การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูงแบบเป็นลำดับชั้น

นางสาวสวรรยา ภู่เงิน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557

A COMPARISON OF VARIABLE SCREENING METHODS FOR HIERARCHICAL TESTING OF HIGH-DIMENSIONAL REGRESSION COEFFICIENTS

Miss Sawanya Poongoen

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 2014

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูงแบบเป็นลำดับชั้น
โดย	นางสาวสวรรยา ภู่เงิน
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

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

.....กรรมการ
(รองศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์)

.....กรรมการภายนอกมหาวิทยาลัย
(ศาสตราจารย์ ดร. แอนโทนี เฮย์เตอร์)

สวรรยา ภู่เงิน : การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูงแบบเป็นลำดับชั้น (A COMPARISON OF VARIABLE SCREENING METHODS FOR HIERARCHICAL TESTING OF HIGH-DIMENSIONAL REGRESSION COEFFICIENTS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. วิฐรา พึ่งพาพงศ์, 98 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระจากวิธี Lasso, Adaptive Lasso และ Elastic Net สำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น ในการจัดกลุ่มตัวแปรตามความสัมพันธ์ของตัวแปรอิสระ จากนั้นจึงใช้วิธีการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง เพื่อหาค่า p-value ของกลุ่มสัมประสิทธิ์การถดถอยแต่ละกลุ่ม โดยการศึกษานี้จะเปรียบเทียบประสิทธิภาพของวิธีคัดกรองตัวแปรอิสระจากการจำลองข้อมูลและใช้ข้อมูลจริงที่มีขอบเขตต่างๆกัน โดยในส่วนของข้อมูลจำลองมีอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรเท่ากับ 100: 500 และ 100:1000 และความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 , 0.5 และ 0.9 ขณะที่ในส่วนของข้อมูลจริงจะมีความสัมพันธ์ของตัวแปรแปรอิสระเป็น 2 แบบ คือมีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง ทั้งนี้จะใช้อัตราความผิดพลาดรวม และอำนาจการทดสอบเป็นเครื่องมือในการเปรียบเทียบและการวัดประสิทธิภาพ

การศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่าการคัดกรองตัวแปรทั้ง 3 วิธีให้อำนาจการทดสอบต่ำ ซึ่งเมื่อเปรียบเทียบการคัดกรอง 3 วิธีพบว่า วิธี Lasso มีอำนาจการทดสอบมากที่สุด รองลงมาคือวิธี Adaptive Lasso และ วิธี Elastic Net ตามลำดับ แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวม พบว่าวิธี Adaptive Lasso และวิธี Elastic Net มีค่าต่ำที่สุด

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2557

5681598226 : MAJOR STATISTICS

KEYWORDS: HIGH-DIMENSIONAL DATA / MULTI-SPLIT / HIERARCHICAL CLUSTERING

SAWANYA POONGOEN: A COMPARISON OF VARIABLE SCREENING METHODS FOR HIERARCHICAL TESTING OF HIGH-DIMENSIONAL REGRESSION COEFFICIENTS. ADVISOR: VITARA PUNGPAPONG, Ph.D., 98 pp.

This research is aimed to compare the variable screening methods including Lasso, Adaptive Lasso and Elastic Net for hierarchical testing of high-dimensional regression coefficients. Hierarchical Clustering is employed to group independents variables based on their correlations. Multi-split method is then used to obtain p-values for each group of regression coefficients. Simulated data and real data are carried out to compare the performance of variable screening methods. For simulated data, we consider the case when ratios of the sample size and number of independent variables are 100:500 and 100:1000 and the correlation among independent variables are 0.0 , 0.5 and 0.9. For real data sets, normal correlation and high correlation among independent variables are considered here. Family wise error rate and power of the test are computed to compare the performance of variable screening methods.

In this study, we found that all three screening methods have low power. Furthermore, Lasso has the largest power followed by Adaptive Lasso and Elastic Net respectively. However, Adaptive Lasso and Elastic Net has lower family wise error rate than Lasso.

Department: Statistics
Field of Study: Statistics
Academic Year: 2014

Student's Signature .....

Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความช่วยเหลือและความเอาใจใส่จาก อาจารย์ ดร.วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่กรุณาให้คำปรึกษา อบรมสั่งสอน และให้ข้อคิดเห็นต่างๆ ตลอดจนให้ความช่วยเหลือ คำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณท่าน รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์ และ Professor Anthony J. Hayter กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อสอบ ตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ให้โอกาสทางการศึกษา และอบรมสั่งสอนให้ความรู้ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่คอยให้กำลังใจและความห่วงใย ส่งเสริมและสนับสนุนมาโดยตลอด และขอขอบคุณเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือ ให้คำแนะนำและเป็นกำลังใจให้กับผู้วิจัยตลอดมา

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากความเจริญก้าวหน้าทางด้านเทคโนโลยีในโลกยุคปัจจุบันมีความรวดเร็วและสะดวกมากขึ้น ในการเก็บข้อมูลเพื่อใช้ในการวิจัยสามารถทำได้ง่าย อีกทั้งยังเก็บข้อมูลในปริมาณที่มากได้ในระยะเวลาอันสั้น ทำให้ในปัจจุบันเราต้องพบเจอกับข้อมูลที่มีขนาดใหญ่ขึ้นกว่าในอดีตมาก ซึ่งส่งผลให้มีผู้สนใจที่จะวิเคราะห์เพื่อศึกษาข้อมูลที่มีขนาดใหญ่เพิ่มมากขึ้น โดยในการวิเคราะห์การถดถอยเพื่อดูความสัมพันธ์ของตัวแปรอิสระ (Independent Variables) และตัวแปรตาม (Dependent Variables) จะสามารถประมาณค่าสัมประสิทธิ์การถดถอยได้โดยใช้วิธีกำลังสองน้อยที่สุด(Ordinary Least Squares(OLS)) โดยวิธีนี้จะสามารถใช้ได้กับข้อมูลที่จำนวนตัวแปรน้อยกว่าขนาดตัวอย่างเท่านั้น สำหรับข้อมูลที่ตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่าง จะเรียกข้อมูลลักษณะนี้ว่า ข้อมูลที่มีมิติสูง (High-Dimensional Data) ซึ่งจะไม่สามารถใช้วิธีการวิเคราะห์การถดถอยแบบดั้งเดิมได้ นอกจากนี้ ในข้อมูลที่มีมิติสูง ยังอาจจะต้องเผชิญกับปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันสูง (Multicollinearity) จนไม่สามารถเลือกได้ว่าตัวแปรอิสระตัวใดมีความสำคัญมากกว่ากัน ดังนั้นการทดสอบความสำคัญของสัมประสิทธิ์การถดถอยทีละตัวอาจไม่เหมาะสม ในกรณีนี้เราควรทดสอบความสำคัญของตัวแปรอิสระเป็นกลุ่มตามระดับความสัมพันธ์ของตัวแปรอิสระแทน ซึ่ง Mandozzi and Bühlmann (2013) ได้เสนอการทดสอบสมมติฐานแบบเป็นลำดับชั้นของกลุ่มสัมประสิทธิ์การถดถอย โดยที่กลุ่มสัมประสิทธิ์การถดถอยจะถูกจัดกลุ่มตามความสัมพันธ์ของตัวแปรอิสระ โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) ในการวิเคราะห์การถดถอยในแต่ละลำดับชั้นของการแบ่งกลุ่มตัวแปรอิสระ

ในปัจจุบันได้มีการเสนอวิธี Penalized Likelihood Estimator เพื่อใช้ในการหาตัวแบบการถดถอยที่เหมาะสมสำหรับข้อมูลที่มีมิติสูง โดยเป็นการเพิ่ม Likelihood ด้วย Penalty Function ซึ่งเป็นฟังก์ชันในรูปของสัมประสิทธิ์และหาค่าสัมประสิทธิ์ที่ทำให้ค่า Penalized Likelihood ดังกล่าวมีค่าสูงสุด หากเราเลือกใช้ Penalty Function ที่เหมาะสม จะทำให้ค่าสัมประสิทธิ์บางตัวมีค่าเท่ากับศูนย์ โดยที่วิธี Penalized Likelihood Estimator จะทำการประมาณค่าสัมประสิทธิ์ไปพร้อมๆกับ

การคัดเลือกตัวแปรเข้ามายังตัวแบบได้ ซึ่งในการวิจัยครั้งผู้วิจัยสนใจทำการเปรียบเทียบวิธีการคัดกรองตัวแปร 3 วิธี ดังนี้

วิธี Least Absolute Shrinkage and Selection Operator (Lasso) นำเสนอโดย Tibshirani (1996) ในวิธี Lasso จะมีการใช้ ℓ_1 – norm สำหรับ Penalty Function ในการปรับค่าผลรวมความคลาดเคลื่อนกำลังสองให้มีค่าน้อยที่สุด ซึ่งวิธี Lasso มีข้อดีคือจะทำการประมาณค่าสัมประสิทธิ์แบบต่อเนื่อง (Continuous Shrinkage) ไปคราวเดียวกันกับการคัดเลือกตัวแปรเข้ามายังตัวแบบ และวิธี Lasso จะมีประสิทธิภาพสูงในการพยากรณ์ เมื่อตัวแบบจริงมีจำนวนตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมีจำนวนไม่มากนัก แต่วิธี Lasso ยังมีข้อเสียคือ ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง วิธี Lasso มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวแปรเดียวจากกลุ่มตัวแปรอิสระที่มีความสัมพันธ์กันสูงเข้ามายังตัวแบบ จึงทำให้ตัวประมาณค่าสัมประสิทธิ์ที่ได้มีความเอนเอียง (Bias)

วิธี Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso) นำเสนอโดย Hui Zou (2006) เป็นวิธีที่พัฒนามาจากวิธี Lasso แต่มีการเพิ่มเงื่อนไขโดยการให้ค่าน้ำหนัก (Weight) กับพารามิเตอร์แต่ละตัวแตกต่างกันใน Penalty Function ซึ่งการเพิ่มค่าน้ำหนักนี้จะช่วยแก้ไขปัญหาความไม่คงเส้นคงวาที่ทำให้เกิดความเอนเอียงในวิธี Lasso ได้อีกด้วย ซึ่งการใช้ Penalty Function ดังกล่าวจะทำให้ได้ตัวประมาณที่มีคุณสมบัติ Oracle ของตัวประมาณค่ากล่าวคือ เมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ วิธี Adaptive Lasso จะมีความสามารถในการเลือกตัวแปรได้เสมือนกับว่าทราบตัวแบบที่แท้จริง (True Model) ซึ่งคุณสมบัตินี้ไม่มีในวิธี Lasso

วิธี Elastic Net (EN) นำเสนอโดย Hui Zou and Trevor (2005) เป็นวิธีที่แก้ไขข้อจำกัดของวิธี Lasso เนื่องจากมีคุณสมบัติช่วยลดค่าสัมประสิทธิ์ของตัวแปรแบบต่อเนื่องและทำการคัดเลือกตัวแปรไปพร้อมๆกัน โดยวิธี EN มีการให้ Penalty Function อยู่ในรูปของ ℓ_1 – norm และ ℓ_2 – norm และวิธี EN จะมีประสิทธิภาพที่ดีกว่าวิธี Lasso ในกรณีที่ตัวพยากรณ์ในแต่ละตัวแบบมีความสัมพันธ์กันมาก โดยเฉพาะอย่างยิ่งในกรณีที่จำนวนตัวแปรอิสระมีขนาดใหญ่กว่าขนาดของตัวอย่างมากๆ เนื่องจาก วิธี EN สามารถคัดกรองตัวแปรอิสระและประมาณค่าได้ในคราวเดียวกัน

อย่างไรก็ตามวิธีดังกล่าวให้เพียงค่าประมาณของสัมประสิทธิ์โดยไม่ได้รายงานค่า p-value ซึ่งเป็นค่าที่บ่งบอกถึงความสำคัญของตัวแปรและมีความจำเป็นต่อผู้วิเคราะห์ผล ดังนั้นการคัดกรองตัวแปร จึงอาศัยเพียงค่าสัมประสิทธิ์การถดถอยของตัวแปรเท่านั้น ต่อมา ในปี 2009 Meinshausen, Meier, and Bühlmann (2009) ได้เสนอวิธีการหาค่า p-value สำหรับการทดสอบสัมประสิทธิ์การ

ถดถอยกรณีที่ข้อมูลมีมิติสูงโดยใช้วิธีการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง (Multi-split) ซึ่งจะทำการทดสอบสัมประสิทธิ์การถดถอยทีละตัว โดยวิธีการดังกล่าวจะทำการสุ่มแบ่งข้อมูลออกเป็นสองส่วนเท่าๆกัน จากนั้นใช้ข้อมูลเพียงส่วนแรกในการคัดกรองตัวแปรโดยวิธี Penalized Likelihood และใช้ข้อมูลส่วนที่สองในการคำนวณหาค่า p-value สำหรับตัวแปรที่ผ่านการคัดกรองแล้วเท่านั้น ซึ่งจำนวนตัวแปรที่ผ่านการคัดกรองจะมีจำนวนน้อยกว่าครึ่งหนึ่งของขนาดตัวอย่าง ดังนั้น การหาค่า p-value สามารถทำได้ปกติโดยไม่ติดปัญหาเรื่องมิติที่สูงของข้อมูลอีกต่อไป จากนั้นจึงทำการสุ่มแบ่งข้อมูลหลายๆครั้งและทำซ้ำกระบวนการข้างต้น

ต่อมา ศุภวัฒน์ อังคะสี (2556) ได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของวิธี Penalized Likelihood ต่างๆในขั้นตอนการคัดกรองตัวแปรสำหรับวิธีการหาค่า p-value โดยวิธีการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง โดยทำการเปรียบเทียบวิธี Lasso (Tibshirani (1996)), วิธี Adaptive Lasso (Hui Zou (2006)), Elastic Net (EN) (Hui Zou and Trevor (2005)) และ Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li (2001)) และใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของแต่ละวิธีสามเกณฑ์ ได้แก่ ความผิดพลาดในการตรวจจับเชิงบวก ความผิดพลาดในการตรวจจับเชิงลบ และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับศูนย์ ศุภวัฒน์ อังคะสี (2556) พบว่าวิธี Adaptive Lasso และวิธี SCAD มีประสิทธิภาพดีกว่าอีกสองวิธีในแง่ของความผิดพลาดในการตรวจจับเชิงบวกและความผิดพลาดในการตรวจจับเชิงลบ แต่หากพิจารณาจำนวนสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์ พบว่าวิธี Lasso และวิธี Elastic Net จะให้จำนวนสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับศูนย์ใกล้เคียงกับตัวแบบที่แท้จริงมากกว่า

จากการใช้วิธีการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง และหาค่า p-value ของสัมประสิทธิ์การถดถอยทีละตัวเราจะทำการหาค่า p-value ของกลุ่มสัมประสิทธิ์การถดถอยแทน แล้วนำค่า p-value ที่ได้ในแต่ละครั้งมาทำการปรับค่าโดยยังไม่มีการรวมค่า หลังจากนั้นจึงนำค่า p-value ที่ทำการปรับค่าแล้วทั้งหมดมาทำการรวมค่าอีกครั้งด้วยฟังก์ชันควอนไทล์ (Quantile Function) เมื่อได้ค่า p-value สำหรับการทดสอบกลุ่มสัมประสิทธิ์การถดถอยจากวิธี Multi-Split เป็นที่เรียบร้อยแล้วจึงนำค่า p-value ดังกล่าวมาใช้ในการเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระว่าส่งผลต่อการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยหรือไม่ และวิธีการคัดกรองตัวแปรอิสระแบบใดที่มีประสิทธิภาพในการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยมากที่สุด โดยใช้อัตราความ

ผิดพลาดรวม (Family Wise Error Rate (FWER)) และค่าอำนาจการทดสอบ (Power of Test) ของค่า p-value ที่ได้จากการทดสอบกลุ่มสัมประสิทธิ์การถดถอยของข้อมูลที่มีมิติสูงแบบเป็นลำดับชั้น

ในการศึกษาครั้งนี้ ผู้วิจัยมีความสนใจในการทดสอบกลุ่มสัมประสิทธิ์การถดถอยของข้อมูลที่มีมิติสูงแบบเป็นลำดับชั้นและต้องการจะเปรียบเทียบว่าวิธีการคัดกรองตัวแปรที่ดีที่สุด จากวิธี Lasso, Adaptive Lasso และ Elastic Net แล้วนำตัวแปรที่ได้จากการคัดกรองจากทั้ง 3 วิธีที่กล่าวมานั้นมาใช้ในการหาค่า p-value และนำค่า p-value ที่ได้นี้มาใช้ในหาอัตราความผิดพลาดรวม (Family Wise Error Rate (FWER)) และค่าอำนาจการทดสอบ (Power of Test) เพื่อเปรียบเทียบว่าวิธีการคัดกรองตัวแปรอิสระแบบใดที่มีประสิทธิภาพในการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยมากที่สุด

1.2 วัตถุประสงค์

เพื่อศึกษาและเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระในแบบจำลองเชิงเส้นตรง เพื่อหาค่า p-value สำหรับทดสอบกลุ่มสัมประสิทธิ์การถดถอย

1.3 ข้อตกลงเบื้องต้น

ในการศึกษาครั้งนี้จะทำการเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระในแบบจำลองเชิงเส้นตรง เพื่อหาค่า p-value สำหรับทดสอบกลุ่มสัมประสิทธิ์การถดถอย โดยใช้ข้อมูลจำลองจากตัวแบบข้างล่างดังนี้

$$\mathrm{Y} = \mathbf{X}\beta + \varepsilon \tag{1.1}$$

เมื่อ $\mathrm{Y} = (\mathrm{y}_1, \mathrm{y}_2, \dots, \mathrm{y}_\mathrm{n})'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x 1 ที่เป็นตัวแปรสุ่มมาตรฐาน

$\mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n})'$ เป็นเมทริกซ์ตัวแปรอิสระขนาด n x p ที่เป็นตัวแปรสุ่มมาตรฐาน โดยที่ $\mathrm{X}_\mathrm{i} = \begin{bmatrix} \mathrm{x}_{\mathrm{i}1} \\ \vdots \\ \mathrm{x}_{\mathrm{ip}} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \dots, \beta_\mathrm{p})'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_\mathrm{n})'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1 และ $\varepsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_\mathrm{n})$

โดยค่าแปรปรวนของค่าความคลาดเคลื่อน σ^2 จะถูกเลือกเพื่อให้ได้อัตราส่วนสัญญาณต่อสัญญาณรบกวนอยู่ในระดับปานกลาง (Moderate Signal-to-noise Ratio) ดังสมการต่อไปนี้

$$\mathrm{SNR} = \sqrt{\frac{(\beta^0)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta^0}{n\sigma^2}} \qquad (1.2)$$

