การเปรียบเทียบวิธีการแบ่งข้อมูลอย่างสุ่ม และวิธีบูตสแตรปในการปรับค่าพี-แวลูของสัมประสิทธิ์การถดถอยที่มีมิติสูง

นางสาวบงกชพร เนาวนัติ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

A COMPARISON ON P-VALUE ADJUSTMENT BETWEEN RANDOM – SPLIT AND BOOTSTRAP METHODS IN HIGH DIMENSIONAL REGRESSION

Miss Bongkotchaporn Nouwanut

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการแบ่งข้อมูลอย่างสุ่ม และวิธีบูตสแตรปในการปรับค่าพี-แวลูของสัมประสิทธิ์การถดถอยที่มีมิติสูง
โดย	นางสาวบงกชพร เนาวนัติ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร. วิฐรา พึ่งพาพงศ์)

.....กรรมการ
(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

.....กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. อรุณี กำลัง)

บงกชพร เนาวนัติ : การเปรียบเทียบวิธีการแบ่งข้อมูลอย่างสุ่ม และวิธีบูตสแตรปในการปรับค่าพี-แวลูของสัมประสิทธิ์การถดถอยที่มีมิติสูง. (A COMPARISON ON P-VALUE ADJUSTMENT BETWEEN RANDOM – SPLIT AND BOOTSTRAP METHODS IN HIGH DIMENSIONAL REGRESSION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. วิฐรา พึ่งพาพงศ์, 80 หน้า.

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบแนวทางในการเลือกใช้วิธี Random Split และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง อีกทั้งเพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรระหว่างวิธี Random Split และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง ซึ่งเกณฑ์ที่ใช้ในการเปรียบเทียบ คือจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว โดยข้อมูลที่ใช้ในการศึกษาได้จากการจำลองข้อมูลโดยมีขนาดตัวอย่างต่อจำนวนตัวแปรอิสระเป็น 10:20, 10:50, 10:100, 100:200, 100:500, 100:1,000, 200:400, 200:1,000 และ 200:2,000 ตามลำดับด้วยจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ 0.1 เท่า, 0.25 เท่า และ 0.45 เท่าของขนาดตัวอย่างที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9

จากผลการศึกษาโดยเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก พบว่าการแบ่งข้อมูลด้วยวิธี Random Split มีประสิทธิภาพในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงมากกว่าการแบ่งข้อมูลด้วยวิธีบูตสแตรป แต่ในแง่ของจำนวนความผิดพลาดในการตรวจจับเชิงลบและจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว พบว่ากรณีส่วนใหญ่การแบ่งข้อมูลด้วยวิธีบูตสแตรปจะมีประสิทธิภาพในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงมากกว่าการแบ่งข้อมูลด้วยวิธี Random Split

ภาควิชา สถิติ
สาขาวิชา สถิติ
ปีการศึกษา 2556

ลายมือชื่อนิสิต .....
ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

5581561926 : MAJOR STATISTICS
KEYWORDS: HIGH DIMENSIONAL DATA / FALSE DISCOVERY RATE / MULTI-SPLITTING / BOOTSTRAP

BONGKOTCHAPORN NOUWANUT: A COMPARISON ON P-VALUE ADJUSTMENT BETWEEN RANDOM – SPLIT AND BOOTSTRAP METHODS IN HIGH DIMENSIONAL REGRESSION. ADVISOR: VITARA PUNGPAPONG, Ph.D., 80 pp.

The objective of this research is to study and compare on p-value adjustment between Random – Split and Bootstrap methods in high dimensional regression, include studying and comparing efficiency in variable selection on p-value adjustment between Random – Split and Bootstrap methods in high dimensional regression. The number of false positive, the number of false negative and the number of nonzero coefficient are three criteria using for comparison. The data in this study under several situations which are the ratio of sample size to the number of independent variables are 10:20, 10:50, 10:100, 100:200, 100:500, 100:1,000, 200:400, 200:1,000 and 200:2,000 with true nonzero coefficients are 0.1, 0.25 and 0.45 of sample size which correlation level of independent variables are 0, 0.5 and 0.9

Based on the simulation results by comparing the number of false positive show that data splitting with Random – Split method is more efficient than Bootstrap method on p-value adjustment in high dimensional regression. However, the number of false negative and the number of nonzero coefficients, overall, data splitting with Bootstrap method is more efficient than Random – Split method on p-value adjustment in high dimensional regression.

Department: Statistics Student's Signature .....

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

Academic Year: 2013

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของอาจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ คำปรึกษา ตลอดจนช่วยเหลือแก้ไขข้อบกพร่องต่าง ๆ จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ ผู้วิจัยขอกราบขอบพระคุณและสำนึกในพระคุณเป็นอย่างยิ่ง

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.ธีระพร วีระถาวร ในฐานะประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ในฐานะกรรมการสอบวิทยานิพนธ์ และอาจารย์ ดร. อรุณี กำลัง ในฐานะกรรมการภายนอกสอบวิทยานิพนธ์ ที่กรุณาตรวจแก้วิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ที่ให้โอกาสทางการศึกษาและให้ความรู้แก่ผู้วิจัยจนกระทั่งสำเร็จการศึกษา

สุดท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา และครอบครัว ซึ่งสนับสนุนด้านการศึกษาและให้กำลังใจแก่ผู้วิจัยเสมอมาจนกระทั่งสำเร็จการศึกษา ตลอดจนเพื่อน ๆ ทุกคนที่ให้คำปรึกษาและเป็นกำลังให้ด้วยดีมาโดยตลอด

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันงานวิจัยด้านต่าง ๆ เกือบทุกแขนงทั้งด้านสังคมศาสตร์ วิทยาศาสตร์ หรือทางการแพทย์ โดยทั่วไปมักมีชุดข้อมูลขนาดใหญ่อยู่เป็นจำนวนมาก อีกทั้งประสิทธิภาพในการจัดเก็บข้อมูลที่สูงขึ้นประกอบกับต้นทุนในการจัดเก็บที่ต่ำลง ทั้งในด้านฮาร์ดแวร์และระบบจัดการฐานข้อมูล ซึ่งในบางกรณีข้อมูลดังกล่าวอาจถูกจัดเก็บโดยที่จำนวนตัวแปร (p) มากกว่าขนาดตัวอย่าง (n) ตัวอย่างเช่น Microarray (Heller, 2002) จะเห็นได้จากด้านการแพทย์ที่ต้องการศึกษาการพัฒนาเทคโนโลยีดีเอ็นเอนไมโครอาร์เรย์ซึ่งมีตัวอย่างที่ต้องการทดลองเท่ากับ 43 ตัวอย่างการทดลองและมีจำนวนยีนซึ่งเป็นตัวแปรอิสระเท่ากับ 4205 ยีน จากตัวอย่างที่กล่าวมาข้างต้นจะเห็นว่าข้อมูลที่ได้มาเป็นข้อมูลที่จำนวนยีนมากกว่าตัวอย่างการทดลอง หรืออีกแง่หนึ่งคือข้อมูลที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างซึ่งในที่นี้จะเรียกข้อมูลลักษณะดังกล่าวว่าข้อมูลที่มีมิติสูง (High-dimensional data)

เนื่องมาจากในปัจจุบันประสิทธิภาพในการจัดเก็บข้อมูลสูงขึ้นทำให้การนำข้อมูลที่มีมิติสูงมาใช้ในการวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) เป็นปัญหาที่ได้รับความสนใจเป็นอย่างมาก กล่าวคือโดยทั่วไปการประมาณค่าสัมประสิทธิ์ในสมการการถดถอยสามารถทำได้โดยวิธีการประมาณกำลังสองน้อยที่สุด (Ordinary Least Squares Estimation; OLS) ซึ่งจำเป็นที่จะต้องมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ จึงจะสามารถหาตัวประมาณด้วยวิธีนี้ได้ แต่เนื่องจากในที่นี้จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง ทำให้ไม่สามารถใช้วิธีการประมาณกำลังสองน้อยที่สุดประมาณค่าสัมประสิทธิ์ในสมการการถดถอยได้ ด้วยเหตุนี้จึงมีนักสถิติหลายคนได้เสนอวิธีการประมาณค่าสัมประสิทธิ์ในสมการการถดถอยสำหรับข้อมูลที่มีมิติสูง อาทิเช่น

วิธี Least absolute shrinkage and selection operator (Lasso) นำเสนอโดย Tibshirani (1996) กล่าวว่า โดยทั่วไปการใช้วิธีการประมาณกำลังสองน้อยที่สุดในการประมาณค่าจะได้ตัวประมาณค่าที่มีความเอนเอียง (Bias) ต่ำ แต่มีความแปรปรวน (Variance) ค่อนข้างสูง แต่การประมาณด้วยวิธี Lasso จะไปเพิ่มความเอนเอียงให้มากขึ้นเพื่อทำให้ความแปรปรวนลดลง ด้วยเหตุนี้ตัวประมาณวิธีนี้จึงมีความเอนเอียงทำให้ค่าสัมประสิทธิ์ส่วนใหญ่ที่ได้จากวิธีนี้มีค่าเป็นศูนย์ ดังนั้นวิธี Lasso จึงเป็นวิธีที่สามารถเลือกตัวแปรและประมาณค่าสัมประสิทธิ์การถดถอยได้ในขณะเดียวกัน

จากข้อเสียของตัวประมาณ Lasso ที่มีความเอนเอียง Zou (2006) ได้นำเสนอวิธี Adaptive Least absolute shrinkage and selection operator (Adaptive Lasso) ซึ่งเป็นวิธีที่พัฒนามาจากวิธี Lasso เพียงแต่เพิ่มค่าถ่วงน้ำหนักที่แตกต่างกันของพารามิเตอร์แต่ละตัวทำให้ตัวประมาณที่ได้จากวิธีนี้ไม่มีความเอนเอียงและมีคุณสมบัติ Oracle (Oracle Property) กล่าวคือคุณสมบัติที่ตัวประมาณสามารถคัดเลือกตัวแปรได้ถูกต้องเสมือนกับว่าทราบตัวแบบที่แท้จริงกรณีที่ขนาดตัวอย่างเข้าใกล้อนันต์

วิธีการประมาณค่าสัมประสิทธิ์ทั้งสองวิธีข้างต้นจะได้เฉพาะค่าประมาณของสัมประสิทธิ์การถดถอยเท่านั้นไม่สามารถตรวจสอบนัยสำคัญของตัวแปรได้ ซึ่งในการศึกษาครั้งนี้ผู้วิจัยมีความสนใจในการตรวจสอบนัยสำคัญของตัวแปรโดยการคำนวณค่า p-value ของสัมประสิทธิ์การถดถอยแต่ละตัว

ในปี 2008 Wasserman and Roeder ได้ทำการศึกษาการคำนวณค่า p-value โดยการแบ่งข้อมูลเพียงครั้งเดียว ซึ่งแบ่งข้อมูลออกเป็น 2 ส่วนเท่า ๆ กัน ส่วนแรกใช้ในการคัดกรองตัวแปร (Variable Screening) ซึ่งสามารถลดมิติของตัวแปรได้และส่วนที่สองใช้ในการคำนวณค่า p-value โดยใช้การประมาณกำลังสองน้อยที่สุด

ต่อมา Meinshausen, Meier et al. (2009) พบว่าการแบ่งข้อมูลเพียงครั้งเดียวอาจทำให้ค่า p-value ที่ได้มาเกิดความเอนเอียง Meinshausen et al. จึงได้เสนอวิธี Multi – Split ในการคำนวณค่า p-value สำหรับสัมประสิทธิ์การถดถอยที่มีมิติสูง ซึ่งวิธีดังกล่าวมีประสิทธิภาพมากกว่าการแบ่งข้อมูลเพียงครั้งเดียว โดยวิธีนี้จะแบ่งข้อมูลออกเป็น 2 ชุดอย่างสุ่ม (Random Split) แต่ละชุดมีขนาดตัวอย่างเท่ากันกล่าวคือขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ $\frac{n}{2}$ โดยที่ข้อมูลชุดแรกจะใช้ในการคัดกรองตัวแปรซึ่งจะสามารถลดมิติของตัวแปรได้ สำหรับข้อมูลชุดที่สองจะใช้ในการคำนวณค่า p–value โดยใช้การประมาณกำลังสองน้อยที่สุดสำหรับตัวแปรที่ผ่านการคัดกรอง ทั้งนี้ Meinshausen et al. เสนอให้แบ่งข้อมูลออกเป็น 2 ชุดหลาย ๆ ครั้ง รวมถึงได้นำเสนอวิธีในการรวม p–value ของสัมประสิทธิ์การถดถอยแต่ละตัวที่ได้จากการแบ่งข้อมูลหลาย ๆ ครั้ง (Multi – Split)

อย่างไรก็ตามวิธีการแบ่งข้อมูลข้างต้น พบว่ามีข้อจำกัดซึ่งอาจทำให้ประสิทธิภาพในการแบ่งข้อมูลลดลงดังนั้นในการศึกษาครั้งนี้ผู้วิจัยได้ทำการศึกษาวิธีการแบ่งข้อมูลเพื่อขจัดข้อจำกัดดังกล่าว โดยมีงานวิจัยที่ศึกษา ดังนี้

วิธีบูตสแตรป นำเสนอโดย Efron (1979)Efron (1979) ซึ่งเป็นวิธีการหนึ่งที่มีผู้นำไปใช้กันอย่างแพร่หลายโดยมีหลักเกณฑ์ คือตัวอย่างที่ถูกเก็บรวบรวมเปรียบเสมือนประชากรจริงจึงทำการสุ่มตัวอย่างด้วยจำนวนครั้งที่มากพอโดยที่แต่ละหน่วยตัวอย่างมีโอกาสถูกเลือกเท่า ๆ กันซึ่งวิธีบูตสแตรปสามารถแบ่งออกเป็นแบบใช้พารามิเตอร์และไม่ใช้พารามิเตอร์ กล่าวคือ

วิธีบูตสแตรปแบบใช้พารามิเตอร์จะต้องทราบการแจกแจงของประชากรที่สุ่มตัวอย่างมา แต่แบบไม่ใช้พารามิเตอร์จะจำลองค่าสังเกตโดยใช้ฟังก์ชันการแจกแจงเชิงประจักษ์ นั่นคือเป็นการสุ่มตัวอย่างแบบคืนที่ (Resampling with replacement) โดยที่มีหน่วยตัวอย่างซ้ำกันได้

ในการศึกษาครั้งนี้ผู้วิจัยต้องการเปรียบเทียบวิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรปในการปรับค่า p–value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยใช้การจำลองข้อมูลที่มีอัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระที่ต่างกัน และจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ในกรณีที่แตกต่างกันตามสัดส่วนของจำนวนตัวแปรอิสระ เพื่อตรวจสอบดูว่าวิธีในการแบ่งข้อมูลแบบใดให้ผลดีกว่ากัน

1.2 วัตถุประสงค์

1. เพื่อศึกษาและเปรียบเทียบแนวทางในการเลือกใช้วิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
2. เพื่อศึกษาและเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรระหว่างวิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

1.3 ขอบเขตของการวิจัย

1. ตัวแบบที่ใช้ในการศึกษาคือ ตัวแบบการถดถอยเชิงเส้น หมายถึง สมการที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระที่อยู่ในรูป

$$\mathrm{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathrm{Y} = \begin{bmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \vdots \\ \mathrm{Y}_n \end{bmatrix}, \mathbf{X} = \begin{pmatrix} \mathrm{X}_1 \\ \vdots \\ \mathrm{X}_n \end{pmatrix} \text{เมื่อ } \mathrm{X}_i = (\mathrm{X}_{i1} \quad \cdots \quad \mathrm{X}_{ip});\ i = 1,\ldots,n,\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{X} \sim \mathrm{N}(0, \sum) \text{ และ } \varepsilon \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I_n}) \text{ เมื่อ } \sigma^2 = 1$$

$$\text{โดยที่ } \sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix} \text{ และ } \rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$$

เมื่อ n แทน ขนาดตัวอย่าง

p แทน จำนวนตัวแปรอิสระ

Y แทน เวกเตอร์ของตัวแปรตามขนาด $n\times 1$

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n\times p$

β แทน เวกเตอร์ของพารามิเตอร์ในสมการถดถอยขนาด $p\times 1$

ε แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $n\times 1$

2. ศึกษาภายใต้อัตราส่วนของขนาดตัวอย่าง (n) กับจำนวนตัวแปรอิสระ (p)

กรณีที่ 1: $n=10$

เปรียบเทียบอัตราส่วน 1 : 2 = 10 : 20

เปรียบเทียบอัตราส่วน 1 : 5 = 10 : 50

เปรียบเทียบอัตราส่วน 1 : 10 = 10 : 100

กรณีที่ 2 : $n=100$

เปรียบเทียบอัตราส่วน 1 : 2 = 100 : 200

เปรียบเทียบอัตราส่วน 1 : 5 = 100 : 500

เปรียบเทียบอัตราส่วน 1 : 10 = 100 : 1,000

กรณีที่ 3 : $n=200$

เปรียบเทียบอัตราส่วน 1 : 2 = 200 : 400

เปรียบเทียบอัตราส่วน 1 : 5 = 200 : 1,000

เปรียบเทียบอัตราส่วน 1 : 10 = 200 : 2,000

หมายเหตุ: การที่กำหนดอัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระเป็น 1:2, 1:5 และ 1:10 เพื่อแบ่งระดับเป็นขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ตามลำดับ

3. ศึกษาภายใต้เงื่อนไขจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์คือ 0.1 เท่า, 0.25 เท่าและ 0.45 เท่าของขนาดตัวอย่าง

4. ศึกษาภายใต้เงื่อนไขระดับความสัมพันธ์ของตัวแปรอิสระ

$$\mathbf{X}\sim \mathrm{N}(0,\sum) \text{ เมื่อ } \sigma^2=1$$

โดยที่ $\sum=\begin{pmatrix}\rho_{11} & \cdots & \rho_{1p}\\ \vdots & \ddots & \vdots\\ \rho_{p1} & \cdots & \rho_{pp}\end{pmatrix}$ และ $\rho_{ij}=\begin{cases}1; i=j\\ \rho^{|i-j|}; i\neq j\end{cases}$

โดยจะศึกษาภายใต้ค่า ρ 3 กรณี ดังนี้

กรณีที่ 1 $\rho=0$

กรณีที่ 2 $\rho=0.5$

กรณีที่ 3 $\rho=0.9$

โดยจะแบ่งระดับความสัมพันธ์ ดังนี้

ระดับความสัมพันธ์ $\rho = 0$ กรณีตัวแปรอิสระเป็น 20, 50, 100, 200, 400, 500, 1000 และ 2000 จะได้ว่า $\rho_{ij} = 0$ เท่ากันในทุกกรณี

