การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูง

นายทวีศักดิ์ เล็กตระกูลชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

A COMPARISON OF VARIABLE SCREENING METHODS IN HIGH-DIMENSION DATA

Mr. Taweesak Lektrakulchai

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 2016

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูง
โดย	นายทวีศักดิ์ เล็กตระกูลชัย
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

.....กรรมการ

(อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.ฐิตารีย์ รุ่งรัตน์เกษม)

ทวีศักดิ์ เล็กตระกูลชัย : การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูง (A COMPARISON OF VARIABLE SCREENING METHODS IN HIGH-DIMENSION DATA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.วิฐรา พึ่งพาพงศ์, 57 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีคัดกรองตัวแปรอิสระจากวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ สำหรับข้อมูลที่มีมิติสูง โดยการจำลองข้อมูลที่มีขอบเขตต่างๆ กัน โดยที่กำหนดจำนวนตัวแปรอิสระเป็น 1000 , 2000 และ 4000 ซึ่งความสัมพันธ์ของตัวแปรอิสระเป็น 0.5 และ 0.9 ทั้งนี้จะใช้ค่าความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซตตัวแปรอิสระที่ผ่านการคัดกรอง ที่ทำให้เซตตัวแปรอิสระที่แท้จริงเป็นสับเซตของเซตตัวแปรอิสระที่ผ่านการคัดกรอง เป็นเครื่องมือในการเปรียบเทียบและวัดประสิทธิภาพ

จากการศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่าวิธีลาสโซ สามารถคัดกรองตัวแปรได้มีประสิทธิภาพมากที่สุด รองลงมาคือวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญกับวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่างมีความสามารถเท่าเทียมกัน และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำเป็นวิธีที่มีประสิทธิภาพที่น้อยที่สุด

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559

5881527926 : MAJOR STATISTICS
KEYWORDS: HIGH-DIMENSIONAL DATA / VARIABLE SCREENING / MULTIPLE LINEAR REGRESSION ANALYSIS / SURE INDEPENDENCE SCREENING / LASSO / DISTANCE CORRELATION / ITERATIVELY THRESHOLDED RIDGE REGRESSION SCREENING

TAWEESAK LEKTRAKULCHAI: A COMPARISON OF VARIABLE SCREENING METHODS IN HIGH-DIMENSION DATA. ADVISOR: ASST. PROF. VITARA PUNGPAPONG, Ph.D., 57 pp.

This research aims to compare the variable screening of Multiple Linear Regression Analysis , Least Absolute Shrinkage And Selection Operator (LASSO) , Sure Independence Screening (SIS) , Distance Correlation Sure Independence Screening (DC-SIS) and Iteratively Thresholded Ridge Regression Screener (ITRRS) for high dimensional data. Here we use simulation data to compare the performance of variable screening methods. we set numbers of explanatory variables are 1000 , 2000 and 4000 which the correlation among explanatory variables are 0.5 and 0.9. The performance are compared in terms of the accuracy of variable screening , mean and standard deviation of the smallest number of sets variable screening when set true variable is a subset of variable screening.

In this study, we found that LASSO has the best performance followed by Multiple Linear Regression Analysis , SIS and DC-SIS have same result and ITRRS has the worst performance.

Department: Statistics Student's Signature .....
Field of Study: Statistics Advisor's Signature .....
Academic Year: 2016

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความช่วยเหลือและเอาใจใส่จากผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่กรุณาสละเวลาให้คำปรึกษา อบรมสั่งสอน และให้ข้อคิดเห็นต่าง ๆ ตลอดจน ให้ความช่วยเหลือ คำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน จนกระทั้งวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ผู้วิจัยของกราบขอบพระคุณท่าน รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช และ อาจารย์ ดร.ฐิตารีย์ รุ่งรัตน์เกษม กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อเป็นกรรมการสอบครั้งนี้ ตลอดจนช่วยตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ได้ให้โอกาสทางการศึกษา และอบรมสั่งสอนให้ความรู้ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษาในครั้งนี้