เมื่อ SNR เป็นอัตราส่วนสัญญาณต่อสัญญาณรบกวนที่กำหนดตามจำนวนของตัวแปร

n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอิสระ

และเวกเตอร์ค่าสัมประสิทธิ์การถดถอย β จะได้จากการสุ่ม มาเป็นจำนวน 20 ค่าที่ไม่เท่ากับ 0 ในตำแหน่งของ β ดังนี้

$$\begin{aligned}\{\ &\beta_1 = \beta_2 \ldots = \beta_8 = 1,\\ &\beta_{101} = \beta_{102} = \cdots = \beta_{104} = 1,\\ &\beta_{201} = \beta_{202} = \cdots = \beta_{204} = 1,\\ &\beta_{301} = \beta_{302} = 1,\\ &\beta_{401} = \beta_{402} = 1\ \}\end{aligned}$$

1.4 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาในส่วนของข้อมูลจำลองและข้อมูลจริง ภายใต้ขอบเขตการวิจัยดังนี้

1.4.1 ข้อมูลจำลอง

ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระที่ต่างกัน (n:p) 2 ระดับ คือ 100:500 และ 100:1000 และระดับความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ

ระดับที่ 1 : ρ = 0.0

ระดับที่ 2 : ρ = 0.5

ระดับที่ 3 : ρ = 0.9

สำหรับเมทริกซ์ $\mathbf{X}$ จะใช้การจำลอง $X_i \sim N(0, \Sigma)$

โดยที่ $\Sigma = \begin{bmatrix} \begin{pmatrix} 1 & \cdots & \rho_{1,100} \\ \vdots & \ddots & \vdots \\ \rho_{100,1} & \cdots & 1 \end{pmatrix} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \begin{pmatrix} 1 & \cdots & \rho_{p-100,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,p-100} & \cdots & 1 \end{pmatrix} \end{bmatrix}$

นอกจากนี้ในส่วนของค่าแปรปรวนของค่าความคลาดเคลื่อน σ^2 จะถูกเลือกเพื่อให้ได้อัตราส่วนสัญญาณต่อสัญญาณรบกวน(SNR) เท่ากับ 12 เมื่อ p = 500 และ 24 เมื่อ p = 1,000

1.4.2 ข้อมูลจริง

1.4.2.1 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันปกติ

ข้อมูลที่ได้มาจาก (Van't Veer et al., 2002) เป็นข้อมูลไมโครแอเรย์ของดีเอ็นเอเนื้องอกในเต้านมของผู้ป่วย 337 คน โดยตัวแปรอิสระแต่ละตัว คือ ระดับการแสดงออกของยีน (Gene Expression) แต่ละตัวซึ่งมีทั้งหมดมากกว่า 20,000 ยีนในข้อมูลชุดนี้ โดยที่ข้อมูลยีนที่เก็บจะเรียงลำดับตามตำแหน่งทางพันธุกรรม (เช่น จากโครโมโซมที่ 1 ถึง 22) จากนั้นทำการสุ่มตัวแปรอิสระขึ้นมาตามจำนวนที่ต้องการและทำการสุ่มคนไข้มาเพียง 100 คน

โดยเปรียบเทียบอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ที่ 100:500 , 100:1000

1.4.2.2 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันสูง

ซึ่งจะใช้ข้อมูลจากชุดแรกแต่จะแตกต่างกันในกระบวนการสุ่มที่จะทำการสุ่มเป็นชุดๆ คือ จะสุ่มตัวแปรอิสระทีละ 100 ตัวแปรที่อยู่ติดกันจากตัวแปรทั้งหมด เนื่องจากข้อมูลยีนมีการจัดเรียงลำดับตามตำแหน่งทางพันธุกรรม ดังนั้นตัวแปรที่อยู่ติดกันจึงมีความสัมพันธ์กันสูงกว่าตัวแปรที่อยู่ห่างกัน และทำการสุ่มคนไข้มาเพียง 100 คน

โดยเปรียบเทียบอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ที่ 100:500 , 100:1000

1.5 คำจำกัดความที่ใช้ในงานวิจัย

ข้อมูลที่มีมิติสูง (High-Dimensional Data)

คือ ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง

การแบ่งข้อมูลโดยสุ่มหลายๆครั้ง (The Multi - Split)

คือ การแบ่งข้อมูลออกเป็น 2 ส่วนโดยการสุ่มหลายๆครั้ง และนำข้อมูลดังกล่าวไปใช้ในการวิเคราะห์ผลทางสถิติ

อัตราความผิดพลาดรวม (Family Wise Error Rate (FWER))

คือ อัตราส่วนของจำนวนการจำลองที่เกิดความผิดพลาดประเภทที่ 1 อย่างน้อย 1 ครั้ง ต่อจำนวนครั้งที่ทำการจำลองข้อมูล

ค่าอำนาจการทดสอบ (Power of Test)

คือ การวัดจำนวนที่เกิดความถูกต้องจากข้อสรุปที่ว่าค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นตรงมีค่าไม่เท่ากับศูนย์เมื่อค่าประมาณสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการคัดกรองตัวแปรอิสระวิธีใดเหมาะสมในการคัดกรองตัวแปรและมีประสิทธิภาพมากที่สุด สำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง ในการทดสอบสมมติฐาน $H_0^C : \beta_j = 0$ สำหรับ ทุกๆ $j \in C$ และ $H_A^C : \beta_j \neq 0$ สำหรับ j อย่างน้อย 1 ตัวที่ $j \in C$ เราจะทำการปฏิเสธสมมติฐานก็ต่อเมื่อ $P_h^C < \alpha$ เมื่อ α คือระดับนัยสำคัญ

1. อัตราความผิดพลาดรวม (Family Wise Error Rate (FWER))

อัตราความผิดพลาดรวม คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 1 (Type I error rate) อย่างน้อย 1 ครั้ง โดยความผิดพลาดประเภทที่ 1 เกิดจากการปฏิเสธสมมติฐานว่าง H_0 เมื่อสมมติฐานว่าง H_0 เป็นจริง ในข้อมูลจำลองแต่ละกรณี ซึ่งเป็นโอกาสของการกระทำความผิดพลาดประเภทที่ 1 อย่างน้อย 1 ครั้งของชุดการเปรียบเทียบจำนวน 1 ชุด ดังนั้นอัตราความผิดพลาดรวมจะหาได้จากสมการต่อไปนี้

$$\text{FWER} = \frac{\text{จำนวนการจำลองที่เกิดความผิดพลาดประเภทที่ } \mathbf{1} \text{ อย่างน้อย } \mathbf{1} \text{ ครั้ง}}{\text{จำนวนการจำลองข้อมูล}} \qquad (1.3)$$

สำหรับการวิจัยในครั้งนี้ทางผู้วิจัยจะทำการจำลองข้อมูล 50 ครั้ง ดังนั้นอัตราความผิดพลาดรวมจะหาได้จาก

$$\text{FWER} = \frac{\text{จำนวนการจำลองที่เกิดความผิดพลาดประเภทที่ } \mathbf{1} \text{ อย่างน้อย } \mathbf{1} \text{ ครั้ง}}{\mathbf{50}} \tag{1.4}$$

โดยทฤษฎีที่ 1 ใน Mandozzi and Bühlmann (2013) ระบุว่าหากเราทำการปฏิเสธ $\mathrm{H}_0^{\mathrm{C}}$ ก็ต่อเมื่อ $\mathrm{P}_{\mathrm{h}}^{\mathrm{C}} < \alpha$ แล้ว อัตราความผิดพลาดรวมจะเท่ากับ α ด้วย ดังนั้นอัตราความผิดพลาดรวม (FWER) ที่คำนวณได้จึงไม่ควรเกิน α

2. ค่าอำนาจการทดสอบ (Power of Test)

อำนาจการทดสอบ หมายถึง ความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่าง H_0 เมื่อสมมติฐานว่าง H_0 เป็นเท็จ ซึ่งในการทดสอบทางสถิติที่มีค่าอำนาจการทดสอบยิ่งมากจะแสดงว่าการทดสอบนั้นยิ่งดี ค่าอำนาจการทดสอบสามารถคำนวณได้ดังนี้

$$\text{Power} = \frac{\text{จำนวนครั้งที่ปฏิเสธ } \mathbf{H_0} \text{ เมื่อ } \mathbf{H_0} \text{ เป็นเท็จ}}{\text{จำนวน } \mathbf{H_0} \text{ ที่เป็นเท็จ}} \tag{1.5}$$

โดยที่ถ้าวิธีใดให้ค่าวัดประสิทธิภาพ FWER ต่ำที่สุด และให้ค่าวัดประสิทธิภาพ Power สูงที่สุด จะถือว่าวิธีนั้นมีความเหมาะสมที่จะใช้ในการคัดกรองตัวแปรอิสระและมีความเหมาะสมที่จะใช้ในการประมาณค่ากลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง

1.7 วิธีการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง
2. กำหนดการจำลองข้อมูล
   2.1 กำหนดค่าเริ่มต้นและรูปแบบของตัวแบบที่ใช้จำลองข้อมูลโดยจะทำการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และพารามิเตอร์จำนวน p ตัว โดยใช้อัตราส่วน n:p ดังนี้
      - 100:500
      - 100:1000

2.2 ทำการจำลองข้อมูลในส่วนของข้อมูลจำลองให้มีการแจกแจงแบบนอร์มอล (Normal Distribution)

$$\boxed{Y = \mathbf{X}\beta + \varepsilon} \quad \text{โดยที่ } X_i \sim N(0, \Sigma) \text{ , } \varepsilon \sim N(0, \sigma^2 I_n)$$

สำหรับเมทริกซ์ $\mathbf{X}$ จะใช้การจำลอง $X_i \sim N(0, \Sigma)$

โดยที่ $\Sigma = \begin{bmatrix} \begin{pmatrix} 1 & \cdots & \rho_{1,100} \\ \vdots & \ddots & \vdots \\ \rho_{100,1} & \cdots & 1 \end{pmatrix} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \begin{pmatrix} 1 & \cdots & \rho_{p-100,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,p-100} & \cdots & 1 \end{pmatrix} \end{bmatrix}$

เมื่อ $Y = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x 1 ที่เป็นตัวแปรสุ่มมาตรฐาน

$\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ เป็นเมทริกซ์ตัวแปรอิสระขนาด n x p ที่เป็นตัวแปรสุ่มมาตรฐาน โดยที่ $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1 และ $\varepsilon \sim N(0, \sigma^2 I_n)$

โดยในส่วนของค่าแปรปรวนของค่าความคลาดเคลื่อน σ^2 จะถูกเลือกเพื่อให้ได้อัตราส่วนสัญญาณต่อสัญญาณรบกวน(SNR) เท่ากับ 12 เมื่อ p = 500 และ 24 เมื่อ p = 1,000

n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอิสระ

และเวกเตอร์ค่าสัมประสิทธิ์การถดถอย β จะได้จากการสุ่ม มาเป็นจำนวน 20 ค่าที่ไม่เท่ากับ 0 ในตำแหน่งของ β ดังนี้

$$\begin{aligned} \{\ & \beta_1 = \beta_2 \ldots = \beta_8 = 1, \\ & \beta_{101} = \beta_{102} = \cdots = \beta_{104} = 1, \\ & \beta_{201} = \beta_{202} = \cdots = \beta_{204} = 1, \\ & \beta_{301} = \beta_{302} = 1, \\ & \beta_{401} = \beta_{402} = 1 \} \end{aligned}$$

2.3 ทำการสุ่มข้อมูลจากข้อมูลจริงที่ได้เลือกมาศึกษาดังนี้ ในอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรที่เท่ากับข้อมูลจำลอง

2.3.1 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันปกติ

ข้อมูลที่ได้มาจาก (Van't Veer et al., 2002) เป็นข้อมูลไมโครแอเรย์ของดีเอ็นเอเนื้องอกในเต้านมของผู้ป่วย 337 คน โดยตัวแปรอิสระแต่ละตัว คือ ระดับการแสดงออกของยีน (Gene Expression) แต่ละตัวซึ่งมีทั้งหมดมากกว่า 20,000 ยีนในข้อมูลชุดนี้ โดยที่ข้อมูลยีนที่เก็บจะเรียงลำดับตามตำแหน่งทางพันธุกรรม (เช่น จากโครโมโซมที่ 1 ถึง 22) จากนั้นทำการสุ่มตัวแปรอิสระขึ้นมาตามจำนวนที่ต้องการและทำการสุ่มคนไข้มาเพียง 100 คน

2.3.2 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันสูง

ซึ่งจะใช้ข้อมูลจากชุดแรกแต่จะแตกต่างกันในกระบวนการสุ่มที่จะทำการสุ่มเป็นชุดๆ คือ จะสุ่มตัวแปรอิสระทีละ 100 ตัวแปรที่อยู่ติดกันจากตัวแปรทั้งหมด เนื่องจากข้อมูลยีนมีการจัดเรียงลำดับตามตำแหน่งทางพันธุกรรม ดังนั้นตัวแปรที่อยู่ติดกันจึงมีความสัมพันธ์กันสูงกว่าตัวแปรที่อยู่ห่างกัน และทำการสุ่มคนไข้มาเพียง 100 คน

3. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคัดกรองตัวแปรอิสระสำหรับวิธีการจัดกลุ่มแบบเป็นลำดับชั้นของตัวแปรอิสระ และคำนวณค่า p-value

3.1 วิธี Lasso

3.2 วิธี Adaptive Lasso

3.3 วิธี Elastic Net

4. ทำการเปรียบเทียบประสิทธิภาพของ วิธี Lasso, Adaptive Lasso และ Elastic Net โดยใช้เกณฑ์อัตราความผิดพลาดรวม (Family Wise Error Rate (FWER)) และ ค่าอำนาจการทดสอบ (Power of Test) ในการวิเคราะห์ข้อมูลและสถิติที่ได้

1.8 ประโยชน์ที่คาดว่าจะได้รับ

สามารถหาวิธีการคัดกรองตัวแปรอิสระในแบบจำลองเชิงเส้นตรง เพื่อหาค่า p-value สำหรับทดสอบกลุ่มสัมประสิทธิ์การถดถอยของข้อมูลที่มีมิติสูงแบบเป็นลำดับชั้นได้อย่างเหมาะสม

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การประมาณค่าสัมประสิทธิ์การถดถอย (β) ของข้อมูลในตัวแบบเพื่อคัดเลือกตัวแปรเข้ามายังตัวแบบในกรณีที่ข้อมูลมีจำนวนตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ (n>p) สามารถทำได้โดยวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares : OLS) แต่ในปัจจุบันฐานข้อมูลส่วนใหญ่มีขนาดใหญ่ขึ้น ดังนั้นจึงมีข้อมูลที่มีจำนวนของตัวแปรอิสระมากกว่าขนาดตัวอย่าง (p>n) เกิดขึ้นและเรียกข้อมูลประเภทนี้ว่า “ข้อมูลที่มีมิติสูง (High-Dimensional Data)” โดยการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระในกรณีเป็นข้อมูลที่มีมิติสูงนั้นไม่สามารถหาจากวิธีดังกล่าวได้ นอกจากนี้ในข้อมูลที่มีมิติสูง เรายังจะต้องเผชิญกับปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันสูง จนไม่สามารถเลือกได้ว่าตัวแปรอิสระตัวใดมีความสำคัญมากกว่ากันได้ ดังนั้นในงานวิจัยนี้จะกล่าวถึงการทดสอบสมมติฐานแบบเป็นลำดับชั้นของกลุ่มสัมประสิทธิ์การถดถอย โดยที่กลุ่มสัมประสิทธิ์การถดถอยจะถูกจัดกลุ่มตามความสัมพันธ์ของตัวแปรอิสระ โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) และใช้วิธีการคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูง คือ วิธี Lasso วิธี Adaptive Lasso และวิธี Elastic Net แต่เนื่องจากทั้ง 3 วิธีข้างต้นไม่ได้รายงานค่า p-value ที่มีความสำคัญต่อการวิเคราะห์ผลทางสถิติ ดังนั้นจึงจะกล่าวถึงวิธี Multi–Split ซึ่งเป็นวิธีที่ใช้ในการหาค่า p-value สำหรับข้อมูลที่มีมิติสูง และเกณฑ์ที่ใช้ในการตัดสินใจ คือ อัตราความผิดพลาดรวม (Family Wise Error Rate) และค่าอำนาจการทดสอบ (Power of Test) เพื่อใช้ในการวิเคราะห์ข้อมูลและสถิติที่ได้

2.1 การทดสอบสมมติฐานแบบเป็นลำดับชั้น (Mandozzi and Bühlmann (2013))

Mandozzi and Bühlmann (2013) เสนอการหาค่า p-value สำหรับการทดสอบกลุ่มสัมประสิทธิ์การถดถอยโดยพัฒนามาจากแนวคิดการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง (Multi-split) ของ Meinshausen, Meier, and Bühlmann (2009) โดยขั้นตอนหลักในการทดสอบสมมติฐานแบบเป็นลำดับชั้นสำหรับสมมติฐาน $H_0^C : \beta_j = 0$ สำหรับ ทุกๆ $j \in C$ และ $H_A^C : \beta_j \neq 0$ สำหรับ j อย่างน้อย 1 ตัวที่ $j \in C$

มีทั้งหมด 3 ขั้นตอน ดังนี้

2.1.1 การจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) ของตัวแปรอิสระ

2.1.2 การหาค่า p-value ในการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยโดยวิธีการแบ่งข้อมูลแบบสุ่ม B ครั้ง

2.1.3 การรวมค่า p-value B ค่า สำหรับการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยแต่ละกลุ่ม

สำหรับรายละเอียดในแต่ละขั้นตอน เป็นดังต่อไปนี้

2.1.1 การจัดกลุ่มแบบเป็นลำดับชั้นของตัวแปรอิสระ

ในการจัดกลุ่มแบบเป็นลำดับชั้นของตัวแปรอิสระ ผู้วิจัยสนใจที่จะเลือกใช้วิธี Agglomerative Hierarchical Method และใช้ Complete Linkage ในการคำนวณค่าความต่างของตัวแปรอิสระแต่ละคู่ ในการจัดกลุ่มแบบเป็นลำดับชั้นของตัวแปรอิสระทั้งหมด p ตัว วิธี Agglomerative Hierarchical Method จะเริ่มต้นด้วยกลุ่ม p กลุ่ม ซึ่งแต่ละกลุ่มจะมีตัวแปรอิสระเพียงตัวเดียว จากนั้นจึงทำการจับกลุ่มตัวแปรอิสระที่มีความคล้ายคลึงกันมากที่สุด และทำแบบนี้ต่อไปเรื่อยๆ จนกระทั่งเหลือกลุ่มเพียงกลุ่มเดียวเท่านั้น สำหรับขั้นตอนการจัดกลุ่มแบบเป็นลำดับชั้นมีรายละเอียดดังต่อไปนี้

