงานวิจัยนี้จะพิจารณาตามการจำลองข้อมูลที่แตกต่างกัน จำนวน 8 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว โดยสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 ผลจากข้อมูลจำนวน 8 ชุด

งานวิจัยนี้พิจารณาตามข้อมูลจำลองที่แตกต่างกัน จำนวน 8 ชุด โดยมีการกำหนดขอบเขตดังต่อไปนี้

1) กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$
2) สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ กำหนด $\sum$ จาก 2 วิธี ได้แก่ Toeplitz และ Equal Correlation ดังสมการต่อไปนี้
   - วิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ด้วย $\rho = 0.5, 0.9$
   - วิธี Equal Correlation: $\sum_{ij} = \rho$ ด้วย $\rho = 0.5, 0.9$
3) ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จาก 2 วิธี ดังนี้
   - วิธีที่ 1 (Hard Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2, \ldots, p$
   - วิธีที่ 2 (Weak Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0, 0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1,2, \ldots, p$
4) สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

โดยกำหนดให้ข้อมูลแต่ละชุดมีการจำลองข้อมูล ดังต่อไปนี้

- ข้อมูลชุดที่ 1 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.5$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity
- ข้อมูลชุดที่ 2 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.5$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity
- ข้อมูลชุดที่ 3 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.9$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity
- ข้อมูลชุดที่ 4 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ที่ $\rho = 0.9$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity
- ข้อมูลชุดที่ 5 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ ที่ $\rho = 0.5$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity
- ข้อมูลชุดที่ 6 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ ที่ $\rho = 0.5$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity
- ข้อมูลชุดที่ 7 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ ที่ $\rho = 0.9$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity
- ข้อมูลชุดที่ 8 : สร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Equal Correlation: $\sum_{ij} = \rho$ ที่ $\rho = 0.9$ และตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity

Table 9 แสดงวิธีการสร้างช่วงความเชื่อมั่นที่เหมาะสมที่สุดสำหรับสัมประสิทธิ์การถดถอยลอจิสติก เมื่อพิจารณาค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว โดยจำแนกตามชุดข้อมูลทั้งหมด 8 ชุด

วิธีการสร้างช่วงความเชื่อมั่น	ชุดข้อมูล							
	1	2	3	4	5	6	7	8
วิธี Parametric Bootstrap Lasso+MLE								
วิธี Parametric Bootstrap Lasso+Partial Ridge	✓	✓	✓	✓	✓	✓	✓	✓
วิธี Paired Bootstrap Lasso+MLE								
วิธี Paired Bootstrap Lasso+Partial Ridge	✓	✓		✓	✓	✓		✓

หมายเหตุ: ✓ หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุด โดยพิจารณาจากเกณฑ์ทั้งหมด

จากตาราง Table 9 สามารถสรุปได้ว่า เมื่อพิจารณาค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไว โดยจำแนกตามชุดข้อมูลทั้งหมด 8 ชุด จะพบว่าวิธีที่มีประสิทธิภาพมากที่สุดในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติก คือ วิธี Parametric Bootstrap Lasso+Partial Ridge ซึ่งสามารถทำงานได้ดีในทุกชุดข้อมูล และวิธี Paired Bootstrap Lasso+Partial Ridge ก็มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลชุดที่ 1,2,4,5,6,8 ด้วยเช่นกัน ซึ่งจะสังเกตว่าวิธี Paired Bootstrap Lasso+Partial Ridge ไม่มีประสิทธิภาพเพียงพอในการวิเคราะห์ข้อมูลชุดที่ 3 และ 7 เนื่องจากข้อมูลทั้งสองชุดนี้มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity กล่าวคือ การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+Partial Ridge มีข้อจำกัดในการวิเคราะห์ข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity

5.1.2 ผลจากความแตกต่างระหว่างวิธีการสร้างตัวแปรอิสระ $\boldsymbol{X}$

Table 10 แสดงค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรอิสระ

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความกว้างเฉลี่ย AW			
	วิธี Toeplitz		วิธี Equal Correlation	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
วิธี Parametric Bootstrap Lasso+MLE	3.94e+12	7.15e+12	6.41e+12	2.49e+13
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.07**	0.04	**0.04**	**0.05**
วิธี Paired Bootstrap Lasso+MLE	1.56e+12	1.37e+12	1.20e+12	5.94e+12
วิธี Paired Bootstrap Lasso+Partial Ridge	0.10	**0.01**	0.17	0.22

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 10 สามารถสรุปได้ว่า ค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge มีค่าน้อยหรือแคบกว่าวิธี Parametric Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge อย่างชัดเจน กล่าวคือ วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี

Lasso+Partial Ridge จะให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรอิสระที่แตกต่างกัน

Table 11 แสดงค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรอิสระ

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความน่าจะเป็นครอบคลุม			
	วิธี Toeplitz		วิธี Equal Correlation	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
วิธี Parametric Bootstrap Lasso+MLE	0.51	0.89	0.50	0.76
วิธี Parametric Bootstrap Lasso+Partial Ridge	0.98	0.98	**0.98**	**0.98**
วิธี Paired Bootstrap Lasso+MLE	**1.00**	**1.00**	0.92	0.81
วิธี Paired Bootstrap Lasso+Partial Ridge	0.98	0.98	**0.98**	**0.98**