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฌ
บทที่ 1 บทนำ 1
1.1 ความเป็นมาและความสำคัญของปัญหา 1
1.2 วัตถุประสงค์การวิจัย 2
1.3 ข้อตกลงเบื้องต้น 2
1.4 ขอบเขตของการวิจัย 3
1.5 คำจำกัดความที่ใช้ในงานวิจัย 4
1.6 เกณฑ์ที่ใช้ในการตัดสินใจ 4
1.7 วิธีการดำเนินการวิจัย 6
1.8 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย 8
บทที่ 2 ทฤษฎีและตัวสถิติที่เกี่ยวข้อง 9
2.1 การวิเคราะห์การถดถอยพหุเชิงเส้น (Multiple Linear Regression Analysis) (สุพล ดุรงค์วัฒนา, 2015) 9
2.2 วิธีลาสโซ (Least Absolute Shrinkage And Selection Operator : LASSO) (Tibshirani, 2011) 11
2.3 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ (Sure Independence Screening : SIS) (Fan & Lv, 2008; Li, Zhong, & Zhu, 2012) 12

2.4 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง (Distance Correlation Sure Independence Screening : DC-SIS) (Li et al., 2012) 13

2.5 วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ (Iteratively Thresholded Ridge Regression Screener : ITRRS) (Fan & Lv, 2008) 14

2.6 เกณฑ์ที่ใช้ในการตัดสินใจ 15

บทที่ 3 วิธีการดำเนินการศึกษา 16

3.1 ขอบเขตของการวิจัย 16

3.2 ขั้นตอนในการดำเนินการศึกษา 18

3.3 ขั้นตอนการทำงานของโปรแกรม 20

บทที่ 4 ผลการวิจัย 22

4.1 ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร จากการคัดกรองตัวแปร ทั้ง 5 วิธี 24

4.2 ผลเปรียบเทียบค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ทั้ง 5 วิธี 28

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 34

5.1 สรุปผลการวิจัย 34

5.2 สรุปผลโดยรวม 37

5.3 ข้อเสนอแนะ 37

รายการอ้างอิง 39

ภาคผนวก 40

ประวัติผู้เขียนวิทยานิพนธ์ 57

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ 4.1 แสดงค่าความถูกต้องในการคัดกรองตัวแปรของแต่ละสถานการณ์ จากข้อมูลจำลอง 100 แบบจำลอง สำหรับตัวแบบเป็นแบบเชิงเส้น 25

ตารางที่ 4.2 แสดงค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ของแต่ละสถานการณ์ จากข้อมูลจำลอง 100 แบบจำลอง สำหรับตัวแบบเป็นแบบเชิงเส้น 29

ตารางที่ 5.1.1 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุดสำหรับตัวแบบเชิงเส้น เมื่อพิจารณาความถูกต้องในการคัดกรองตัวแปร โดยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 200 โดยจำแนกตามอัตราส่วนขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) , ความสัมพันธ์ของตัวแปรอิสระ และ อัตราส่วนสัญญาณต่อสัญญาณรบกวน 34

ตารางที่ 5.1.2 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุดสำหรับตัวแบบเชิงเส้น เมื่อพิจารณาจากค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ โดยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 200 โดยจำแนกตามอัตราส่วนขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) , ความสัมพันธ์ของตัวแปรอิสระ และ อัตราส่วนสัญญาณต่อสัญญาณรบกวน 35