ระดับความสัมพันธ์	จำนวนตัวแปรอิสระ	$\rho_{ij}; i \neq j$
$\rho = 0.5$	20	$\rho_{ij} \in [1.9\times10^{-6}, 0.5]$
	50	$\rho_{ij} \in [1.7\times10^{-15}, 0.5]$
	100	$\rho_{ij} \in [1.6\times10^{-30}, 0.5]$
	200	$\rho_{ij} \in [1.2\times10^{-60}, 0.5]$
	400	$\rho_{ij} \in [0, 0.5]$
	500	$\rho_{ij} \in [0, 0.5]$
	1000	$\rho_{ij} \in [0, 0.5]$
	2000	$\rho_{ij} \in [0, 0.5]$
$\rho = 0.9$	20	$\rho_{ij} \in [0.14, 0.9]$
	50	$\rho_{ij} \in [0.0057, 0.9]$
	100	$\rho_{ij} \in [0.000029, 0.9]$
	200	$\rho_{ij} \in [7.8\times10^{-10}, 0.9]$
	400	$\rho_{ij} \in [5.5\times10^{-19}, 0.9]$
	500	$\rho_{ij} \in [1.4\times10^{-23}, 0.9]$
	1000	$\rho_{ij} \in [1.9\times10^{-46}, 0.9]$
	2000	$\rho_{ij} \in [3.4\times10^{-92}, 0.9]$

5. ในการศึกษาครั้งนี้จะจำลองข้อมูลตามขอบเขตงานวิจัยข้างต้น ซึ่งผู้วิจัยจะประมวลผลโดยใช้โปรแกรม R เวอร์ชั่น 3.0.2 โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำของข้อมูลแต่ละกรณีไว้ที่จำนวน 100 รอบ

1.4 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินว่าวิธีการแบ่งข้อมูลวิธีใดมีความเหมาะสมในการหาค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงระหว่างวิธี Random Split และวิธีบูตสแตรป คือจำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive), จำนวนความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน โดยใช้ข้อมูลที่จำลองขึ้นมาโดยทั้ง 3 เกณฑ์ข้างต้นจะทำการเปรียบเทียบระหว่างสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัวโดยควบคุม FDR

กำหนดให้ $S = \{j; \beta_j \neq 0\}$

$\hat{S} = \{j;$ ปฏิเสธ H_0: $\beta_j = 0\}$

(1.) จำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive) คือจำนวนที่เกิดความผิดพลาดจากค่าสัมประสิทธิ์การถดถอยที่มีค่าไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์ ซึ่งสามารถหาได้จาก

$$\text{False Positive} = \left|\hat{S} \cap S^c\right|^1$$

(2.) จำนวนความผิดพลาดในการตรวจจับเชิงลบ (False Negative) คือจำนวนที่เกิดความผิดพลาดจากค่าสัมประสิทธิ์การถดถอยที่มีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์ ซึ่งสามารถหาได้จาก

$$\text{False Negative} = \left|\hat{S}^c \cap S\right|^*$$

(3.) จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว หรือ $\left|\hat{s}\right|$

1.5 คำจำกัดความที่ใช้ในการวิจัย

1. ข้อมูลที่มีมิติสูง (High - Dimensional Data) คือ ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าจำนวนขนาดตัวอย่าง $(p > n)$
2. การแบ่งข้อมูลอย่างสุ่ม (Random Split) คือ การแบ่งข้อมูลออกเป็นสองชุดอย่างสุ่มแต่ละชุดมีขนาดตัวอย่างเท่ากัน
3. การแบ่งข้อมูลหลาย ๆ ครั้ง (Multi Split) คือ การแบ่งข้อมูลออกเป็นสองชุดโดยจะทำซ้ำหลาย ๆ รอบ
4. อัตราส่วนความผิดพลาดที่เกิดขึ้น (False Discovery Rate : FDR) คือ ค่าคาดหวังของสัดส่วนของการปฏิเสธที่ผิดพลาดระหว่างการปฏิเสธทั้งหมด
5. ความผิดพลาดในการตรวจจับเชิงบวก (False Positive) คือ การปฏิเสธ สิ่งที่เป็นจริง
6. ความผิดพลาดในการตรวจจับเชิงลบ (False Negative) คือการไม่ปฏิเสธ สิ่งที่เป็นเท็จ

[1] |A| คือ จำนวนสมาชิกในเซต A

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อใช้เป็นแนวทางในการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรระหว่างวิธีการแบ่งข้อมูลอย่างสุ่ม และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
2. เพื่อใช้เป็นแนวทางในการเลือกใช้วิธี Random Split และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงได้อย่างเหมาะสม

1.7 วิธีดำเนินการวิจัย

1. ศึกษาค้นคว้าเอกสารต่าง ๆ ทั้งทฤษฎีที่เกี่ยวข้อง และงานวิจัยที่เกี่ยวกับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
2. จำลองข้อมูลตามขอบเขตที่ต้องการศึกษา
   - สร้างข้อมูลที่มีขนาดตัวอย่าง (n) และจำนวนตัวแปรอิสระ (p) ตามที่กำหนด
   - กำหนดระดับความสัมพันธ์ของตัวแปรอิสระ

$$\mathbf{X} \sim \mathrm{N}(0, \sum)$$

$$\text{โดยที่ } \sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix} \text{ และ } \rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$$

เมื่อ $\rho = 0$, $\rho = 0.5$ และ $\rho = 0.9$

- กำหนดจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1 เท่า, 0.25 เท่า และ 0.45 เท่าของขนาดตัวอย่างแบบสุ่ม
- กำหนดค่า β ที่ไม่เท่ากับศูนย์ โดยสุ่มจำนวน β ที่ไม่เท่ากับศูนย์จากการแจกแจงยูนิฟอร์ม $\mathrm{U} \sim [0.5, 5]$
- สร้างข้อมูลที่มีการแจกแจงปกติ

$$\mathrm{Y} = \mathbf{X}\beta + \varepsilon \ ; \mathbf{X} \sim \mathrm{N}(0, \sum) \text{ และ } \varepsilon \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I_n}) \text{ เมื่อ } \sigma^2 = 1$$

$$\text{โดยที่ } \sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix} \text{ และ } \rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$$

เมื่อ n แทน ขนาดตัวอย่าง

p แทน จำนวนตัวแปรอิสระ

Y แทน เวกเตอร์ของตัวแปรตามขนาด $n \times 1$

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n \times p$

β แทน เวกเตอร์ของพารามิเตอร์ในสมการถดถอยขนาด $p \times 1$

ε แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $n \times 1$

3. นำข้อมูลที่ได้ในแต่ละชุดมาหาค่า p-value โดยวิธี Multi–Split ซึ่งจะแบ่งข้อมูลออกเป็น 2 ชุดหลาย ๆ ครั้ง ซึ่งในที่นี้จะทำการแบ่งข้อมูลจำนวน 50 รอบ โดยแต่ละรอบจะแบ่งข้อมูลโดยใช้วิธี

- Random Split โดยแบ่งข้อมูลออกเป็นสองชุดแต่ละชุดมีขนาดตัวอย่างเท่ากัน กล่าวคือขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ $\frac{n}{2}$
- วิธีบูตสแตรป โดยแบ่งข้อมูลออกเป็นสองชุดแต่ละชุดมีขนาดตัวอย่างเท่ากัน กล่าวคือขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ n เนื่องจากเป็นการสุ่มแบบไม่คืนที่ ควบคุม FDR สำหรับค่า p-value ที่ได้เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

4. หลังจากนั้นนำข้อมูลที่ได้จากการแบ่งข้อมูลมาคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และหาค่า p-value แล้วจึงปรับค่า p-value และรวมค่า p-value จากการทำซ้ำ

5. ควบคุม FDR สำหรับค่า p-value ที่ได้เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

6. คำนวณค่าความผิดพลาดในการตรวจจับเชิงบวก, ความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว

7. วิเคราะห์และสรุปผลการศึกษา

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

โดยทั่วไปการประมาณค่าสัมประสิทธิ์ในสมการการถดถอยสามารถทำได้โดยวิธีการประมาณกำลังสองน้อยที่สุด (Ordinary Least Square: OLS) ซึ่งจำเป็นที่จะต้องมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ แต่เนื่องจากในปัจจุบันประสิทธิภาพในการจัดเก็บข้อมูลสูงขึ้นประกอบกับต้นทุนในการจัดเก็บที่ต่ำลงทำให้ในบางกรณีข้อมูลดังกล่าวอาจถูกเก็บโดยที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง ทำให้ไม่สามารถใช้วิธีการประมาณกำลังสองน้อยที่สุดประมาณค่าสัมประสิทธิ์ในสมการการถดถอยได้สำหรับข้อมูลลักษณะดังกล่าว ด้วยเหตุนี้จึงใช้วิธีการประมาณค่าสัมประสิทธิ์ในสมการการถดถอยสำหรับข้อมูลที่ลักษณะตามที่กล่าวข้างต้น นั่นคือ วิธี Lasso และวิธี Adaptive Lasso แต่เนื่องจากวิธีทั้งสองจะได้เฉพาะค่าประมาณของสัมประสิทธิ์การถดถอยเท่านั้นไม่สามารถตรวจสอบนัยสำคัญของตัวแปรได้ ดังนั้นในบทนี้กล่าวถึงรายละเอียดและลักษณะทั่วไปของการแบ่งข้อมูลในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงด้วย 2 วิธีคือ การแบ่งข้อมูลโดยใช้วิธี Random Split และการแบ่งข้อมูลโดยใช้วิธีบูตสแตรปภายใต้การหาค่า p-value โดยวิธี Multi-Split ภายใต้การควบคุม False Discovery Rate (FDR) เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

2.1 ปัญหาการทดสอบสมมติฐานของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูง

การวิเคราะห์การถดถอย หมายถึงการศึกษาเกี่ยวกับความสัมพันธ์เชิงเส้นระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป สำหรับวิธีการประมาณสัมประสิทธิ์การถดถอยที่นิยมนำมาใช้กันมากที่สุด คือ วิธีการประมาณกำลังสองน้อยที่สุด

2.1.1 วิธีการประมาณกำลังสองน้อยที่สุด (Ordinary Least Square: OLS)

วิธีกำลังสองน้อยที่สุด ถูกนำเสนอโดย Friedrich Gauss โดยมีแนวคิดคือการสร้างสมการที่ใช้พยากรณ์ค่าของตัวแปรตอบสนองสำหรับค่าของตัวแปรพยากรณ์ที่สนใจบางค่า ดังนั้นจึงจำเป็นต้องหาตัวแบบที่เหมาะสมกับค่าที่สังเกตของ Y กับค่าของ X_i ที่ทราบค่าแล้ว นั่นคือ เราต้องหาค่าของสัมประสิทธิ์การถดถอย β ที่สอดคล้องกับข้อมูลที่มีอยู่

ให้ $\tilde{\mathbf{X}} = (1_n, X)$ โดยที่ 1_n คือ เวกเตอร์ที่มีค่า 1 ขนาด n

ในความเป็นจริงเราไม่สามารถหาค่าพารามิเตอร์ $\beta_0, \beta_1, ..., \beta_p$ ได้ แต่สามารถประมาณค่าพารามิเตอร์เหล่านี้ได้โดยใช้ค่าสังเกตของตัวอย่างนั่นคือ ถ้า $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ เป็นค่าประมาณของ $\beta_0, \beta_1, ..., \beta_p$ ตามลำดับ ซึ่งการหาค่าประมาณ $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ โดยวิธีกำลังสองน้อยที่สุดทำได้โดยการหาอนุพันธ์ของ $\sum e_i^2 = e'e$ เทียบกับ $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ แล้วกำหนดให้เท่ากับศูนย์ ซึ่งทำให้ได้สมการปกติ (Normal Equation)

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\hat{\beta} = \tilde{\mathbf{X}}'Y \tag{2.1}$$

ในกรณีที่ $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ ไม่ใช่เมตริกซ์เอกฐาน (Nonsingular Matrix) นั่นคือสามารถหา $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ ได้ ดังนั้น

$$\hat{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Y \tag{2.2}$$

2.1.2 การทดสอบสมมติฐานของสัมประสิทธิ์การถดถอย

เป็นการทดสอบว่าสัมประสิทธิ์การถดถอยแต่ละตัวเท่ากับศูนย์หรือไม่

สมมติฐานเพื่อการทดสอบ คือ

H_0: $\beta_j = 0$ เมื่อ $j = 1, ..., p$

H_1: $\beta_j \neq 0$ เมื่อ $j = 1, ..., p$

2.1.3 ตัวสถิติทดสอบ

การทดสอบ t (t-test) เป็นวิธีการนำเสนอโดยวิลเลี่ยม สตีลเลย์ ก็อตเซท ในปี 1908 เพื่อหาข้อสรุปของค่าเฉลี่ยของประชากรจากกลุ่มตัวอย่าง ดังนั้นการแจกแจงแบบที (t-distribution) จึงได้ชื่อว่า การแจกแจงแบบทีของสติวเดนต์ จะได้ว่า

$$t = \frac{\hat{\beta}_j - 0}{s\sqrt{c_{jj}}} \tag{2.3}$$

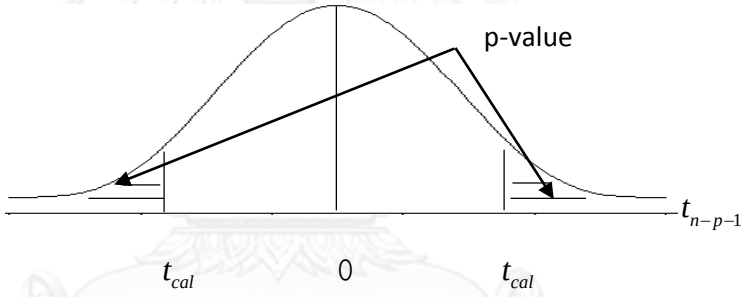
มีองศาอิสระ $n - p - 1$

โดย $s = \sqrt{\frac{e'e}{n-p-1}}$ และ c_{jj} ค่าตำแหน่งที่ j ในแนวทแยงของเมทริกซ์ $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$

เมื่อ $e = \mathrm{Y} - \hat{\mathrm{Y}}$ เป็นเวกเตอร์ของเศษตกค้าง

2.1.4 P-Value

โดยทั่วไป เมื่อเราต้องการสรุปผลการทดสอบสมมติฐานนั้น เราจะสนใจว่าสมมติฐานว่างถูกปฏิเสธหรือไม่ ถ้าไม่ปฏิเสธสมมติฐานว่าง นั่นก็แปลว่าสมมติฐานว่างเป็นจริง โดย p-value คือความน่าจะเป็นที่ค่าสถิติทดสอบมีค่าเกินกว่าสถิติทดสอบที่คำนวณได้จากข้อมูลตัวอย่างภายใต้สมมติฐานว่างที่เป็นจริง ในการทดสอบ H_0: $\beta_j = 0$ และ H_1: $\beta_j \neq 0$ ซึ่งเป็นการทดสอบสมมติฐานแบบสองทาง ดังนั้นเราจะต้องคำนวณ p-value ของการทดสอบแบบสองทาง ซึ่งคือความน่าจะเป็นที่ t_{n-p-1} จะมีค่ามากกว่าค่าสถิติทดสอบ และความน่าจะเป็นที่ t_{n-p-1} จะมีค่าน้อยกว่าค่าติดลบของสถิติทดสอบ (กัลยา วานิชย์บัญชา, 2553) หรือ p-value = P($t_{n-p-1} > |t_{cal}|$) ดังรูปที่ 1



รูปที่ 1 แสดงพื้นที่ของการปฏิเสธสมมติฐานว่างในกรณีที่เป็นการทดสอบสองทาง และการแจกแจงเป็นแบบปกติ

เนื่องจากการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณกำลังสองน้อยที่สุด จะใช้ในกรณีที่จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง จึงจะสามารถหา $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ ได้ แต่ในกรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างทำให้ $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$ หาค่าไม่ได้ เนื่องจาก $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ เป็นเมตริกซ์เอกฐาน ด้วยเหตุนี้จึงต้องหาวิธีการประมาณค่าสัมประสิทธิ์ในสมการการถดถอยสำหรับข้อมูลที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างด้วยวิธีอื่น เช่นวิธี Penalized Likelihood Estimator

2.2 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Penalized Likelihood Estimator

ค่าสัมประสิทธิ์การถดถอยสามารถหาได้จากการหาค่าสัมประสิทธิ์ที่ทำให้ Penalized Likelihood มีค่าสูงสุด หรือคือ

$$\hat{\beta} = \arg\min_{\beta}[-l(\beta) + P_{\lambda}(\beta)], \lambda \geq 0 \tag{2.4}$$

เมื่อ $l(\beta)$ คือ –loglikelihood

$P_{\lambda}(\beta)$ คือ Penalty Function

λ คือ Tuning Parameter

จากสมการ (2.4) ข้างต้นหากเลือก Penalty Function ที่เหมาะสมจะสามารถทำให้สัมประสิทธิ์บางตัวเท่ากับศูนย์

2.2.1 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Least Absolute Shrinkage and Selection Operator (LASSO)

ในปี 1996 Tibshirani ได้เสนอวิธี Lasso โดย Lasso เป็นรูปแบบหนึ่งของ Penalized Likelihood Estimator ซึ่งจะมี Penalty Function ดังนี้

$$P_{\lambda}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \text{ โดยที่ } \lambda > 0 \tag{2.5}$$

จากสมการ (2.5) การใช้ $|\beta_j|$ จะได้สัมประสิทธิ์การถดถอยบางตัวเป็นศูนย์ แต่เนื่องจากวิธี Lasso มีความเอนเอียงทำให้ในปี 2006 Zou ได้เสนอวิธี Adaptive Least absolute shrinkage and selection operator (Adaptive Lasso)

2.2.2 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Adaptive Least absolute shrinkage and selection operator (Adaptive Lasso)

ปี 2006 Zou ได้เสนอวิธี Adaptive Lasso ซึ่งเป็นวิธีที่พัฒนามาจากวิธี Lasso เพียงแต่เพิ่มค่าถ่วงน้ำหนักที่แตกต่างกันของพารามิเตอร์แต่ละตัว โดย Adaptive Lasso ก็เป็นรูปแบบหนึ่งของ Penalized Likelihood Estimation ซึ่งจะมี Penalty Function ดังนี้

$$P_{\lambda}(\beta) = \lambda \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right| \tag{2.6}$$

$$\text{โดย } \hat{w}_j = \begin{cases} \dfrac{1}{\left|\hat{\beta}_{OLS}\right|}, n > p \\ \dfrac{1}{\left|\hat{\beta}_{Ridge}\right|}, n < p \end{cases}$$