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 11 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรอิสระด้วยวิธี Toeplitz นั้น พบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE จะให้ค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมสูงที่สุด และวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความน่าจะเป็นครอบคลุมน้อยกว่าวิธี Paired Bootstrap Lasso เล็กน้อย แต่เนื่องจากข้อมูลในตารางที่ 5.1.2.1 แสดงให้เห็นว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ และสำหรับชุดข้อมูลที่มีการสร้างตัวแปรอิสระด้วยวิธี Equal Correlation จะพบว่าค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมที่ได้จากทั้งวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+PartialRidge มีค่าสูงกว่าวิธี Parametric Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge อย่างชัดเจน

Table 12 แสดงค่าเฉลี่ยของค่าความแม่นยำที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรอิสระ

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความแม่นยำ			
	วิธี Toeplitz		วิธี Equal Correlation	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
วิธี Parametric Bootstrap Lasso+MLE	0.04	0.02	0.02	0.03
วิธี Parametric Bootstrap Lasso+Partial Ridge	**1.00**	**1.00**	**1.00**	**1.00**
วิธี Paired Bootstrap Lasso+MLE	0.64	0.48	0.47	0.47
วิธี Paired Bootstrap Lasso+Partial Ridge	**1.00**	0.51	0.99	0.56

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 12 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรอิสระด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.5$ และ $\rho = 0.9$ นั้น พบว่า ค่าเฉลี่ยของค่าความแม่นยำที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าเท่ากับหนึ่งในทุกข้อมูล กล่าวคือ การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพมากที่สุดในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรอิสระที่แตกต่างกัน และสำหรับวิธี Paired Bootstrap Lasso+Partial Ridge พบว่าวิธีนี้ให้ค่าเฉลี่ยของค่าความแม่นยำใกล้เคียงกับหนึ่ง ในชุดข้อมูลที่มีการสร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.5$ และให้ค่าเฉลี่ยของค่าความแม่นยำมากกว่า 0.50 ในชุดข้อมูลที่มีการสร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.9$ แต่ในทางตรงข้าม พบว่า วิธี Paired Bootstrap Lasso+MLE มีค่าเฉลี่ยของค่าความแม่นยำ เท่ากับ 0.63 สำหรับข้อมูลที่มีการสร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz ที่ $\rho = 0.5$ แต่หากพิจารณาร่วมกับข้อมูลชุดอื่น ๆ จะพบว่าค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ทั้งหมด นอกจากนี้ยังพบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+MLE ให้ค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ในทุกชุดข้อมูล จึงสรุปได้ว่า วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความแม่นยำสูงกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรอิสระที่แตกต่างกัน

Table 13 แสดงค่าเฉลี่ยของค่าความไวที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรอิสระ

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความไว			
	วิธี Toeplitz		วิธี Equal Correlation	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
วิธี Parametric Bootstrap Lasso+MLE	0.36	0.10	0.32	0.34
วิธี Parametric Bootstrap Lasso+Partial Ridge	**1.00**	**1.00**	**1.00**	**0.99**
วิธี Paired Bootstrap Lasso+MLE	0.49	0.33	0.35	0.45
วิธี Paired Bootstrap Lasso+Partial Ridge	**1.00**	0.61	0.94	0.56

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 13 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรอิสระด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.5$ และ $\rho = 0.9$ นั้น พบว่า ค่าเฉลี่ยของค่าความไวที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าเท่ากับหนึ่งในทุกข้อมูล กล่าวคือ การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพมากที่สุดในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรอิสระที่แตกต่างกัน และสำหรับวิธี Paired Bootstrap Lasso+Partial Ridge พบว่าวิธีนี้ให้ค่าเฉลี่ยของค่าความไวใกล้เคียงกับหนึ่ง ในชุดข้อมูลที่มีการสร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.5$ และให้ค่าเฉลี่ยของค่าความไวมากกว่า 0.50 ในชุดข้อมูลที่มีการสร้างตัวแปรอิสระ $\boldsymbol{X}$ ด้วยวิธี Toeplitz และวิธี Equal Correlation ที่ $\rho = 0.9$ แต่ในทางตรงข้าม พบว่าทั้งวิธี Paired Bootstrap Lasso+MLE และวิธี Parametric Bootstrap Lasso+MLE มีค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ในทุกชุดข้อมูล จึงสรุปได้ว่า วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความไวสูงกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรอิสระที่แตกต่างกัน

ดังนั้นเมื่อพิจารณาข้อสรุปที่ได้จากตาราง Table 10 ถึงตาราง Table 13 ทำให้สรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+Partial Ridge มีประสิทธิภาพมากกว่าวิธี Lasso+MLE เมื่อจำแนกตามวิธีการสร้างตัวแปรอิสระ $\boldsymbol{X}$

5.1.3 ผลจากความแตกต่างระหว่างวิธีการสร้างตัวแปรตาม Y

Table 14 แสดงค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรตาม

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความกว้างเฉลี่ย	
	วิธี Hard Sparsity	วิธี Weak Sparsity
วิธี Parametric Bootstrap Lasso+MLE	6.14e+12	1.50e+13
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.06**	**0.04**
วิธี Paired Bootstrap Lasso+MLE	1.49e+12	3.55e+12
วิธี Paired Bootstrap Lasso+Partial Ridge	0.10	0.18

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 14 สามารถสรุปได้ว่า ค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge มีค่าน้อยหรือแคบกว่าวิธี Parametric Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge อย่างชัดเจน กล่าวคือ วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรตามที่แตกต่างกัน