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องด้วยความก้าวหน้าทางด้านเทคโนโลยีในโลกปัจจุบันที่มีความเร็วและสะดวกมากขึ้น ข้อมูลต่างๆที่สามารถนำมาวิเคราะห์ก็มีอยู่มากมายและถูกจัดเก็บไว้ในรูปอิเล็กทรอนิกส์ที่ง่ายต่อการเข้าถึง วิธีในการวิเคราะห์ข้อมูลที่ใช้กันอย่างแพร่หลายก็คือ วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น (Multiple Linear Regression Analysis) ซึ่งเป็นการวิเคราะห์ความสัมพันธ์ของตัวแปรอิสระ (Explanatory Variable) กับตัวแปรตาม (Response Variable) โดยวิธีนี้จะสามารถใช้ได้กับข้อมูลที่มีจำนวนตัวแปรอิสระน้อยกว่าจำนวนขนาดตัวอย่าง ($p < n$) เท่านั้น แต่ข้อมูลในปัจจุบันที่ตัวแปรมีจำนวนมากเราต้องเผชิญกับกรณีที่ตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่างด้วย ซึ่งข้อมูลลักษณะนี้เรียกว่า ข้อมูลที่มีมิติสูง (High – Dimensional Data) การวิเคราะห์การถดถอยพหุเชิงเส้นไม่สามารถนำมาใช้กับข้อมูลที่มีมิติสูงได้ เนื่องจากเกิดปัญหา 3 ประการ 1. ไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีกำลังสองน้อยสุด (Ordinary Least Square (OLS)) 2. ปัญหาที่ตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นสูง (Multicollinearity) ซึ่งจะส่งผลต่อการประมาณค่าที่ไม่เสถียรและไม่มีประสิทธิภาพ และ 3. การแปลผลลัพธ์ของตัวแบบมีความยากและสลับซับซ้อน(วิฐรา พึ่งพาพงศ์, 2015)

ในการวิเคราะห์การถดถอยพหุเชิงเส้นนั้นจะมีการคัดเลือกตัวแปร (Variable Selection) เพื่อเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามเข้ามาในตัวแบบ ด้วยวิธี Stepwise Regression ก็เป็นวิธีหนึ่งในการคัดเลือกตัวแปรซึ่งเป็นวิธีที่ดัดแปลงมาจากวิธี Forward Selection คือการคัดเลือกตัวแปรเข้าสู่ตัวแบบ และทำการ Backward Elimination คือการคัดเลือกตัวแปรออกจากตัวแบบ ซึ่งจะมีเกณฑ์ในการวัดตัวแบบที่มีความเหมาะสม 2 เกณฑ์ คือ Akaike's Information Criteria (AIC) และ Bayes Information Criteria (BIC) แต่การคัดเลือกตัวแปรจะไม่สามารถใช้กับข้อมูลที่มีมิติสูงได้

Tibshirani (1996) เสนอวิธี LASSO โดยมีวัตถุประสงค์เพื่อให้เป็นวิธีที่สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์ β ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงได้ในคราวเดียวกัน ซึ่ง β ที่ประมาณส่วนใหญ่จะมีค่าเป็นศูนย์และบางตัวไม่เท่ากับศูนย์ แต่ไม่สามารถทำการทดสอบสมมติฐานที่ว่า $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$ ได้ จึงไม่สามารถบอกได้ว่าตัวแปรอิสระที่มีค่า $\beta_j \neq 0$ นั้นจะมีความสัมพันธ์กับตัวแปรตามหรือไม่ จึงมีการเสนอวิธีคัดกรองตัวแปร (Variable Screening) โดยมีวัตถุประสงค์เพื่อกรองตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดมา d

ตัว โดยที่ d < n ดังนั้นเราจึงสามารถใช้วิธีการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิมกับตัวแปรที่ผ่านการคัดกรองแล้วตลอดจนสามารถใช้วิธีการคัดเลือกตัวแปรเพื่อให้ได้ตัวแบบในขั้นสุดท้ายได้

ในงานวิจัยนี้มีวัตถุประสงค์ในการนำเสนอทางเลือกในการลดจำนวนตัวแปรอิสระให้มีขนาดเล็กลง รวมถึงนำเสนอบทวิเคราะห์ในการเลือกใช้เครื่องมือสำหรับการคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูงให้มีความเหมาะสมในการประยุกต์ใช้งาน

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำในการคัดกรองตัวแปรจากข้อมูลที่มีมิติสูง