โดยที่ $\hat{\beta}_{OLS}$ คือ การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณกำลังสองน้อยที่สุด

$\hat{\beta}_{Ridge}$ คือ การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีริดจ์ ซึ่งการใช้การถดถอยริดจ์จะทำให้ได้สัมประสิทธิ์การถดถอยที่เป็นศูนย์

ตัวประมาณที่ได้จากวิธี Adaptive Lasso นี้จะไม่มีความเอนเอียงและมีคุณสมบัติ Oracle (Oracle Property) กล่าวคือคุณสมบัติที่ตัวประมาณสามารถคัดเลือกตัวแปรได้ถูกต้องเสมือนกับว่าทราบตัวแบบที่แท้จริงกรณีที่ขนาดตัวอย่างเข้าใกล้อนันต์

จากวิธีการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Adaptive Lasso จะใช้ในขั้นตอนการหาเซตของดัชนีที่ตัวประมาณสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์ในวิธี Multi-Split

เนื่องจากวิธีการประมาณค่าสัมประสิทธิ์ทั้งสองวิธีข้างต้นหาได้เฉพาะการประมาณค่าพารามิเตอร์เท่านั้น ไม่สามารถตรวจสอบนัยสำคัญของตัวแปรได้ ทั้งนี้มีความสนใจในการตรวจสอบนัยสำคัญของตัวแปรโดยการคำนวณค่า p-value ของสัมประสิทธิ์การถดถอยแต่ละตัว ด้วยเหตุนี้จึงหาค่า p-value โดยวิธี Multi-Split

2.3 การหาค่า p-value ในการทดสอบสมมติฐานของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูง โดยวิธี Multi-Split

ในปี 2009 Meinshausen et al. ได้เสนอวิธีการแบ่งข้อมูลหลาย ๆ ครั้ง (Multi – Split) ในการคำนวณค่า p-value ซึ่งวิธี Multi – Split ก็เป็นการแบ่งข้อมูลออกเป็น 2 ชุดอย่างสุ่ม คือ $D_{in} = (X_{in}, Y_{in})$ และ $D_{out} = (X_{out}, Y_{out})$ แต่ละชุดมีขนาดตัวอย่างเท่ากัน กล่าวคือขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ $\frac{n}{2}$ โดยที่ข้อมูลใน D_{in} จะใช้ในการคัดกรองตัวแปร (Variable Screening) ซึ่งจะสามารถลดมิติของตัวแปรอิสระได้ สำหรับข้อมูลใน D_{out} จะใช้ในการคำนวณค่า p–value โดยใช้การประมาณกำลังสองน้อยที่สุด (Ordinary Least Squares Estimation) บนเซตของดัชนีที่ตัวประมาณสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์สำหรับตัวแปรที่ผ่านการคัดกรอง และตั้งค่า p-value แต่ละตัวเท่ากับ 1 สำหรับทุกค่าของจำนวนตัวแปรอิสระที่ไม่อยู่ในเซตของดัชนีที่ตัวประมาณสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์ โดยวิธี Multi-Split มีขั้นตอนดังต่อไปนี้

ให้ $S = \{j; \beta_j \neq 0\}$ และ

$\tilde{S} = \{j; \hat{\beta}_j \neq 0\}$

สำหรับ $b = 1, ..., B$ เมื่อ B คือ จำนวนรอบของการแบ่งข้อมูล

1. สุ่มแยกข้อมูลเป็น 2 ชุด คือ $D_{in}{}^{(b)}$ และ $D_{out}{}^{(b)}$ ด้วยขนาดที่เท่ากัน คือ $\frac{n}{2}$
2. ใช้ข้อมูล $D_{in}{}^{(b)}$ ในการประมาณเซต S ด้วย $\tilde{S}^{(b)}$
3. ใช้ข้อมูล $D_{out}{}^{(b)}$ ในการคำนวณค่า p-value ดังนี้

(ก.) หาก $j \in \tilde{S}$ จะประมาณค่าสัมประสิทธิ์การถดถอยโดยใช้วิธีกำลังสองน้อยที่สุด และคำนวณค่า p-value

(ข.) หาก $j \notin \tilde{S}$ จะกำหนดให้ p-value ซึ่ง $\tilde{P}_j^{(b)} = 1$

โดยที่ $\tilde{P}_j^{(b)}$ คือ p-value ที่ได้ในแต่ละครั้ง

4. ปรับค่า p-value โดยให้ค่า p-value ใหม่เป็น

$$P_j^{(b)} = \min(\tilde{P}_j^{(b)} \left|\tilde{S}^{(b)}\right|, 1),\ j = 1, ..., p \qquad (2.7)$$

โดยที่ $\left|\tilde{S}^{(b)}\right|$ คือ จำนวนสมาชิกในเซต $\tilde{S}^{(b)}$

จากขั้นตอนข้อ 1 – 4 จะได้ค่า p-value สำหรับสัมประสิทธิ์การถดถอยแต่ละตัวที่มีการปรับค่าแล้ว B ค่า ดังนั้น Meinshausen et al. (2009) จึงนำเสนอวิธีการรวม p-value B ค่า ให้เป็นค่าเดียว โดยใช้ฟังก์ชันควอนไทล์เชิงประจักษ์ ดังนี้

$$Q_j(\gamma) = \min\{1, q_\gamma(\{P_j^{(b)} / \gamma; b = 1,..., B\})\} \quad (2.8)$$

โดย $q_\gamma(.)$ เป็นฟังก์ชันควอนไทล์เชิงประจักษ์ (empirical quantile function) และ $\gamma \in (0,1)$

สำหรับแต่ละตัวแปรอิสระ $j = 1,..., p$ ค่า p-value ถูกกำหนดโดย $Q_j(\gamma)$ ดังนั้นจึงมีการเสนอให้ใช้ตัวที่ปรับค่าได้แทนการเลือกค่าที่เหมาะสม

ให้ $\gamma_{\min} \in (0,1)$ เป็นขอบเขตล่าง ซึ่ง Meinshausen et al. กำหนดให้ $\gamma_{\min} = 0.05$ โดย

$$P_j = \min\{1, (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma)\} \quad (2.9)$$

2.4 การหาค่า p-value ในการทดสอบสมมติฐานของสัมประสิทธิ์การถดถอยเมื่อข้อมูลมีมิติสูงโดยวิธีบูตสแตรป

วิธีบูตสแตรป เป็นวิธีที่เสนอโดย Efron ในปี 1979 ซึ่งเป็นการสุ่มตัวอย่างแบบคืนที่ (Resampling with Replacement) กล่าวคือมีหน่วยตัวอย่างซ้ำกันได้ ซึ่งมีหลักเกณฑ์ คือตัวอย่างที่ถูกเก็บ รวบรวมเปรียบเสมือนประชากรจริงจึงทำการสุ่มตัวอย่างด้วยจำนวนครั้งที่มากพอโดยที่แต่ละหน่วยตัวอย่างมีโอกาสถูกเลือกเท่า ๆ กัน เพื่อสร้างการแจกแจงของตัวสถิติตัวอย่าง

Efron เสนอให้ใช้วิธีการสุ่มตัวอย่างแบบคืนที่ขนาด n จากตัวอย่างสุ่มชุดเดียวที่มี เพื่อสร้างชุดตัวอย่างขนาด n ที่เป็นไปได้ นั่นคือแทนที่จะสุ่มตัวอย่างซ้ำ ๆ จากประชากรที่มีฟังก์ชันการแจกแจงสะสม (F) โดยตรง จะใช้การสุ่มตัวอย่างจาก Empirical distribution function ของข้อมูลตัวอย่าง

สุ่มตัวอย่าง n ตัว คือ $x_1,..., x_n$ ที่เป็นอิสระกันมาจากประชากรที่มีการแจกแจงแบบต่าง ๆ ให้ θ เป็นพารามิเตอร์ที่ต้องการประมาณในประชากรดังกล่าวนี้ และให้ $\hat{\theta}_B$ เป็นค่าประมาณของพารามิเตอร์ θ ด้วยวิธีบูตสแตรปที่คำนวณจากข้อมูลตัวอย่างขนาด n สร้างฟังก์ชันการแจกแจงโดยให้ความน่าจะเป็นของ $x_t; t = 1, 2,..., n$ เป็น $\frac{1}{n}$ ซึ่งเรียกฟังก์ชันการแจกแจงแบบนี้ว่า Empirical distribution function

ให้ $\theta = T(F)$ เป็นคุณสมบัติที่น่าสนใจของฟังก์ชันการแจกแจงสะสม ตัวอย่างเช่น $T(F) = \int z dF(z)$ เป็นค่าเฉลี่ยของการแจกแจง ให้ $x_1,..., x_n$ เป็นข้อมูลที่สังเกตได้ของตัวแปรสุ่ม

$\mathbf{X_1}, \ldots, \mathbf{X_n} \sim i.i.d.\ F$ โดยที่ F คือฟังก์ชันการแจกแจงสะสม ซึ่งสอดคล้องกับฟังก์ชันความหนาแน่น f

ให้ $\mathbf{X} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ เป็นเซตข้อมูลทั้งหมด

ถ้า $\hat{F}$ เป็นฟังก์ชันการแจกแจงเชิงประจักษ์ (The Empirical distribution function) ของข้อมูลที่สังเกตได้ แล้วตัวประมาณของ θ คือ $\hat{\theta} = T(\hat{F})$

เนื่องจากในบางกรณีเราอาจไม่ทราบการแจกแจง F หรือการแจกแจง F อาจจะอยู่ในรูปแบบที่ยาก ดังนั้นวิธีบูตสแตรปจึงใช้ฟังก์ชันการแจกแจงเชิงประจักษ์ $\hat{F}$ แทน ดังนี้

ให้ $\mathbf{X}^*$ แสดงตัวอย่างบูตสแตรปของข้อมูลเทียม (pseudo-data) ซึ่งจะเรียกว่าเซตข้อมูลเทียม (pseudo-dataset) สมาชิกของ $\mathbf{X}^* = \{\mathbf{X_1^*}, \ldots, \mathbf{X_n^*}\}$ เป็นตัวแปรสุ่ม i.i.d. ด้วยการแจกแจง $\hat{F}$ วิธีบูตสแตรปจะทำการสุ่มข้อมูลแบบคืนที่จากข้อมูลเทียม $\mathbf{X}^*$ ด้วยขนาด n มาจำนวน B ครั้ง ทำให้ได้ข้อมูลบูตสแตรปตัวอย่าง (Bootstrap Sample) จำนวน B เซต จากนั้นจึงนำข้อมูลบูตสแตรปตัวอย่างดังกล่าวมาใช้ในการหาค่า p-value ต่อไป

หลังจากคำนวณค่า p-value แล้วจะต้องมีการนำค่า p-value ดังกล่าวมาใช้ในการคัดเลือกตัวแปรในขั้นตอนสุดท้าย โดยใช้การควบคุม False Discovery Rate (FDR)

2.5 การควบคุม False Discovery Rate (FDR) หลังจากได้ค่า p-value จากวิธี Multi-Split แล้ว

การทดสอบสมมติฐานทางสถิติจะมีความผิดพลาดอยู่เสมอ ซึ่งอาจจะเนื่องจากการใช้ข้อมูลตัวอย่างมาสรุปผลการทดสอบเพื่ออ้างอิงถึงประชากร ทำให้ผลการทดสอบปฏิเสธสมมติฐาน H_0 ทั้งที่ H_0 เป็นจริง หรือผลการทดสอบสรุปว่าไม่ปฏิเสธ H_0 ทั้งที่ H_0 ไม่จริง ดังนั้นความผิดพลาดนี้จึงแบ่งได้ 2 ประเภทคือ

1. ความผิดพลาดประเภทที่ 1 (Type I Error) เป็นความผิดพลาดที่เกิดขึ้นเนื่องจากผู้วิจัยสรุปว่าสมมติฐานว่างไม่จริง (ปฏิเสธ H_0) ทั้งที่ในความเป็นจริงนั้นสมมติฐาน H_0 จริง เรียกความผิดพลาดชนิดนี้ว่า ระดับนัยสำคัญ (Level of significance)

α = P (ปฏิเสธ H_0 โดยที่ H_0 เป็นจริง)

α = โอกาสที่ผู้วิจัยจะสรุปผิด คือสรุปว่า H_0 ไม่จริงทั้งที่ความจริงสมมติฐาน H_0 จริง

2. ความผิดพลาดประเภทที่ 2 (Type II Error) เป็นความผิดพลาดที่เกิดขึ้นจากการที่ผู้วิจัยไม่ปฏิเสธ H_0 จริง โดยที่ในความเป็นจริงนั้น H_0 ไม่จริง

β = P (ไม่ปฏิเสธ H_0 โดยที่ H_0 ไม่จริง)

β = โอกาสที่ผู้วิจัยจะสรุปผิด โดยสรุปว่า H_0 จริงทั้งที่ความจริง H_0 ไม่จริง

ในการทดสอบสมมติฐานแต่ละครั้ง ผู้ทดสอบย่อมต้องการที่จะให้มีการผิดพลาดทั้งสองประเภทน้อยที่สุด แต่ถ้าลด α จะทำให้ β เพิ่มขึ้น ในทำนองเดียวกัน ถ้าลด β จะทำให้ α เพิ่มขึ้น ดังนั้นการที่จะลดค่าทั้งสองก็ต้องเพิ่มขนาดตัวอย่าง โดยทั่วไปผู้ทดสอบจะกำหนดค่า α หรือกำหนดระดับความเชื่อมั่น $1-\alpha$ คือโอกาสที่จะไม่ปฏิเสธ H_0 โดยที่ H_0 จริง

ตารางที่ 1

แสดงผลการทดสอบและความผิดพลาดในการทดสอบ

ผลการทดสอบ	ความเป็นจริง	
	H_0 จริง	H_0 ไม่จริง
ไม่ปฏิเสธ H_0	ผลการทดสอบถูกต้อง	ความผิดพลาดประเภทที่ 2 (β)
ปฏิเสธ H_0	ความผิดพลาดประเภทที่ 1 (α)	ผลการทดสอบถูกต้อง

ในปี 1995 Benjamini และ Hochberg เสนอการควบคุมค่าคาดหวังของสัดส่วนการปฏิเสธที่เป็นเท็จ หรือเรียกว่า False Discovery Rate (FDR) โดยพิจารณาปัญหาของการทดสอบสมมติฐานว่าง m สมมติฐาน โดยสามารถแสดงการควบคุมอัตราของค่าความคลาดเคลื่อนที่เกิดขึ้นของการปฏิเสธสมมติฐานว่างเปรียบเทียบกับผลการทดสอบและความผิดพลาดในการทดสอบดังตารางต่อไปนี้

ตารางที่ 2

แสดงผลของจำนวนการทดสอบและความผิดพลาดในการทดสอบสมมติฐานว่าง m สมมติฐาน

	ไม่มีนัยสำคัญ	มีนัยสำคัญ	ผลรวม
สมมติฐานว่างจริง	U	V	m_0
สมมติฐานว่างไม่จริง	T	S	$m-m_0$
ผลรวม	m – R	R	m

โดย m คือ จำนวนการทดสอบสมมติฐานทั้งหมด

m_0 คือ ผลรวมจำนวนของสมมติฐานว่างจริง

$m - m_0$ คือ ผลรวมจำนวนของสมมติฐานว่างไม่จริง

U คือ ผลการทดสอบถูกต้อง เมื่อสมมติฐานว่างจริง

V คือ ความผิดพลาดที่เกิดขึ้น เมื่อสมมติฐานว่างจริง

T คือ ความผิดพลาดที่เกิดขึ้น เมื่อสมมติฐานว่างไม่จริง

S คือ ผลการทดสอบถูกต้อง เมื่อสมมติฐานว่างไม่จริง

R คือ ผลรวมจำนวนของการปฏิเสธสมมติฐานว่าง

เมื่อพิจารณาปัญหาของการทดสอบสมมติฐานว่าง m ที่เกิดขึ้นซึ่ง m_0 เป็นสมมติฐานว่างจริง และ R เป็นผลรวมจำนวนของการปฏิเสธสมมติฐานว่าง จากตารางที่ 2 จะได้ว่าสมมติฐาน m ถูกสันนิษฐานว่าเป็นตัวแปรที่ทราบค่า R เป็นตัวแปรสุ่มที่สามารถสังเกตได้ ส่วน U, V, S และ T เป็นตัวแปรสุ่มที่ไม่สามารถสังเกตได้ ถ้าแต่ละสมมติฐานว่างถูกทดสอบแบบแยกที่ระดับ α แล้ว R = R (α) จะเพิ่มขึ้นใน α

ค่าความผิดพลาดของสัดส่วนที่ได้กระทำโดยสมมติฐานว่างที่ปฏิเสธเป็นเท็จสามารถแสดงด้วยตัวแปรสุ่ม Q – V/(V + S) – สัดส่วนของสมมติฐานว่างที่ถูกปฏิเสธซึ่งถูกปฏิเสธอย่างไม่ถูกต้อง โดยปกติจะนิยามว่า Q = 0 เมื่อ V + S = 0 ซึ่งเป็นค่าความผิดพลาดของการปฏิเสธที่เป็นเท็จที่สามารถกระทำได้ ซึ่ง Q เป็นตัวแปรสุ่มที่ไม่สามารถสังเกตได้ หรือเป็นตัวแปรสุ่มที่ไม่ทราบ โดยทั่วไป FDR นิยามได้ว่า

$$Q_e = E(Q) = E\{V/(V + S)\} = E(V/R) \qquad (2.9)$$

โดยที่ Q_e คือค่าคาดหวังของ Q

<u>ขั้นตอนการควบคุม False Discovery Rate</u>

ในที่นี้การควบคุม FDR แบ่งเป็น 2 กรณีดังนี้

(ก.) การควบคุม FDR ในการทดสอบหลาย ๆ การทดสอบภายใต้การทดสอบทางสถิติที่เป็นอิสระกัน (Benjamini and & Hochberg 1995)

พิจารณาการทดสอบ $H_{0,1}$, $H_{0,2}$, ..., $H_{0,m}$ โดยใช้การคำนวณค่า p-value $P_1, P_2, ..., P_m$ ให้ $P_{(1)} \le P_{(2)} \le ...P_{(m)}$ เป็นการเรียงลำดับ p-value และสมมติฐานว่าง $H_{(0,i)}$ สอดคล้องกับ $P_{(i)}$

ให้ k เป็น ค่า i ที่มากที่สุดที่ $P_{(i)} \le \frac{i}{m} q$ (2.10)