Table 15 แสดงค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรตาม

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความน่าจะเป็นครอบคลุม	
	วิธี Hard Sparsity	วิธี Weak Sparsity
วิธี Parametric Bootstrap Lasso+MLE	0.38	0.93
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.98**	0.98
วิธี Paired Bootstrap Lasso+MLE	0.86	**0.99**
วิธี Paired Bootstrap Lasso+Partial Ridge	**0.98**	0.98

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 15 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity นั้น พบว่า ค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมที่ได้จากวิธี Parametric Bootstrap Lasso +Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge มีค่าสูงกว่าวิธี Parametric Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge อย่างชัดเจน แต่สำหรับชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity จะพบว่า ค่าเฉลี่ยของค่าความน่าจะเป็นครอบคลุมที่ได้จากวิธี Paired Bootstrap Lasso+MLE มีค่ามากที่สุด และวิธี Parametric Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความน่าจะเป็นครอบคลุมน้อยกว่าวิธี Paired Bootstrap Lasso เล็กน้อย แต่เนื่องจากข้อมูลในตารางที่ 5.1.3.1 แสดงให้เห็นว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งส่งผลทำให้ค่าความน่าจะเป็นครอบคลุมมีค่าสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ นอกจากนี้พบว่าการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+MLE เป็นวิธีที่ให้ค่าเฉลี่ยของความน่าจะเป็นครอบคลุมต่ำที่สุดในทุกชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ แตกต่างกัน

Table 16 แสดงค่าเฉลี่ยของค่าความแม่นยำที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรตาม

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความแม่นยำ	
	วิธี Hard Sparsity	วิธี Weak Sparsity
วิธี Parametric Bootstrap Lasso+MLE	0.02	0.02
วิธี Parametric Bootstrap Lasso+Partial Ridge	**1.00**	**1.00**
วิธี Paired Bootstrap Lasso+MLE	0.08	0.95
วิธี Paired Bootstrap Lasso+Partial Ridge	0.53	**1.00**

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 16 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity และวิธี Weak Sparsity พบว่า ค่าเฉลี่ยของค่าความแม่นยำที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าเท่ากับหนึ่งในทุกข้อมูล กล่าวคือ การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพมากที่สุดในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ที่แตกต่างกัน และ

สำหรับวิธี Paired Bootstrap Lasso+Partial Ridge พบว่าวิธีนี้ให้ค่าเฉลี่ยของค่าความแม่นยำเท่ากับหนึ่ง ในชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity และให้ค่าเฉลี่ยของค่าความแม่นยำมากกว่า 0.50 ในชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity แต่ในทางตรงข้าม พบว่า วิธี Paired Bootstrap Lasso+MLE มีค่าเฉลี่ยของค่าความแม่นยำ เท่ากับ 0.95 สำหรับข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity แต่หากพิจารณาร่วมกับข้อมูลชุดอื่น ๆ จะพบว่าค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ทั้งหมด นอกจากนี้ยังพบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+MLE ให้ค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ในทุกชุดข้อมูล จึงสรุปได้ว่า วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความแม่นยำสูงกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ที่แตกต่างกัน

Table 17 แสดงค่าเฉลี่ยของค่าความไวที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี โดยจำแนกตามวิธีการสร้างตัวแปรตาม

วิธีการสร้างช่วงความเชื่อมั่น	ค่าความไว	
	วิธี Hard Sparsity	วิธี Weak Sparsity
วิธี Parametric Bootstrap Lasso+MLE	0.50	0.09
วิธี Parametric Bootstrap Lasso+Partial Ridge	**1.00**	**1.00**
วิธี Paired Bootstrap Lasso+MLE	0.11	0.68
วิธี Paired Bootstrap Lasso+Partial Ridge	0.56	0.99

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 17 สามารถสรุปได้ว่า สำหรับชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity และวิธี Weak Sparsity พบว่า ค่าเฉลี่ยของค่าความไวที่ได้จากวิธี Parametric Bootstrap Lasso+Partial Ridge มีค่าเท่ากับหนึ่งในทุกข้อมูล กล่าวคือ การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพมากที่สุดในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ที่แตกต่างกัน และสำหรับวิธี Paired Bootstrap Lasso+Partial Ridge พบว่าวิธีนี้ให้ค่าเฉลี่ยของค่าความไวใกล้เคียงกับหนึ่ง ในชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity และให้ค่าเฉลี่ยของค่าความไวมากกว่า 0.50 ในชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity แต่ในทางตรงข้าม พบว่าวิธี Paired Bootstrap Lasso+MLE ให้ค่าเฉลี่ยของค่าความไวมากกว่า 0.50 ในชุดข้อมูลที่มี

การสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Weak Sparsity แต่ให้ค่าเฉลี่ยของความไวต่ำกว่า 0.50 ในชุดข้อมูลที่มีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ด้วยวิธี Hard Sparsity นอกจากนี้ยังพบว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Parametric Bootstrap Lasso+MLE มีค่าเฉลี่ยของความแม่นยำต่ำกว่า 0.50 ในทุกชุดข้อมูล จึงสรุปได้ว่า วิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ค่าเฉลี่ยของความไวสูงกว่าวิธี Lasso+MLE ในทุกชุดข้อมูลที่ถูกสร้างด้วยวิธีการสร้างตัวแปรตาม $\boldsymbol{Y}$ ที่แตกต่างกัน

ดังนั้นเมื่อพิจารณาข้อสรุปที่ได้จากตารางที่ 5.1.3.1 ถึงตารางที่ 5.1.3.4 ทำให้สรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+Partial Ridge มีประสิทธิภาพมากกว่าวิธี Lasso+MLE เมื่อจำแนกตามวิธีการสร้างตัวแปรตาม $\boldsymbol{Y}$