1.3 ข้อตกลงเบื้องต้น

ศึกษาตัวแปรภายใต้รูปแบบความสัมพันธ์แบบเชิงเส้น

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

จาก $X_i \sim N(0, \Sigma)$

โดยที่ $\Sigma = \begin{pmatrix} \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \cdots & 0 & 0 \\ \vdots & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \vdots & \vdots \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} \end{pmatrix}_{p \times p}$

เมื่อ X_{ij} คือ ตัวแปรอิสระทุก $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$

Y_i คือ ตัวแปรตามทุก $i = 1, 2, \ldots, n$

β_0, β_j คือ พารามิเตอร์ที่ไม่ทราบค่าจริง เรียกว่า สัมประสิทธิ์ความถดถอย (Coefficients of Regression) ทุก $j = 1, 2, \ldots, p$

ε_i คือ ค่าความคลาดเคลื่อน โดยที่ $\varepsilon \overset{iid}{\sim} N(0, \sigma^2 I_n)$

1.4 ขอบเขตของการวิจัย

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 200:1000 , 200:2000 และ 200:4000

2. ศึกษาภายใต้ความสัมพันธ์ของ Correlation ของตัวแปรอิสระ 2 ระดับ คือ

ระดับที่ 1 : $\rho = 0.5$

ระดับที่ 2 : $\rho = 0.9$

กล่าวคือ สำหรับ $i = 1,2,\ldots,n \quad X_i \sim N(0,\Sigma)$

โดยที่ $\Sigma = \begin{pmatrix} \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \cdots & 0 & 0 \\ \vdots & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \vdots & \vdots \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} \end{pmatrix}_{p \times p}$

3. ศึกษาภายใต้ค่าสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 โดยกำหนดให้มีค่าเป็น 1 กับ -1 ด้วยความน่าจะเป็นเท่ากัน

4. กำหนดให้ $s_i = \{j : \beta_j \neq 0\}$ เป็นเซตของดัชนีสัมประสิทธิ์การถดถอยที่แท้จริง β_j ไม่เท่ากับ 0 (สมมติว่าค่าในเซต s_i เรียงลำดับจากน้อยไปมาก) ศึกษาภายใต้จำนวนสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 ดังนี้

กรณีที่ p=1,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 6$

กรณีที่ p=2,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 10$

กรณีที่ p=4,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 20$

5. ศึกษาภายใต้ค่าความแปรปรวนของค่าความคลาดเคลื่อน σ^2 โดยที่กำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio (SNR)) = 2 , 8 (Bühlmann & Mandozzi, 2014)

$$SNR = \sqrt{\frac{\beta^T X^T X \beta}{n\sigma^2}}$$

6. ศึกษาภายใต้ตัวแบบเชิงเส้น : ความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามเป็นแบบเชิงเส้น

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad ; \varepsilon \overset{iid}{\sim} N\left(0, \sigma^2 I_n\right)$$

7. ศึกษาภายใต้การจำลองข้อมูลในแต่กรณีจำนวน 100 ครั้ง

1.5 คำจำกัดความที่ใช้ในงานวิจัย

ข้อมูลที่มีมิติสูง (High Dimensional Data)

คือ ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าจำนวนของขนาดตัวอย่าง (p > n)

ค่า p-value

คือ ค่าความน่าจะเป็นที่แสดงถึงความเสี่ยงในการปฏิเสธสมมติฐานว่าง เมื่อสมมติฐานว่างเป็นจริง

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการคัดกรองตัวแปรวิธีใดเหมาะสมในการคัดกรองตัวแปรของข้อมูลที่มีมิติสูงมากที่สุด โดยจะพิจารณาจากความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

กำหนดให้ $s_i = \{j : \beta_j \neq 0\}$ เป็นเซตของดัชนีสัมประสิทธิ์การถดถอยที่แท้จริง β_j ไม่เท่ากับ 0 ในการจำลองข้อมูลครั้งที่ i (สมมติว่าค่าในเซต s_i เรียงลำดับจากน้อยไปมาก)