แล้วจะปฏิเสธ $H_{(0,i)}$ ทั้งหมด เมื่อ $i = 1, 2, ..., k$

จากขั้นตอนข้างต้นหากต้องการควบคุม FDR ที่ระดับ q เมื่อ $0 \le q \le 1$

(ข.) การควบคุม FDR ในการทดสอบหลาย ๆ การทดสอบภายใต้การทดสอบทางสถิติที่ไม่เป็นอิสระกัน (Benjamini and Yekutieli 2001)

จากแนวคิดของ Benjamini and Hochberg ที่แสดงว่าเมื่อการทดสอบทางสถิติเป็นอิสระต่อกันจะควบคุม FDR ที่ระดับ q แต่หากเป็นการทดสอบภายใต้ความไม่เป็นอิสระต่อกัน ขั้นตอนข้างต้นจะไม่สามารถควบคุม FDR ที่ระดับ q อีกต่อไป แต่จะควบคุม FDR ที่ระดับ $q\sum_{i=1}^{p} i^{-1}$

ให้ $P_{(1)} \le P_{(2)} \le ...P_{(\mathrm{m})}$ เป็นการเรียงลำดับ p-value และสมมติฐานว่าง $H_{(0,i)}$ สอดคล้องกับ $P_{(i)}$

ให้ k เป็น ค่า i ที่มากที่สุดที่ $P_{(i)} \le \frac{i}{m} q$ (2.11)

และปฏิเสธ $H_{(0,i)}$ ทั้งหมดเมื่อ $i = 1, 2, ..., k$

จากขั้นตอนข้างต้นควบคุม FDR ที่ระดับ $q\sum_{i=1}^{p} i^{-1}$

เนื่องจากค่า p-value ที่ได้จากวิธี Multi-Split เป็นค่า p-value ที่ปรับค่าแล้วจากการทดสอบสมมติฐาน p สมมติฐานภายใต้ความไม่เป็นอิสระกัน ดังนั้นวิธีการคัดเลือกตัวแปรโดยใช้ FDR จะเปลี่ยนไปโดยไม่มีการหารด้วย m จะได้ว่า

$$h = \max\{i : P_{(i)} \le iq\} \quad (2.12)$$

ซึ่งเซตที่ถูกเลือกของตัวแปรจะแสดงด้วยค่า h โดย

$$\hat{S} = \{j; \text{ปฏิเสธ } H_0: \beta_j = 0\} = \{j; P_j \le P_{(h)}\} \quad (2.13)$$

ถ้าไม่มีการปฏิเสธสมมติฐานว่าง แสดงว่า $\hat{S} = \phi$ แต่ถ้า $P_{(i)} > iq$ จะปฏิเสธสมมติฐานว่างสำหรับ $i = 1, 2, ..., p$ ซึ่งจะควบคุม FDR ได้ที่ระดับเดียวกับของ Benjamini and Yekutieli นั่นคือควบคุม FDR ที่ระดับ $q\sum_{i=1}^{p} i^{-1}$

2.6 เกณฑ์ที่ใช้ในการพิจารณาประสิทธิภาพของการแบ่งข้อมูล

เกณฑ์ที่ใช้ในการตัดสินว่าวิธีการแบ่งข้อมูลวิธีใดมีความเหมาะสมในการหาค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงระหว่างวิธี Random Split และวิธีบูตสแตรป คือจำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive), ความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยใช้ข้อมูลที่จำลองขึ้นมาโดยทั้ง 3 เกณฑ์ข้างต้นจะทำการเปรียบเทียบระหว่างสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัวโดยควบคุม FDR

กำหนดให้ $S = \{j; \beta_j \neq 0\}$

$\hat{S} = \{j;$ ปฏิเสธ H_0: $\beta_j = 0\}$

(1.) จำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive) คือจำนวนที่เกิดความผิดพลาดจากค่าสัมประสิทธิ์การถดถอยที่มีค่าไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์ ซึ่งสามารถหาได้จาก

$$\text{False Positive} = \left|\hat{S} \cap S^c\right|^*$$

(2.) จำนวนความผิดพลาดในการตรวจจับเชิงลบ (False Negative) คือจำนวนที่เกิดความผิดพลาดจากค่าสัมประสิทธิ์การถดถอยที่มีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์ ซึ่งสามารถหาได้จาก

$$\text{False Negative} = \left|\hat{S}^c \cap S\right|^*$$

(3.) จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว หรือ $|\hat{s}|$

เกณฑ์ที่ใช้วัดว่าวิธีใดเหมาะสมกับการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงคือ

- ถ้า False Positive ต่ำ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
- ถ้า False Negative ต่ำ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
- ถ้า $|\hat{s}|$ มีค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

* |A| คือ จำนวนสมาชิกในเซต A

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาเปรียบเทียบวิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง พร้อมทั้งเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรระหว่างวิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรป เมื่อใช้การจำลองข้อมูลที่มีอัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระที่ต่างกัน และจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ในกรณีที่แตกต่างกันตามสัดส่วนของจำนวนตัวแปรอิสระ ในการเปรียบเทียบว่าวิธีในการแบ่งข้อมูลแบบใดให้ผลดีกว่ากันจะพิจารณาจากจำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive), ความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของข้อมูลที่จำลองขึ้นมา

การจำลองข้อมูลในแต่ละสถานการณ์ผู้วิจัยทำงานด้วยโปรแกรม R เวอร์ชั่น 3.0.2 ในบทนี้จะกล่าวถึงแผนการดำเนินการวิจัย ขั้นตอนในการดำเนินการวิจัย และขั้นตอนการทำงานของโปรแกรม ซึ่งมีรายละเอียด ดังนี้

3.1 แผนการดำเนินการวิจัย

ในการวิจัยครั้งนี้ได้กำหนดสถานการณ์ต่าง ๆ ที่จะทำการศึกษาดังนี้

1. ตัวแบบที่ใช้ในการศึกษาคือ ตัวแบบการถดถอยเชิงเส้น หมายถึง สมการที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระที่อยู่ในรูป

$$\mathrm{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathrm{Y} = \begin{bmatrix} \mathrm{Y}_1 \\ \mathrm{Y}_2 \\ \vdots \\ \mathrm{Y}_n \end{bmatrix}, \mathbf{X} = \begin{pmatrix} \mathrm{X}_1 \\ \vdots \\ \mathrm{X}_n \end{pmatrix} \text{เมื่อ } \mathrm{X_i} = (\mathrm{X_{i1}} \quad \cdots \quad \mathrm{X_{ip}});\ i = 1,\ldots,n\ , \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_\mathrm{p} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\mathrm{n} \end{bmatrix}$$

$$\mathbf{X} \sim \mathrm{N}(0, \sum) \text{ และ } \varepsilon \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I_n}) \text{ เมื่อ } \sigma^2 = 1$$

$$\text{โดยที่ } \sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix} \text{ และ } \rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$$

เมื่อ n แทน ขนาดตัวอย่าง

p แทน จำนวนตัวแปรอิสระ

Y แทน เวกเตอร์ของตัวแปรตามขนาด $n\times 1$

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n\times p$

β แทน เวกเตอร์ของพารามิเตอร์ในสมการถดถอยขนาด $p\times 1$

ε แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $n\times 1$

2. ศึกษาภายใต้อัตราส่วนของขนาดตัวอย่าง (n) กับจำนวนตัวแปรอิสระ (p)

กรณีที่ 1: $n=10$

เปรียบเทียบอัตราส่วน 1 : 2 = 10 : 20

เปรียบเทียบอัตราส่วน 1 : 5 = 10 : 50

เปรียบเทียบอัตราส่วน 1 : 10 = 10 : 100

กรณีที่ 2 : $n=100$

เปรียบเทียบอัตราส่วน 1 : 2 = 100 : 200

เปรียบเทียบอัตราส่วน 1 : 5 = 100 : 500

เปรียบเทียบอัตราส่วน 1 : 10 = 100 : 1,000

กรณีที่ 3 : $n=200$

เปรียบเทียบอัตราส่วน 1 : 2 = 200 : 400

เปรียบเทียบอัตราส่วน 1 : 5 = 200 : 1,000

เปรียบเทียบอัตราส่วน 1 : 10 = 200 : 2,000

3. ศึกษาภายใต้เงื่อนไขจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์คือ 0.1 เท่า, 0.25 เท่า และ 0.45 เท่าของขนาดตัวอย่าง

4. ศึกษาภายใต้เงื่อนไขระดับความสัมพันธ์ของตัวแปรอิสระ

$\mathbf{X}\sim \mathrm{N}(0,\sum)$ เมื่อ $\sigma^2=1$

โดยที่ $\sum=\begin{pmatrix}\rho_{11} & \cdots & \rho_{1p}\\ \vdots & \ddots & \vdots\\ \rho_{p1} & \cdots & \rho_{pp}\end{pmatrix}$ และ $\rho_{ij}=\begin{cases}1; i=j\\ \rho^{|i-j|}; i\neq j\end{cases}$

โดยจะศึกษาภายใต้ค่า ρ 3 กรณี ดังนี้

กรณีที่ 1 $\rho=0$

กรณีที่ 2 $\rho=0.5$

กรณีที่ 3 $\rho=0.9$

โดยจะแบ่งระดับความสัมพันธ์ ดังนี้

ระดับความสัมพันธ์ $\rho = 0$ กรณีตัวแปรอิสระเป็น 20, 50, 100, 200, 400, 500, 1000 และ 2000 จะได้ว่า $\rho_{ij} = 0$ เท่ากันในทุกกรณี

ระดับความสัมพันธ์	จำนวนตัวแปรอิสระ	$\rho_{ij}; i \neq j$
$\rho = 0.5$	20	$\rho_{ij} \in [1.9\times10^{-6}, 0.5]$
	50	$\rho_{ij} \in [1.7\times10^{-15}, 0.5]$
	100	$\rho_{ij} \in [1.6\times10^{-30}, 0.5]$
	200	$\rho_{ij} \in [1.2\times10^{-60}, 0.5]$
	400	$\rho_{ij} \in [0, 0.5]$
	500	$\rho_{ij} \in [0, 0.5]$
	1000	$\rho_{ij} \in [0, 0.5]$
	2000	$\rho_{ij} \in [0, 0.5]$
$\rho = 0.9$	20	$\rho_{ij} \in [0.14, 0.9]$
	50	$\rho_{ij} \in [0.0057, 0.9]$
	100	$\rho_{ij} \in [0.000029, 0.9]$
	200	$\rho_{ij} \in [7.8\times10^{-10}, 0.9]$
	400	$\rho_{ij} \in [5.5\times10^{-19}, 0.9]$
	500	$\rho_{ij} \in [1.4\times10^{-23}, 0.9]$
	1000	$\rho_{ij} \in [1.9\times10^{-46}, 0.9]$
	2000	$\rho_{ij} \in [3.4\times10^{-92}, 0.9]$

5. ในการศึกษาครั้งนี้จะจำลองข้อมูลตามขอบเขตงานวิจัยข้างต้น ซึ่งผู้วิจัยจะประมวลผลโดยใช้โปรแกรม R เวอร์ชั่น 3.0.2 โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำของข้อมูลแต่ละกรณีไว้ที่จำนวน 100 รอบ

3.2 ขั้นตอนในการดำเนินการวิจัย

สำหรับการดำเนินการวิจัยมีดังนี้

1. ศึกษาค้นคว้าเอกสารต่าง ๆ ทั้งทฤษฎีที่เกี่ยวข้อง และงานวิจัยที่เกี่ยวกับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง
2. จำลองข้อมูลตามขอบเขตที่ต้องการศึกษา
   - สร้างข้อมูลที่มีขนาดตัวอย่าง (n) และจำนวนตัวแปรอิสระ (p) ตามที่กำหนด
   - กำหนดระดับความสัมพันธ์ของตัวแปรอิสระ

     $\mathbf{X} \sim \mathrm{N}(0, \sum)$ เมื่อ

โดยที่ $\sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix}$ และ $\rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$

เมื่อ $\rho = 0$, $\rho = 0.5$ และ $\rho = 0.9$

- กำหนดจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1 เท่า, 0.25 เท่า และ 0.45 เท่าของขนาดตัวอย่างแบบสุ่ม
- กำหนดค่า β ที่ไม่เท่ากับศูนย์ โดยสุ่มจำนวน β ที่ไม่เท่ากับศูนย์จากการแจกแจงยูนิฟอร์ม $\mathrm{U} \sim [0.5,5]$
- สร้างข้อมูลที่มีการแจกแจงปกติ

$$\mathrm{Y} = \mathbf{X}\beta + \varepsilon \ ; \mathbf{X} \sim \mathrm{N}(0, \sum) \text{ และ } \varepsilon \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I_n}) \text{ เมื่อ } \sigma^2 = 1$$

โดยที่ $\sum = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{pmatrix}$ และ $\rho_{ij} = \begin{cases} 1; i = j \\ \rho^{|i-j|}; i \neq j \end{cases}$

เมื่อ n แทน ขนาดตัวอย่าง

p แทน จำนวนตัวแปรอิสระ

Y แทน เวกเตอร์ของตัวแปรตามขนาด $n \times 1$

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n \times p$

β แทน เวกเตอร์ของพารามิเตอร์ในสมการถดถอยขนาด $p \times 1$

ε แทน เวกเตอร์ของความคลาดเคลื่อนขนาด $n \times 1$

3. ข้อมูลที่ได้ในแต่ละชุดมาหาค่า p-value โดยวิธี Multi – Split ซึ่งจะแบ่งข้อมูลออกเป็น 2 ชุดหลาย ๆ ครั้ง ซึ่งในที่นี้จะทำการแบ่งข้อมูลจำนวน 50 รอบ โดยแต่ละรอบจะแบ่งข้อมูลโดยใช้วิธี

- Random Split โดยแบ่งข้อมูลออกเป็นสองชุดแต่ละชุดมีขนาดตัวอย่างเท่ากัน กล่าวคือ ขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ $\frac{n}{2}$
- วิธีบูตสแตรป โดยแบ่งข้อมูลออกเป็นสองชุดแต่ละชุดมีขนาดตัวอย่างเท่ากัน กล่าวคือขนาดตัวอย่างของข้อมูลแต่ละชุดเท่ากับ n เนื่องจากเป็นการสุ่มแบบไม่คืนที่ ควบคุม FDR สำหรับค่า p-value ที่ได้เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

4. หลังจากนั้นนำข้อมูลที่ได้จากการแบ่งข้อมูลมาคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และหาค่า p-value แล้วจึงปรับค่า p-value และรวมค่า p-value จากการทำซ้ำ

5. ควบคุม FDR สำหรับค่า p-value ที่ได้เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

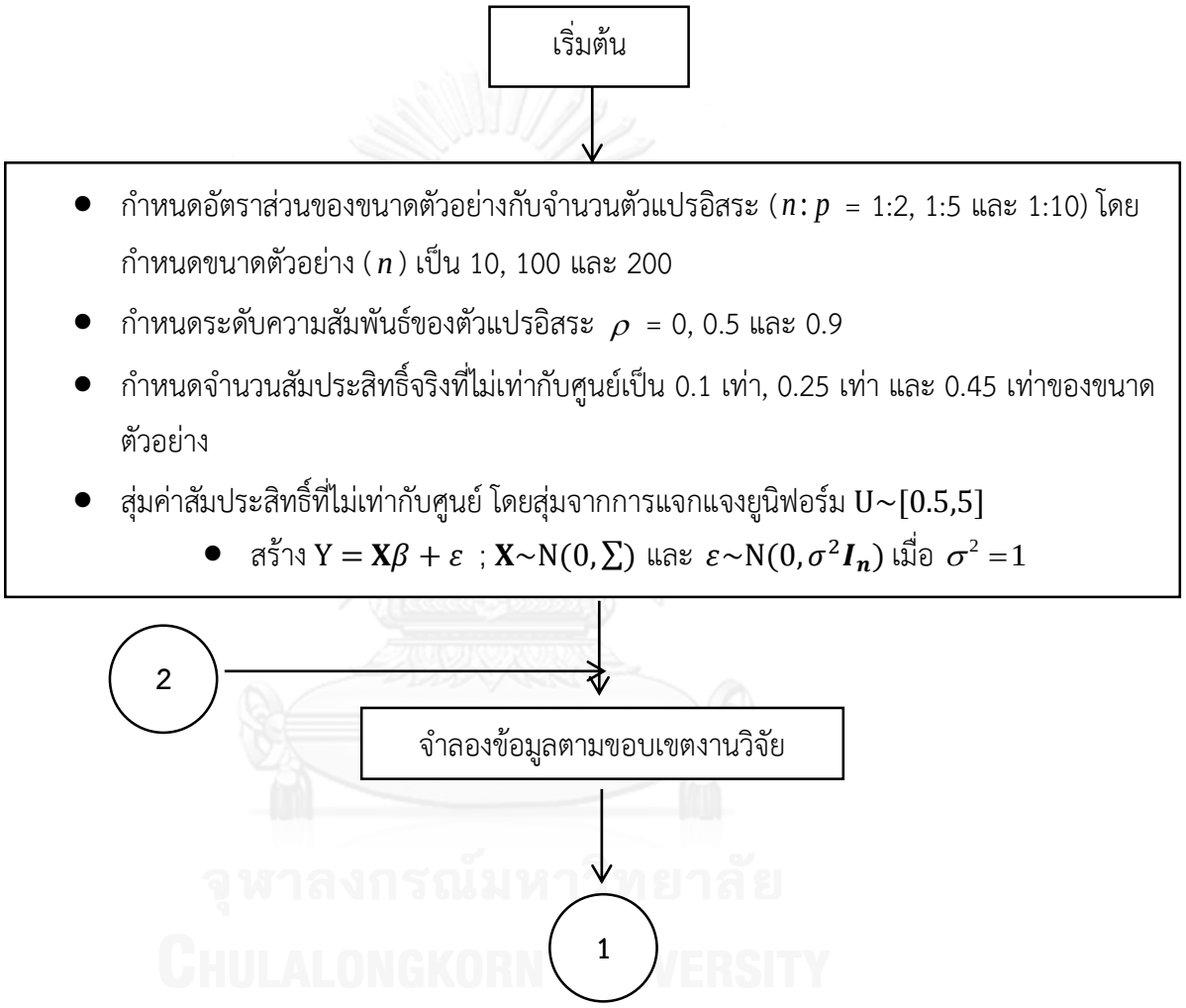
6. คำนวณจำนวนความผิดพลาดในการตรวจจับเชิงบวก, ความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว

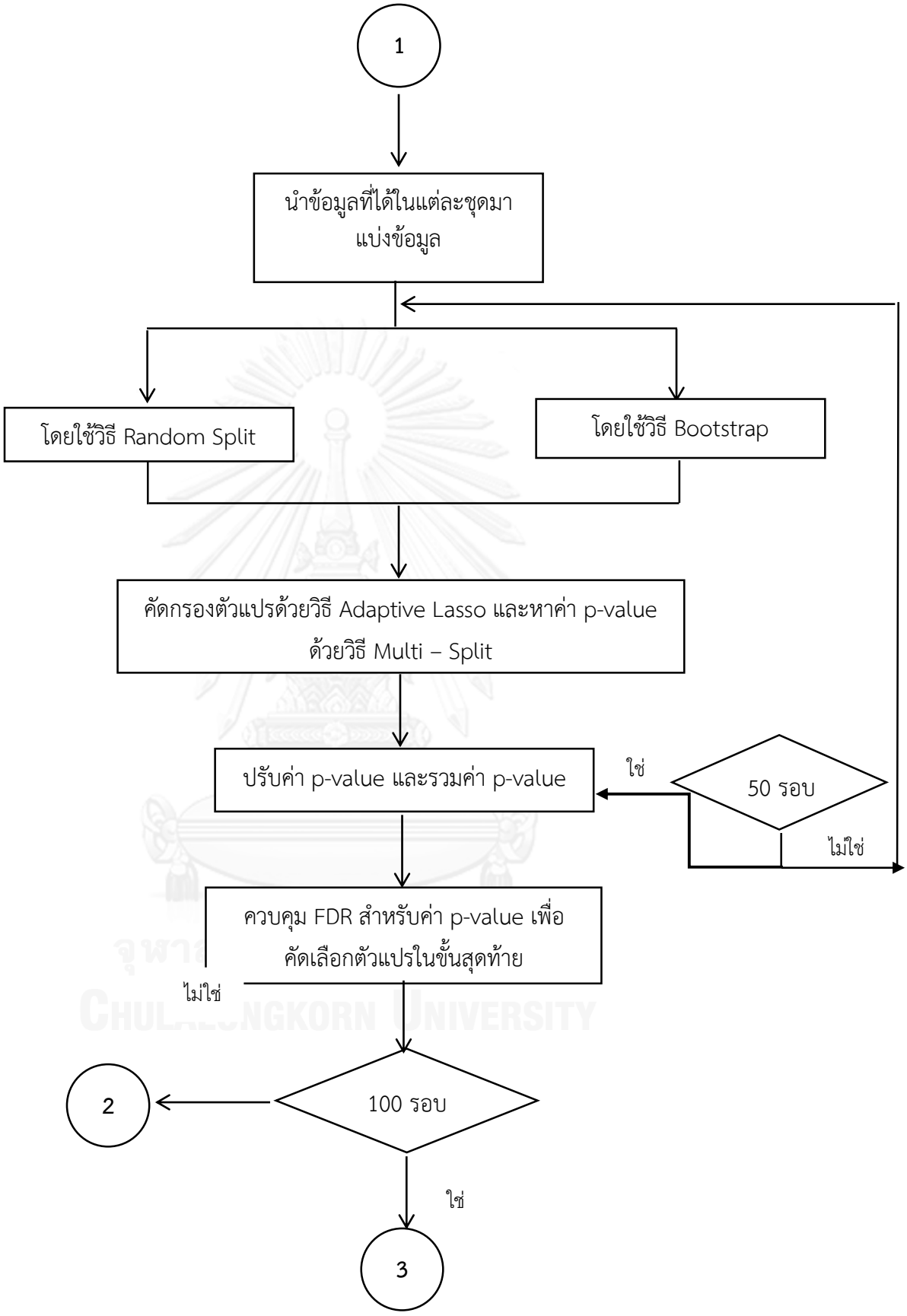
7. วิเคราะห์และสรุปผลการศึกษา

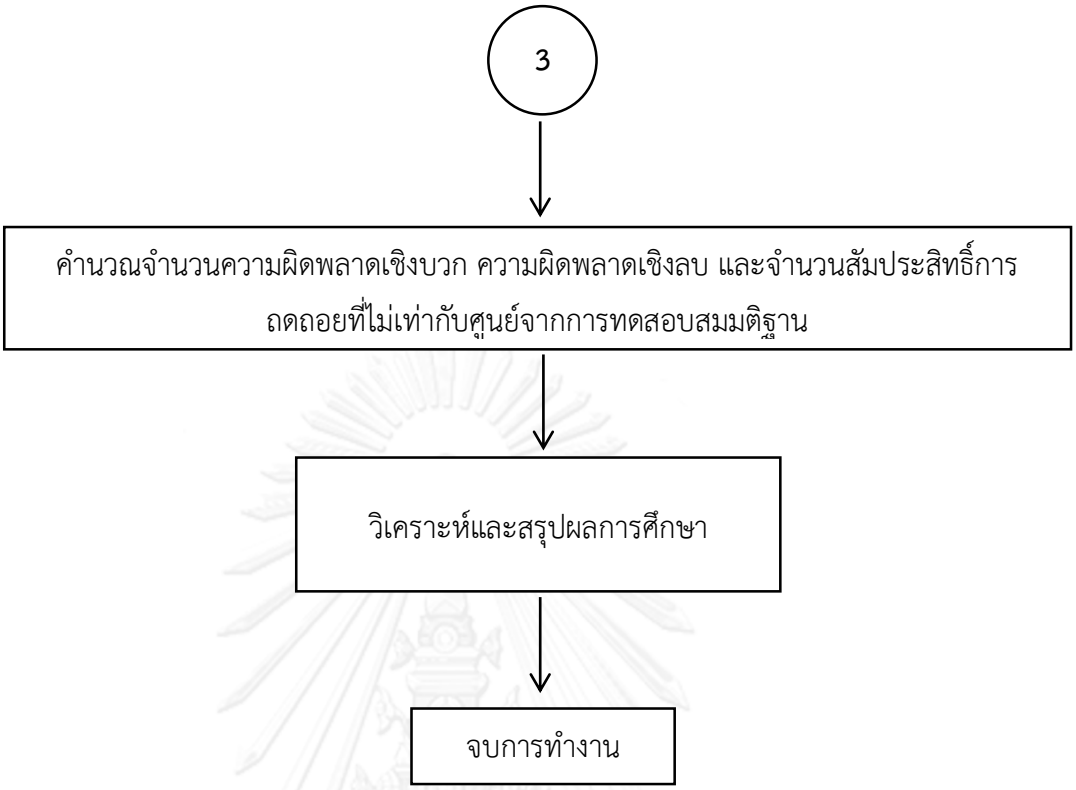
ทำการเปรียบเทียบสำหรับแต่ละวิธีการแบ่งข้อมูลแล้วทำการสรุปผลว่าวิธีการแบ่งข้อมูลใดเหมาะสมสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

3.3 ขั้นตอนการทำงานของโปรแกรม

ภาพที่ 1 แสดงแผนผังขั้นตอนการวิจัย




1
นำข้อมูลที่ได้ในแต่ละชุดมา
แบ่งข้อมูล
โดยใช้วิธี Random Split
โดยใช้วิธี Bootstrap
คัดกรองตัวแปรด้วยวิธี Adaptive Lasso และหาค่า p-value
ด้วยวิธี Multi – Split
ปรับค่า p-value และรวมค่า p-value
ใช่
50 รอบ
ไม่ใช่
ควบคุม FDR สำหรับค่า p-value เพื่อ
คัดเลือกตัวแปรในขั้นสุดท้าย
ไม่ใช่
2
100 รอบ
ใช่
3



หมายเหตุ: ทำซ้ำ 100 รอบ ตั้งแต่ขั้นตอนการจำลองข้อมูลจนกระทั่งถึงขั้นตอนการควบคุม FDR ที่ระดับ q-level = 0.1 เพื่อคัดเลือกตัวแปรในขั้นสุดท้าย

บทที่ 4

ผลการวิเคราะห์ข้อมูล

การศึกษางานวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบแนวทางในการเลือกใช้วิธีการแบ่งข้อมูลอย่างสุ่มและวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง พร้อมทั้งเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรระหว่างวิธีการแบ่งข้อมูลอย่างสุ่ม และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยทำการศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระด้วยจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 ของขนาดตัวอย่าง และศึกษาภายใต้เงื่อนไขระดับความสัมพันธ์ของตัวแปรอิสระ จากการศึกษาในครั้งนี้เกณฑ์การพิจารณาเปรียบเทียบคือ จำนวนความผิดพลาดในการตรวจจับเชิงบวก (False Positive), จำนวนความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบ สมมติฐานโดยใช้ข้อมูลจำลอง

ในการนำเสนอผลการวิจัยจะแสดงในรูปแบบของตาราง โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่าง ๆ ดังนี้

n	แทน ขนาดของตัวอย่าง
p	แทน จำนวนตัวแปรอิสระ
$n:p$	แทน ขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ
$\|s\|$	แทน จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์
Random Split	แทน การแบ่งข้อมูลอย่างสุ่ม
Bootstrap	แทน การแบ่งข้อมูลด้วยวิธีบูตสแตรป
FP	แทน จำนวนความผิดพลาดในการตรวจจับเชิงบวก
FN	แทน จำนวนความผิดพลาดในการตรวจจับเชิงลบ
$\|\hat{s}\|$	แทน จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของสัมประสิทธิ์แต่ละตัว

ในการนำเสนอผลการเปรียบเทียบวิธีการแบ่งข้อมูลทั้งสองวิธี จะแบ่งการนำเสนอออกเป็น 3 ส่วนด้วยกัน โดยที่ส่วนแรกจะเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป กรณีที่ $n=10$ ส่วนที่ 2 จะเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิง บวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป กรณีที่ $n=100$ และส่วนสุดท้ายจะเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป กรณีที่ $n=200$

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป เมื่อพิจารณาในกรณีที่ $n=10$

1.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคือ 10:20, 10:50 และ 10:100
1.2 เมื่อให้จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 ของขนาดตัวอย่าง
1.3 เมื่อให้ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9

<u>ส่วนที่ 2</u> ผลการเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป เมื่อพิจารณาในกรณีที่ $n=100$

2.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคือ 100:200, 100:500 และ 100:1,000

2.2 เมื่อให้จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 ของขนาดตัวอย่าง
2.3 เมื่อให้ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9

<u>ส่วนที่ 3</u> ผลการเปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานระหว่างวิธี Random Split และวิธีบูตสแตรป เมื่อพิจารณาในกรณีที่ $n=200$

3.1 เมื่อให้อัตราส่วนระหว่างขนาดตัวอย่างต่อตัวแปรอิสระคือ 200:400, 200:1,000 และ 200:2,000

3.2 เมื่อให้จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 ของขนาดตัวอย่าง
3.3 เมื่อให้ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9

หมายเหตุ: ในแต่ละกรณีจะจำลองข้อมูล 100 ชุด

นอกจากการเปรียบเทียบทั้งสามวิธีข้างต้น เรายังเปรียบเทียบประสิทธิภาพของวิธี Random Split และวิธีบูตสแตรปโดยใช้ Wilcoxon's Signed-Rank Test เหตุที่เราใช้ Wilcoxon's Signed-Rank Test เนื่องจากข้อมูลที่ได้ไม่เป็นการแจกแจงแบบปกติ โดยมีสมมติฐานดังนี้

H_0: วิธีการแบ่งข้อมูลทั้งสองวิธีไม่แตกต่างกัน
H_1: วิธีการแบ่งข้อมูลทั้งสองวิธีแตกต่างกัน

ถ้าปฏิเสธ H_0 แสดงว่าวิธีการแบ่งข้อมูลทั้งสองวิธีแตกต่างกันอย่างมีนัยสำคัญที่ $\alpha = 0.05$

4.1 ผลการเปรียบเทียบข้อมูลจำลองขนาด 10 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป โดยจะเปรียบเทียบดังนี้

4.1.1 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.1.2 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.1.3 การเปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

ตารางที่ 4.1.1 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n : p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s \rvert$)	จำนวนความผิดพลาดเชิงบวก		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho = 0$	10:20	1	0.000 (0.000)	7.470 (1.367)	< 0.00001*
		2	0.010 (0.100)	2.380 (1.953)	< 0.00001*
		4	0.000 (0.000)	3.780 (2.604)	< 0.00001*
	10:50	1	0.000 (0.000)	1.430 (1.647)	< 0.00001*
		2	0.000 (0.000)	2.260 (2.596)	< 0.00001*
		4	0.000 (0.000)	3.280 (2.839)	< 0.00001*
	10:100	1	0.000 (0.000)	0.560 (1.242)	< 0.00001*
		2	0.000 (0.000)	1.040 (1.974)	< 0.00001*
		4	0.000 (0.000)	3.110 (2.856)	< 0.00001*
$\rho = 0.5$	10:20	1	0.000 (0.000)	2.180 (2.245)	< 0.00001*
		2	0.000 (0.000)	2.800 (2.507)	< 0.00001*
		4	0.000 (0.000)	1.560 (2.467)	< 0.00001*
	10:50	1	0.000 (0.000)	1.940 (2.103)	< 0.00001*
		2	0.000 (0.000)	1.160 (2.049)	< 0.00001*
		4	0.000 (0.000)	2.220 (2.710)	< 0.00001*
	10:100	1	0.000 (0.000)	1.240 (1.799)	< 0.00001*
		2	0.000 (0.000)	2.440 (2.599)	< 0.00001*
		4	0.000 (0.000)	3.150 (2.638)	< 0.00001*
$\rho = 0.9$	10:20	1	0.000 (0.000)	0.640 (1.685)	0.000307*
		2	0.000 (0.000)	0.880 (2.119)	< 0.00001*
		4	0.040 (0.197)	4.430 (2.815)	< 0.00001*
	10:50	1	0.010 (0.100)	1.830 (2.065)	< 0.00001*
		2	0.010 (0.100)	2.350 (2.869)	< 0.00001*
		4	0.000 (0.000)	5.340 (3.613)	< 0.00001*
	10:100	1	0.000 (0.000)	1.140 (1.928)	< 0.00001*
		2	0.000 (0.000)	0.710 (2.189)	0.001597*
		4	0.000 (0.000)	2.320 (2.902)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha = 0.05$

จากตารางที่ 4.1.1 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยคำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

ตารางที่ 4.1.2 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n:p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(\lvert s\rvert)$	จำนวนความผิดพลาดเชิงลบ		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho=0$	10:20	1	0.860 (0.349)	0.000 (0.000)	< 0.00001*
		2	2.000 (0.000)	0.000 (0.000)	< 0.00001*
		4	4.000 (0.000)	1.560 (1.104)	< 0.00001*
	10:50	1	0.780 (0.416)	0.140 (0.349)	< 0.00001*
		2	1.510 (0.859)	0.710 (0.701)	< 0.00001*
		4	4.000 (0.000)	2.560 (0.988)	< 0.00001*
	10:100	1	0.400 (0.492)	0.100 (0.302)	< 0.00001*
		2	0.880 (0.998)	0.400 (0.667)	< 0.00001*
		4	4.000 (0.000)	3.080 (0.861)	< 0.00001*
$\rho=0.5$	10:20	1	0.760 (0.429)	0.130 (0.338)	< 0.00001*
		2	1.970 (0.171)	0.520 (0.674)	< 0.00001*
		4	1.720 (1.990)	0.700 (1.115)	< 0.00001*
	10:50	1	0.940 (0.239)	0.260 (0.441)	< 0.00001*
		2	0.840 (0.992)	0.360 (0.644)	< 0.00001*
		4	2.680 (1.890)	1.610 (1.413)	< 0.00001*
	10:100	1	0.700 (0.461)	0.230 (0.423)	< 0.00001*
		2	1.990 (0.100)	1.050 (0.730)	< 0.00001*
		4	4.000 (0.000)	3.080 (0.825)	< 0.00001*
$\rho=0.9$	10:20	1	0.140 (0.349)	0.030 (0.171)	0.002602*
		2	0.500 (0.870)	0.130 (0.393)	< 0.00001*
		4	3.970 (0.171)	1.310 (0.929)	< 0.00001*
	10:50	1	0.890 (0.314)	0.280 (0.451)	< 0.00001*
		2	1.270 (0.962)	0.520 (0.659)	< 0.00001*
		4	4.000 (0.000)	2.330 (1.215)	< 0.00001*
	10:100	1	0.600 (0.492)	0.250 (0.435)	< 0.00001*
		2	0.320 (0.737)	0.180 (0.479)	0.001222*
		4	2.520 (1.941)	1.870 (1.600)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha=0.05$

จากตารางที่ 4.1.2 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยคำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

ตารางที่ 4.1.3 แสดงจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n:p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(\lvert s \rvert)$	จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho=0$	10:20	1	0.140 (0.349)	8.470 (1.367)	< 0.00001*
		2	0.010 (0.100)	3.820 (2.204)	< 0.00001*
		4	0.000 (0.000)	6.220 (3.024)	< 0.00001*
	10:50	1	0.090 (0.288)	2.160 (1.824)	< 0.00001*
		2	0.010 (0.100)	3.070 (3.006)	< 0.00001*
		4	0.000 (0.000)	4.720 (3.340)	< 0.00001*
	10:100	1	0.040 (0.197)	0.900 (1.528)	< 0.00001*
		2	0.000 (0.000)	1.520 (2.480)	< 0.00001*
		4	0.000 (0.000)	4.030 (3.362)	< 0.00001*
$\rho=0.5$	10:20	1	0.240 (0.429)	3.050 (2.380)	< 0.00001*
		2	0.030 (0.171)	4.280 (2.756)	< 0.00001*
		4	0.000 (0.000)	2.580 (3.635)	< 0.00001*
	10:50	1	0.060 (0.239)	2.680 (2.188)	< 0.00001*
		2	0.000 (0.000)	1.640 (2.657)	< 0.00001*
		4	0.000 (0.000)	3.290 (3.540)	< 0.00001*
	10:100	1	0.060 (0.239)	1.770 (2.004)	< 0.00001*
		2	0.010 (0.100)	3.390 (2.737)	< 0.00001*
		4	0.000 (0.000)	4.070 (2.989)	< 0.00001*
$\rho=0.9$	10:20	1	0.050 (0.219)	0.800 (1.985)	0.0002164*
		2	0.000 (0.000)	1.250 (2.672)	< 0.00001*
		4	0.070 (0.256)	7.120 (3.189)	< 0.00001*
	10:50	1	0.120 (0.327)	2.550 (2.213)	< 0.00001*
		2	0.020 (0.141)	3.110 (3.360)	< 0.00001*
		4	0.000 (0.000)	7.010 (4.152)	< 0.00001*
	10:100	1	0.020 (0.141)	1.510 (2.186)	< 0.00001*
		2	0.000 (0.000)	0.850 (2.492)	0.000714*
		4	0.000 (0.000)	2.970 (3.526)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha=0.05$