1.) เริ่มต้นจากกลุ่มทั้งหมด p กลุ่ม ซึ่งแต่ละกลุ่มจะมีตัวแปรอิสระเพียงตัวเดียว และเมทริกซ์ความต่าง (Distance Matrix) $D = \{d_{jk}\}$ โดยที่ d_{jk} คือระยะห่างแบบ Euclidean ของกลุ่มที่ j และ k

2.) พิจารณาเมทริกซ์ความต่าง โดยหาระยะห่างของสองกลุ่มที่น้อยที่สุด สมมติให้กลุ่ม U และ V มีระยะห่างที่น้อยที่สุด ซึ่งเท่ากับ d_{UV}

3.) ทำการรวมกลุ่ม U และ V เข้าด้วยกันเป็นกลุ่ม (UV) และปรับค่าระยะห่างในเมทริกซ์ความต่าง ดังนี้

a. ลบแถวและคอลัมน์ที่เกี่ยวกับกลุ่ม U และ V ทั้งหมด

b. เพิ่มแถวและคอลัมน์ และคำนวณค่าความต่างของกลุ่ม (UV) กับกลุ่มอื่นๆ ที่เหลือโดยวิธี Complete Linkage เช่น กลุ่ม W ซึ่งสามารถเขียนในรูปของสมการได้ดังนี้

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\} \quad (2.1)$$

ซึ่ง d_{UW} และ d_{VW} จะใช้เป็นค่าความต่างที่มากที่สุดระหว่างกลุ่ม U กับกลุ่ม W และกลุ่ม V กับกลุ่ม W ตามลำดับ

4.) ทำซ้ำขั้นตอนที่ 2.) และ 3.) จำนวน p - 1 ครั้งจนกระทั่งเหลือกลุ่มเพียงกลุ่มเดียวซึ่งประกอบไปด้วยตัวแปรอิสระทั้งหมด

2.1.2 การหาค่า p-value ในการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยโดยวิธีการแบ่งข้อมูลแบบสุ่ม B ครั้ง

กำหนดให้ $b = 1,2,\dots,B$ เมื่อ B คือจำนวนครั้งในการจำลองข้อมูล และมีขั้นตอนในการหาค่า p-value ดังนี้

1.) ทำการแบ่งข้อมูลออกเป็นสองส่วน เรียกว่า $D_{in}^{(b)}$ และ $D_{out}^{(b)}$ ซึ่งมีขนาด $N_{in}^{(b)}$ และ $N_{out}^{(b)}$ ตามลำดับ โดยที่ $N_{in}^{(b)} + N_{out}^{(b)} = n$ และ $N_{in}^{(b)} = N_{out}^{(b)}$ เมื่อ n เป็นเลขคู่ หรือ $N_{in}^{(b)} + 1 = N_{out}^{(b)}$ เมื่อ n เป็นเลขคี่

2.) ใช้เฉพาะข้อมูล $D_{in}^{(b)}$ ในการคัดกรองตัวแปร ซึ่งจะให้ตัวประมาณ $\hat{S}^{(b)} = \left\{j : \hat{\beta}_j^{(b)} \neq 0\right\}$ โดยมีเงื่อนไขว่า $\left|\hat{S}^{(b)}\right| < \frac{n}{2}$

3.) ใช้เฉพาะข้อมูล $D_{out}^{(b)}$ ในการคำนวณหาค่า p-value ของการทดสอบสมมติฐานว่าง $H_0^{C \cap \hat{S}^{(b)}}: \beta_j = 0$ สำหรับ ทุกๆ $j \in C \cap \hat{S}^{(b)}$ โดยใช้การทดสอบแบบเอฟ

- **การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ของกลุ่มสัมประสิทธิ์การถดถอย**

การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_j สำหรับ ทุกๆ $j \in C$ เมื่อ j = 1, 2, 3, ...,p เป็นการทดสอบว่า $\beta_j = 0$ หรือไม่ ถ้า $\beta_j = 0$ แล้ว จะหมายความว่าไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ X_j สำหรับ ทุกๆ $j \in C$ กับ Y แต่ถ้า $\beta_j \neq 0$ นั่นหมายความว่า X_j สำหรับทุกๆ $j \in C$ กับ Y มีความสัมพันธ์เชิงเส้นต่อกัน ซึ่งรวมไปถึงว่าในการรู้ค่าของ X_j จะช่วยให้การคาดคะเนค่า Y ได้ถูกต้องเพิ่มมากขึ้น การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_j จึงเป็นการทดสอบสมมติฐาน

$$H_0^C : \beta_j = 0 \text{ สำหรับ ทุกๆ } j \in C$$

$$H_A^C : \beta_j \neq 0 \text{ สำหรับ } j \text{ อย่างน้อย 1 ตัวที่ } j \in C$$

ถ้ามีตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม (Y) p ตัว คือ $X_1, X_2, \ldots, X_p$ สมการถดถอยเชิงซ้อน คือ

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e \quad (2.2)$$

โดยทั่วไปค่าพารามิเตอร์ $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ เราจะไม่สามารถทราบค่าได้ดังนั้นจึงทำการสุ่มตัวอย่างมาจากประชากรแล้วจึงนำมาหาสมการถดถอยเชิงเส้นของกลุ่มตัวอย่าง

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ldots + \hat{\beta}_p x_p \quad (2.3)$$

โดยที่ $\hat{\beta}_0$ คือ ระยะตัดแกน Y ซึ่ง $\beta_1, \beta_2, \ldots, \beta_p$ หมายถึงเมื่อกำหนดให้ $X_1 = X_2 = \cdots = X_p = 0$

$\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ เป็นค่าประมาณของสัมประสิทธิ์การตัดสินใจเชิงส่วนซึ่งมีหน่วยเหมือนกันและมีความหมาย ดังนี้

$\hat{\beta}_1$ เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่าง Y และ X_1 หมายถึง ถ้า X_1 เพิ่มขึ้น 1 หน่วย จะทำให้ Y เปลี่ยนไป $\hat{\beta}_1$ หน่วย (ขึ้นอยู่กับเครื่องหมายของ $\hat{\beta}_1$) โดยที่กำหนดให้ตัวแปรอิสระอื่นๆ มีค่าคงที่

$\hat{\beta}_2$ เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่าง Y และ X_2 หมายถึง ถ้า X_2 เพิ่มขึ้น 1 หน่วย จะทำให้ Y เปลี่ยนไป $\hat{\beta}_2$ หน่วย (ขึ้นอยู่กับเครื่องหมายของ $\hat{\beta}_2$) โดยที่กำหนดให้ตัวแปรอิสระอื่นๆ มีค่าคงที่

.
.
.

$\hat{\beta}_p$ เป็นค่าที่แสดงถึงความสัมพันธ์ระหว่าง Y และ X_p หมายถึง ถ้า X_p เพิ่มขึ้น 1 หน่วย จะทำให้ Y เปลี่ยนไป $\hat{\beta}_p$ หน่วย (ขึ้นอยู่กับเครื่องหมายของ $\hat{\beta}_p$) โดยที่กำหนดให้ตัวแปรอิสระอื่นๆ มีค่าคงที่

เราจะใช้ค่าสถิติ $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ที่ได้จากการสุ่มตัวอย่างไปทดสอบว่าพารามิเตอร์ $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ของประชากรว่ามีค่าเท่ากับ 0 หรือไม่

การประมาณค่า $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ด้วยค่า $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ เพื่อทำให้ผลบวกของค่าคลาดเคลื่อนกำลังสองมีค่าน้อยสุด โดยใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Square) นั่นคือ การหาค่า $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ ที่ทำให้ $\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$ มีค่าต่ำที่สุด

ดังนั้น ค่าความคลาดเคลื่อนในการประมาณค่า Y_j ด้วย $\hat{Y}_j$ คือ $e_j = Y_j - \hat{Y}_j$

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (multiple linear regression analysis) เป็นเทคนิคการวิเคราะห์การถดถอยที่เกี่ยวข้องกับตัวแปรอิสระที่มากกว่าหนึ่งตัวแปร ซึ่งการเพิ่มตัวแปรอิสระที่เกี่ยวข้องเข้ามาในการวิเคราะห์จะทำให้ความถูกต้องของการวิเคราะห์เพิ่มมากขึ้นและค่าคลาดเคลื่อนมาตรฐานของตัวประมาณค่า (standard error of estimates) ลดลง

จากสมการความถดถอยเชิงเส้นพหุคูณ

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_i \quad (2.4)$$

สมการถดถอยเชิงเส้นพหุคูณมีข้อตกลงที่เกี่ยวข้อง ดังนี้

1. ความคลาดเคลื่อน ε_i มีการแจกแจงแบบปกติ โดยมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่เท่ากับ σ^2 หรือ $\varepsilon_i \sim N(0, \sigma^2)$
2. ε_i แต่ละค่าเป็นอิสระต่อกัน

จากข้อตกลงดังกล่าวทำให้ค่าของตัวแปรตามมีการแจกแจงแบบปกติ หรือสามารถเขียนได้ในรูปของ $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}, \sigma^2)$

หากมีตัวแปรอิสระ p - 1 ตัว แล้วตัวแบบการถดถอยเชิงเส้นพหุคูณในรูปทั่วไปสามารถแสดงได้ด้วยสมการนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_{p-1} X_{i\,p-1} + \varepsilon_i \quad (2.5)$$

โดย $\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_{p-1}$ = ค่าสัมประสิทธิ์การถดถอย

$X_{i1}, X_{i2}, X_{i3}, \ldots, X_{i\,p-1}$ = ค่าของตัวแปรอิสระ p – 1ตัวในการเก็บข้อมูลครั้งที่ i

ฟังก์ชันระหว่างตัวแปรอิสระ p - 1 ตัวกับตัวแปรตามจะเป็นแนวระนาบหลายระดับ (hyperplane) ในกรณีของสมการ (2.5) เป็นสมการที่ตัวแปรอิสระไม่มีปฏิสัมพันธ์กัน (interaction) หรือเรียกว่ามีอิทธิพลแบบรวมกัน (additive effect) คือการที่ตัวแปรอิสระตัวหนึ่งมีผลต่อการเปลี่ยนแปลงค่าของตัวแปรตามโดยไม่ได้รับอิทธิพลของตัวแปรอิสระอีกตัวหนึ่ง

หากตัวแปรอิสระทั้งสองมีปฏิสัมพันธ์กันแล้วการเปลี่ยนแปลงค่าของตัวแปรตามที่เกิดจากตัวแปรอิสระตัวหนึ่งจะเปลี่ยนค่าไปมากหรือน้อยขึ้นอยู่กับระดับของตัวแปรอิสระอีกตัวหนึ่ง ซึ่งสามารถแสดงได้ด้วยสมการดังนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \cdots + \varepsilon_i \quad (2.6)$$

โดย β_3 = ค่าสัมประสิทธิ์การถดถอยที่เกิดจากปฏิสัมพันธ์

ค่าแปรปรวนของ Y = ค่าความแปรปรวนที่เกิดจากอิทธิพลของ $X_1, X_2, X_3, \dots, X_p$ + ค่าแปรปรวนอย่างสุ่ม หรือ SST = SSR - SSE

โดยที่ SST (Sum Square of Total) คือ ค่าความแปรปรวนทั้งหมดของ Y ความผันแปรทั้งหมด หรือ $SST = \sum_{i=1}^{n}(Y_i - \bar{y})^2$ เมื่อ i คือครั้งของการเก็บข้อมูล ; $i = 1,2, \dots$,จำนวนครั้งที่เก็บข้อมูล

SSR (Sum Square of Regression) คือ ค่าความแปรปรวนของ Y เนื่องจากอิทธิพลของ $X_1, X_2, X_3, \dots, X_p$

SSE (Sum Square of Error) คือ ค่าความแปรปรวนของ Y เนื่องจากอิทธิพลอื่นๆ หรือเรียกว่า ค่าแปรปรวนอย่างสุ่ม

ตารางที่ 2.1 การวิเคราะห์ความแปรปรวนสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

แหล่งแปรปรวน (SV)	ผลบวกกำลังสอง (Sum of Squares)	องศาอิสระ (df)	ผลบวกกำลังสองเฉลี่ย (Mean Square)	F
ความถดถอย (Regression)	SSR	p	MSR = SSR/p	MSR/ MSE
ความคลาดเคลื่อน (Error)	SSE	n-p-1	MSE = SSE/(n-p-1)	
Total	SST	n-1		

โดยมีสูตรการคำนวณ ดังนี้

$$SST = \sum_{i=1}^{n}(Y_i - \bar{y})^2$$

$$SSR = \beta' \mathbf{X}' Y - n\bar{y}^2$$

$$SSE = \sum_{i=1}^{n}[Y_i - (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)]^2$$

หรือ $SSE = SST - SSR = Y'Y - \beta'\mathbf{X}'Y$

โดยที่ $Y = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x 1 ที่เป็นตัวแปรสุ่มมาตรฐาน

$\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ เป็นเมทริกซ์ตัวแปรอิสระขนาด n x p ที่เป็นตัวแปรสุ่มมาตรฐาน เมื่อ $X_i = \begin{bmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

จากตารางการวิเคราะห์ความแปรปรวนจะใช้ในการทดสอบสมมติฐานเกี่ยวกับความสัมพันธ์ระหว่าง Y และ $X_1, X_2, X_3, \ldots, X_p$ โดยมีการตั้งสมมติฐาน ดังนี้

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$$
$$H_A : \text{มี } \beta_i \text{ อย่างน้อย 1 ค่าที่ } \neq 0 \; ; i = 1,2, \ldots, p$$

สถิติทดสอบ $F = \frac{MSR}{MSE} = \frac{(\beta'\mathbf{X}'Y - n\bar{y}^2)/p}{(Y'Y - \beta'\mathbf{X}'Y)/(n-p-1)}$ (2.7)

เขตปฏิเสธ จะปฏิเสธสมมติฐาน H_0 ถ้า $F > F_{p,\, n-p-1\,;\,1-\alpha}$

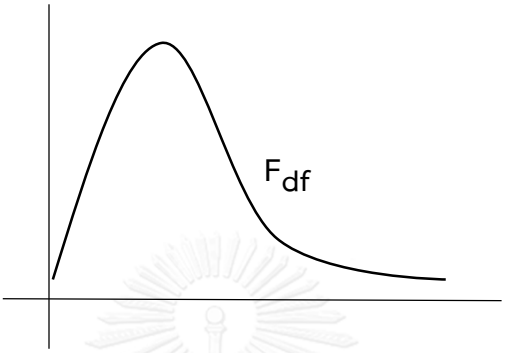
ผลของการทดสอบสมมติฐาน อาจจะเป็น

ก. ไม่ปฏิเสธสมมติฐาน $H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$ ซึ่งสามารถสรุปได้ว่า Y ไม่มีความสัมพันธ์กับ $X_1, X_2, X_3, \ldots, X_p$ ในรูปเชิงเส้น

ข. ปฏิเสธสมมติฐาน H_0 หรือ ไม่ปฏิเสธสมมติฐาน H_A ซึ่งสามารถสรุปได้ว่า มี X_i อย่างน้อย 1 ตัว ที่มีความสัมพันธ์กับ Y ในรูปเชิงเส้น จึงต้องทำการทดสอบต่อไปว่า X_i ตัวใดที่มีความสัมพันธ์กับ Y โดยใช้สถิติทดสอบที

ลักษณะเส้นโค้งของการแจกแจงแบบเอฟ มีการแจกแจงแบบเบ้ทางบวก และค่าของ F จะเป็นค่าบวกเสมอ นั่นคือ $0 < F < \infty$ และลักษณะของเส้นโค้งจะขึ้นกับ degree of freedom

ดังรูป



รูปภาพ แสดงลักษณะเส้นโค้งของการแจกแจงแบบเอฟ

- **ค่า P – Value**

ในการทดสอบสมมติฐานจะสนใจว่าสมมติฐานว่างนั้นจะถูกปฏิเสธหรือไม่ หากสมมติฐานว่างไม่ถูกปฏิเสธนั่นไม่ได้หมายความว่าสมมติฐานว่างนั้นเป็นจริงเสมอ แต่ยังไม่มีหลักฐานที่เพียงพอที่จะปฏิเสธสมมติฐานว่าง ในการตัดสินใจว่าจะปฏิเสธสมมติฐานว่างหรือไม่นั้นเราจำเป็นต้องกำหนดระดับนัยสำคัญ α โดยที่ α = P(ปฏิเสธ H_0 | H_0 เป็นจริง) การทดสอบสมมติฐานนั้นสามารถทำได้สองวิธี คือ การเปรียบเทียบขอบเขตวิกฤต และ การพิจารณาค่า p-value

ค่า p –value คือ ค่าความน่าจะเป็นที่แสดงถึงหลักฐานที่จะนำมาแย้งกับสมมติฐานว่างว่าเหตุการณ์นั้นมีโอกาสเกิดขึ้นได้มากน้อยเพียงไร ภายใต้สมมติฐานว่างนั้นเป็นจริงที่ได้จากการคำนวณชุดของข้อมูลตัวอย่างที่เก็บรวบรวมมา ซึ่งมีค่ามากหรือน้อยขึ้นอยู่กับว่าค่าที่กำหนด (Cut off point) ภายใต้สมมติฐานว่างที่เป็นจริง ในการคำนวณค่า p-value อาศัยทฤษฎี Integration เพื่อที่จะหาพื้นที่ใต้กราฟ (ในทางสถิติ พื้นที่ใต้กราฟคือค่าความน่าจะเป็น) ในการทดสอบสมมติฐานจะปฏิเสธสมมติฐานว่างก็ต่อเมื่อค่า p-value< α

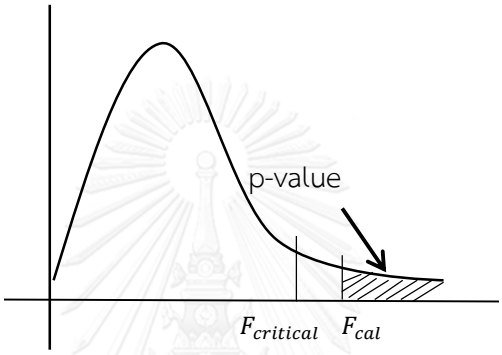
การคำนวณค่า P-Value

กรณีการทดสอบ F-test

หลักการพิจารณาการทดสอบสมมติฐานหางเดียวด้านขวาสามารถคำนวณ p-value ได้ดังนี้

P-value = $\int_{F_{cal}}^{\infty} f(t)\,dt$ โดยขอบเขตของการอินทิเกรต คือ $[F_{cal}, \infty]$

เมื่อ f(t) คือฟังก์ชันการแจกแจงเอฟ (Student's F-distribution)



แสดงการหาพื้นที่ค่า p-value ของการแจกแจงแบบเอฟ

เมื่อ F_{cal} คือ ค่าสถิติ F ที่ได้จากการคำนวณจากสูตรที่ (2.7)