5.1.4 ผลจากความแตกต่างระหว่างการบูตสแตรปด้วยวิธี Parametric Bootstrap และวิธี Paired Bootstrap

Table 18 แสดงค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไวที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+MLE โดยคำนวณจากข้อมูลทั้งหมด 8 ชุด และจำแนกตามวิธีบูตสแตรปสำหรับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ระหว่างวิธี Parametric Bootstrap และวิธี Paired Bootstrap

วิธีการสร้างช่วงความเชื่อมั่น	เกณฑ์การเปรียบเทียบประสิทธิภาพ			
	AW	CP	Precision	Recall
วิธี Parametric Bootstrap Lasso+MLE	1.06e+13	0.67	0.03	0.29
วิธี Paired Bootstrap Lasso+MLE	**2.52e+12**	**0.93**	**0.51**	**0.39**

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 18 สามารถสรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+MLE โดยใช้การประมาณค่าด้วยวิธี Paired Bootstrap มีค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Parametric Bootstrap และให้ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไวสูงกว่าวิธี Parametric Bootstrap กล่าวคือ การประมาณค่าด้วยวิธี Paired Bootstrap เหมาะสมกับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+MLE มากกว่าการประมาณค่าด้วยวิธี Parametric Bootstrap

Table 19 แสดงค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไวที่ได้จากวิธีการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge โดยคำนวณจากข้อมูลทั้งหมด 8 ชุด และจำแนกตามวิธีบูตสแตรปสำหรับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติกส์ระหว่างวิธี Parametric Bootstrap และวิธี Paired Bootstrap

วิธีการสร้างช่วงความเชื่อมั่น	เกณฑ์การเปรียบเทียบประสิทธิภาพ			
	AW	CP	Precision	Recall
วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.05**	**0.98**	**1.00**	**1.00**
วิธี Paired Bootstrap Lasso+Partial Ridge	0.14	0.97	0.77	0.78

หมายเหตุ: ตัวอักษรหนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละวิธี

จากตาราง Table 19 สามารถสรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+Partial Ridge พบว่า การประมาณค่าสัมประสิทธิ์ด้วยวิธี Parametric Bootstrap มีค่าเฉลี่ยของค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Paired Bootstrap และให้ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำและค่าความไวสูงกว่าวิธี Paired Bootstrap กล่าวคือ การประมาณค่าด้วยวิธี Parametric Bootstrap เหมาะสมกับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+Partial Ridge มากกว่าการประมาณค่าด้วยวิธี Paired Bootstrap

ดังนั้นเมื่อพิจารณาข้อสรุปที่ได้จากตารางที่ 5.1.4.1 ถึงตารางที่ 5.1.4.2 ทำให้สรุปได้ว่า การประมาณค่าด้วยวิธี Paired Bootstrap เหมาะสมกับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+MLE และการประมาณค่าด้วยวิธี Parametric Bootstrap เหมาะสมกับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+Partial Ridge

5.2 สรุปผลโดยรวม

จากการวิจัยเพื่อเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+Partial Ridge ซึ่งทำการวิจัยโดยจำลองชุดข้อมูลที่แตกต่างกัน จำนวน 8 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4

วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso
+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความแม่นยำ และค่าความไว ผลปรากฏว่า วิธี Parametric Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นมากที่สุด รองลงมาคือ วิธี Paired Bootstrap Lasso+Partial Ridge และวิธี Paired Bootstrap Lasso+MLE ตามลำดับ ซึ่งวิธี Paired Bootstrap Lasso+MLE ถูกสรุปให้มีประสิทธิภาพน้อยกว่าวิธี Paired Bootstrap Lasso+Partial Ridge เนื่องจากการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Paired Bootstrap Lasso+MLE มีค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่นกว้างมาก ซึ่งจะส่งผลทำให้มีค่าความน่าจะเป็นครอบคลุมสูงแต่ช่วงความเชื่อมั่นที่คำนวณได้จะไม่มีประสิทธิภาพ และให้ค่าความแม่นยำและค่าความไวต่ำกว่าวิธี Paired Bootstrap Lasso+Partial Ridge อย่างชัดเจน กล่าวคือ แม้ว่าวิธี Paired Bootstrap Lasso+MLE จะให้ค่าความน่าจะเป็นครอบคลุมสูงกว่าวิธีอื่น ๆ แต่เมื่อพิจารณาร่วมกับเหตุผลที่กล่าวมาข้างต้น จึงแสดงให้เห็นว่า วิธี Paired Bootstrap Lasso+Partial Ridge มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นมากกว่าวิธี Paired Bootstrap Lasso+MLE อย่างไรก็ตาม ผลการวิจัยของเราพบว่า การสร้างช่วงความเชื่อมั่นด้วยวิธี Paired Bootstrap Lasso+Partial Ridge มีข้อจำกัดในการวิเคราะห์ข้อมูลที่มีลักษณะ Hard Sparsity เนื่องจากให้ค่าความแม่นยำและค่าความไวต่ำกว่า 0.50 และวิธีที่มีประสิทธิภาพในการสร้างช่วงความเชื่อมั่นน้อยที่สุด ก็คือ วิธี Parametric Bootstrap Lasso+ML