จากตารางที่ 4.1.3 ซึ่งแสดงผลของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=10$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน แต่ตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวน สัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้น เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้น ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

สรุปผลส่วนที่ 4.1 ผลการเปรียบเทียบข้อมูลจำลองขนาด 10 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป

พิจารณาตารางที่ 4.1.1 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงบวกของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Random Split เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.1.2 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงลบของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.1.3 เปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะเห็นว่าวิธี Random Split และวิธี Bootstrap มีแนวโน้มที่ต่างกัน อีกทั้งจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีค่าใกล้เคียงจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี Random Split โดยส่วนใหญ่ ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

จากผลการศึกษาตารางที่ 4.1.1 – 4.1.3 กรณีที่ $n=10$ แม้ว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap แต่จำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ส่วนใหญ่มีค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี Random Split

โดยรวมแล้วยังสรุปไม่ได้ว่าวิธีการแบ่งข้อมูลในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง วิธีใดมีประสิทธิภาพสูงสุด แต่เมื่อพิจารณาอำนาจการทดสอบพบว่าวิธี Bootstrap มีประสิทธิภาพในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงมากกว่าวิธี Random Split เนื่องจากวิธี Bootstrap มีขนาดตัวอย่างของข้อมูลแต่ละชุดมากกว่าวิธี Random Split ทำให้อำนาจการทดสอบของวิธี Bootstrap มีค่ามากกว่าวิธี Random Split

4.2 ผลการเปรียบเทียบข้อมูลจำลองขนาด 100 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป โดยจะเปรียบเทียบดังนี้

4.2.1 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.2.2 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.2.3 การเปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

ตารางที่ 4.2.1 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n:p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s\rvert$)	จำนวนความผิดพลาดเชิงบวก		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho=0$	100:200	10	0.000 (0.000)	0.140 (0.377)	0.0005371*
		25	0.000 (0.000)	0.400 (0.696)	< 0.00001*
		45	0.000 (0.000)	0.660 (0.831)	< 0.00001*
	100:500	10	0.000 (0.000)	0.100 (0.302)	0.001904*
		25	0.000 (0.000)	0.580 (0.806)	< 0.00001*
		45	0.000 (0.000)	0.710 (0.844)	< 0.00001*
	100:1,000	10	0.000 (0.000)	0.190 (0.465)	< 0.00001*
		25	0.000 (0.000)	0.450 (0.672)	< 0.00001*
		45	0.000 (0.000)	0.570 (0.728)	< 0.00001*
$\rho=0.5$	100:200	10	0.000 (0.000)	0.240 (0.495)	< 0.00001*
		25	0.000 (0.000)	0.660 (0.819)	< 0.00001*
		45	0.000 (0.000)	1.390 (1.163)	< 0.00001*
	100:500	10	0.000 (0.000)	0.210 (0.456)	< 0.00001*
		25	0.000 (0.000)	0.810 (0.895)	< 0.00001*
		45	0.000 (0.000)	0.790 (0.856)	< 0.00001*
	100:1,000	10	0.000 (0.000)	0.170 (0.403)	0.0001082*
		25	0.000 (0.000)	0.460 (0.809)	< 0.00001*
		45	0.000 (0.000)	0.770 (0.874)	< 0.00001*
$\rho=0.9$	100:200	10	0.630 (0.720)	1.430 (1.312)	< 0.00001*
		25	0.100 (0.333)	7.770 (2.562)	< 0.00001*
		45	0.040 (0.197)	8.900 (2.809)	< 0.00001*
	100:500	10	1.190 (1.098)	2.260 (1.784)	< 0.00001*
		25	0.060 (0.239)	8.510 (2.827)	< 0.00001*
		45	0.000 (0.000)	6.160 (2.684)	< 0.00001*
	100:1,000	10	1.470 (1.150)	3.730 (2.224)	< 0.00001*
		25	0.030 (0.171)	5.850 (2.199)	< 0.00001*
		45	0.010 (0.100)	3.930 (1.976)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha=0.05$

จากตารางที่ 4.2.1 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยคำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ยกเว้นในกรณีที่ $\rho=0.9$ ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split มีแนวโน้มลดลง ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

ตารางที่ 4.2.2 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n:p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s\rvert$)	จำนวนความผิดพลาดเชิงลบ		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho=0$	100:200	10	0.660 (0.768)	0.000 (0.000)	< 0.00001*
		25	23.990 (1.685)	5.880 (2.931)	< 0.00001*
		45	44.960 (0.197)	32.730 (3.856)	< 0.00001*
	100:500	10	1.580 (1.609)	0.110 (0.373)	< 0.00001*
		25	24.750 (0.520)	15.400 (3.461)	< 0.00001*
		45	44.990 (0.100)	40.790 (1.908)	< 0.00001*
	100:1,000	10	3.350 (2.171)	0.420 (0.806)	< 0.00001*
		25	24.850 (0.411)	19.130 (2.525)	< 0.00001*
		45	44.990 (0.100)	42.740 (1.330)	< 0.00001*
$\rho=0.5$	100:200	10	0.620 (1.033)	0.040 (0.197)	< 0.00001*
		25	22.780 (3.221)	3.460 (2.052)	< 0.00001*
		45	44.990 (0.100)	30.150 (3.608)	< 0.00001*
	100:500	10	1.340 (1.578)	0.100 (0.333)	< 0.00001*
		25	24.800 (0.550)	14.050 (4.001)	< 0.00001*
		45	44.930 (0.256)	39.860 (2.547)	< 0.00001*
	100:1,000	10	3.540 (2.267)	0.460 (0.658)	< 0.00001*
		25	24.920 (0.307)	18.190 (2.707)	< 0.00001*
		45	44.990 (0.100)	42.440 (1.493)	< 0.00001*
$\rho=0.9$	100:200	10	0.910 (0.900)	0.240 (0.534)	< 0.00001*
		25	19.440 (3.154)	4.630 (1.851)	< 0.00001*
		45	44.760 (0.534)	25.320 (3.393)	< 0.00001*
	100:500	10	1.110 (0.952)	0.250 (0.520)	< 0.00001*
		25	24.190 (1.187)	10.340 (2.705)	< 0.00001*
		45	44.970 (0.171)	36.030 (2.976)	< 0.00001*
	100:1,000	10	2.330 (1.615)	0.470 (0.627)	< 0.00001*
		25	24.630 (0.837)	16.170 (2.857)	< 0.00001*
		45	44.980 (0.141)	41.000 (2.184)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha=0.05$

จากตารางที่ 4.2.2 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยคำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

ตารางที่ 4.2.3 แสดงจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n:p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s\rvert$)	จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho=0$	100:200	10	9.340 (0.768)	10.140 (0.377)	< 0.00001*
		25	1.010 (1.685)	19.520 (2.787)	< 0.00001*
		45	0.040 (0.197)	12.930 (3.740)	< 0.00001*
	100:500	10	8.420 (1.609)	9.990 (0.460)	< 0.00001*
		25	0.250 (0.520)	10.180 (3.362)	< 0.00001*
		45	0.010 (0.100)	4.920 (2.111)	< 0.00001*
	100:1,000	10	6.650 (2.171)	9.770 (0.827)	< 0.00001*
		25	0.150 (0.411)	6.320 (2.506)	< 0.00001*
		45	0.010 (0.100)	2.830 (1.436)	< 0.00001*
$\rho=0.5$	100:200	10	9.380 (1.033)	10.200 (0.492)	< 0.00001*
		25	2.220 (3.221)	22.200 (1.803)	< 0.00001*
		45	0.010 (0.100)	16.240 (3.970)	< 0.00001*
	100:500	10	8.660 (1.578)	10.110 (0.490)	< 0.00001*
		25	0.200 (0.550)	11.760 (3.903)	< 0.00001*
		45	0.070 (0.256)	5.930 (2.595)	< 0.00001*
	100:1,000	10	6.460 (2.267)	9.710 (0.686)	< 0.00001*
		25	0.080 (0.307)	7.270 (2.670)	< 0.00001*
		45	0.010 (0.100)	3.330 (1.724)	< 0.00001*
$\rho=0.9$	100:200	10	9.720 (1.198)	11.190 (1.361)	< 0.00001*
		25	5.660 (3.242)	28.140 (2.486)	< 0.00001*
		45	0.280 (0.587)	28.580 (3.729)	< 0.00001*
	100:500	10	10.080 (1.489)	12.010 (1.749)	< 0.00001*
		25	0.870 (1.220)	23.170 (3.525)	< 0.00001*
		45	0.030 (0.171)	15.130 (4.074)	< 0.00001*
	100:1,000	10	9.140 (2.103)	13.260 (2.286)	< 0.00001*
		25	0.400 (0.853)	14.680 (3.604)	< 0.00001*
		45	0.030 (0.171)	7.930 (2.945)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha=0.05$

จากตารางที่ 4.2.3 ซึ่งแสดงผลของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n=100$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวน สัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha=0.05$

สรุปผลส่วนที่ 4.2 ผลการเปรียบเทียบข้อมูลจำลองขนาด 100 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป

พิจารณาตารางที่ 4.2.1 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น ยกเว้นในกรณีที่ $\rho = 0.9$ ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split มีแนวโน้มลดลง อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงบวกของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Random Split เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.2.2 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงลบของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.2.3 เปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานมีแนวโน้มไม่สม่ำเสมอ บางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ แต่เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีค่าใกล้เคียงจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี

Random Split ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

จากผลการศึกษาตารางที่ 4.2.1 – 4.2.3 กรณีที่ $n=100$ แม้ว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap แต่จำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี Random Split

โดยรวมแล้วยังสรุปไม่ได้ว่าวิธีการแบ่งข้อมูลในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง วิธีใดมีประสิทธิภาพสูงสุด แต่เมื่อพิจารณาอำนาจการทดสอบพบว่าวิธี Bootstrap มีประสิทธิภาพในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงมากกว่าวิธี Random Split เนื่องจากวิธี Bootstrap มีขนาดตัวอย่างของข้อมูลแต่ละชุดมากกว่าวิธี Random Split ทำให้อำนาจการทดสอบของวิธี Bootstrap มีค่ามากกว่าวิธี Random Split

4.3 ผลการเปรียบเทียบข้อมูลจำลองขนาด 200 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป โดยจะเปรียบเทียบดังนี้

4.3.1 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.3.2 การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

4.3.3 การเปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยจากข้อมูลจำลอง 100 ชุดระหว่างวิธี Random Split และวิธีบูตสแตรป

ตารางที่ 4.3.1 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n : p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s \rvert$)	จำนวนความผิดพลาดเชิงบวก		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho = 0$	200:400	20	0.000 (0.000)	0.100 (0.302)	0.001904*
		50	0.000 (0.000)	0.360 (0.674)	< 0.00001*
		90	0.000 (0.000)	0.730 (0.790)	< 0.00001*
	200:1,000	20	0.000 (0.000)	0.040 (0.197)	0.07186
		50	0.000 (0.000)	0.580 (0.912)	< 0.00001*
		90	0.000 (0.000)	1.110 (1.100)	< 0.00001*
	200:2,000	20	0.000 (0.000)	0.070 (0.256)	0.01073*
		50	0.000 (0.000)	0.580 (0.755)	< 0.00001*
		90	0.000 (0.000)	0.740 (0.883)	< 0.00001*
$\rho = 0.5$	200:400	20	0.010 (0.100)	0.170 (0.451)	0.0003141*
		50	0.000 (0.000)	1.110 (0.952)	< 0.00001*
		90	0.000 (0.000)	1.580 (1.281)	< 0.00001*
	200:1,000	20	0.000 (0.000)	0.110 (0.373)	0.006008*
		50	0.000 (0.000)	1.080 (1.032)	< 0.00001*
		90	0.000 (0.000)	1.360 (1.210)	< 0.00001*
	200:2,000	20	0.000 (0.000)	0.040 (0.197)	0.07186
		50	0.000 (0.000)	0.740 (0.812)	< 0.00001*
		90	0.000 (0.000)	0.600 (0.899)	< 0.00001*
$\rho = 0.9$	200:400	20	1.520 (1.337)	1.880 (1.665)	0.0365*
		50	0.400 (0.667)	15.520 (3.653)	< 0.00001*
		90	0.060 (0.239)	15.110 (3.890)	< 0.00001*
	200:1,000	20	3.020 (1.875)	3.030 (2.153)	0.7649
		50	0.240 (0.474)	16.980 (3.567)	< 0.00001*
		90	0.000 (0.000)	11.070 (3.468)	< 0.00001*
	200:2,000	20	3.600 (2.445)	4.720 (3.556)	0.0009008*
		50	0.030 (0.171)	10.140 (5.312)	< 0.00001*
		90	0.010 (0.100)	6.290 (2.479)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha = 0.05$

จากตารางที่ 4.3.1 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยคำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ส่วนใหญ่มีค่าเป็นศูนย์ ยกเว้นในกรณีที่ $\rho = 0.9$ ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split มีแนวโน้มลดลง ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้น

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 โดยส่วนใหญ่ นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha = 0.05$ ยกเว้นกรณีที่ขนาดตัวอย่างต่อจำนวนตัวแปรอิสระเป็น 200:1,000 ด้วยจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 20 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0 และ 0.9 ตามลำดับ อีกทั้งกรณีที่ขนาดตัวอย่างต่อจำนวนตัวแปรอิสระเป็น 200:2,000 ด้วยจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 20 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0.5 พบว่าค่า p-value มากกว่า 0.05 นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap ไม่แตกต่างกัน

ตารางที่ 4.3.2 แสดงจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n : p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ ($\lvert s \rvert$)	จำนวนความผิดพลาดเชิงลบ		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho = 0$	200:400	20	0.190 (0.465)	0.000 (0.000)	0.0001937*
		50	48.240 (3.105)	10.950 (3.737)	< 0.00001*
		90	89.960 (0.197)	69.120 (4.152)	< 0.00001*
	200:1,000	20	1.900 (1.772)	0.040 (0.197)	< 0.00001*
		50	49.740 (0.485)	31.600 (4.699)	< 0.00001*
		90	89.960 (0.197)	82.370 (2.766)	< 0.00001*
	200:2,000	20	5.980 (2.853)	0.380 (0.616)	< 0.00001*
		50	49.880 (0.327)	40.330 (3.493)	< 0.00001*
		90	89.950 (0.219)	86.330 (1.615)	< 0.00001*
$\rho = 0.5$	200:400	20	0.180 (0.435)	0.000 (0.000)	0.0001499*
		50	47.870 (2.845)	7.340 (3.497)	< 0.00001*
		90	89.980 (0.141)	60.770 (5.325)	< 0.00001*
	200:1,000	20	1.480 (1.432)	0.020 (0.141)	< 0.00001*
		50	49.510 (0.904)	29.610 (4.634)	< 0.00001*
		90	89.960 (0.197)	81.110 (2.937)	< 0.00001*
	200:2,000	20	0.790 (2.124)	0.050 (0.219)	0.0006552*
		50	49.810 (0.419)	38.650 (3.888)	< 0.00001*
		90	55.760 (43.874)	53.130 (41.837)	< 0.00001*
$\rho = 0.9$	200:400	20	0.770 (0.815)	0.260 (0.441)	< 0.00001*
		50	38.340 (4.768)	9.790 (2.599)	< 0.00001*
		90	89.240 (0.933)	50.460 (4.766)	< 0.00001*
	200:1,000	20	1.250 (1.095)	0.150 (0.386)	< 0.00001*
		50	47.960 (1.699)	20.630 (3.894)	< 0.00001*
		90	89.890 (0.345)	73.690 (4.182)	< 0.00001*
	200:2,000	20	2.790 (2.275)	0.370 (0.734)	< 0.00001*
		50	43.570 (16.192)	29.780 (11.520)	< 0.00001*
		90	89.990 (0.099)	83.150 (2.559)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha = 0.05$

จากตารางที่ 4.3.2 ซึ่งแสดงผลของจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ย คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Random Split และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha = 0.05$

ตารางที่ 4.3.3 แสดงจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย (ค่าเบี่ยงเบนมาตรฐาน) คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป

ระดับความสัมพันธ์ของตัวแปรอิสระ	$n : p$	จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(\lvert s \rvert)$	จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน		2-sided p-value จาก Signed Rank Test
			Random Split	Bootstrap	
$\rho = 0$	200:400	20	19.810 (0.465)	20.100 (0.302)	< 0.00001*
		50	1.760 (3.105)	39.410 (3.596)	< 0.00001*
		90	0.040 (0.197)	21.610 (4.112)	< 0.00001*
	200:1,000	20	18.100 (1.772)	20.000 (0.284)	< 0.00001*
		50	0.260 (0.485)	18.980 (4.656)	< 0.00001*
		90	0.040 (0.197)	8.740 (2.977)	< 0.00001*
	200:2,000	20	14.020 (2.853)	19.690 (0.662)	< 0.00001*
		50	0.120 (0.327)	10.250 (3.500)	< 0.00001*
		90	0.050 (0.219)	4.410 (1.837)	< 0.00001*
$\rho = 0.5$	200:400	20	19.830 (0.451)	20.170 (0.451)	< 0.00001*
		50	2.130 (2.845)	43.770 (3.464)	< 0.00001*
		90	0.020 (0.141)	30.810 (5.504)	< 0.00001*
	200:1,000	20	18.520 (1.432)	20.090 (0.404)	< 0.00001*
		50	0.490 (0.904)	21.470 (4.435)	< 0.00001*
		90	0.040 (0.197)	10.250 (3.322)	< 0.00001*
	200:2,000	20	2.410 (5.657)	3.190 (7.351)	0.0006859*
		50	0.190 (0.419)	12.090 (3.944)	< 0.00001*
		90	0.040 (0.197)	3.270 (3.168)	< 0.00001*
$\rho = 0.9$	200:400	20	20.750 (1.513)	21.620 (1.674)	0.0001064*
		50	12.060 (4.936)	55.730 (3.933)	< 0.00001*
		90	0.820 (0.968)	54.650 (5.406)	< 0.00001*
	200:1,000	20	21.770 (2.169)	22.880 (2.119)	0.000653*
		50	2.280 (1.875)	46.350 (4.755)	< 0.00001*
		90	0.110 (0.345)	27.380 (5.403)	< 0.00001*
	200:2,000	20	17.410 (8.347)	20.950 (9.931)	< 0.00001*
		50	0.460 (0.915)	24.360 (10.592)	< 0.00001*
		90	0.020 (0.199)	13.140 (3.685)	< 0.00001*