หมายเหตุ ในทางปฏิบัติการคำนวณค่า p-value โดยอาศัยทฤษฎี Integration มีความยุ่งยากและซับซ้อนในการคำนวณ ดังนั้นนักสถิติจึงใช้ตารางสถิติ F เข้ามาช่วยในการทดสอบ หากค่าที่ได้จากการคำนวณ F_{cal} ไม่ปรากฏในตาราง นักสถิติมักใช้วิธีการเทียบบัญญัติไตรยางค์

จากเงื่อนไขในขั้นตอนที่ 2.) ที่ว่า $|\hat{S}^{(b)}| < \frac{n}{2}$ ทำให้ผู้วิเคราะห์ไม่ต้องประสบกับปัญหาข้อมูลที่มีมิติสูงในการทดสอบแบบเอฟอีกต่อไปจากนั้นทำการกำหนดค่า p-value สำหรับกลุ่มสัมประสิทธิ์การถดถอย C ดังนี้

$$p^{C,(b)} = \begin{cases} p_{F-test}^{C \cap \hat{S}^{(b)}} \text{ based on } \mathbf{Y}_{N_{out}^{(b)}}, \mathbf{X}_{N_{out}^{(b)}, \hat{S}^{(b)}} & , \text{เมื่อ } C \cap \hat{S}^{(b)} \neq \emptyset \\ 1 & , \text{เมื่อ } C \cap \hat{S}^{(b)} = \emptyset \end{cases} \quad (2.8)$$

4.) ทำการปรับค่า p-value ที่ได้ในขั้นตอนที่ 3.) ดังนี้

$$p_{adj}^{C,(b)} = \begin{cases} \min(p^{C,(b)}\frac{|\hat{S}^{(b)}|}{|C \cap \hat{S}^{(b)}|}, 1) & , \text{เมื่อ } C \cap \hat{S}^{(b)} \neq \emptyset \\ 1 & , \text{เมื่อ } C \cap \hat{S}^{(b)} = \emptyset \end{cases} \quad (2.9)$$

การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดสามารถทำงานได้ภายใต้เงื่อนไขที่ขนาดของตัวอย่างใหญ่กว่าจำนวนตัวแปรอิสระ (n > p) แต่ในกรณีที่ขนาดของตัวอย่างเล็กกว่าจำนวนตัวแปรอิสระ (n < p) จะไม่สามารถหาค่า $(\mathbf{X}'\mathbf{X})^{-1}$ ได้เนื่องจาก $\mathbf{X}'\mathbf{X}$ ไม่เป็นเมตริกซ์เอกฐาน (Nonsingular Matrix) ดังนั้นในกรณีที่ n < p จะสามารถทำการประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธี Penalized Likelihood

2.1.3 การรวมค่า p-value B ค่าสำหรับกลุ่มสัมประสิทธิ์การถดถอยแต่ละกลุ่ม

จากขั้นตอนที่ 2.1.2 จะได้ค่า p-value ของการทดสอบสมมติฐานของสัมประสิทธิ์การถดถอยในแต่ละกลุ่ม C จำนวน B ค่า ดังนั้นจึงต้องทำการรวม p-value B ค่าให้เป็นค่าเดียวโดยใช้ควอนไทล์เชิงประจักษ์ (Empirical Quantile) ดังนี้

$$Q^C(\gamma) = \min\left\{1, q_\gamma(\{\frac{p_{adj}^{C,(b)}}{\gamma}; b = 1, \dots, B\})\right\} \quad (2.10)$$

โดยที่ $q_\gamma(\cdot)$ คือ ฟังก์ชันควอนไทล์เชิงประจักษ์ เมื่อ $\gamma \in (0,1)$ จากนั้นจึงปรับค่า p-value ของกลุ่มสัมประสิทธิ์ในแต่ละกลุ่มในลำดับชั้น τ โดยนิยามค่า Hierarchically Adjusted Aggregated P-values ดังนี้

$$Q_h^C(\gamma) = \max_{D \epsilon \tau : C \subseteq D} Q^D(\gamma) \quad (2.11)$$

จากการปรับค่าในสมการที่ (2.11) จะได้ว่าค่า $Q_h^C(\gamma)$ สำหรับกลุ่มสัมประสิทธ์ C จะมีค่ามากกว่า $Q_h^{A(C)}(\gamma)$ เสมอ เมื่อ $A(C)$ คือกลุ่มสัมประสิทธิ์ที่เป็นบรรพบุรุษของกลุ่ม C

Mandozzi and Bühlmann (2013) กล่าวว่า การหาค่า γ ที่ดีที่สุดในสมการ (2.10) อาจไม่มีความเหมาะสม กล่าวคือ ค่า γ ที่ดีที่สุดอาจทำให้ไม่สามารถควบคุมความผิดพลาดประเภทที่หนึ่งตามระดับที่เราต้องการได้ ดังนั้น Mandozzi and Bühlmann (2013) จึงเสนอให้ใช้สมการ (2.12) แทนสมการ (2.10) ดังนี้

$$P^C = \min\{1, (1 - \log \gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q^C(\gamma)\} \tag{2.12}$$

โดยที่ $\gamma_{\min} = 0.05$

และใช้สมการ (2.13) แทนสมการ (2.11) ดังนี้

$$P_h^C = \max_{D \epsilon \tau \,:\, C \subseteq D} P^D. \tag{2.13}$$

ซึ่งค่า P_h^C คือค่า p-value ที่เราต้องการในการทดสอบสมมติฐานว่าง H_0^C สำหรับ $c \in \tau$

2.2 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Penalized Regression

การหาค่าประมาณของสัมประสิทธิ์การถดถอย ที่ทำให้ Penalized Likelihood มีค่าสูงสุด นั่นคือ

$$\hat{\beta} = \arg\min_{\beta} \left(-\ell(\beta) + P_\lambda(\beta)\right), \ \lambda \geq 0 \tag{2.14}$$

โดยที่ $-\ell(\beta)$ = - Log Likelihood Function

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

λ คือ Tuning Parameter โดยที่ $\lambda \geq 0$

สมการข้างต้นถือเป็นการคัดกรองตัวแปรเข้าในตัวแบบ หากเราเลือกใช้ Penalty Function ที่เหมาะสมจะสามารถคัดกรองตัวแปรเข้าในตัวแบบได้ กล่าวคือ Penalty Function จะทำให้ค่าสัมประสิทธิ์บางตัวมีค่าเท่ากับศูนย์ ซึ่งในการวิจัยครั้งผู้วิจัยสนใจทำการเปรียบเทียบวิธีการคัดกรองตัวแปร 3 วิธี ดังนี้

2.2.1 Penalty Function ของวิธี Least Absolute Shrinkage and Selection Operator (Lasso)

วิธี Least Absolute Shrinkage and Selection Operator (Lasso) เสนอโดย Tibshirani (1996) ในวิธี Lasso จะใช้ ℓ_1 – norm สำหรับ Penalty Function $\left(P_\lambda(\beta)\right)$ ในการปรับค่าผลรวมความคลาดเคลื่อนกำลังสองให้มีค่าน้อยที่สุด

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{P} |\beta_j| \quad \text{โดยที่ } \lambda > 0 \tag{2.15}$$

วิธี Lasso มีข้อดีคือจะทำการประมาณสัมประสิทธิ์แบบต่อเนื่อง (Continuous Shrinkage) ไปคราวเดียวกันกับการคัดกรองตัวแปรเข้ามายังตัวแบบ และวิธี Lasso จะมีประสิทธิภาพสูงในการพยากรณ์ เมื่อตัวแบบจริงมีจำนวนตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมีจำนวนไม่มากนัก แต่วิธี Lasso ยังมีข้อเสียคือ ตัวประมาณค่าสัมประสิทธิ์ที่ได้มีความเอนเอียง (Bias)

2.2.2 Penalty Function ของวิธี Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso)

วิธี Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso) เสนอโดย Hui Zou (2006) เป็นวิธีที่พัฒนามาจากวิธี Lasso แต่มีการเพิ่มเงื่อนไขโดยการให้ค่าน้ำหนัก (Weight) กับพารามิเตอร์แต่ละตัวแตกต่างกันใน Penalty Function $\left(P_\lambda(\beta)\right)$ ซึ่งการเพิ่มค่าน้ำหนักนี้จะช่วยแก้ไขปัญหาความไม่คงเส้นคงวาที่ทำให้เกิดความเอนเอียงในวิธี Lasso ได้อีกด้วย

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{P} \widehat{W}_j |\beta_j| \tag{2.16}$$

โดยที่ $\widehat{W}_j = \begin{cases} \frac{1}{|\hat{\beta}_{OLS}|} & ; n > p \\ \frac{1}{|\hat{\beta}_{Ridge}|} & ; n < p \end{cases}$

การใช้ Penalty Function ดังกล่าวจะทำให้ได้ตัวประมาณที่มีคุณสมบัติ Oracle ของตัวประมาณค่า กล่าวคือ เมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ วิธี Adaptive Lasso จะมีความสามารถในการเลือกตัวแปรได้เสมือนกับว่าทราบตัวแบบที่แท้จริง (True Model) ซึ่งคุณสมบัตินี้ไม่มีในวิธี Lasso

2.2.3 Penalty Function ของวิธี Elastic Net (EN)

วิธี Elastic Net (EN) เสนอโดย Hui Zou and Trevor (2005) วิธี Elastic Net มีคุณสมบัติช่วยลดค่าสัมประสิทธิ์ของตัวแปรแบบต่อเนื่องและทำการคัดกรองตัวแปรไปพร้อมๆกัน โดยวิธี Elastic Net มีการให้ Penalty Function $\left(\mathrm{P}_{\lambda}(\beta)\right)$ อยู่ในรูปของ ℓ_1 – norm และ ℓ_2 – norm

$$\mathrm{P}_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \quad \text{โดยที่ } \lambda_1 > 0 \text{ และ } \lambda_2 > 0 \tag{2.17}$$

เนื่องจาก วิธี Elastic Net สามารถคัดกรองตัวแปรอิสระและประมาณค่าได้ในคราวเดียวกัน ดังนั้นวิธีนี้จึงเหมาะสมในการวิเคราะห์ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมากๆ และมีตัวแปรอิสระที่มีความสัมพันธ์กันสูง

2.3 อัตราความผิดพลาดรวม (Family Wise Error Rate (FWER))

อัตราความผิดพลาดรวม คือ ความน่าจะเป็นที่จะเกิดความผิดพลาดประเภทที่ 1 (Type I error rate) อย่างน้อย 1 ครั้ง โดยความผิดพลาดประเภทที่ 1 เกิดจากการปฏิเสธสมมติฐานว่าง H_0 เมื่อสมมติฐานว่าง H_0 เป็นจริง ในข้อมูลจำลองแต่ละกรณี ซึ่งเป็นโอกาสของการกระทำความผิดพลาดประเภทที่ 1 อย่างน้อย 1 ครั้งของชุดการเปรียบเทียบจำนวน 1 ชุด ดังนั้นอัตราความผิดพลาดรวมจะหาได้จากสมการต่อไปนี้

$$\mathrm{FWER} = \frac{\text{จำนวนการจำลองที่เกิดความผิดพลาดประเภทที่ 1 อย่างน้อย 1 ครั้ง}}{\text{จำนวนการจำลองข้อมูล}} \tag{2.18}$$

โดยทฤษฎีที่ 1 ใน Mandozzi and Bühlmann (2013) ระบุว่าหากเราทำการปฏิเสธ $\mathrm{H}_0^{\mathrm{C}}$ ก็ต่อเมื่อ $\mathrm{P}_{\mathrm{h}}^{\mathrm{C}} < \alpha$ แล้ว อัตราความผิดพลาดรวมจะเท่ากับ α ด้วย ดังนั้นอัตราความผิดพลาดรวม (FWER) ที่คำนวณได้จึงไม่ควรเกิน α

2.4 ค่าอำนาจการทดสอบ (Power of Test)

อำนาจการทดสอบ หมายถึง ความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่าง $\mathbf{H_0}$ เมื่อสมมติฐานว่าง $\mathbf{H_0}$ เป็นเท็จ ซึ่งในการทดสอบทางสถิติที่มีค่าอำนาจการทดสอบยิ่งมากจะแสดงว่าการทดสอบนั้นยิ่งดี ค่าอำนาจการทดสอบสามารถคำนวณได้ดังนี้

$$\text{Power} = \frac{\text{จำนวนครั้งที่ปฏิเสธ } \mathbf{H_0} \text{ เมื่อ } \mathbf{H_0} \text{ เป็นเท็จ}}{\text{จำนวน } \mathbf{H_0} \text{ ที่เป็นเท็จ}} \tag{2.19}$$

โดยที่ถ้าวิธีใดให้ค่าวัดประสิทธิภาพ FWER ต่ำที่สุด และให้ค่าวัดประสิทธิภาพ Power สูงที่สุด จะถือว่าวิธีนั้นมีความเหมาะสมที่จะใช้ในการคัดกรองตัวแปรอิสระและมีความเหมาะสมที่จะใช้ในการประมาณค่ากลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง

บทที่ 3

วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระและทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) ในการจัดกลุ่มตัวแปร ระหว่างวิธี Lasso วิธี Adaptive Lasso และวิธี Elastic Net โดยมีการจำลองข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) และใช้ข้อมูลจริงที่ได้จาก (Van't Veer et al., 2002) ซึ่งเป็นข้อมูลไมโครแอเรย์ของดีเอ็นเอเนื้องอกในเต้านม หลังจากนั้นทำการจัดกลุ่มตัวแปรอิสระตามความสัมพันธ์แล้วจึงทำการประมาณค่า p-value ให้กับกลุ่มของกลุ่มสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่ได้รับการคัดเลือกเข้ามายังตัวแบบจากวิธีข้างต้น โดยใช้เกณฑ์อัตราความผิดพลาดรวม (Family Wise Error Rate (FWER)) และพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าอำนาจการทดสอบ (Power of Test) จากการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอย โดยทำการวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชัน 3.1.3 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการวิจัย

ในการศึกษาครั้งนี้จะทำการศึกษาในส่วนของข้อมูลจำลองและข้อมูลจริง ภายใต้ขอบเขตการวิจัยดังนี้

3.1.1 ข้อมูลจำลอง

ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระที่ต่างกัน (n:p) 2 ระดับ คือ 100:500 และ 100:1000 และระดับความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ

ระดับที่ 1 : ρ = 0.0

ระดับที่ 2 : ρ = 0.5

ระดับที่ 3 : ρ = 0.9

สำหรับเมทริกซ์ $\mathbf{X}$ จะใช้การจำลอง $\mathrm{X_i} \sim \mathrm{N}(0, \Sigma)$

โดยที่ $\Sigma = \begin{bmatrix} \begin{pmatrix} 1 & \cdots & \rho_{1,100} \\ \vdots & \ddots & \vdots \\ \rho_{100,1} & \cdots & 1 \end{pmatrix} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \begin{pmatrix} 1 & \cdots & \rho_{p-100,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,p-100} & \cdots & 1 \end{pmatrix} \end{bmatrix}$

นอกจากนี้ในส่วนของค่าแปรปรวนของค่าความคลาดเคลื่อน σ^2 จะถูกเลือกเพื่อให้ได้อัตราส่วนสัญญาณต่อสัญญาณรบกวน(SNR) เท่ากับ 12 เมื่อ p = 500 และ 24 เมื่อ p = 1,000

3.1.2 ข้อมูลจริง

3.1.2.1 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันปกติ

ข้อมูลที่ได้มาจาก (Van't Veer et al., 2002) เป็นข้อมูลไมโครแอเรย์ของดีเอ็นเอเนื้องอกในเต้านมของผู้ป่วย 337 คน โดยตัวแปรอิสระแต่ละตัว คือ ระดับการแสดงออกของยีน (Gene Expression) แต่ละตัวซึ่งมีทั้งหมดมากกว่า 20,000 ยีนในข้อมูลชุดนี้ โดยที่ข้อมูลยีนที่เก็บจะเรียงลำดับตามตำแหน่งทางพันธุกรรม (เช่น จากโครโมโซมที่ 1 ถึง 22) จากนั้นทำการสุ่มตัวแปรอิสระขึ้นมาตามจำนวนที่ต้องการและทำการสุ่มคนไข้มาเพียง 100 คน

โดยเปรียบเทียบอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ที่ 100:500 , 100:1000

3.1.2.2 ชุดข้อมูลเกี่ยวกับผู้ป่วยโรคมะเร็งเต้านมที่มีความสัมพันธ์กันสูง

ซึ่งจะใช้ข้อมูลจากชุดแรกแต่จะแตกต่างกันในกระบวนการสุ่มที่จะทำการสุ่มเป็นชุดๆ คือ จะสุ่มตัวแปรอิสระทีละ 100 ตัวแปรที่อยู่ติดกันจากตัวแปรทั้งหมด เนื่องจากข้อมูลยีนมีการจัดเรียงลำดับตามตำแหน่งทางพันธุกรรม ดังนั้นตัวแปรที่อยู่ติดกันจึงมีความสัมพันธ์กันสูงกว่าตัวแปรที่อยู่ห่างกัน และทำการสุ่มคนไข้มาเพียง 100 คน

โดยเปรียบเทียบอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ที่ 100:500 , 100:1000

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง

2. กำหนดการจำลองข้อมูล

2.1 กำหนดค่าเริ่มต้นและรูปแบบของตัวแบบที่ใช้จำลองข้อมูลโดยจะทำการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และพารามิเตอร์จำนวน p ตัว โดยใช้อัตราส่วน n:p ดังนี้

- 100:500
- 100:1000

2.2 จำลองข้อมูลที่มีการแจกแจงแบบนอร์มอล (Normal Distribution)

$$\boxed{Y = \mathbf{X}\beta + \varepsilon} \quad \text{โดยที่} \quad X_i \sim N(0, \Sigma) \ , \ \varepsilon \sim N(0, \sigma^2 I_n)$$

สำหรับเมทริกซ์ $\mathbf{X}$ จะใช้การจำลอง $X_i \sim N(0, \Sigma)$

$$\text{โดยที่} \quad \Sigma = \begin{bmatrix} \begin{pmatrix} 1 & \cdots & \rho_{1,100} \\ \vdots & \ddots & \vdots \\ \rho_{100,1} & \cdots & 1 \end{pmatrix} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \begin{pmatrix} 1 & \cdots & \rho_{p-100,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,p-100} & \cdots & 1 \end{pmatrix} \end{bmatrix}$$

เมื่อ $Y = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x 1 ที่เป็นตัวแปรสุ่มมาตรฐาน

$\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ เป็นเมทริกซ์ตัวแปรอิสระขนาด n x p ที่เป็นตัวแปรสุ่มมาตรฐาน โดยที่ $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1 และ $\varepsilon \sim N(0, \sigma^2 I_n)$