$T_i = \{j :$ ตัวแปรอิสระตัวที่ j ที่ผ่านการคัดกรองตัวแปร โดยเรียงลำดับตามความสัมพันธ์ของตัวแปรอิสระจากค่าที่มีความสัมพันธ์มากไปหาค่าที่มีความสัมพันธ์น้อย ซึ่งจะมีความแตกต่างกันในแต่ละวิธี $\}$

1. ค่าความถูกต้องในการคัดกรองตัวแปร

คือ การพิจารณาเซตของดัชนีสัมประสิทธิ์การถดถอยที่ (s_i) เป็นซับเซตของเซตที่ได้มาจากวิธีคัดกรองตัวแปร (T_i) นั่นก็คือ $u_i = \begin{cases} 1_{\{s_i \subset T_i\}} \\ 0_{\{s_i \not\subset T_i\}} \end{cases}$ ในการจำลองครั้งที่ i ซึ่งจากการจำลองข้อมูลทั้งหมด 100 ครั้ง เพราะฉะนั้นค่าความถูกต้องในการคัดกรองตัวแปร คือ $U = \sum_{i=1}^{100} u_i$ โดยที่ดูค่าความถูกต้องในการคัดกรองตัวแปรที่มีค่ามากที่สุด

2. ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

ในการหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ จะสามารถคำนวณได้ก็ต่อเมื่อ $s_i \subset T_i$ เท่านั้น ดังนั้นจึงกำหนดให้ค่าที่ได้จากการนับตำแหน่งของตัวแปรอิสระของเซต T_i ตั้งแต่ตัวแรกจนกว่าจะมีตัวแปรอิสระในเซต s_i ครบทุกตัวจึงหยุด โดยที่ค่าของตำแหน่งที่นับจนครบนั้นมีค่าเป็น D_i ในการจำลองครั้งที่ i ซึ่งจากการจำลองข้อมูลทั้งหมด 100 ครั้ง จึงดูค่าเฉลี่ยและค่าเบี่ยงเบงมาตรฐานที่มีค่าน้อยที่สุด ซึ่งหาได้จาก $M = \frac{\sum D_i}{U}$ และ $S.D. = \sqrt{\frac{\sum (D_i - M)^2}{U - 1}}$ ตามลำดับ

<u>ตัวอย่างเช่น</u>

ให้ $A = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}$ เป็นเซตของตัวแปรอิสระที่ยังไม่ผ่านการคัดกรอง

ให้ $s = \{2,9,14\}$ เป็นเซตของตัวแปรอิสระที่แท้จริงหรือมีค่าสัมประสิทธิ์ไม่เท่ากับ 0

1. $T_1 = \{11,2,4,12,8,7,3,1\}$ คือตัวแปรอิสระที่ได้จากการคัดกรอง จะสังเกตว่า เซต T_1 นั้นมีตัวแปรอิสระเพียงตัวเดียวที่ผ่านการคัดกรองมาได้ถูกต้องนั้นคือ 2 ดังนั้นจึงไม่สามารถหาค่า D_1 ได้เนื่องจาก s ไม่เป็นสับเซตของ T_1

2. $T_2 = \{15,2,6,9,14,3,7,8\}$ คือตัวแปรอิสระที่ได้จากการคัดกรอง จะสังเกตว่า เซต T_2 นั้นมีตัวแปรอิสระผ่านการคัดกรองมาได้ถูกต้องครบทั้งสามตัวคือ 2 , 9 และ 14 ดังนั้นจึงสามารถหาค่า D_2 ได้ โดยทำการนับตำแหน่งของเซต T_2 จากตัวแรกไปจนกว่าจะมีตัวแปรอิสระในเซต s ครบ

ทุกตัว นั้นคือ 15 , 2 , 6 , 9 และ 14 จะเห็นว่าเมื่อนับไปห้าตัวพบว่าเซต T_2 นั้นมีตัวแปรอิสระในเซต s ครบทุกตัวแล้ว เพราะฉะนั้นค่า $D_2 = 5$