ดังนั้นจึงสรุปได้ว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+Partial Ridge มีประสิทธิภาพมากกว่าวิธี Lasso+MLE ซึ่งสอดคล้องกับงานวิจัยของ Liu et al. (2020) ที่กล่าวไว้ว่า วิธี bootstrap LPR หรือวิธี Lasso+ Partial Ridge มีความกว้างของช่วงความเชื่อมั่นสั้นที่สุดและให้ค่าความน่าจะเป็นครอบคลุมที่ดี สำหรับข้อมูลที่มีค่าสัมประสิทธิ์ขนาดเล็กแต่ไม่เท่ากับศูนย์ ดังนั้นหากผู้ปฏิบัติงานเน้นที่ช่วงความเชื่อมั่นสำหรับค่าสัมประสิทธิ์การถดถอยขนาดเล็ก เช่น กรณีตัวประมาณมากเลขศูนย์ (Sparse Estimator) พวกเราขอแนะนำวิธี bootstrap LPR หรือวิธี Lasso+Partial Ridge ในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธ์การถดถอย

นอกจากนี้งานวิจัยยังแสดงข้อสรุปที่น่าสนใจอีกหนึ่งประการ คือ การประมาณค่าด้วยวิธี Paired Bootstrap เหมาะสมกับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+MLE และการประมาณค่าด้วยวิธี Parametric Bootstrap เหมาะสม

กับการใช้งานร่วมกับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธี Lasso+Partial Ridge

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ผู้ที่สนใจอาจจะนำไปศึกษาต่อได้ในอีกเรื่องของ

1. วิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติก เนื่องจากงานวิจัยนี้ได้นำมาศึกษาเพียง 4 วิธีเท่านั้น ซึ่งในความจริงแล้วอาจยังมีอีกหลายวิธีที่น่าสนใจและมีประสิทธิภาพมากกว่าทั้ง 4 วิธีที่ได้กล่าวมาแล้วข้างต้น
2. ขอบเขตของงานวิจัย เช่น ในเรื่องของการกำหนดขนาดตัวอย่าง ขนาดตัวแปรอิสระ จำนวนทำซ้ำของบูตสแตรป และการกำหนดพารามิเตอร์ที่มีการปรับค่าแล้ว (λ_1,λ_2) ด้วยวิธี fold cross validation เป็นต้น
3. กรณีที่ Y ไม่ได้มีการแจกแจงแบบทวินาม (Binomial Distribution)
4. การทำซ้ำ (Replication) ซึ่งในการรันโปรแกรมด้วยจำนวนทำซ้ำ 50 รอบ อาจไม่ใช่ค่าที่ดีที่สุด

บรรณานุกรม

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical science*, 533-558.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*: CRC press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Liu, H., Xu, X., & Li, J. J. (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.

Liu, H., & Yu, B. (2013). Asymptotic properties of Lasso+ mLS and Lasso+ Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics, 7*, 3124-3169.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106): Sage.

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

วิฐรา พึ่งพาพงศ์. (2015). บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. *Thai Science and Technology Journal*, 212-223.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่างการเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยมีการสร้างช่วงความเชื่อมั่นด้วยวิธี

- Parametric Bootstrap Lasso+MLE
- Parametric Bootstrap Lasso+Partial Ridge
- Paired Bootstrap Lasso+MLE
- Paired Bootstrap Lasso+Partial Ridge

ภายใต้ปัจจัยของการสร้างข้อมูลชุดที่ 1 ดังต่อไปนี้

1. กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

2. สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ โดยกำหนด $\sum$ ด้วยวิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ และให้ $\rho = 0.5$

3. ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จากวิธี Hard Sparsity ซึ่งจะสุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2,\ldots,p$

4. สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$


```
library(glmnet)
library(Rcpp)
library(mvtnorm)

n <- 200
p <- 500
r <- 50 #number of Replication
B <- 1000 #number of Bootstrap
alpha <- 0.05
set.seed(42) #for reproducibility
```

```
############# Simulate Sigma_X : 1 time #############
#---- Toeplitz with rho = 0.5 ----#
rho1 <- 0.5
sigma1_X <- matrix(NA,nrow=p,ncol=p,byrow=TRUE)

for(j in 1:p){
  for(i in 1:p){
    if(i!=j){sigma1_X[i,j] = rho1^abs(i-j)}
    else{sigma1_X[i,j] = 1}
  }}

############# Simulate Beta : 1 time #############
#---- Hard Sparsity ----#
num_nonzero_beta <- 10
beta_hsp <-  rep(0, 500)
pos_nonzero <- sample(1:p,num_nonzero_beta,replace=FALSE)
beta_hsp[pos_nonzero] <- runif(num_nonzero_beta,1/3,1)

############# Simulate X : 1 time #############
library(mvtnorm)
Med_X <- matrix(0,nrow=p,ncol=1)
X1 <- rmvnorm(n,Med_X,sigma1_X)

############# Simulate Y : r times #############
#----Y1_H = X1 and Hard Sparsity ----#
all_Y <-  matrix(NA,nrow=n,ncol=r) #matrix for keep all simulated Y

for(b in 1:r){
  BetaH_X1 <- X1 %*% as.matrix(beta_hsp)
  exp_betaHX1 <- exp(BetaH_X1)
  pi <- exp_betaHX1/(1+exp_betaHX1)
```

```
  Y1_H <- matrix(NA,nrow=n,ncol=1,byrow=TRUE)
  for(i in 1:n){
    Y1_H[i] = rbinom(1,size=1,prob=pi[i])}
  all_Y[,b] <- Y1_H}

########################################
####### Parametric Bootstrap Lasso+MLE #######
########################################
set.seed(42) #for reproducibility
beta_vec <-  rep(0, p) #beta vector for Parametric Bootstra
beta_vec_boot <-  rep(0, p) #beta vector for Lasso+MLE
all_beta <- matrix(ncol = p, nrow = B)
keep= array(NA, dim = c(p, B, r)) #keep all beta in 3 dimensions
all_CI <- array(NA, dim = c(p, 2, r)) #keep all Confidence interval in 3 dimensions

for(b in 1:r){
  data1 <- data.frame(all_Y[,b],X1)
  X <- data.matrix(data1[,-1])
  Y <- data1[,1]
  cv_model <- cv.glmnet(X, Y, family = "binomial", type.measure='mse', nfolds = 5,
alpha = 1) #perform 5-fold cross-validation to find optimal lambda value
  best_model <- glmnet(X, Y, family = "binomial", type.measure='mse', alpha = 1,
lambda = cv_model$lambda.min) #find coefficients of best model
  post <- (which(coef(best_model)[-1] != 0)) #position of X which has non-zero beta
  x <- X[,post]
  glm2 <- glm(Y~x, family = "binomial",maxit=50) #MLE
  beta_vec <-  rep(0, 500) #build beta vector
  beta_vec[post] = coef(glm2)[-1]
  Beta_X <- coef(glm2)[1] + X %*% as.matrix(beta_vec) #estimate pi hat and simulate
Y_star
  exp_betaX <- exp(Beta_X)
```

```
  pi <- exp_betaX/(1+exp_betaX)
  Y_star <- matrix(NA,nrow=n,ncol=B,byrow=TRUE)
  for(i in 1:n){
    Y_star[i,] = rbinom(B,size=1,prob=pi[i])}
  for (i in seq_len(B)) {
    x1 <- X
    y1 <- Y_star[,i]
   cv_mod <- cv.glmnet(x1, y1, family = "binomial", type.measure='mse', nfolds = 5,
alpha = 1) #lasso regression
    best_mod <- glmnet(x1, y1, family = "binomial", type.measure='mse', alpha = 1,
lambda = cv_mod$lambda.min)
    pos <- which(coef(best_mod)[-1] != 0) #select position of x1 which has non-zero
beta
    x2 <- x1[,pos]
    ndata <- data.frame(y1,x2)
    glm <- glm(y1~.,data=ndata, family = "binomial",maxit=50) #set maxtit = 50 >>
converge
    beta_vec_boot[pos] = coef(glm)[-1]
    all_beta[i,] <- beta_vec_boot}
  keep[,,b] <- t(all_beta) #>>>>>save keep<<<<<
  CI <- apply(t(keep[,,b]), 2, quantile, probs = c(alpha/2, 1-(alpha/2)), na.rm = TRUE)
  all_CI[,1,b] <- CI[1,] #Lower bound
  all_CI[,2,b] <- CI[2,] #Upper bound
}

# Save an object to a file
all_CI1 <- all_CI
saveRDS(all_CI1, file ="ParaBLM_CI1.rds")
```

```
#####################################################
######### Parametric Bootstrap Lasso+Partial Ridge #########
#####################################################
set.seed(42) #for reproducibility
beta_vec <-  rep(0, p) #build beta vector
all_beta <- matrix(ncol = p, nrow = B)
keep= array(NA, dim = c(p, B, r)) #keep all beta in 3 dimensions
all_CI <- array(NA, dim = c(p, 2, r)) #keep all Confidence interval in 3 dimensions

for(b in 1:r){
  data1 <- data.frame(all_Y[,b],X1)
  X <- data.matrix(data1[,-1])
  Y <- data1[,1]
  cv_model <- cv.glmnet(X, Y, family = "binomial", type.measure='mse', nfolds = 5,
alpha = 1) #perform 5-fold cross-validation to find optimal lambda value
  best_model <- glmnet(X, Y, family = "binomial", type.measure='mse', alpha = 1,
lambda = cv_model$lambda.min) #find coefficients of best model
  post <- (which(coef(best_model)[-1] != 0)) #position of X which has non-zero beta
  x <- X[,post]