โดยค่าแปรปรวนของค่าความคลาดเคลื่อน σ^2 จะถูกเลือกเพื่อให้ได้อัตราส่วนสัญญาณต่อสัญญาณรบกวน(SNR) เท่ากับ 12 เมื่อ p = 500 และ 24 เมื่อ p = 1,000

n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอิสระ

และเวกเตอร์ค่าสัมประสิทธิ์การถดถอย β จะได้จากการสุ่ม มาเป็นจำนวน 20 ค่าที่ไม่เท่ากับ 0 ในตำแหน่งของ β ดังนี้

$$\begin{aligned}\{\ & \beta_1 = \beta_2 \ldots = \beta_8 = 1, \\ & \beta_{101} = \beta_{102} = \cdots = \beta_{104} = 1, \\ & \beta_{201} = \beta_{202} = \cdots = \beta_{204} = 1, \\ & \beta_{301} = \beta_{302} = 1, \\ & \beta_{401} = \beta_{402} = 1\ \}\end{aligned}$$

2.3 ทำการจำลองข้อมูลขึ้นมาตามอัตราส่วน (n:p) ตามที่กำหนดไว้ข้างต้น และทำการสุ่มข้อมูลจากข้อมูลจริงที่ได้เลือกมาศึกษาดังกล่าวในอัตราส่วนที่เท่ากับข้อมูลจำลอง

3. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคัดกรองตัวแปรอิสระสำหรับวิธีการจัดกลุ่มแบบเป็นลำดับชั้นของตัวแปรอิสระ และคำนวณค่า p-value

3.1 วิธี Lasso

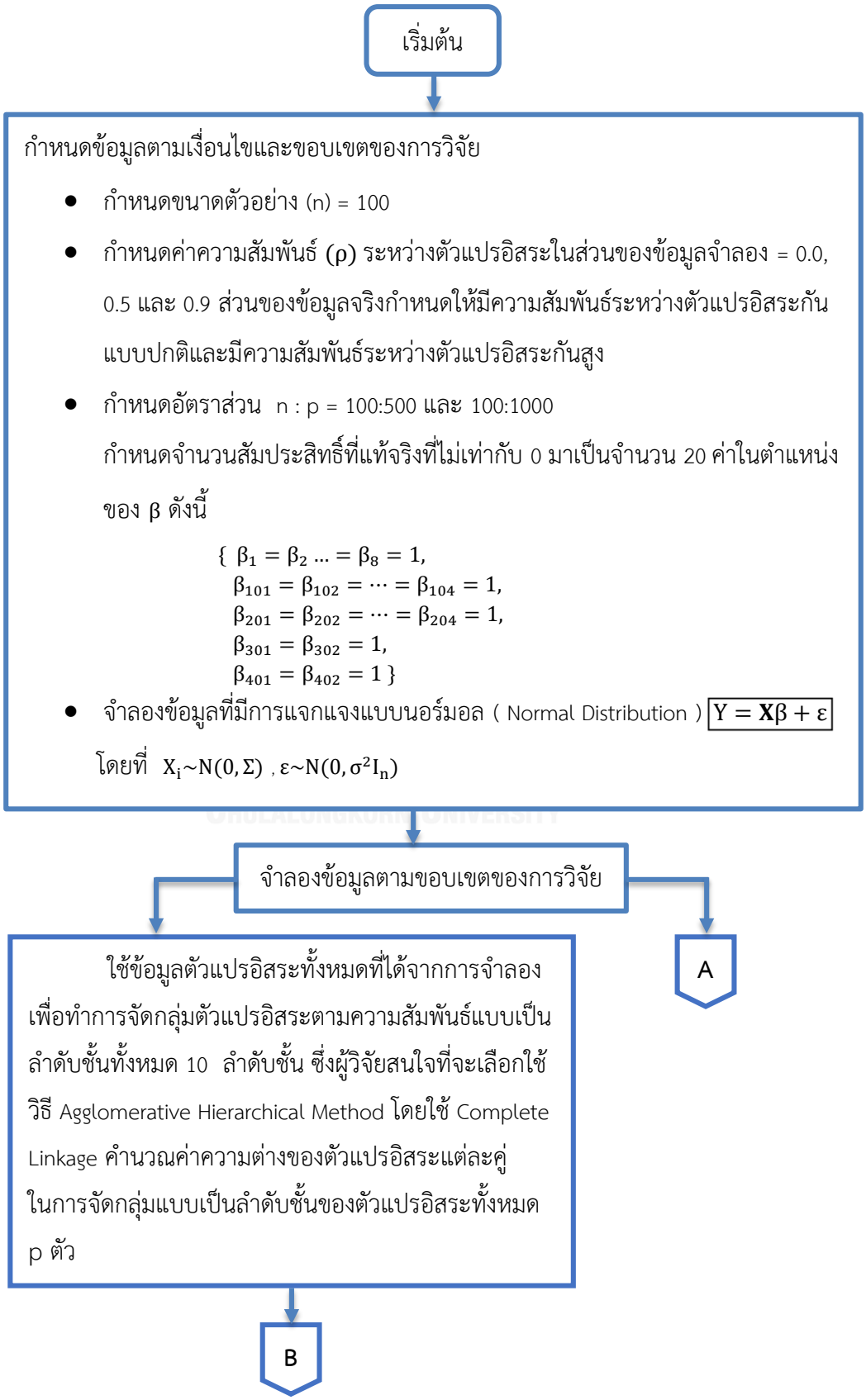
3.2 วิธี Adaptive Lasso

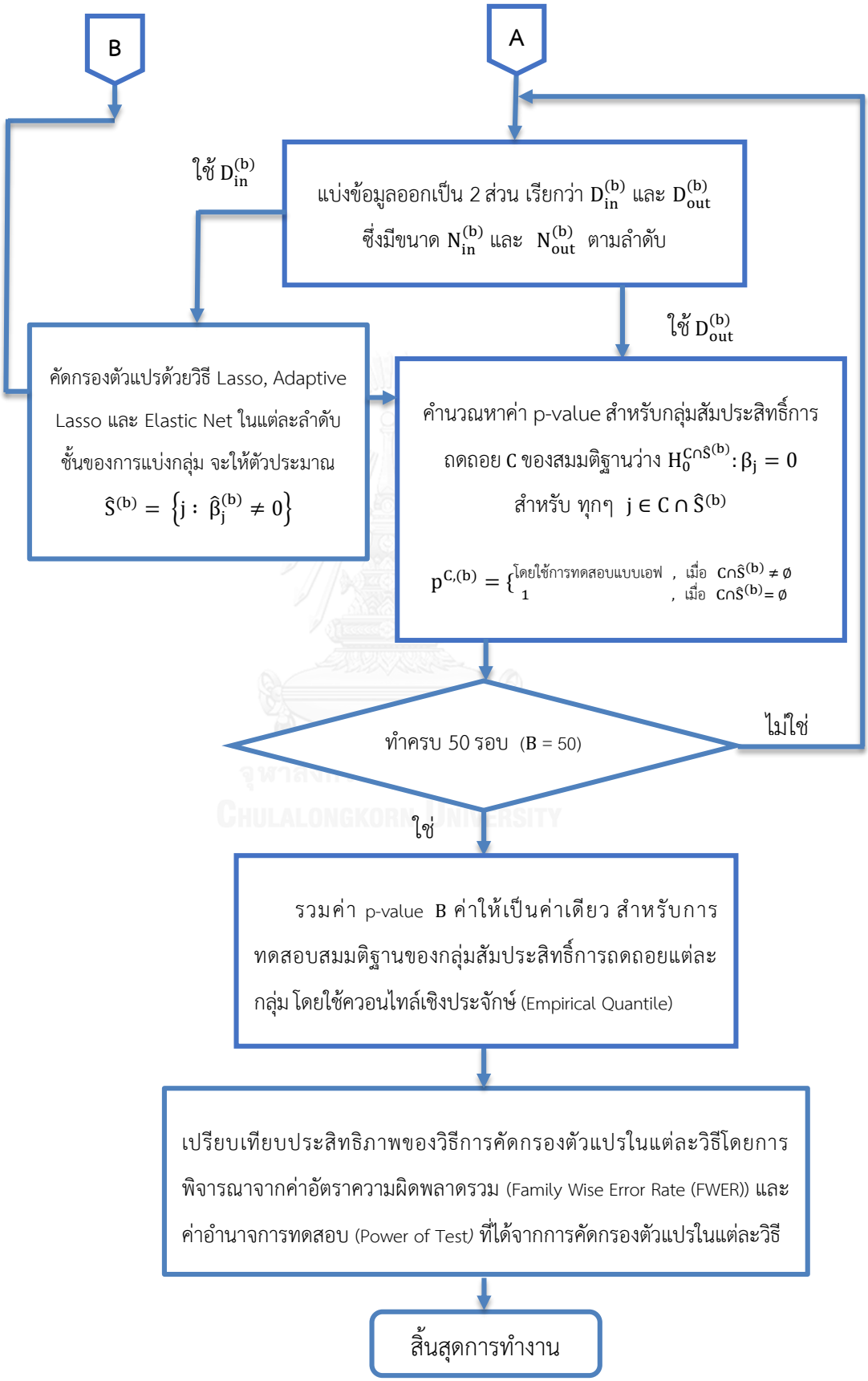
3.3 วิธี Elastic Net

4. นำข้อมูลที่ได้จากข้อ 3. มาหาค่าอัตราความผิดพลาดรวม(Family Wise Error Rate (FWER)) และค่าอำนาจการทดสอบ (Power of Test) ของค่า p-value ที่ได้มาจากการจำลองข้อมูลขึ้นมาทั้งหมด 50 ชุดข้อมูล

5. วิเคราะห์ผลลัพธ์โดยทำการเปรียบเทียบค่าอัตราความผิดพลาดรวม (Family Wise Error Rate (FWER)) และ ค่าอำนาจการทดสอบ (Power of Test) ที่ได้จากการคัดกรองตัวแปรในแต่ละวิธี โดยจำแนกตามประเภทของข้อมูล, อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) และความสัมพันธ์ (Correlation) ของตัวแปรอิสระในแต่ละวิธีการคัดกรองตัวแปร

3.3 ขั้นตอนการทำงานของโปรแกรม




A
B
แบ่งข้อมูลออกเป็น 2 ส่วน เรียกว่า $D_{in}^{(b)}$ และ $D_{out}^{(b)}$ ซึ่งมีขนาด $N_{in}^{(b)}$ และ $N_{out}^{(b)}$ ตามลำดับ
ใช้ $D_{in}^{(b)}$
ใช้ $D_{out}^{(b)}$
คัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ Elastic Net ในแต่ละลำดับชั้นของการแบ่งกลุ่ม จะให้ตัวประมาณ $\hat{S}^{(b)} = \{j : \hat{\beta}_j^{(b)} \neq 0\}$
คำนวณหาค่า p-value สำหรับกลุ่มสัมประสิทธิ์การถดถอย C ของสมมติฐานว่าง $H_0^{C \cap \hat{S}^{(b)}} : \beta_j = 0$ สำหรับ ทุกๆ $j \in C \cap \hat{S}^{(b)}$
$p^{C,(b)} = \{$ โดยใช้การทดสอบแบบเอฟ , เมื่อ $C \cap \hat{S}^{(b)} \neq \emptyset$; 1 , เมื่อ $C \cap \hat{S}^{(b)} = \emptyset$
ทำครบ 50 รอบ (B = 50)
ไม่ใช่
ใช่
รวมค่า p-value B ค่าให้เป็นค่าเดียว สำหรับการทดสอบสมมติฐานของกลุ่มสัมประสิทธิ์การถดถอยแต่ละกลุ่ม โดยใช้ควอนไทล์เชิงประจักษ์ (Empirical Quantile)
เปรียบเทียบประสิทธิภาพของวิธีการคัดกรองตัวแปรในแต่ละวิธีโดยการพิจารณาจากค่าอัตราความผิดพลาดรวม (Family Wise Error Rate (FWER)) และค่าอำนาจการทดสอบ (Power of Test) ที่ได้จากการคัดกรองตัวแปรในแต่ละวิธี
สิ้นสุดการทำงาน

บทที่ 4

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการคัดกรองตัวแปรอิสระระหว่างวิธี Lasso, Adaptive Lasso และ Elastic Net สำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น (Hierarchical Clustering) ในการจัดกลุ่มตัวแปรอิสระตามความสัมพันธ์ของตัวแปรอิสระ จากนั้นจึงใช้วิธีการแบ่งข้อมูลแบบสุ่มหลายๆครั้ง (Multi-split) เพื่อหาค่า p-value ของกลุ่มสัมประสิทธิ์การถดถอย โดยการจำลองข้อมูลและใช้ข้อมูลจริงที่มีขอบเขตต่างกัน โดยจะพิจารณาในส่วนของอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรเป็น 100:500 และ 100:1000 และความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 , 0.5 และ 0.9 แต่ในส่วนของข้อมูลจริงจะมีความสัมพันธ์ของตัวแปรแปรอิสระเป็น 2 แบบคือมีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าอัตราความผิดพลาดรวม (Family Wise Error Rate : FWER) และค่าอำนาจการทดสอบ (Power of Test : Power) โดยถ้าวิธีใดให้ค่าอัตราความผิดพลาดรวมต่ำที่สุดหรือมีค่าใกล้ 0 และมีค่าเฉลี่ยของค่าอำนาจการทดสอบมากที่สุด จะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพและมีความเหมาะสมในการคัดเลือกตัวแปรสำหรับข้อมูลที่มีมิติสูง (High-Dimensional Data) มากที่สุด

อักษรย่อและสัญลักษณ์ต่างๆที่ปรากฏในการนำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆแทนความหมายดังนี้

n	แทน ขนาดของตัวอย่าง
p	แทน จำนวนตัวแปรอิสระ
ρ	แทน ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ
$n:p$	แทน ขนาดของตัวอย่างต่อจำนวนตัวแปรอิสระ
Lasso	แทน การคัดเลือกตัวแปรด้วยวิธี Lasso
Adaptive Lasso	แทน การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso

Elastic Net	แทน การคัดเลือกตัวแปรด้วยวิธี Elastic Net
FWER	แทน อัตราความผิดพลาดรวม (Family Wise Error Rate)
Power	แทน อำนาจการทดสอบ (Power of Test)
β	แทน สัมประสิทธิ์การถดถอยของตัวแปรอิสระ
Mean	แทน ค่าเฉลี่ย
S.D.	แทน ค่าเบี่ยงเบนมาตรฐาน

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 2 ส่วน คือ ในส่วนที่ 1 จะเปรียบเทียบอัตราความผิดพลาดรวม (Family Wise Error Rate) ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net และส่วนที่ 2 จะเปรียบเทียบอำนาจการทดสอบ (Power of Test) ระหว่างการคัดกรองตัวแปรจากทั้ง 3 วิธีข้างต้น

โดยผลการวิจัยจะแบ่งออกเป็น 2 ส่วน ดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

เมื่อพิจารณาในกรณี

1. ข้อมูลจำลอง
    - 1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000
    - 1.2 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ ρ = 0.0 ρ = 0.5 และ ρ = 0.9
2. ข้อมูลจริง
    - 2.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000
    - 2.2 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 2 ระดับ คือ มีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง

<u>ส่วนที่ 2</u> ผลเปรียบเทียบอำนาจการทดสอบ (Power of Test) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

เมื่อพิจารณาในกรณี

1. ข้อมูลจำลอง
    - 1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000
    - 1.2 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ ρ = 0.0 ρ = 0.5 และ ρ = 0.9
2. ข้อมูลจริง
    - 2.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000
    - 2.2 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 2 ระดับ คือ มีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง

4.1 ผลการเปรียบเทียบอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, และ Elastic Net และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการคัดกรองตัวแปรแต่ละวิธี ภายใต้ปัจจัยดังต่อไปนี้

1. กรณีของข้อมูลจำลอง

    1.1 อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000

    1.2 ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ ρ = 0.0 , ρ = 0.5 และ ρ = 0.9

2. กรณีของข้อมูลจริง

    2.1 อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000

    2.2 ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 2 ระดับ คือ มีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง

โดยแสดงผลในตารางที่ 4.1.1 - 4.1.2 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ประเภทของข้อมูล	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
FWER	ข้อมูลจำลอง	4.1.1	• อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) • ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ	1. Lasso 2. Adaptive Lasso 3. Elastic Net
	ข้อมูลจริง	4.1.2		

ตารางที่ 4.1.1 แสดงค่าอัตราความผิดพลาดรวม (FWER) ของแต่ละสถานการณ์จากข้อมูลจำลอง 50 ชุด

n : p	อัตราความผิดพลาดรวม								
	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
	$\rho = 0.0$			$\rho = 0.5$			$\rho = 0.9$		
	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net
100:500	**0.00**	**0.00**	**0.00**	0.10	**0.00**	**0.00**	0.06	**0.00**	**0.00**
100:1000	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.1 ซึ่งแสดงผลของอัตราความผิดพลาดรวม (FWER) โดยนับจากข้อมูลจำลอง 50 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ Elastic Net พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์(Correlation) ของตัวแปรอิสระที่ $\rho = 0.0$ การคัดกรองตัวแปรด้วย 3 วิธีข้างต้นหาค่า FWER ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าวิธีคัดกรองตัวแปรทั้ง 3 วิธีเป็นวิธีที่เหมาะสมกับการคัดกรองตัวแปร
- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net หาค่า FWER ได้เท่ากับ 0 ดังนั้นการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net จึงเป็นวิธีที่เหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วย 3 วิธีข้างต้นหาค่า FWER ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าวิธีคัดกรองตัวแปรทั้ง 3 วิธีเป็นวิธีที่เหมาะสมกับการคัดกรองตัวแปร

ตารางที่ 4.1.2 แสดงค่าอัตราความผิดพลาดรวม (FWER) ของแต่ละสถานการณ์โดยสุ่มข้อมูลจากข้อมูลจริงจำนวน 50 ชุด

n : p	อัตราความผิดพลาดรวม					
	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ					
	ปกติ			สูง		
	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net
100:500	0.06	**0.00**	**0.00**	0.04	**0.00**	**0.00**
100:1000	0.04	**0.00**	**0.00**	0.08	**0.00**	**0.00**

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.2 ซึ่งแสดงผลของอัตราความผิดพลาดรวม (FWER) โดยนับจากข้อมูลที่ทำการสุ่มมาจากข้อมูลจริง 50 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ Elastic Net พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติและมีความสัมพันธ์กันสูง การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net หาค่า FWER ได้เท่ากับ 0 ดังนั้นการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net จึงเป็นวิธีที่เหมาะสมที่สุด
- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติ จะเห็นได้ว่าค่า FWER ของการคัดกรองตัวแปรด้วยวิธี Lasso มีค่ามากกว่าค่า FWER เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

นั่นแสดงให้เห็นว่าวิธี Lasso จะมีประสิทธิภาพมากขึ้นเมื่อตัวแปรอิสระมีความสัมพันธ์กันมากขึ้น