หมายเหตุ: * หมายถึง แตกต่างอย่างมีนัยสำคัญที่ $\alpha = 0.05$

จากตารางที่ 4.3.3 ซึ่งแสดงผลของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย คำนวณจากข้อมูลจำลอง 100 ชุด กรณีที่ $n = 200$ ระหว่างวิธี Random Split และวิธีบูตสแตรป พบว่า

1. เมื่อพิจารณาจากจำนวนตัวแปรอิสระ โดยที่ขนาดตัวอย่างเท่ากัน จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น เมื่อตัวแปรอิสระเพิ่มขึ้น

2. เมื่อพิจารณาจากจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ $(|s|)$ จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น ถ้าจำนวนสัมประสิทธิ์จริงที่ไม่เท่าศูนย์เพิ่มขึ้น

3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จะได้ว่าค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split มีแนวโน้มที่จะลดลงเรื่อย ๆ ในทางตรงกันข้ามค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap ไม่สม่ำเสมอ กล่าวคือบางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น

4. เมื่อพิจารณาจากค่า p-value ที่ได้จากวิธี Wilcoxon's Signed Rank Test ในการทดสอบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธี Bootstrap แตกต่างกันหรือไม่ ซึ่งในที่นี้ได้ค่า p-value น้อยกว่า 0.05 ในทุกกรณี นั่นคือจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap แตกต่างกันอย่างมีนัยสำคัญที่ $\alpha = 0.05$

สรุปผลส่วนที่ 4.3 ผลการเปรียบเทียบข้อมูลจำลองขนาด 200 ค่าระหว่างวิธี Random Split และวิธีบูตสแตรป

พิจารณาตารางที่ 4.3.1 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น ยกเว้นในกรณีที่ $\rho = 0.9$ ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split มีแนวโน้มลดลง อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงบวกของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Random Split เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.3.2 เปรียบเทียบจำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบมีแนวโน้มเพิ่มขึ้นเมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนความผิดพลาดในการตรวจจับเชิงลบของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

พิจารณาตารางที่ 4.3.3 เปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ยระหว่างวิธี Random Split และวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานมีแนวโน้มไม่สม่ำเสมอ บางกรณีมีแนวโน้มลดลง บางกรณีมีแนวโน้มเพิ่มขึ้น เมื่อจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์ที่ไม่เท่ากับศูนย์ และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น อีกทั้งจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของทั้งสองวิธีก็แตกต่างกันอย่างมีนัยสำคัญ แต่เมื่อเปรียบเทียบวิธี Random Split กับวิธี Bootstrap พบว่าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีค่าใกล้เคียงจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี

Random Split ในทุกกรณี ซึ่งสามารถกล่าวได้ว่าวิธี Bootstrap เป็นวิธีที่เหมาะสมในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

จากผลการศึกษาตารางที่ 4.3.1 – 4.3.3 กรณีที่ $n = 200$ แม้ว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split ต่ำกว่าวิธี Bootstrap แต่จำนวนความผิดพลาดในการตรวจจับเชิงลบของวิธี Bootstrap ต่ำกว่าวิธี Random Split และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Bootstrap มีค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์มากกว่าวิธี Random Split

โดยรวมแล้วยังสรุปไม่ได้ว่าวิธีการแบ่งข้อมูลในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง วิธีใดมีประสิทธิภาพสูงสุด แต่เมื่อพิจารณาอำนาจการทดสอบพบว่าวิธี Bootstrap มีประสิทธิภาพในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงมากกว่าวิธี Random Split เนื่องจากวิธี Bootstrap มีขนาดตัวอย่างของข้อมูลแต่ละชุดมากกว่าวิธี Random Split ทำให้อำนาจการทดสอบของวิธี Bootstrap มีค่ามากกว่าวิธี Random Split

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษาการเปรียบเทียบวิธี Random Split และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยทำการศึกษาสำหรับข้อมูลที่มีขนาดตัวอย่างต่อจำนวนตัวแปรอิสระเท่ากับ 10:20, 10:50, 10:100, 100:200, 100:500, 100:1,000, 200:400, 200:1,000 และ 200:2,000 ตามลำดับ ด้วยจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 ของขนาดตัวอย่างที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9 เกณฑ์ในการเปรียบเทียบคือ จำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ย จำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ย และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานโดยเฉลี่ย โดยเน้นการเปรียบเทียบความเหมาะสมของวิธีที่จะใช้ในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง สามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

จากการศึกษาเพื่อเปรียบเทียบประสิทธิภาพในการแบ่งข้อมูลระหว่างวิธี Random Split และวิธีบูตสแตรปในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง ถ้าพิจารณาวิธีการแบ่งข้อมูลที่ให้ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวก และค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบต่ำที่สุด จะถือว่าวิธีนั้นมีความเหมาะสมจะใช้ในการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง แต่ถ้าพิจารณาวิธีการแบ่งข้อมูลที่ให้ค่าเฉลี่ยของถ้าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานได้ค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง จากการทดลองสามารถสรุปผลได้ดังนี้

<u>ส่วนที่ 1</u> การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก

เมื่อพิจารณาเปรียบเทียบวิธี Random Split และวิธีบูตสแตรป สิ่งที่ส่งผลต่อจำนวนความผิดพลาดในการตรวจจับเชิงบวกคือ

1. เมื่อจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1 เท่าของขนาดตัวอย่าง จะให้จำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยต่ำกว่ากรณีที่จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.25 เท่าของขนาดตัวอย่าง และจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.45 เท่าของขนาดตัวอย่างตามลำดับ นั่นคือ เมื่อจำนวนข้อมูลที่มีจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เพิ่มมากขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงบวกก็จะเพิ่มขึ้นด้วย
2. อัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อพิจารณาที่ขนาดตัวอย่างเท่ากัน ถ้าตัวแปรอิสระเพิ่มขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงบวกก็จะมีแนวโน้มเพิ่มขึ้น
3. การกำหนดระดับความสัมพันธ์ของตัวแปรอิสระ เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงบวกก็จะเพิ่มขึ้นด้วย

ตารางที่ 5.1.1 แสดงวิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงระหว่างวิธี Random Split และวิธีบูตสแตรปที่ให้ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงบวกต่ำ โดยจำแนกตามอัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อ $|s| = 0.1, 0.25, 0.45$ ของขนาดตัวอย่าง และ $\rho = 0, 0.5, 0.9$

$n : p$	$\|s\|$	ระดับความสัมพันธ์ของตัวแปรอิสระ					
		$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		ความผิดพลาดในการตรวจจับเชิงบวก					
		Random Split	Bootstrap	Random Split	Bootstrap	Random Split	Bootstrap
10:20	1	✓		✓		✓	
	2	✓		✓		✓	
	4	✓		✓		✓	
10:50	1	✓		✓		✓	
	2	✓		✓		✓	
	4	✓		✓		✓	
10:100	1	✓		✓		✓	
	2	✓		✓		✓	
	4	✓		✓		✓	
100:200	10	✓		✓		✓	
	25	✓		✓		✓	
	45	✓		✓		✓	
100:500	10	✓		✓		✓	
	25	✓		✓		✓	
	45	✓		✓		✓	
100:1,000	10	✓		✓		✓	
	25	✓		✓		✓	
	45	✓		✓		✓	
200:400	20	✓		✓		✓	
	50	✓		✓		✓	
	90	✓		✓		✓	
200:1,000	20	✓		✓		✓	
	50	✓		✓		✓	
	90	✓		✓		✓	
200:2,000	20	✓		✓		✓	
	50	✓		✓		✓	
	90	✓		✓		✓	

จากตารางเมื่อพิจารณาจากการเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก โดยพิจารณาเปรียบเทียบวิธี Random Split และวิธีบูตสแตรปว่าวิธีใดเหมาะสมกับการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยถ้าความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยต่ำ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง จะเห็นว่าโดยรวมแล้ววิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง พบว่าวิธี Random Split มีความเหมาะสมที่จะแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

<u>ส่วนที่ 2</u> การเปรียบเทียบความผิดพลาดในการตรวจจับเชิงลบ

เมื่อพิจารณาเปรียบเทียบวิธี Random Split และวิธีบูตสแตรป สิ่งที่ส่งผลต่อจำนวนความผิดพลาดในการตรวจจับเชิงลบคือ

1. เมื่อจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1 เท่าของขนาดตัวอย่างจะให้จำนวนความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยต่ำกว่ากรณีที่จำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.25 เท่าของขนาดตัวอย่าง และจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.45 เท่าของขนาดตัวอย่างตามลำดับ นั่นคือเมื่อจำนวนข้อมูลที่มีจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เพิ่มมากขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงลบก็จะเพิ่มขึ้นด้วย
2. อัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อพิจารณาที่ขนาดตัวอย่างเท่ากัน ถ้าตัวแปรอิสระเพิ่มขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงลบก็จะมีแนวโน้มเพิ่มขึ้น
3. การกำหนดระดับความสัมพันธ์ของตัวแปรอิสระ เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงลบก็จะเพิ่มขึ้นด้วย

ตารางที่ 5.1.2 แสดงวิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงระหว่างวิธี Random Split และวิธีบูตสแตรปที่ให้ค่าเฉลี่ยของจำนวนความผิดพลาดในการตรวจจับเชิงลบต่ำ โดยจำแนกตามอัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อ $|s| = 0.1, 0.25, 0.45$ ของขนาดตัวอย่าง และ $\rho = 0, 0.5, 0.9$

$n:p$	$\|s\|$	ระดับความสัมพันธ์ของตัวแปรอิสระ					
		$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		ความผิดพลาดในการตรวจจับเชิงลบ					
		Random Split	Bootstrap	Random Split	Bootstrap	Random Split	Bootstrap
10:20	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
10:50	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
10:100	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
100:200	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
100:500	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
100:1,000	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
200:400	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓
200:1,000	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓
200:2,000	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓

จากตารางเมื่อพิจารณาจากการเปรียบเทียบความผิดพลาดในการตรวจจับเชิงลบโดยพิจารณาเปรียบเทียบวิธีRandom Split และวิธีบูตสแตรปว่าวิธีใดเหมาะสมกับการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยถ้าความผิดพลาดในการตรวจจับเชิงลบโดยเฉลี่ยต่ำ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง จะเห็นว่าโดยรวมแล้ววิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง พบว่าวิธีบูตสแตรปมีความเหมาะสมที่จะแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

<u>ส่วนที่ 3</u> การเปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน

เมื่อพิจารณาเปรียบเทียบวิธี Random Split และวิธีบูตสแตรป จากผลการทดลองที่เกิดขึ้น เมื่อจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1, 0.25 และ 0.45 เท่าของขนาดตัวอย่าง แม้จะส่งผลต่อจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน แต่ผลก็ไม่เห็นชัดเจนเหมือนกับจำนวนความผิดพลาดในการตรวจจับเชิงบวก และจำนวนความผิดพลาดในการตรวจจับเชิงลบ เช่นเดียวกับอัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อตัวแปรอิสระเพิ่มขึ้น และระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น แม้จะส่งผลต่อจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน แต่ผลก็ไม่เห็นชัดเจนเหมือนกับจำนวนความผิดพลาดในการตรวจจับเชิงบวก และจำนวนความผิดพลาดในการตรวจจับเชิงลบ

ตารางที่ 5.1.3 แสดงวิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูงระหว่างวิธี Random Split และวิธีบูตสแตรปที่ให้ค่าเฉลี่ยของจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานได้ค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ เพื่อเปรียบเทียบของแต่ละวิธี โดยจำแนกตามอัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ เมื่อ $|s| = 0.1, 0.25, 0.45$ ของขนาดตัวอย่าง และ $\rho = 0, 0.5, 0.9$

$n:p$	$\|s\|$	ระดับความสัมพันธ์ของตัวแปรอิสระ					
		$\rho = 0$		$\rho = 0.5$		$\rho = 0.9$	
		จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน					
		Random Split	Bootstrap	Random Split	Bootstrap	Random Split	Bootstrap
10:20	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
10:50	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
10:100	1		✓		✓		✓
	2		✓		✓		✓
	4		✓		✓		✓
100:200	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
100:500	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
100:1,000	10		✓		✓		✓
	25		✓		✓		✓
	45		✓		✓		✓
200:400	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓
200:1,000	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓
200:2,000	20		✓		✓		✓
	50		✓		✓		✓
	90		✓		✓		✓

จากตารางเมื่อพิจารณาจากการเปรียบเทียบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน โดยพิจารณาเปรียบเทียบวิธี Random Split และวิธีบูตสแตรปว่าวิธีใดเหมาะสมกับการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยถ้าจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานได้ค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ จะถือว่ากรณีนั้นมีความเหมาะสมในการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง จะเห็นว่าโดยรวมแล้ววิธีการแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง พบว่าวิธีบูตสแตรปมีความเหมาะสมที่จะแบ่งข้อมูลสำหรับการปรับค่า p-value ของสัมประสิทธิ์การถดถอยที่มีมิติสูง

จากตารางที่ 5.1.1 – 5.1.3 สรุปได้ว่าวิธี Random Split มีประสิทธิภาพในการแบ่งข้อมูลมากกว่าวิธีบูตสแตรปในกรณีจำนวนความผิดพลาดในการตรวจจับเชิงบวก เนื่องจากจำนวนความผิดพลาดในการตรวจจับเชิงบวกโดยเฉลี่ยของวิธี Random Split มีค่าต่ำกว่าวิธีบูตสแตรปในทุกกรณี แต่ในแง่ของจำนวนความผิดพลาดในการตรวจจับเชิงลบกลับพบว่าวิธีบูตสแตรปมีประสิทธิภาพในการแบ่งข้อมูลมากกว่าวิธี Random Split เนื่องจากจำนวนความผิดพลาดในการ ตรวจจับเชิงลบโดยเฉลี่ยของวิธีบูตสแตรปมีค่าต่ำกว่าวิธี Random Split เหมือนกับจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธีบูตสแตรปมีประสิทธิภาพในการแบ่งข้อมูลมากกว่าวิธี Random Split เนื่องจากจำนวนสัมประสิทธิ์การถดถอย ที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานมีค่าใกล้เคียงกับจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับ ศูนย์นั่นเอง

หากจะพิจารณาว่าเกณฑ์ที่ใช้ในการตัดสินใจเกณฑ์ใดมีความสำคัญที่สุด ซึ่งในที่นี้ก็คือ จำนวนความผิดพลาดในการตรวจจับเชิงบวก หรืออีกแง่หนึ่งคือความผิดพลาดประเภทที่ 1 และ จำนวนความผิดพลาดในการตรวจจับเชิงลบ หรืออีกแง่หนึ่งคือความผิดพลาดประเภทที่ 2 มักจะขึ้นอยู่กับบริบทหรือสภาพแวดล้อมนั้น ๆ ตัวอย่างเช่น

- หากพิจารณาว่าความผิดพลาดประเภทที่ 1 เป็นความผิดพลาดที่ร้ายแรงกว่าความผิดพลาดประเภทที่ 2

ถ้าต้องการทดสอบ

H_0: นายเอไม่ได้ฆ่าคน

H_1: นายเอฆ่าคน

ดังนั้นความผิดพลาดประเภทที่ 1 คือ เหตุการณ์ที่นายเอไม่ได้ฆ่าคน แต่ศาลตัดสินว่านายเอเป็นคนฆ่า หรือกล่าวในอีกแง่หนึ่งว่า ศาลตัดสินให้คนบริสุทธิ์ต้องโทษ ส่วนความผิดพลาดประเภทที่ 2 คือ เหตุการณ์ที่นายเอฆ่าคน แต่ศาลตัดสินว่านายเอไม่ได้เป็นคนฆ่า หรือกล่าวในอีกแง่หนึ่ง ว่าศาลตัดสินให้คนผิดลอยนวล ซึ่งในที่นี้เหตุการณ์ที่ศาลตัดสินให้คนบริสุทธิ์ต้องโทษสำคัญกว่า เนื่องจากการทำเช่นนี้อาจทำให้คนบริสุทธิ์เกิดความเสียหายหรือเสียชื่อเสียง

- หากพิจารณาว่าความผิดพลาดประเภทที่ 2 เป็นความผิดพลาดที่ร้ายแรงกว่าความผิดพลาดประเภทที่ 1

ถ้าต้องการทดสอบ

H_0: ยาไม่เป็นอันตรายต่อผู้ใช้

H_1: ยาเป็นอันตรายต่อผู้ใช้

ดังนั้นความผิดพลาดประเภทที่ 1 คือ เหตุการณ์ที่ยาเป็นอันตรายต่อผู้ใช้ ทั้ง ๆ ที่ยาไม่เป็นอันตรายต่อผู้ใช้ ซึ่งพบว่าเหตุการณ์เช่นนี้ไม่เกิดความเสียหายต่อผู้ใช้ ส่วนความผิดพลาดประเภทที่ 2 คือ เหตุการณ์ที่ยาไม่เป็นอันตรายต่อผู้ใช้ ทั้ง ๆ ที่ยาเป็นอันตรายต่อผู้ใช้ ซึ่งในที่นี้เหตุการณ์ที่ ยาไม่เป็นอันตรายต่อผู้ใช้ ทั้ง ๆ ที่ยาเป็นอันตรายต่อผู้ใช้สำคัญกว่า เนื่องจากเหตุการณ์เช่นนี้อาจทำให้ผู้ใช้ถึงแก่ความตายได้

จากตัวอย่างความผิดพลาดประเภทที่ 1 และความผิดพลาดประเภทที่ 2 จะเห็นว่าการที่จะสรุปว่าความผิดพลาดประเภทใดร้ายแรงหรือสำคัญกว่ากัน มักจะขึ้นอยู่กับบริบทหรือสภาพแวดล้อมนั้น ๆ

- <u>เมื่อพิจารณาความคงเส้นคงวา</u>

พบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของวิธี Random Split และวิธีบูตสแตรปให้ผลใกล้เคียงกัน แต่จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธีบูตสแตรปต่างให้ผลที่มีความแตกต่างกันมาก ดังนั้นถ้าพิจารณาความคงเส้นคงวาอาจสรุปได้ว่าทั้ง 3 เกณฑ์ต่างให้แนวโน้มที่เหมือนกันในทุกกรณี

- <u>เมื่อพิจารณาอัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระ</u>

พบว่าเมื่ออัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระยิ่งมาก จำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธีบูตสแตรปจะมีประสิทธิภาพในการทำงานที่ไม่เหมาะสม เพราะเมื่อข้อมูลเปลี่ยนแปลงไปทำให้การตรวจสอบนัยสำคัญของตัวแปรโดยการคำนวณค่า p-value ของสัมประสิทธิ์การถดถอยแต่ละตัวทำได้ยาก