  glm2 <- glm(Y~x, family = "binomial",maxit=50) #MLE
  beta_vec <-  rep(0, 500) #build beta vector
  beta_vec[post] = coef(glm2)[-1]
  Beta_X <- coef(glm2)[1] + X %*% as.matrix(beta_vec) #estimate pi hat and simulate
Y_star
  exp_betaX <- exp(Beta_X)
  pi <- exp_betaX/(1+exp_betaX)
  Y_star <- matrix(NA,nrow=n,ncol=B,byrow=TRUE)
  for(i in 1:n){
    Y_star[i,] = rbinom(B,size=1,prob=pi[i])}
  for (i in seq_len(B)) {
    x1 <- X
```

```
    y1 <- Y_star[,i]
    cv_las <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 1) #Lasso 2 th
    lasso_coef <- coef(cv_las,s=cv_las$lambda.min)[-1]
    cv_ridge <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 0) #ridge 1
th
    ridge_coef <- coef(cv_ridge,s=cv_ridge$lambda.min)[-1]
    idx <- ifelse(lasso_coef==0,1,0) #built index
    rid_cv <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 0,
penalty.factor=idx) #Fit Partial Ridge
    lass_par_rid <- coef(rid_cv,s=rid_cv$lambda.min)[-1]
    all_beta[i,] <- lass_par_rid}
  keep[,,b] <- all_beta
  CI <- apply(t(keep[,,b]), 2, quantile, probs = c(alpha/2, 1-(alpha/2)), na.rm = TRUE)
  all_CI[,1,b] <- CI[1,] #Lower bound
  all_CI[,2,b] <- CI[2,] #Upper bound
}

# Save an object to a file
all_CI2 <- all_CI
saveRDS(all_CI2, file ="ParaBLPR_CI1.rds")

##########################################
######### Paired Bootstrap Lasso+MLE #########
##########################################
set.seed(42) #for reproducibility
beta_vec <-  rep(0, p) #build beta vector
all_beta <- matrix(ncol = p, nrow = B)
keep= array(NA, dim = c(p, B, r)) #keep all beta in 3 dimensions
all_CI <- array(NA, dim = c(p, 2, r)) #keep all Confidence interval in 3 dimensions
```

```
for(b in 1:r){
  data1 <- data.frame(all_Y[,b],X1)
  for (i in seq_len(B)) {
    boot_sam <- data1[sample(n, replace = TRUE),] #bootstrap resample of data
    X <- data.matrix(boot_sam[,-1])
    Y <- boot_sam[,1]
    cv_model <- cv.glmnet(X, Y, family = "binomial", type.measure='mse', nfolds = 5,
alpha = 1) #lasso regression
    best_model <- glmnet(X, Y, family = "binomial", type.measure='mse', alpha = 1,
lambda = cv_model$lambda.min)
    post <- which(coef(best_model)[-1] != 0)
    x <- X[,post]
    newdata <- data.frame(Y,x)
    glm2 <- glm(Y~.,data=newdata, family = "binomial",maxit=50) #prove not converge
with maxit=50
    beta_vec[post] = coef(glm2)[-1]
    all_beta[i,] <- beta_vec}
  keep[,,b] <- t(all_beta)
  CI <- apply(t(keep[,,b]), 2, quantile, probs = c(alpha/2, 1-(alpha/2)), na.rm = TRUE)
  all_CI[,1,b] <- CI[1,] #Lower bound
  all_CI[,2,b] <- CI[2,] #Upper bound
}

# Save an object to a file <DONE>
all_CI3 <- all_CI
saveRDS(all_CI3, file = "PairBLM_CI1.rds")

################################################
#### Paired Bootstrap Lasso+Partial Ridge ####
################################################
set.seed(42) #for reproducibility
```

```
all_beta <- matrix(ncol = p, nrow = B)
keep= array(NA, dim = c(p, B, r)) #keep all beta in 3 dimensions
all_CI <- array(NA, dim = c(p, 2, r)) #keep all Confidence interval in 3 dimensions

for(b in 1:r){
  data1 <- data.frame(all_Y[,b],X1)
  for (i in 1:B) {
    boot_sam <- data1[sample(n, replace = TRUE),] #bootstrap resample of data
    x1 <- data.matrix(boot_sam[,-1])
    y1 <- boot_sam[,1]
    cv_las <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 1) #Lasso 2 th
    lasso_coef <- coef(cv_las,s=cv_las$lambda.min)[-1]
    cv_ridge <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 0) #ridge 1
th
    ridge_coef <- coef(cv_ridge,s=cv_ridge$lambda.min)[-1]
    idx <- ifelse(lasso_coef==0,1,0) #built index
    rid_cv <- cv.glmnet(x1, y1, type.measure='mse', nfolds = 5, alpha = 0,
penalty.factor=idx) #Fit Partial Ridge
    lass_par_rid <- coef(rid_cv,s=rid_cv$lambda.min)[-1]
    all_beta[i,] <- lass_par_rid}
  keep[,,b] <- all_beta
  CI <- apply(t(keep[,,b]), 2, quantile, probs = c(alpha/2, 1-(alpha/2)), na.rm = TRUE)
  all_CI[,1,b] <- CI[1,] #Lower bound
  all_CI[,2,b] <- CI[2,] #Upper bound
}

# Save an object to a file
all_CI4 <- all_CI
saveRDS(all_CI4, file = "PairBLPR_CI1.rds")
```

```
################################################
################## Average width ################
################################################

dat1 <- readRDS("ParaBLM_CI1.rds")
dat2 <- readRDS("ParaBLPR_CI1.rds")
dat3 <- readRDS("PairBLM_CI1.rds")
dat4 <- readRDS("PairBLPR_CI1.rds")

p <- 500 #number of Parameters
r <- 50 #number of Replications

>>>>> Data1 & Parametric Bootstrap Lasso+MLE <<<<<
diff <- matrix(ncol = r, nrow = p)