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติและมีความสัมพันธ์กันสูง การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net หาค่า FWER ได้เท่ากับ 0 ดังนั้นการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี Elastic Net จึงเป็นวิธีที่เหมาะสมที่สุด
- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระกันสูง จะเห็นได้ว่าค่า FWER ของการคัดกรองตัวแปรด้วยวิธี Lasso มีค่ามากกว่าค่า FWER เมื่อตัวแปรอิสระมีความสัมพันธ์แบบปกติ นั่นแสดงให้เห็นว่าในกรณีที่มีจำนวนตัวแปรอิสระมากวิธี Lasso จะมีประสิทธิภาพมากขึ้นเมื่อตัวแปรอิสระมีความสัมพันธ์กันน้อยลง

และจากตารางที่ 4.1.1 และตารางที่ 4.1.2 จะได้ว่าจำนวนของตัวแปรอิสระที่มีขนาดใหญ่ขึ้นจะหาค่า FWER ให้มีค่าเท่ากับ 0 ได้ดีกว่าจำนวนของตัวแปรอิสระที่มีขนาดเล็ก

4.2 ผลเปรียบเทียบอำนาจการทดสอบ (Power of Test) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, และ Elastic Net และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการคัดกรองตัวแปรแต่ละวิธี ภายใต้ปัจจัยดังต่อไปนี้

1. กรณีของข้อมูลจำลอง

1.1 อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000

1.2 ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ ρ = 0.0 , ρ = 0.5 และ ρ = 0.9

2. กรณีของข้อมูลจริง

2.1 อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:500 และ 100:1000

2.2 ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 2 ระดับ คือ มีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง

โดยแสดงผลในตารางที่ 4.2.1 - 4.2.2 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ประเภทของข้อมูล	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
Power	ข้อมูลจำลอง	4.2.1	• อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) • ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ	1. Lasso 2. Adaptive Lasso 3. Elastic Net
	ข้อมูลจริง	4.2.2		

ตารางที่ 4.2.1 แสดงค่าเฉลี่ย (ค่าส่วนเบี่ยงเบนมาตรฐาน) ของค่าอำนาจการทดสอบ (Power) ที่คำนวณจากข้อมูลจำลอง 50 ชุด

n : p	อำนาจการทดสอบ								
	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
	$\rho = 0.0$			$\rho = 0.5$			$\rho = 0.9$		
	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net
100:500	**0.0349 (0.0734)**	0.0099 (0.0256)	0.0041 (0.0127)	**0.1707 (0.1461)**	0.0536 (0.0246)	0.0000 (0.0000)	**0.1374 (0.1356)**	0.0713 (0.0369)	0.0000 (0.0000)
100:1000	**0.0033 (0.0169)**	0.0015 (0.0059)	0.0009 (0.0063)	**0.2164 (0.1402)**	0.0667 (0.0262)	0.0000 (0.0000)	**0.0904 (0.1336)**	0.0756 (0.0364)	0.0000 (0.0000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.1 ซึ่งแสดงผลของค่าอำนาจการทดสอบ (Power) โดยเป็นค่าเฉลี่ยของข้อมูลจำลองขนาด 50 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ Elastic Net พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.0$, $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Lasso สามารถหาค่า Power ได้มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงมีความเหมาะสมที่สุด
- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso จะมีค่ามากที่สุดเมื่อเทียบกับกรณีที่ตัวแปรอิสระมีความสัมพันธ์ $\rho = 0.0$ และ 0.9 ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และมีความสัมพันธ์ของตัวแปรอิสระ $\rho = 0.5$ จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.0$, $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Lasso สามารถหาค่า Power ได้มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงมีความเหมาะสมที่สุด
- เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso จะมีค่ามากที่สุดเมื่อเทียบกับกรณีที่ตัวแปรอิสระมีความสัมพันธ์ $\rho = 0.0$ และ 0.9 ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และมีความสัมพันธ์ของตัวแปรอิสระ $\rho = 0.5$ จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.2.1 ยังสามารถสรุปได้อีกว่า

- ค่า Power ที่ได้ในทุกๆกรณีมีค่าน้อยมาก ซึ่งอาจไม่เหมาะสมกับการใช้งานจริง

ตารางที่ 4.2.2 แสดงค่าเฉลี่ย (ค่าส่วนเบี่ยงเบนมาตรฐาน) ของค่าอำนาจการทดสอบ (Power) ที่คำนวณจากการสุ่มข้อมูลจากข้อมูลจริงจำนวน 50 ชุด

n : p	อำนาจการทดสอบ					
	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ					
	ปกติ			สูง		
	Lasso	Adaptive Lasso	Elastic Net	Lasso	Adaptive Lasso	Elastic Net
100:500	**0.7829 (0.2352)**	0.1265 (0.0467)	0.1113 (0.0574)	**0.7650 (0.2159)**	0.1195 (0.0387)	0.1188 (0.0615)
100:1000	**0.7215 (0.2563)**	0.1216 (0.0463)	0.1013 (0.0611)	**0.8386 (0.2080)**	0.1432 (0.0435)	0.1174 (0.0668)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.2 ซึ่งแสดงผลของค่าอำนาจการทดสอบ (Power) โดยเป็นค่าเฉลี่ยของข้อมูลจากการสุ่มข้อมูลจริงขนาด 50 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ Elastic Net พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติและมีความสัมพันธ์กันสูง การคัดกรองตัวแปรด้วยวิธี Lasso สามารถหาค่า Power ได้มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงมีความเหมาะสมที่สุด

- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติ การคัดกรองตัวแปรด้วยวิธี Lasso จะมีค่า Power มากกว่ากรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และมีความสัมพันธ์ของตัวแปรอิสระแบบปกติ จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระแบบปกติและมีความสัมพันธ์กัน การคัดกรองตัวแปรด้วยวิธี Lasso สามารถหาค่า Power ได้มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงมีความเหมาะสมที่สุด
- เมื่อตัวแปรอิสระมีความสัมพันธ์ของตัวแปรอิสระกันสูง การคัดกรองตัวแปรด้วยวิธี Lasso จะมีค่า Power มากกว่ากรณีที่ตัวแปรอิสระมีความสัมพันธ์แบบปกติ ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และมีความสัมพันธ์ของตัวแปรอิสระกันสูง จึงเหมาะสมที่สุด

และจากตารางที่ 4.2.1 และตารางที่ 4.2.2 จะได้ว่าประเภทของข้อมูลมีผลต่อค่าของอำนาจการทดสอบ ซึ่งจะเห็นได้ว่าข้อมูลจำลองให้ค่า Power ที่น้อยมาก ซึ่งอาจไม่เหมาะกับการใช้งานจริงเลย แต่ข้อมูลจริงให้ค่า Power ที่มากพอที่จะสามารถนำไปใช้งานจริงได้

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระของวิธี Lasso วิธี Adaptive Lasso และวิธี Elastic Net โดยจะพิจารณาอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปร เป็น 100:500 และ 100:1000 และความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 , 0.5 และ 0.9 แต่ในส่วนของข้อมูลจริงจะมีความสัมพันธ์ของตัวแปรแปรอิสระเป็น 2 แบบคือมีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง โดยมีเกณฑ์ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าอัตราความผิดพลาดรวม (Family Wise Error Rate : FWER) และค่าอำนาจการทดสอบ (Power of Test : Power) โดยสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 แบ่งผลการวิจัยออกเป็น 2 ส่วน โดยพิจารณาตามขนาดของตัวอย่าง ดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบอัตราความผิดพลาดรวม (Family Wise Error Rate) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

ตารางที่ 5.1.1 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุด เมื่อพิจารณาอัตราความผิดพลาดรวม (FWER) ระหว่างวิธี Lasso, Adaptive Lasso, และ Elastic Net จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 100 โดยจำแนกตามอัตราส่วน ขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ,ประเภทของข้อมูล และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

n : p	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ				
	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.9$	ปกติ	สูง
	ข้อมูลจำลอง			ข้อมูลจริง	
พิจารณาจากค่า FWER					
100:500	L+AL+EN	AL+EN	AL+EN	AL+EN	AL+EN
100:1000	L+AL+EN	L+AL+EN	L+AL+EN	AL+EN	AL+EN

หมายเหตุ

L หมายถึง วิธี LASSO

AL หมายถึง วิธี Adaptive Lasso

EN หมายถึง วิธี Elastic Net

จากตารางที่ 5.1.1 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง (n) เท่ากับ 100 พิจารณาจากค่า FWER โดยค่าของ FWER จะได้วิธีที่เหมาะสมในการคัดกรองตัวแปรคล้ายกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ Elastic Net จะเหมาะสมมากที่สุด โดยทุกวิธีจะสามารถทำงานได้ดีในทุกระดับความสัมพันธ์ของตัวแปรอิสระและในทุกประเภทของข้อมูล

ส่วนที่ 2 ผลเปรียบเทียบอำนาจการทดสอบ (Power of Test) จากการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยโดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้น ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso และ Elastic Net

ตารางที่ 5.1.2 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุด เมื่อพิจารณาอำนาจการทดสอบ (Power) ระหว่างวิธี Lasso, Adaptive Lasso, และ Elastic Net จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 100 โดยจำแนกตามอัตราส่วน ขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) ,ประเภทของข้อมูล และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

n : p	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ				
	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.9$	ปกติ	สูง
	ข้อมูลจำลอง			ข้อมูลจริง	
พิจารณาจากค่า Power					
100:500	L	L	L	L	L
100:1000	L	L	L	L	L

หมายเหตุ

L หมายถึง วิธี LASSO

AL หมายถึง วิธี Adaptive Lasso

EN หมายถึง วิธี Elastic Net

จากตารางที่ 5.1.2 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง(n) เท่ากับ 100 พิจารณาจากค่า Power โดยค่าของ Power จะได้วิธีที่เหมาะสมในการคัดกรองตัวแปรคล้ายกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Lasso จะเหมาะสมมากที่สุด โดยในกรณีของข้อมูลจำลองจะให้ค่า Power ที่มีค่าน้อยมากจนไม่เหมาะสมกับการใช้งานจริง แต่ในกรณีของข้อมูลจริงจะให้ค่า Power ที่มีค่ามากพอสมควรแก่การใช้งานจริง

5.1.2 ผลจากความแตกต่างระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p)

จากผลที่ได้จะพบว่าเมื่ออัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) ยิ่งมีค่าห่างกันมาก การคัดกรองตัวแปรของวิธี Lasso, Adaptive Lasso และ Elastic Net จะมีประสิทธิภาพลดลงในทุกวิธีกรณีที่ความสัมพันธ์ของตัวแปรอิสระมีความสัมพันธ์กันน้อย แต่ในกรณีที่ความสัมพันธ์ของตัวแปรอิสระมีความสัมพันธ์กันมากขึ้น ขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) ยิ่งมีค่าห่างกันมาก การคัดกรองตัวแปรของวิธี Lasso, Adaptive Lasso และ Elastic Net จะมีประสิทธิภาพเพิ่มขึ้นในทุกวิธี นั่นแสดงว่าทั้ง 3 วิธีข้างต้นจะมีประสิทธิภาพสูงสุดเมื่อขนาดของ n และ p มีค่าใกล้กันมากที่สุด หรือทั้ง 3 วิธีจะมีประสิทธิภาพลดลงในกรณีที่ข้อมูลมีมิติสูงขึ้น

5.1.3 ผลจากความแตกต่างของความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

จากผลการวิจัยที่ได้พบว่าในกรณีของข้อมูลจำลองเมื่อความสัมพันธ์ของตัวแปรอิสระมีค่าเท่ากับ 0.0 และ 0.5 การหาค่า FWER และ Power จะมีประสิทธิภาพ แต่เมื่อเพิ่มขนาดของความสัมพันธ์ของตัวแปรอิสระเป็น 0.9 จะพบว่าประสิทธิภาพในการหาค่า FWER และ Power จะลดลงเมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น ส่วนในกรณีของข้อมูลจริงเมื่อความสัมพันธ์ของตัวแปรอิสระเป็นแบบมีความสัมพันธ์กันปกติ การหาค่า FWER และ Power จะมีประสิทธิภาพ แต่เมื่อเพิ่มขนาดของความสัมพันธ์ของตัวแปรอิสระเป็นแบบมีความสัมพันธ์กันสูง จะพบว่าประสิทธิภาพในการหาค่า FWER และ Power จะลดลงเมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้น

5.2 สรุปผลโดยรวม

จากผลการวิจัยในการเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระจากวิธี Lasso , วิธี Adaptive Lasso และวิธี Elastic Net สำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูง โดยใช้เทคนิคการจัดกลุ่มแบบเป็นลำดับชั้น ในการจัดกลุ่มตัวแปรตามความสัมพันธ์ของตัวแปรอิสระ ผลปรากฏว่าการคัดกรองตัวแปรด้วยวิธี Lasso มีอำนาจการทดสอบมากที่สุด รองลงมาคือวิธี Adaptive Lasso และ วิธี Elastic Net ตามลำดับ แต่เมื่อพิจารณาถึงอัตราความผิดพลาดรวม พบว่าวิธี Adaptive Lasso และวิธี Elastic Net มีค่าต่ำที่สุด ซึ่งเมื่อพิจารณาถึงทั้งสองค่าพร้อมกัน พบว่าวิธีการคัดกรองตัวแปรวิธีใดมีอำนาจการทดสอบมากก็จะมีอัตราความผิดพลาดรวมสูงเช่นกัน

ดังนั้นจึงไม่สามารถสรุปได้ว่าวิธีการในการคัดกรองตัวแปรอิสระวิธีใดเป็นวิธีที่ดีที่สุด อย่างไรก็ตาม ในการนำไปใช้งานจริง ผู้นำไปใช้ควรจะพิจารณาว่าจะให้ความสำคัญกับอัตราความผิดพลาดรวม หรือ อำนาจการทดสอบ มากกว่ากัน และเลือกวิธีการคัดกรองที่เหมาะสม

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ผู้ที่สนใจอาจจะนำไปศึกษาต่อได้อีกในเรื่องของ

1. วิธีการคัดกรองตัวแปร ในงานวิจัยนี้เลือกมาศึกษาเพียง 3 วิธีเท่านั้น ในความเป็นจริงแล้วยังมีอีกหลายวิธีที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการคัดกรองตัวแปรอื่นๆ มาร่วมพิจารณาเพื่อเปรียบเทียบประสิทธิภาพได้อีก
2. ขอบเขตในการวิจัย ในเรื่องของขนาดตัวอย่าง, อัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) , ประเภทของข้อมูล และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ อาจจะมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้นได้
3. กรณีที่ $\mathbf{Y}$ ไม่มีการแจกแจงแบบ Normal
4. กรณีที่รูปแบบความสัมพันธ์ของ $\mathbf{X}$ เปลี่ยนไป
5. ค่าเฉลี่ยของการรันโปรแกรม 50 รอบอาจไม่ใช่ค่าที่ดีที่สุด ดังนั้นการรายงานผลการวิจัยอาจใช้ค่ามัธยฐานแทนหรือค่าเปอร์เซนไทล์ที่ k เมื่อ k มีค่าเล็กเพื่อแสดงถึงค่าที่ดีหรือมีประสิทธิภาพสูงในแต่ละกรณี
6. การแบ่งกลุ่มแบบเป็นลำดับชั้นทั้งหมด 10 ลำดับชั้นอาจจะไม่เพียงพอต่อการพิจารณาในการตัดสินใจ
7. เกณฑ์ที่ใช้ในการตัดสินใจอาจต้องมีการใช้เกณฑ์ในการตัดสินใจว่าวิธีใดมีประสิทธิภาพในหลายวิธีมากกว่านี้

รายการอ้างอิง

ภาษาไทย

ศุภวัฒน์ อังคะสี. (2556). การเปรียบเทียบวิธีการคัดกรองตัวแปรสำหรับวิธีการแบ่งข้อมูลตัวอย่างหลายครั้งในการหาค่าพี-แวลูสำหรับข้อมูลที่มีมิติสูง. (วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต), จุฬาลงกรณ์มหาวิทยาลัย, คณะพาณิชยศาสตร์และการบัญชี

ภาษาอังกฤษ

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association,* 96(456), 1348-1360. doi: 10.1198/016214501753382273

Mandozzi, J., & Bühlmann, P. (2013). Hierarchical Testing in the High-Dimensional Setting with Correlated Variables. *arXiv preprint arXiv:*1312.5556.