- <u>เมื่อพิจารณาจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์</u>

พบว่าเมื่อจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ยิ่งมาก จำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธีบูตสแตรปจะให้ผลไม่ค่อยดี ส่งผลให้สัมประสิทธิ์การถดถอยแต่ละตัวเชื่อถือไม่ได้ เนื่องจากการศึกษาครั้งนี้คัดกรองตัวแปรด้วยวิธี Adaptive Lasso ซึ่งวิธีนี้ต้องการสัมประสิทธิ์ส่วนใหญ่เป็นศูนย์

- <u>เมื่อพิจารณาระดับความสัมพันธ์ของตัวแปรอิสระ</u>

พบว่าเมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น จำนวนความผิดพลาดในการตรวจจับเชิงบวก จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split และวิธีบูตสแตรปจะให้ผลที่ไม่ค่อยดีนัก

<u>สรุป</u>

เมื่อพิจารณาจำนวนความผิดพลาดในการตรวจจับเชิงบวก ความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน จะเห็นว่าปัจจัยทั้งสามมีความคงเส้นคงวา ไม่ว่าจำนวนตัวแปรอิสระจะเพิ่มขึ้น หรือจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์จะมากแค่ไหนก็ไม่ส่งผลกระทบต่อทั้งสามปัจจัย เมื่อพิจารณาวิธี Random Split กับวิธีบูตสแตรปพบว่าจำนวนความผิดพลาดในการตรวจจับเชิงบวกของทั้ง 2 วิธีให้ผลใกล้เคียงกัน แต่จำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของทั้ง 2 วิธีให้ผลที่มีความแตกต่างกันมาก เนื่องจากจำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานของวิธี Random Split ตรวจพบจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์น้อยกว่าวิธีบูตสแตรป ดังนั้นที่อัตราส่วนของขนาดตัวอย่างกับจำนวนตัวแปรอิสระ จำนวนสัมประสิทธิ์จริง และระดับความสัมพันธ์ของตัวแปรอิสระ ส่งผลกระทบต่อจำนวนความผิดพลาดในการตรวจจับเชิงบวกสูงกว่าจำนวนความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐาน โดยเฉพาะอย่างยิ่งจำนวนสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์ยิ่งมีค่าสูง จำนวนความผิดพลาดในการตรวจจับเชิงบวก ความผิดพลาดในการตรวจจับเชิงลบ และจำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์จากการทดสอบสมมติฐานจะแตกต่างจากเดิมอย่างเห็นได้ชัด

5.2 ข้อเสนอแนะ

1. ในการศึกษาครั้งนี้ผู้วิจัยได้ศึกษาการแจกแจงแบบปกติ สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ข้อมูลมีรูปแบบการแจกแจงแบบอื่น ๆ ซึ่งการที่ข้อมูลมีการแจกแจงเปลี่ยนไปอาจทำให้ได้ผลการวิจัยที่ไม่เหมือนเดิม
2. ศึกษาการแบ่งข้อมูลด้วยวิธีบูตสแตรปที่ใช้ในการหาค่า p-value สำหรับข้อมูลที่มีมิติสูงกับวิธีอื่น ๆ เช่น วิธี SCAD วิธี Elastic Net เป็นต้น

รายการอ้างอิง

Benjamini, Y. and Y. & Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological): 289-300.

Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." Annals of statistics: 1165-1188.

Efron, B. (1979). "Bootstrap methods: another look at the jackknife." The annals of Statistics: 1-26.

Meinshausen, N., et al. (2009). "P-values for high-dimensional regression." Journal of the American Statistical Association 104(488).

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

Zou, H. (2006). "The adaptive lasso and its oracle properties." Journal of the American Statistical Association 101: 1418-1429.

ภาคผนวก

ภาคผนวก ก

โปรแกรม R ที่ใช้ในการจำลอง

ในการศึกษาครั้งนี้จะจำลองข้อมูลตามขอบเขตงานวิจัยข้างต้น ซึ่งผู้วิจัยจะประมวลผลโดยใช้โปรแกรม R เวอร์ชั่น 3.0.2 โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำของข้อมูลแต่ละกรณีไว้ที่จำนวน 100 รอบ ในที่นี้จะขอแสดงเฉพาะกรณีที่มีขนาดตัวอย่างต่อจำนวนตัวแปรอิสระเป็น 100:200 ด้วยสัมประสิทธิ์จริงที่ไม่เท่ากับศูนย์เป็น 0.1 เท่าของขนาดตัวอย่างที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0 ในการยกตัวอย่าง

```
library(mvtnorm)
n<-100
p<-200
nonzero_beta<-0.1
rho<-0

falsepositive_split<-rep(0,100)
falsenegative_split<-rep(0,100)
sizeshat_split<-rep(0,100)
falsepositive_bootstrap<-rep(0,100)
falsenegative_bootstrap<-rep(0,100)
sizeshat_bootstrap<-rep(0,100)

equantile<-function(P,gamma)
{
  output<-rep(0, length(gamma))
  for (m in 1:length(gamma))
  {
    output[m]<-(quantile(P/gamma[m], probs=gamma[m]))
    output[m]<-min(1, output[m])
  }
  return(c(output,m))
}

countnz<-function(x)
{
  return(sum(x!=1))
}
```

```
findfpfn<-function(beta.nzindex, shat)
{
  countmatch<-0
  for (j in 1:length(shat))
  {
    tmp<-match(shat[j], beta.nzindex)
    if (!is.na(tmp))
    {
      countmatch<-countmatch+1
    }
  }
  fp<-length(shat)-countmatch
  fn<-length(beta.nzindex)-countmatch
  return(c(fp, fn))
}

aggregate_pvalue_split<-matrix(rep(0, 100*p), ncol=p)
aggregate_pvalue_bootstrap<-matrix(rep(0, 100*p), ncol=p)

for(k in 1:100){

mu_x<-matrix(data=0,nrow=p,ncol=1)
sigma_x<-matrix(,nrow=p,ncol=p)
 for(j in 1:p){
  for (i in 1:p){
  sigma_x[i,j]=rho^abs(i-j)
    }
}
x<-rmvnorm(n,mu_x,sigma_x)

beta_pos<-sample(1:n,50,replace=F)
beta<-matrix(0,p,1)
nzvalue <- round(nonzero_beta*n)
beta[sample(1:p, nzvalue)]<-runif(nzvalue, 0.5, 5)
```

```
mu_e<-matrix(data=0,nrow=p,ncol=1)
sigma_e<-matrix(c(diag(p)),nrow=p,ncol=p,byrow=TRUE)
e<-matrix(rmvnorm(n,mu_e,sigma_e),n,1)

y<-(x%*%beta)+e

library(parcor)
pvalue.raw<-matrix(rep(0, 50*p), ncol=p)
B<-1
while (B<=50){
```

- วิธี Random Split

```
in.index<-sample(1:n,round(n/2))
x.in_Ran<-x[in.index,]
x.out_Ran<-x[-in.index,]
y.in_Ran<-y[in.index,]
y.out_Ran<-y[-in.index,]
```

- วิธี Adaptive Lasso (Random Split)

```
model.adalasso_Ran<-adalasso(x.in_Ran,y.in_Ran,k=10,use.Gram=F)
beta.adalasso_Ran<-matrix (rep(0,p),ncol=1)
beta.adalasso_Ran[1:p,1]<-model.adalasso_Ran$coefficients.adalasso
```


- วิธี Multi-Split

```
library (Matrix)
count.nonze_Ran<-nnzero(beta.adalasso_Ran, na.counted = NA)

STilde_Ran<-which (beta.adalasso_Ran[,1]!=0)
x.screening_Ran<-x.out_Ran[,STilde_Ran]
x.screening<-as.matrix(x.screening_Ran)

if (sum(beta.adalasso_Ran[1:p,1])!=0){
    OLS_Ran<-lm(y.out_Ran~x.screening_Ran)
```

```
      fit_Ran<-summary(OLS_Ran)

if (fit_Ran$coef[2,4]!="NaN"){
      pvalue.IndVar_Ran<-as.matrix(fit_Ran$coef[,4])
      pvalue_Ran<-matrix(1,p,1)
      pval.adalasso_Ran_pos <- as.matrix(STilde_Ran)

for (bb in 1:count.nonze_Ran){
          test<- as.matrix(fit_Ran$coef[bb+1,4])
                  pvalue_Ran[pval.adalasso_Ran_pos[bb,1]]<- test[,1]
          pvalue_Ran<-as.matrix(pvalue_Ran)
}
          pvalue.raw[B,]<-pvalue_Ran[,1]
      B<-B+1
}
if (fit_Ran$coef[2,4]=="NaN"){B<-B}
}
if (sum(beta.adalasso_Ran[1:p,1])==0){
      pvalue_Ran<- matrix(1,p,1)
}
   }

  pvalue.adjust<-matrix(rep(0, 50*p), ncol=p)
  nz<-apply(pvalue.raw, 1, countnz)

  for(l in 1:50)
  {
   pvalue.adjust[l,]<-pvalue.raw[l,]*nz[l]
   pvalue.adjust[l, which(pvalue.adjust[l,]>1)]<-1
  }
```

- ปรับค่า p-value (Random Split)

```
  gammamin<-0.05
  for (a in 1:p)
  {
```

```
    aggregate_pvalue_split[k,a]<-min(1,(1-log(gammamin)) *
min(equantile(as.vector(pvalue.adjust[,a]), seq(gammamin, 1, 0.05))))
  }
```

- Control FDR (Random Split)

```
  pvalue_sort<-sort(aggregate_pvalue_split[k,])
  pvalue_order<-order(aggregate_pvalue_split[k,])
  qlevel<-0.1
  fdrcondition<-qlevel/sum(1/(1:p))
  hcondition<-pvalue_sort<=fdrcondition
  fdrcondition[fdrcondition>1]<-0.999999999
  h<-min(which(hcondition==FALSE))
  if (h==1)
  {
    falsepositive_split[k]<-0
    falsenegative_split[k]<-sum(beta!=0)
    sizeshat_split[k]<-0
  }
  else
  {
    h<-h -1
    shat<-sort(pvalue_order[1:h])
    tmp<-findfpfn(which(beta!=0), shat)
    falsepositive_split[k]<-tmp[1]
    falsenegative_split[k]<-tmp[2]
    sizeshat_split[k]<-length(shat)
  }

#####################################################################
library(parcor)
pvalue.raw<-matrix(rep(0, 50*p), ncol=p)
B<-1
while(B<=50){
```

- วิธี Bootstrap

```
in.index<-sample(1:n, n, replace=T)
x.in_Boot<-x[in.index,]
y.in_Boot<-y[in.index,]
out.index<-sample(1:n, n, replace=T)
x.out_Boot<-x[out.index,]
y.out_Boot<-y[out.index,]
```

- Adaptive Lasso (Bootstrap)

```
model.adalasso_Boot<-adalasso(x.in_Boot,y.in_Boot,k=10,use.Gram=F)
beta.adalasso_Boot<-matrix(rep(0,p),ncol=1)
beta.adalasso_Boot[1:p,1]<-model.adalasso_Boot$coefficients.adalasso
```

- วิธี Multi-Split (Bootstrap)

```
library(Matrix)
count.nonze_Boot<-nnzero(beta.adalasso_Boot, na.counted = NA)

STilde_Boot<-which(beta.adalasso_Boot[,1]!=0)
x.screening_Boot<-x.out_Boot[,STilde_Boot]
x.screening<-as.matrix(x.screening_Boot)
if(sum(beta.adalasso_Boot[1:p,1])!=0){
    OLS_Boot<-lm(y.out_Boot~x.screening_Boot)
    fit_Boot<-summary(OLS_Boot)

if(fit_Boot$coef[2,4]!="NaN"){
    pvalue.IndVar_Boot<-as.matrix(fit_Boot$coef[,4])

    pvalue_Boot<-matrix(1,p,1)
    pval.adalasso_Boot_pos <- as.matrix(STilde_Boot)

for(bb in 1:count.nonze_Boot){
    test<- as.matrix(fit_Boot$coef[bb+1,4])
    pvalue_Boot[pval.adalasso_Boot_pos[bb,1]]<- test[,1]
```

```
      pvalue_Boot<-as.matrix(pvalue_Boot)
}
      pvalue.raw[B,]<-pvalue_Boot[,1]
      B<-B+1
}

if(fit_Boot$coef[2,4]=="NaN"){B<-B}
}

if(sum(beta.adalasso_Boot[1:p,1])==0){
      pvalue_Boot<- matrix(1,p,1)
}
   }
      pvalue.adjust<-matrix(rep(0, 50*p), ncol=p)
      nz<-apply(pvalue.raw, 1, countnz)

  for(l in 1:50)
  {
   pvalue.adjust[l,]<-pvalue.raw[l,]*nz[l]
   pvalue.adjust[l, which(pvalue.adjust[l,]>1)]<-1
  }
```

- ปรับค่า p-value (Bootstrap)

```
  gammamin<-0.05
  for (a in 1:p)
  {
     aggregate_pvalue_bootstrap[k,a]<-min(1,(1-log(gammamin)) *
min(equantile(as.vector(pvalue.adjust[,a]), seq(gammamin, 1, 0.05))))
  }
```

- Control FDR (Bootstrap)

```
      pvalue_sort<-sort(aggregate_pvalue_bootstrap[k,])
      pvalue_order<-order(aggregate_pvalue_bootstrap[k,])
      qlevel<-0.1
      fdrcondition<-qlevel/sum(1/(1:p))
```

```
      fdrcondition[fdrcondition>1]<-0.999999999
      hcondition<-pvalue_sort<=fdrcondition
      h<-min(which(hcondition==FALSE))
      if (h==1)
  {
    falsepositive_bootstrap[k]<-0
    falsenegative_bootstrap[k]<-sum(beta!=0)
    sizeshat_bootstrap[k]<-0
  }
  else
  {
    h<-h - 1
    shat<-sort(pvalue_order[1:h])
    tmp<-findfpfn(which(beta!=0), shat)
    falsepositive_bootstrap[k]<-tmp[1]
    falsenegative_bootstrap[k]<-tmp[2]
    sizeshat_bootstrap[k]<-length(shat)
  }

}  # End k

output<-data.frame(falsepositive_split, falsenegative_split, sizeshat_split,
falsepositive_bootstrap, falsenegative_bootstrap, sizeshat_bootstrap)
write.table(output, 'D:/tmp/output.csv', quote=F, row.names=F, col.names=T, sep=",")
```

- การหาค่า Wilcoxon test

```
 fp<-read.table('F:/result/falsepositive.csv', sep=",", header=T)

diff<-fp[,1]-fp[,2]

boxplot(diff)

wilcox.test(fp[,1], fp[,2], paired=TRUE, exact=FALSE)


fn<-read.table('F:/result/falsenegative.csv', sep=",", header=T)

diff<-fn[,1]-fn[,2]
```

```
boxplot(diff)

wilcox.test(fn[,1], fn[,2], paired=TRUE, exact=FALSE)

size<-read.table('F:/result/sizeshat.csv', sep=",", header=T)

diff<-size[,1]-size[,2]

boxplot(diff)

wilcox.test(size[,1], size[,2], paired=TRUE, exact=FALSE)
```

ภาคผนวก ข

Wilcoxon's Signed – Rank Test

สถิติทดสอบ Wilcoxon's Signed – Rank Test เป็นสถิติที่ไม่ใช้พารามิเตอร์ ใช้ทดสอบความแตกต่างระหว่างประชากร 2 ชุดที่ไม่อิสระกัน โดยนำเอาปริมาณความมากน้อยของความแตกต่างเข้ามาพิจารณาด้วย (มัลลิกา บุนนาค, 2555) ดังนั้นข้อมูลที่ใช้ในการทดสอบของวิลคอกซันนี้อย่างน้อยที่สุดจะต้องวัดในมาตรฐานแสดงอันตรภาค (interval scale) หรือวัดในมาตราเรียงอันดับ (ordinal scale)

วิธีการทดสอบ

1. กำหนดสมมติฐาน

   H_0: มัธยฐานของประชากร 2 ประชากรไม่แตกต่างกัน หรือ $M_1 = M_2$

   H_1: มัธยฐานของประชากร 2 ประชากรแตกต่างกัน หรือ $M_1 \neq M_2$
2. หาค่าผลต่างระหว่างข้อมูล 2 ชุดทีละคู่ คือค่า $D = X_1 - X_2$
3. ให้ลำดับที่แก่ค่า $|X_1 - X_2|$ จากน้อยไปมาก ถ้าผลต่างเท่ากันให้ใช้ลำดับที่เฉลี่ย (mean rank) ถ้าผลต่างเป็น 0 ให้ตัดทิ้งไป
4. ให้เครื่องหมายบวกลบแก่ลำดับที่ในข้อ 3 ตามเครื่องหมายเดิม
5. หาผลรวมของลำดับที่ที่มีเครื่องหมาย + ให้เป็น $T+$ และหาผลรวมของลำดับที่ที่มีเครื่องหมาย - ให้เป็น $T-$
6. หาค่า $T = \min(T+, T-)$
7. นำค่า T ไปเทียบกับค่าวิกฤต T_α จากตารางแสดงค่าวิกฤตสำหรับการทดสอบโดยใช้ลำดับที่ของวิลคอกซัน และจะปฏิเสธสมมติฐาน H_0 ถ้า $T \leq T_\alpha$
8. ถ้า $n > 30$ T จะมีการแจกแจงโดยประมาณใกล้เคียงการแจกแจงแบบปกติ โดยมี

   ค่าเฉลี่ย $(\mu_T) = \dfrac{n(n+1)}{4}$

   ค่าแปรปรวน $(\sigma_T^2) = \dfrac{n(n+1)(2n+1)}{24}$

ดังนั้น ตัวสถิติสำหรับการทดสอบ คือ

$$Z = \frac{T - \mu_T}{\sigma_T}$$

และนำค่า Z ที่คำนวณได้นี้ไปเทียบกับตารางการแจกแจงแบบปกติมาตรฐาน

เนื่องจากในที่นี้เป็นการทดสอบความไม่เท่ากันจึงเป็นการทดสอบแบบสองด้าน (2-sided p-value) นั่นคือ จะปฏิเสธ H_0 เมื่อ Z ที่คำนวณได้น้อยกว่า $-Z_\alpha$ หรือมากกว่า Z_α

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวบงกชพร เนาวนัติ เกิดวันจันทร์ที่ 27 มิถุนายน พ.ศ. 2531 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ในปีการศึกษา 2553 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555