all_wid <- rep(0, r)
dat <- dat1

for(i in 1:r){
  diff[,i] <- dat[,2,i] - dat[,1,i]
  all_wid[i] <- sum(diff[,i]) / p}

wid_ParaBLM_1 <- all_wid
```

```
>>>>> Data1 & Parametric Bootstrap Lasso+Partial Ridge <<<<<
diff <- matrix(ncol = r, nrow = p)
all_wid <- rep(0, r)
dat <- dat2

for(i in 1:r){
  diff[,i] <- dat[,2,i] - dat[,1,i]
  all_wid[i] <- sum(diff[,i]) / p}

wid_ParaBLPR_1 <- all_wid

>>>>> Data1 & Paired Bootstrap Lasso+MLE <<<<<
diff <- matrix(ncol = r, nrow = p)
all_wid <- rep(0, r)
dat <- dat3

for(i in 1:r){
  diff[,i] <- dat[,2,i] - dat[,1,i]
  all_wid[i] <- sum(diff[,i]) / p}

wid_PairBLM_1 <- all_wid

>>>>> Data1 & Paired Bootstrap Lasso+Partial Ridge <<<<<
diff <- matrix(ncol = r, nrow = p)
all_wid <- rep(0, r)
dat <- dat4

for(i in 1:r){
  diff[,i] <- dat[,2,i] - dat[,1,i]
  all_wid[i] <- sum(diff[,i]) / p}
```

```
wid_PairBLPR_1 <- all_wid

################################################
############### Coverage probability ###############
################################################
r <- 50
#---- Hard Sparsity ----# Simulate Beta : 1 time
set.seed(42) #for reproducibility
num_nonzero_beta <- 10
beta_hsp <-  rep(0, 500)
pos_nonzero <- sample(1:p,num_nonzero_beta,replace=FALSE)
beta_hsp[pos_nonzero] <- runif(num_nonzero_beta,1/3,1)

>>>>>> Data1 & Parametric Bootstrap Lasso+MLE <<<<<
dat <- dat1
cpvec <- rep(0,r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  cpvec[i] <- sum(all_lo[,i] <= beta_hsp & beta_hsp <= all_up[,i]) / p
  }

cp_ParaBLM_1 <- cpvec

>>>>> Data1 & Parametric Bootstrap Lasso+Partial Ridge <<<<<
dat <- dat2
cpvec <- rep(0,r)

for(i in 1:r){
  all_lo <- dat[,1,]
```

```
  all_up <- dat[,2,]
  cpvec[i] <- sum(all_lo[,i] <= beta_hsp & beta_hsp <= all_up[,i]) / p}

cp_ParaBLPR_1 <- cpvec
>>>>> Data1 & Paired Bootstrap Lasso+MLE <<<<<
dat <- dat3
cpvec <- rep(0,r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  cpvec[i] <- sum(all_lo[,i] <= beta_hsp & beta_hsp <= all_up[,i]) / p
}

cp_PairBLM_1 <- cpvec

>>>>> Data1 & Paired Bootstrap Lasso+Partial Ridge <<<<<
dat <- dat4
cpvec <- rep(0,r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  cpvec[i] <- sum(all_lo[,i] <= beta_hsp & beta_hsp <= all_up[,i]) / p
}

cp_PairBLPR_1 <- cpvec
```

```
################################################
################ Precision & Recall ##############
################################################

#---- Hard Sparsity ----# Simulate Beta : 1 time
set.seed(42)
p <- 500
r <- 50
num_nonzero_beta <- 10
beta_hsp <-  rep(0, 500)
pos_nonzero <- sample(1:p,num_nonzero_beta,replace=FALSE)
beta_hsp[pos_nonzero] <- runif(num_nonzero_beta,1/3,1)

>>>>>> Data1 & Parametric Bootstrap Lasso+MLE <<<<<
dat <- dat1
all_pos <- 1:500
s <- pos_nonzero
Precis <- rep(0, r)
Recall <- rep(0, r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  shat <- which(all_up[,i] < 0 | all_lo[,i] > 0)
  Precis[i] <- length(intersect(shat, s))/length(shat)
  Recall[i] <- length(intersect(shat, s))/length(s)
}

Precis_ParaBLM_1 <- Precis
Recall_ParaBLM_1 <- Recall
```

```
>>>>>> Data1 & Parametric Bootstrap Lasso+Partial Ridge <<<<<
dat <- dat2
all_pos <- 1:500
s <- pos_nonzero
Precis <- rep(0, r)
Recall <- rep(0, r)
for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  shat <- which(all_up[,i] < 0 | all_lo[,i] > 0)
  Precis[i] <- length(intersect(shat, s))/length(shat)
  Recall[i] <- length(intersect(shat, s))/length(s)
}

Precis_ParaBLPR_1 <- Precis
Recall_ParaBLPR_1 <- Recall

>>>>>> Data1 & Paired Bootstrap Lasso+MLE <<<<<
dat <- dat3
all_pos <- 1:500
s <- pos_nonzero
Precis <- rep(0, r)
Recall <- rep(0, r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  shat <- which(all_up[,i] < 0 | all_lo[,i] > 0)
  Precis[i] <- length(intersect(shat, s))/length(shat)
  Recall[i] <- length(intersect(shat, s))/length(s)
}
```

```
Precis_PairBLM_1 <- Precis
Recall_PairBLM_1 <- Recall

>>>>>> Data1 & Paired Bootstrap Lasso+Partial Ridge <<<<<
dat <- dat4
all_pos <- 1:500
s <- pos_nonzero
Precis <- rep(0, r)
Recall <- rep(0, r)

for(i in 1:r){
  all_lo <- dat[,1,]
  all_up <- dat[,2,]
  shat <- which(all_up[,i] < 0 | all_lo[,i] > 0)
  Precis[i] <- length(intersect(shat, s))/length(shat)
  Recall[i] <- length(intersect(shat, s))/length(s)
}

Precis_PairBLPR_1 <- Precis
Recall_PairBLPR_1 <- Recall
```

ประวัติผู้เขียน

ชื่อ-สกุล	ณิชากร ไทยวงษ์
วัน เดือน ปี เกิด	24 มีนาคม 2539
สถานที่เกิด	ขอนแก่น
วุฒิการศึกษา	ปริญญาตรี
ที่อยู่ปัจจุบัน	ห้อง 401 ไดมอนด์ บางกอก 1/1 ซอยกิ่งเพชร ถนนเพชรบุรี แขวงถนนเพชรบุรี เขตราชเทวี กรุงเทพ 10400