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). p-Values for High-Dimensional Regression. *Journal of the American Statistical Association,* 104(488), 1671-1681. doi: 10.1198/jasa.2009.tm08647

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* 58(1), 267-288.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., . . . Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature,* 415(6871), 530-536.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association,* 101(476), 1418-1429. doi: 10.1198/016214506000000735

Zou, H., & Trevor , H. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 67(2), 301-320. doi: 10.1111/j.1467-9868.2005.00503.x

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่างกรณีสุ่มข้อมูลจากข้อมูลจำลองที่มีขนาดตัวอย่างเท่ากับ 100 และจำนวนตัวแปรอิสระเท่ากับ 500 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0.0 โดยมีการคัดกรองตัวแปรด้วยวิธี

- Lasso
- Adaptive Lasso
- EN

```
library(mvtnorm)
library(lars)
library(Matrix)
library(parcor)
library(elasticnet)
library(ncvreg)
```



```
###define function equantile###
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))
}
########################################
######## Case n=100 ; p=500 rho=0.0 #########
########################################
n<-100
p<-500
rho<-0.0
########################################
########### Mean and Sigma for  X ##########
########################################
mean_X<-matrix(c(numeric(p)),nrow=p,ncol=1)
Sigma_X<-diag(p)

for(i in 1:(p/100)){
        Sigma_X[(i*100-99):(i*100),(i*100-99):(i*100)]<-rho
                }
diag(Sigma_X)<-1
############################################################
```

```
power.Lasso<-matrix(0,50,1)

type1error.Lasso<-matrix(0,50,1)


power.AdapLasso<-matrix(0,50,1)

type1error.AdapLasso<-matrix(0,50,1)


power.EN<-matrix(0,50,1)

type1error.EN<-matrix(0,50,1)


for(dd in 1:50){


########################################

############ Simulation Data X ############

########################################

X<-rmvnorm(n,mean_X,Sigma_X)

############################################################
```

```
#######################################

############ Simulation Beta #############

#######################################

beta_matrix<-matrix(0,p,1)

fixBeta<-c(1:8,101:104,201:204,301:302,401:402)

beta_matrix[fixBeta,]<-1

Beta<-as.matrix(beta_matrix)

#############################################################



#######################################

############ Simulation Error #############

#######################################

if(p==500){ SNR<-12

        sigmasquar <- (t(Beta[fixBeta, ])%*%t(X[ ,fixBeta])%*%X[
,fixBeta]%*%Beta[fixBeta, ])/((SNR^2)*n) }



if(p==1000){ SNR<-24

        sigmasquar <- (t(Beta[fixBeta, ])%*%t(X[ ,fixBeta])%*%X[
,fixBeta]%*%Beta[fixBeta, ])/((SNR^2)*n) }
```

```
sigmasq<-as.numeric(sigmasquar)

error<-matrix(rnorm(n, mean=0, sd=sqrt(sigmasq)), ncol=1)



#######################################

########  Calculate Y = X(Beta) + error #######

#######################################

Y <- X%*%Beta + error

############################################################

B<-50

pval.lasso.1group <- matrix(1,B,1)

pval.lasso.2group <- matrix(1,B,2)

pval.lasso.3group <- matrix(1,B,3)

pval.lasso.4group <- matrix(1,B,4)

pval.lasso.5group <- matrix(1,B,5)

pval.lasso.6group <- matrix(1,B,6)

pval.lasso.7group <- matrix(1,B,7)

pval.lasso.8group <- matrix(1,B,8)

pval.lasso.9group <- matrix(1,B,9)

pval.lasso.10group <- matrix(1,B,10)
```

```
pval.AdapLasso.1group <- matrix(1,B,1)

pval.AdapLasso.2group <- matrix(1,B,2)

pval.AdapLasso.3group <- matrix(1,B,3)

pval.AdapLasso.4group <- matrix(1,B,4)

pval.AdapLasso.5group <- matrix(1,B,5)

pval.AdapLasso.6group <- matrix(1,B,6)

pval.AdapLasso.7group <- matrix(1,B,7)

pval.AdapLasso.8group <- matrix(1,B,8)

pval.AdapLasso.9group <- matrix(1,B,9)

pval.AdapLasso.10group <- matrix(1,B,10)



pval.EN.1group <- matrix(1,B,1)

pval.EN.2group <- matrix(1,B,2)

pval.EN.3group <- matrix(1,B,3)

pval.EN.4group <- matrix(1,B,4)

pval.EN.5group <- matrix(1,B,5)

pval.EN.6group <- matrix(1,B,6)

pval.EN.7group <- matrix(1,B,7)

pval.EN.8group <- matrix(1,B,8)
```

```
pval.EN.9group <- matrix(1,B,9)

pval.EN.10group <- matrix(1,B,10)



for(b in 1:50){

in.index <- sample(1:n, round(n/2))

 Xin <- X[in.index,]

 Xout <- X[-in.index,]

 Yin <- as.matrix(Y[in.index,])

 Yout <- as.matrix(Y[-in.index,])



##########################################

############  Hierarchical Clustering ############

##########################################

fit <- hclust(dist(t(X)),method = "complete" )

#####plot(fit)#####

betaJ <- matrix(NA,10,10)

aa <- vector("list", 55)



for(k in 1:10){
```

```
        groups <- cutree(fit,k = k)


for(j in 1:k)
{
        indexclust<-which(groups==j)
        BB <- Beta[which(groups==j,)]
        betaJ[k,j] <- nnzero(BB, na.counted = NA)


        index.betaJ <- (which(groups==j))
        aa[[((k*(k+1))/2)-k+j)]] <- index.betaJ


#####################################
################ Lasso ################
#####################################
Lassomodel    <-
lars(Xin,Yin,type="lasso",use.Gram=FALSE,normalize=TRUE,intercept=TRUE)
cvlars   <- cv.lars(Xin,as.numeric(Yin),K=10,type='lasso',plot.it=FALSE)
sAtbest<- cvlars$index[which.min(cvlars$cv)]
tmp    <-   predict.lars(Lassomodel, type="coefficients", s=sAtbest, mode="fraction")
```

```
rtmp <- as.matrix(tmp$coefficients)

beta_lasso       <- matrix(rep(0,p),ncol=1)

beta_lasso[1:p,] <- rtmp

count.lasso.nonzero <- nnzero(beta_lasso, na.counted = NA)



if(count.lasso.nonzero>=49){

          border <- order((abs(beta_lasso[,1])), decreasing = TRUE)

          beta_lasso[border[-(1:48)],1] <- 0

beta_lasso[1:p,] <- beta_lasso

count.lasso.nonzero <- nnzero(beta_lasso, na.counted = NA)

                                    }

########################################

########### Multi  Sample Split ############

########################################

indexXlout<-which(beta_lasso[,1]!=0)

Xlout <- intersect(indexXlout,indexclust)

Xlasso_out <-matrix(Xout[,Xlout],nrow=n/2,)



if(length(Xlout)!=0){     lasso_model <- lm(Yout ~  Xlasso_out)
```

```
        sum_lasso <- summary(lasso_model)

        pVal.lasso <- as.matrix( 1-pf(sum_lasso$f[1],sum_lasso$f[2],sum_lasso$f[3]))


        if (k==1){ pval.lasso.1group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==2){ pval.lasso.2group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==3){ pval.lasso.3group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==4){ pval.lasso.4group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==5){ pval.lasso.5group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==6){ pval.lasso.6group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==7){ pval.lasso.7group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==8){ pval.lasso.8group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==9){ pval.lasso.9group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}
```

```
        if (k==10){ pval.lasso.10group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

                        }


#######################################

############ Adaptive lasso ##############

#######################################

AdapLassomodel      <- adalasso(Xin, Yin,k=10, use.Gram=FALSE)

beta_AdapLasso      <- matrix(rep(0,p), ncol=1)

beta_AdapLasso[1:p,] <- AdapLassomodel$coefficients.adalasso

count.AdapLasso.nonzero <- nnzero(beta_AdapLasso, na.counted = NA)


if(count.AdapLasso.nonzero>=49){

        border <- order((abs(beta_AdapLasso[,1])), decreasing = TRUE)

        beta_AdapLasso[border[-(1:48)],1] <- 0

beta_AdapLasso[1:p,] <- beta_AdapLasso

count.AdapLasso.nonzero <- nnzero(beta_AdapLasso, na.counted = NA)

                        }
```

```
#######################################

########### Multi  Sample Split ###########

#######################################

indexXadaplout<-which(beta_AdapLasso[,1]!=0)

Xadaplout <- intersect(indexXadaplout,indexclust)

Xadaplasso_out <- matrix(Xout[,Xadaplout],nrow=(n/2),)


if(length(Xadaplout)!=0){       AdapLasso_model <- lm(Yout ~ Xadaplasso_out)

        sum_AdapLasso <- summary(AdapLasso_model)

        pVal.aLasso <- as.matrix( 1-
pf(sum_AdapLasso$f[1],sum_AdapLasso$f[2],sum_AdapLasso$f[3]))


        if (k==1){ pval.AdapLasso.1group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==2){ pval.AdapLasso.2group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==3){ pval.AdapLasso.3group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==4){ pval.AdapLasso.4group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}
```

```
        if (k==5){ pval.AdapLasso.5group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==6){ pval.AdapLasso.6group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==7){ pval.AdapLasso.7group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==8){ pval.AdapLasso.8group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==9){ pval.AdapLasso.9group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==10){ pval.AdapLasso.10group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}
```



```
#######################################

############## Elastic Net ###############

#######################################

cvEN    <- cv.glmnet(x=Xin, y=Yin, family="gaussian", nfolds=10)

fitElasticNetmodel      <-  glmnet(x=Xin, y=Yin, family="gaussian", alpha=0.5,
lambda=cvEN$lambda.min)

beta_EN <- matrix(fitElasticNetmodel$beta)

count.EN.nonzero <- nnzero(beta_EN, na.counted = NA)
```

```
if(count.EN.nonzero>=49){
        border <- order((abs(beta_EN[,1])), decreasing = TRUE)
        beta_EN[border[-(1:48)],1] <- 0
beta_EN[1:p,] <- beta_EN
count.EN.nonzero <- nnzero(beta_EN, na.counted = NA)
                        }
#######################################
########### Multi  Sample Split ###########
#######################################
indexXENout<-which(beta_EN[,1]!=0)
XENout <- intersect(indexXENout,indexclust)
XEN_out <- matrix(Xout[,XENout],nrow=(n/2),)

if(length(XENout)!=0){  EN_model <- lm(Yout ~ XEN_out)
        sum_EN <- summary(EN_model)
        pVal.EN <- as.matrix( 1-pf(sum_EN$f[1],sum_EN$f[2],sum_EN$f[3]))

        if (k==1){ pval.EN.1group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}
        if (k==2){ pval.EN.2group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}
```

```
        if (k==3){ pval.EN.3group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==4){ pval.EN.4group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==5){ pval.EN.5group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==6){ pval.EN.6group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==7){ pval.EN.7group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==8){ pval.EN.8group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==9){ pval.EN.9group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==10){ pval.EN.10group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

                    }


}

}

}

write.table(betaJ, paste("./p500rho0.0/p500rho0.0_BetaJ","_",dd,".csv"), sep = ","
,col.names = FALSE ,row.names = FALSE )


pval.lasso.1group[pval.lasso.1group>1] <- 1

pval.lasso.2group[pval.lasso.2group>1] <- 1

pval.lasso.3group[pval.lasso.3group>1] <- 1
```

```
pval.lasso.4group[pval.lasso.4group>1] <- 1

pval.lasso.5group[pval.lasso.5group>1] <- 1

pval.lasso.6group[pval.lasso.6group>1] <- 1

pval.lasso.7group[pval.lasso.7group>1] <- 1

pval.lasso.8group[pval.lasso.8group>1] <- 1

pval.lasso.9group[pval.lasso.9group>1] <- 1

pval.lasso.10group[pval.lasso.10group>1] <- 1



pval.AdapLasso.1group[pval.AdapLasso.1group>1] <- 1

pval.AdapLasso.2group[pval.AdapLasso.2group>1] <- 1

pval.AdapLasso.3group[pval.AdapLasso.3group>1] <- 1

pval.AdapLasso.4group[pval.AdapLasso.4group>1] <- 1

pval.AdapLasso.5group[pval.AdapLasso.5group>1] <- 1

pval.AdapLasso.6group[pval.AdapLasso.6group>1] <- 1

pval.AdapLasso.7group[pval.AdapLasso.7group>1] <- 1

pval.AdapLasso.8group[pval.AdapLasso.8group>1] <- 1

pval.AdapLasso.9group[pval.AdapLasso.9group>1] <- 1

pval.AdapLasso.10group[pval.AdapLasso.10group>1] <- 1
```

```
pval.EN.1group[pval.EN.1group>1] <- 1

pval.EN.2group[pval.EN.2group>1] <- 1

pval.EN.3group[pval.EN.3group>1] <- 1

pval.EN.4group[pval.EN.4group>1] <- 1

pval.EN.5group[pval.EN.5group>1] <- 1

pval.EN.6group[pval.EN.6group>1] <- 1

pval.EN.7group[pval.EN.7group>1] <- 1

pval.EN.8group[pval.EN.8group>1] <- 1

pval.EN.9group[pval.EN.9group>1] <- 1

pval.EN.10group[pval.EN.10group>1] <- 1

########################################

########## aggregate_pvalue_split ##########

########################################

##### call function #####

######################

gammamin <- 0.05

aggregate_pvalue_split_Lasso <- matrix(0,10,10)

aggregate_pvalue_split_AdapLasso <- matrix(0,10,10)

aggregate_pvalue_split_EN <- matrix(0,10,10)
```

```
for(c in 1:10)
{       tmp1<-matrix(0,10,10)
        tmp1<-get(paste("pval.lasso.",c,"group",sep=""))
        tmp2<-matrix(0,10,10)
        tmp2<-get(paste("pval.AdapLasso.",c,"group",sep=""))
        tmp3<-matrix(0,10,10)
        tmp3<-get(paste("pval.EN.",c,"group",sep=""))
                for(d in 1:c){
                aggregate_pvalue_split_Lasso[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp1[,d]), seq(gammamin, 1, 0.05))))
                aggregate_pvalue_split_AdapLasso[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp2[,d]), seq(gammamin, 1, 0.05))))
                aggregate_pvalue_split_EN[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp3[,d]), seq(gammamin, 1, 0.05))))
                }
}

pvalue.Lasso.h <- matrix(0,10,10)
pvalue.AdapLasso.h <- matrix(0,10,10)
pvalue.EN.h <- matrix(0,10,10)
```

```
        pvalue.Lasso.h[1,1] <- aggregate_pvalue_split_Lasso[1,1]

        pvalue.AdapLasso.h[1,1] <- aggregate_pvalue_split_AdapLasso[1,1]

        pvalue.EN.h[1,1] <- aggregate_pvalue_split_EN[1,1]


for(g in 2:10){

        for(f in 1:g){      maxx1 <- 0

                            maxx2 <- 0

                            maxx3 <- 0

        for(e in 1:(g-1)){ if( all(aa[[((g*(g+1))/2)-g+f)]] %in% aa[[(((g-1)*g)/2)-(g-1)+e)]]))

                            { if(aggregate_pvalue_split_Lasso[g,f] < pvalue.Lasso.h[g-1,e])

                                    { maxx1<-pvalue.Lasso.h[g-1,e] }

                                     else{ maxx1<-aggregate_pvalue_split_Lasso[g,f] }

                                        }

                    { if(aggregate_pvalue_split_AdapLasso[g,f] < pvalue.AdapLasso.h[g-1,e])

                                    { maxx2<-pvalue.AdapLasso.h[g-1,e] }

                                     else{ maxx2<-aggregate_pvalue_split_AdapLasso[g,f] }

                                        }

                    { if(aggregate_pvalue_split_EN[g,f] < pvalue.EN.h[g-1,e])

                                    { maxx3<-pvalue.EN.h[g-1,e] }
```

```
                                    else{ maxx3<-aggregate_pvalue_split_EN[g,f] }

                                }

        pvalue.Lasso.h[g,f] <- maxx1

        pvalue.AdapLasso.h[g,f] <- maxx2

        pvalue.EN.h[g,f] <- maxx3

                                        }

        }}


write.table(pvalue.Lasso.h,
paste("./p500rho0.0/p500rho0.0_pvalue.Lasso.h","_",dd,".csv"), sep = "," ,col.names =
FALSE ,row.names = FALSE )

write.table(pvalue.AdapLasso.h,
paste("./p500rho0.0/p500rho0.0_pvalue.AdapLasso.h","_",dd,".csv"), sep = ","
,col.names = FALSE ,row.names = FALSE )

write.table(pvalue.EN.h, paste("./p500rho0.0/p500rho0.0_pvalue.EN.h","_",dd,".csv"),
sep = "," ,col.names = FALSE ,row.names = FALSE )
```

```
#######################################

######### POWER AND TYPE I ERROR ########

#######################################

Pow.L<-0

Pow.A<-0

Pow.E<-0

for(u in 1:10){

        for(v in 1:u){    {if(betaJ[u,v]!=0 && pvalue.Lasso.h[u,v] < 0.05)

                                        Pow.L <- Pow.L + 1 }

                        {if(betaJ[u,v]!=0 && pvalue.AdapLasso.h[u,v] < 0.05)

                                        Pow.A <- Pow.A + 1 }

                         {if(betaJ[u,v]!=0 && pvalue.EN.h[u,v] < 0.05)

                                        Pow.E <- Pow.E + 1 }

                         }

        }

power.Lasso[dd,] <- Pow.L/nnzero(as.numeric(betaJ), na.counted = FALSE)

power.AdapLasso[dd,] <- Pow.A/nnzero(as.numeric(betaJ), na.counted = FALSE)

power.EN[dd,]<- Pow.E/nnzero(as.numeric(betaJ), na.counted = FALSE)
```

```
Type1.L <- 0

Type1.A <- 0

Type1.E <- 0

for(s in 1:10){

        for(t in 1:s){     {if(betaJ[s,t]==0 && pvalue.Lasso.h[s,t] < 0.05)

                                        Type1.L <- Type1.L + 1  }

                           {if(betaJ[s,t]==0 && pvalue.AdapLasso.h[s,t] < 0.05)

                                        Type1.A <- Type1.A + 1  }

                           {if(betaJ[s,t]==0 && pvalue.EN.h[s,t] < 0.05)

                                        Type1.E <- Type1.E + 1  }

                            }

        }

type1error.Lasso[dd,] <- Type1.L

type1error.AdapLasso[dd,] <- Type1.A

type1error.EN[dd,] <- Type1.E



}

power<- data.frame(power.Lasso= power.Lasso, power.AdapLasso= power.AdapLasso,
power.EN= power.EN)
```

```
write.table(power, paste("./p500rho0.0/p500rho0.0_power.test",".csv"), sep = ","
,col.names = TRUE ,row.names = FALSE )

TypeIerror<-data.frame(type1error.Lasso= type1error.Lasso, type1error.AdapLasso=
type1error.AdapLasso, type1error.EN= type1error.EN)

write.table(TypeIerror,paste("./p500rho0.0/p500rho0.0_TypeIerror",".csv"), sep = ","
,col.names = TRUE ,row.names = FALSE )


####################################

###############  FWER  ###############

####################################

FWER.Lasso <- nnzero(as.numeric(type1error.Lasso), na.counted = FALSE)/50

FWER.AdapLasso <- nnzero(as.numeric(type1error.AdapLasso), na.counted =
FALSE)/50

FWER.EN <- nnzero(as.numeric(type1error.EN), na.counted = FALSE)/50


FWER<-data.frame(FWER.Lasso=FWER.Lasso , FWER.AdapLasso=FWER.AdapLasso ,
FWER.EN=FWER.EN )

write.table(FWER, paste("./p500rho0.0/p500rho0.0_FWER.csv"), sep = "," ,col.names =
TRUE ,row.names = FALSE )

#################################################################
```

ตัวอย่างกรณีสุ่มข้อมูลจากข้อมูลจริงที่มีขนาดตัวอย่างเท่ากับ 100 และจำนวนตัวแปรอิสระเท่ากับ 500 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็นความสัมพันธ์กันสูง โดยมีการคัดกรองตัวแปรด้วยวิธี

- Lasso
- Adaptive Lasso
- EN

```
library(mvtnorm)

library(lars)

library(Matrix)

library(parcor)

library(elasticnet)

library(ncvreg)

nki<-read.table("C:/nki_new.csv",sep=",",header=T)

############################################################

####RealData####

nki<-read.table('./nki_new.csv',sep=',',header=T)



###define function equantile###

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))
}
############ Biobase  Data ############
########################################
######## Case n=100 ;p=500 High Rho ########
########################################
n<-100
p<-500
############################################################
power.Lasso <- matrix(0,50,1)
type1error.Lasso <- matrix(0,50,1)
power.AdapLasso <- matrix(0,50,1)
type1error.AdapLasso <- matrix(0,50,1)
power.EN <- matrix(0,50,1)
type1error.EN <- matrix(0,50,1)
```

```
for(dd in 1:50){

########################################

############ Simulation Data X ############

########################################

########## sample index for  X ##########

####################################

######### Highrho #########

gp<-p/100

startindex<-rep(NA,gp)

startindex[1]<-sample(1:(ncol(nki)-100), 1)

i<-2

while (i<=gp)

{

  tmp<-sample(1:(ncol(nki)-100), 1)

  if (sum(abs(startindex[1:(i-1)]-tmp)>100)==(i-1))

  {

    startindex[i]<-tmp

    i<-i+1

  }
```

```
}
for (i in 1:length(startindex))
{
  if (i ==1)
  {
    x<-nki[seq(startindex[i], startindex[i]+99)]
  } else {
    x<-cbind(x, nki[seq(startindex[i], startindex[i]+99)])
  }
}
Nindex<-sample(1:nrow(nki),n)
x<-x[Nindex,]
X<-as.matrix(x)
############################################################
#######################################
############## Simulation Beta #############
#######################################
beta_matrix     <- matrix(0,p,1)
fixBeta <- c(1:8,101:104,201:204,301:302,401:402)
```

```
beta_matrix[fixBeta,]<- 1

Beta<- as.matrix(beta_matrix)

############################################################


########################################

############# Simulation Error ############

########################################

if(p==500){ SNR<-12

        sigmasquar <- (t(Beta[fixBeta, ])%*%(t(X[ ,fixBeta])%*%X[
,fixBeta])%*%Beta[fixBeta, ])/((SNR^2)*n) }


if(p==1000){ SNR<-24

        sigmasquar <- (t(Beta[fixBeta, ])%*%(t(X[ ,fixBeta])%*%X[
,fixBeta])%*%Beta[fixBeta, ])/((SNR^2)*n) }


sigmasq<- as.numeric(sigmasquar)

error<- matrix(rnorm(n, mean=0, sd=sqrt(sigmasq)), ncol=1)

############################################################
```

```
#######################################

######## Calculate Y = X(Beta) + error #######

#######################################

Y         <-  X%*%Beta + error

################################################

B<-50

pval.lasso.1group <- matrix(1,B,1)

pval.lasso.2group <- matrix(1,B,2)

pval.lasso.3group <- matrix(1,B,3)

pval.lasso.4group <- matrix(1,B,4)

pval.lasso.5group <- matrix(1,B,5)

pval.lasso.6group <- matrix(1,B,6)

pval.lasso.7group <- matrix(1,B,7)

pval.lasso.8group <- matrix(1,B,8)

pval.lasso.9group <- matrix(1,B,9)

pval.lasso.10group <- matrix(1,B,10)



pval.AdapLasso.1group <- matrix(1,B,1)

pval.AdapLasso.2group <- matrix(1,B,2)
```

```
pval.AdapLasso.3group <- matrix(1,B,3)

pval.AdapLasso.4group <- matrix(1,B,4)

pval.AdapLasso.5group <- matrix(1,B,5)

pval.AdapLasso.6group <- matrix(1,B,6)

pval.AdapLasso.7group <- matrix(1,B,7)

pval.AdapLasso.8group <- matrix(1,B,8)

pval.AdapLasso.9group <- matrix(1,B,9)

pval.AdapLasso.10group <- matrix(1,B,10)



pval.EN.1group <- matrix(1,B,1)

pval.EN.2group <- matrix(1,B,2)

pval.EN.3group <- matrix(1,B,3)

pval.EN.4group <- matrix(1,B,4)

pval.EN.5group <- matrix(1,B,5)

pval.EN.6group <- matrix(1,B,6)

pval.EN.7group <- matrix(1,B,7)

pval.EN.8group <- matrix(1,B,8)

pval.EN.9group <- matrix(1,B,9)

pval.EN.10group <- matrix(1,B,10)
```

```
for(b in 1:50){

in.index <- sample(1:n, round(n/2))

 Xin <- X[in.index,]

 Xout <- X[-in.index,]

 Yin <- as.matrix(Y[in.index,])

 Yout <- as.matrix(Y[-in.index,])



##########################################

############ Hierarchical Clustering ############

##########################################

fit <- hclust(dist(t(X)),method = "complete" )

##### plot(fit) #####

betaJ <- matrix(NA,10,10)

aa <- vector("list", 55)

for(k in 1:10){

        groups <- cutree(fit,k = k)

for(j in 1:k)

{       indexclust<-which(groups==j)

        BB <- Beta[which(groups==j,)]
```

```
        betaJ[k,j] <- nnzero(BB, na.counted = NA)

        index.betaJ <- (which(groups==j))

        aa[[((k*(k+1))/2)-k+j)]] <- index.betaJ


#####################################

################ Lasso  ###############

#####################################

Lassomodel    <-
lars(Xin,Yin,type="lasso",use.Gram=FALSE,normalize=TRUE,intercept=TRUE)

cvlars   <- cv.lars(Xin,as.numeric(Yin),K=10,type='lasso',plot.it=FALSE)

sAtbest<- cvlars$index[which.min(cvlars$cv)]

tmp              <- predict.lars(Lassomodel, type="coefficients", s=sAtbest,
mode="fraction")

rtmp <- as.matrix(tmp$coefficients)


beta_lasso      <- matrix(rep(0,p),ncol=1)

beta_lasso[1:p,] <- rtmp

count.lasso.nonzero <- nnzero(beta_lasso, na.counted = NA)
```

```
if(count.lasso.nonzero>=49){
        border <- order((abs(beta_lasso[,1])), decreasing = TRUE)
        beta_lasso[border[-(1:48)],1] <- 0
beta_lasso[1:p,] <- beta_lasso
count.lasso.nonzero <- nnzero(beta_lasso, na.counted = NA)
                        }
########################################
########### Multi  Sample Split ###########
########################################
Xlout <- Xout[,which(beta_lasso[,1]!=0)]
indexXlout<-which(beta_lasso[,1]!=0)
Xlout <- intersect(indexXlout,indexclust)
Xlasso_out <-matrix(Xout[,Xlout],nrow=n/2,)
if(length(Xlout)!=0){     lasso_model <- lm(Yout ~  Xlasso_out)
        sum_lasso <- summary(lasso_model)
        pVal.lasso <- as.matrix( 1-pf(sum_lasso$f[1],sum_lasso$f[2],sum_lasso$f[3]))

        if (k==1){ pval.lasso.1group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}
```

```
        if (k==2){ pval.lasso.2group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==3){ pval.lasso.3group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==4){ pval.lasso.4group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==5){ pval.lasso.5group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==6){ pval.lasso.6group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==7){ pval.lasso.7group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==8){ pval.lasso.8group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==9){ pval.lasso.9group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

        if (k==10){ pval.lasso.10group[b,j] <-
pVal.lasso*(count.lasso.nonzero/length(Xlout))}

                }
```

```
#######################################

############# Adaptive lasso #############

#######################################

AdapLassomodel        <- adalasso(Xin, Yin,k=10, use.Gram=FALSE)

beta_AdapLasso        <- matrix(rep(0,p), ncol=1)

beta_AdapLasso[1:p,] <- AdapLassomodel$coefficients.adalasso

count.AdapLasso.nonzero <- nnzero(beta_AdapLasso, na.counted = NA)



if(count.AdapLasso.nonzero>=49){

        border <- order((abs(beta_AdapLasso[,1])), decreasing = TRUE)

        beta_AdapLasso[border[-(1:48)],1] <- 0

beta_AdapLasso[1:p,] <- beta_AdapLasso

count.AdapLasso.nonzero <- nnzero(beta_AdapLasso, na.counted = NA)

                                }

#######################################

########### Multi  Sample Split ###########

#######################################

Xadaplout <- Xout[,which(beta_AdapLasso[,1]!=0)]

indexXadaplout<-which(beta_AdapLasso[,1]!=0)
```

```
Xadaplout <- intersect(indexXadaplout,indexclust)

Xadaplasso_out <- matrix(Xout[,Xadaplout],nrow=(n/2),)

if(length(Xadaplout)!=0){        AdapLasso_model <- lm(Yout ~ Xadaplasso_out)

        sum_AdapLasso <- summary(AdapLasso_model)

        pVal.aLasso <- as.matrix( 1-
pf(sum_AdapLasso$f[1],sum_AdapLasso$f[2],sum_AdapLasso$f[3]))



        if (k==1){ pval.AdapLasso.1group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==2){ pval.AdapLasso.2group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==3){ pval.AdapLasso.3group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==4){ pval.AdapLasso.4group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==5){ pval.AdapLasso.5group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==6){ pval.AdapLasso.6group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==7){ pval.AdapLasso.7group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}
```

```
        if (k==8){ pval.AdapLasso.8group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==9){ pval.AdapLasso.9group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

        if (k==10){ pval.AdapLasso.10group[b,j] <-
pVal.aLasso*(count.AdapLasso.nonzero/length(Xadaplout))}

                }



#######################################

############## Elastic Net ###############

#######################################

cvEN    <- cv.glmnet(x=Xin, y=Yin, family="gaussian", nfolds=10)

fitElasticNetmodel      <-  glmnet(x=Xin, y=Yin, family="gaussian", alpha=0.5,
lambda=cvEN$lambda.min)

beta_EN <- matrix(fitElasticNetmodel$beta)

count.EN.nonzero <- nnzero(beta_EN, na.counted = NA)



if(count.EN.nonzero>=49){

        border <- order((abs(beta_EN[,1])), decreasing = TRUE)

        beta_EN[border[-(1:48)],1] <- 0
```

```
beta_EN[1:p,] <- beta_EN

count.EN.nonzero <- nnzero(beta_EN, na.counted = NA)

                        }

#######################################

############ Multi  Sample Split ###########

#######################################

XENout <- Xout[,which(beta_EN[,1]!=0)]

indexXENout<-which(beta_EN[,1]!=0)

XENout <- intersect(indexXENout,indexclust)

XEN_out <- matrix(Xout[,XENout],nrow=(n/2),)

if(length(XENout)!=0){  EN_model <- lm(Yout ~ XEN_out)

        sum_EN <- summary(EN_model)

        pVal.EN <- as.matrix( 1-pf(sum_EN$f[1],sum_EN$f[2],sum_EN$f[3]))


        if (k==1){ pval.EN.1group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==2){ pval.EN.2group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==3){ pval.EN.3group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==4){ pval.EN.4group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==5){ pval.EN.5group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}
```

```
        if (k==6){ pval.EN.6group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==7){ pval.EN.7group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==8){ pval.EN.8group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==9){ pval.EN.9group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

        if (k==10){ pval.EN.10group[b,j] <- pVal.EN*(count.EN.nonzero/length(XENout))}

                        }

} }

}

write.table(betaJ, paste("./p500nki_high/p500nki_high_BetaJ","_",dd,".csv"), sep = ","
,col.names = FALSE ,row.names = FALSE )



pval.lasso.1group[pval.lasso.1group>1] <- 1

pval.lasso.2group[pval.lasso.2group>1] <- 1

pval.lasso.3group[pval.lasso.3group>1] <- 1

pval.lasso.4group[pval.lasso.4group>1] <- 1

pval.lasso.5group[pval.lasso.5group>1] <- 1

pval.lasso.6group[pval.lasso.6group>1] <- 1

pval.lasso.7group[pval.lasso.7group>1] <- 1

pval.lasso.8group[pval.lasso.8group>1] <- 1
```

```
pval.lasso.9group[pval.lasso.9group>1] <- 1

pval.lasso.10group[pval.lasso.10group>1] <- 1


pval.AdapLasso.1group[pval.AdapLasso.1group>1] <- 1

pval.AdapLasso.2group[pval.AdapLasso.2group>1] <- 1

pval.AdapLasso.3group[pval.AdapLasso.3group>1] <- 1

pval.AdapLasso.4group[pval.AdapLasso.4group>1] <- 1

pval.AdapLasso.5group[pval.AdapLasso.5group>1] <- 1

pval.AdapLasso.6group[pval.AdapLasso.6group>1] <- 1

pval.AdapLasso.7group[pval.AdapLasso.7group>1] <- 1

pval.AdapLasso.8group[pval.AdapLasso.8group>1] <- 1

pval.AdapLasso.9group[pval.AdapLasso.9group>1] <- 1

pval.AdapLasso.10group[pval.AdapLasso.10group>1] <- 1


pval.EN.1group[pval.EN.1group>1] <- 1

pval.EN.2group[pval.EN.2group>1] <- 1

pval.EN.3group[pval.EN.3group>1] <- 1

pval.EN.4group[pval.EN.4group>1] <- 1

pval.EN.5group[pval.EN.5group>1] <- 1
```

```
pval.EN.6group[pval.EN.6group>1] <- 1

pval.EN.7group[pval.EN.7group>1] <- 1

pval.EN.8group[pval.EN.8group>1] <- 1

pval.EN.9group[pval.EN.9group>1] <- 1

pval.EN.10group[pval.EN.10group>1] <- 1


########################################

########## aggregate_pvalue_split ##########

########################################

### call function ###

##################

gammamin <- 0.05

aggregate_pvalue_split_Lasso <- matrix(0,10,10)

aggregate_pvalue_split_AdapLasso <- matrix(0,10,10)

aggregate_pvalue_split_EN <- matrix(0,10,10)

for(c in 1:10)

{       tmp1<-matrix(0,10,10)

        tmp1<-get(paste("pval.lasso.",c,"group",sep=""))

        tmp2<-matrix(0,10,10)
```

```
        tmp2<-get(paste("pval.AdapLasso.",c,"group",sep=""))
        tmp3<-matrix(0,10,10)
        tmp3<-get(paste("pval.EN.",c,"group",sep=""))
                for(d in 1:c){
                aggregate_pvalue_split_Lasso[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp1[,d]), seq(gammamin, 1, 0.05))))
                aggregate_pvalue_split_AdapLasso[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp2[,d]), seq(gammamin, 1, 0.05))))
                aggregate_pvalue_split_EN[c,d] <- min(1,(1-
log(gammamin))*min(equantile(as.vector(tmp3[,d]), seq(gammamin, 1, 0.05))))
                }
}
pvalue.Lasso.h <- matrix(0,10,10)
pvalue.AdapLasso.h <- matrix(0,10,10)
pvalue.EN.h <- matrix(0,10,10)

        pvalue.Lasso.h[1,1] <- aggregate_pvalue_split_Lasso[1,1]
        pvalue.AdapLasso.h[1,1] <- aggregate_pvalue_split_AdapLasso[1,1]
        pvalue.EN.h[1,1] <- aggregate_pvalue_split_EN[1,1]
```

```
for(g in 2:10){

        for(f in 1:g){      maxx1 <- 0

                            maxx2 <- 0

                            maxx3 <- 0

        for(e in 1:(g-1)){ if( all(aa[[((((g*(g+1))/2)-g+f)]] %in% aa[[(((((g-1)*g)/2)-(g-1)+e)]]))

                { if(aggregate_pvalue_split_Lasso[g,f] < pvalue.Lasso.h[g-1,e])

                                    { maxx1<-pvalue.Lasso.h[g-1,e] }

                                    else{ maxx1<-aggregate_pvalue_split_Lasso[g,f] }

                                    }

                { if(aggregate_pvalue_split_AdapLasso[g,f] < pvalue.AdapLasso.h[g-1,e])

                                    { maxx2<-pvalue.AdapLasso.h[g-1,e] }

                                else{ maxx2<-aggregate_pvalue_split_AdapLasso[g,f] }

                                    }

                { if(aggregate_pvalue_split_EN[g,f] < pvalue.EN.h[g-1,e])

                                    { maxx3<-pvalue.EN.h[g-1,e] }

                                    else{ maxx3<-aggregate_pvalue_split_EN[g,f] }

                                    }

        pvalue.Lasso.h[g,f] <- maxx1

        pvalue.AdapLasso.h[g,f] <- maxx2
```

```
        pvalue.EN.h[g,f] <- maxx3

                                                }

        }}


write.table(pvalue.Lasso.h,
paste("./p500nki_high/p500nki_high_pvalue.Lasso.h","_",dd,".csv"), sep = "," ,col.names
= FALSE ,row.names = FALSE )

write.table(pvalue.AdapLasso.h,
paste("./p500nki_high/p500nki_high_pvalue.AdapLasso.h","_",dd,".csv"), sep = ","
,col.names = FALSE ,row.names = FALSE )

write.table(pvalue.EN.h,
paste("./p500nki_high/p500nki_high_pvalue.EN.h","_",dd,".csv"), sep = "," ,col.names =
FALSE ,row.names = FALSE )
```



```
########################################

########POWER AND TYPE I ERROR########

########################################

Pow.L<-0

Pow.A<-0

Pow.E<-0
```

```
for(u in 1:10){

        for(v in 1:u){    {if(betaJ[u,v]!=0 && pvalue.Lasso.h[u,v] < 0.05)

                                        Pow.L <- Pow.L + 1 }

                          {if(betaJ[u,v]!=0 && pvalue.AdapLasso.h[u,v] < 0.05)

                                        Pow.A <- Pow.A + 1 }

                           {if(betaJ[u,v]!=0 && pvalue.EN.h[u,v] < 0.05)

                                        Pow.E <- Pow.E + 1 }

                          }

        }

power.Lasso[dd,] <- Pow.L/nnzero(as.numeric(betaJ), na.counted = FALSE)

power.AdapLasso[dd,] <- Pow.A/nnzero(as.numeric(betaJ), na.counted = FALSE)

power.EN[dd,]<- Pow.E/nnzero(as.numeric(betaJ), na.counted = FALSE)


Type1.L <- 0

Type1.A <- 0

Type1.E <- 0

for(s in 1:10){

        for(t in 1:s){    {if(betaJ[s,t]==0 && pvalue.Lasso.h[s,t] < 0.05)

                                        Type1.L <- Type1.L + 1  }
```

```
                    {if(betaJ[s,t]==0 && pvalue.AdapLasso.h[s,t] < 0.05)
                                    Type1.A <- Type1.A + 1  }
                    {if(betaJ[s,t]==0 && pvalue.EN.h[s,t] < 0.05)
                                    Type1.E <- Type1.E + 1  }
                     }
          }
type1error.Lasso[dd,] <- Type1.L
type1error.AdapLasso[dd,] <- Type1.A
type1error.EN[dd,] <- Type1.E
}

power<- data.frame(power.Lasso= power.Lasso, power.AdapLasso= power.AdapLasso,
power.EN= power.EN)
write.table(power, paste("./p500nki_high/p500nki_high_power.test",".csv"), sep = ","
,col.names = TRUE ,row.names = FALSE )
TypeIerror<-data.frame(type1error.Lasso= type1error.Lasso, type1error.AdapLasso=
type1error.AdapLasso, type1error.EN= type1error.EN)
write.table(TypeIerror,paste("./p500nki_high/p500nki_high_TypeIerror",".csv"), sep = ","
,col.names = TRUE ,row.names = FALSE )
```

```
####################################

###############  FWER  ###############

####################################

FWER.Lasso <- nnzero(as.numeric(type1error.Lasso), na.counted = FALSE)/50

FWER.AdapLasso <- nnzero(as.numeric(type1error.AdapLasso), na.counted =
FALSE)/50

FWER.EN <- nnzero(as.numeric(type1error.EN), na.counted = FALSE)/50



FWER<-data.frame(FWER.Lasso=FWER.Lasso , FWER.AdapLasso=FWER.AdapLasso ,
FWER.EN=FWER.EN )

write.table(FWER, paste("./p500nki_high/p500nki_high_FWER",".csv"), sep = ","
,col.names = TRUE ,row.names = FALSE )
```



```
####################################################################
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวสวรรยา ภู่เงิน เกิดวันเสาร์ที่ 29 กันยายน พ.ศ. 2533 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ในปีการศึกษา 2555 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556