3. $T_3 = \{9,14,2,4,8\}$ คือตัวแปรอิสระที่ได้จากการคัดกรอง จะสังเกตว่า เซต T_3 นั้นมีตัวแปรอิสระผ่านการคัดกรองมาได้ถูกต้องครบทั้งสามตัวคือ 2 , 9 และ 14 ดังนั้นจึงสามารถหาค่า D_3 ได้ โดยทำการนับตำแหน่งของเซต T_3 จากตัวแรกไปจนกว่าจะมีตัวแปรอิสระในเซต s ครบทุกตัว นั้นคือ 9 , 14 และ 2 จะเห็นว่าเมื่อนับไปสามตัวพบว่าเซต T_3 นั้นมีตัวแปรอิสระในเซต s ครบทุกตัวแล้ว เพราะฉะนั้นค่า $D_3 = 3$

1.7 วิธีการดำเนินการวิจัย

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง

2. กำหนดการจำลองข้อมูล

   2.1 กำหนดค่าเริ่มต้นโดยการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และจำนวนพารามิเตอร์ p ตัว โดยใช้อัตราส่วน n:p ดังนี้ 200:1000 , 200:2000 และ 200:4000

   2.2 กำหนดให้จำนวนสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 ดังนี้ 6 , 10 และ 20

   2.3 กำหนดให้ค่าสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 มีค่าเป็น 1 กับ -1 ด้วยความน่าจะเป็นที่เท่ากัน

   2.4 จำลองข้อมูลภายใต้รูปแบบความสัมพันธ์แบบเชิงเส้น

3. จำลองข้อมูลจากค่าเริ่มต้นที่กำหนด

   3.1 จำลองตัวแปรอิสระ X ให้มีการแจกแจงแบบปกติ $X_i \sim N(0,\Sigma)$

โดยที่ $\Sigma = \begin{pmatrix} \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \cdots & 0 & 0 \\ \vdots & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \vdots & \vdots \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} \end{pmatrix}_{p \times p}$

3.2 จำลองค่าความคลาดเคลื่อน ε ให้มีการแจกแจงแบบปกติ $\varepsilon \overset{iid}{\sim} N\left(0, \sigma^2 I_n\right)$

3.3 จำลองข้อมูลตัวแปรตามภายใต้รูปแบบความสัมพันธ์เชิงเส้น

4. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคัดกรองตัวแปรอิสระ

4.1 วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น

4.2 วิธีลาสโซ

4.3 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ

4.4 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง

4.5 วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ

5. นำผลที่ได้จากข้อที่ 4. มาหาค่าความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

6. วิเคราะห์และสรุปผลจากทั้ง 5 วิธี

1.8 ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

เพื่อเป็นแนวทางในการเลือกใช้วิธีการคัดกรองตัวแปรสำหรับลดมิติของตัวแปรให้มีขนาดเล็กลงและสามารถนำไปวิเคราะห์ข้อมูลต่อไป

บทที่ 2
ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การประมาณค่าสัมประสิทธิ์การถดถอยของข้อมูลในตัวแบบเพื่อคัดเลือกตัวแปรเข้ามายังตัวแบบในกรณีที่ข้อมูลมีจำนวนตัวแปรอิสระน้อยกว่าจำนวนขนาดตัวอย่าง (p < n) สามารถทำได้โดยวิธีกำลังสองน้อยสุด (Ordinary Least Square (OLS)) แต่เนื่องจากปัจจุบันข้อมูลมีขนาดใหญ่ขึ้น ดังนั้นจึงมีข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าจำนวนขนาดตัวอย่าง (p > n) ข้อมูลลักษณะนี้เรียกว่า ข้อมูลที่มีมิติสูง (High – Dimensional Data) ซึ่งไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุด (OLS) ได้ ดังนั้นในงานวิจัยนี้จะกล่าวถึงวิธีคัดกรองตัวแปรอิสระสำหรับข้อมูลที่มีมิติสูง คือ วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ โดยเกณฑ์ที่ใช้ในการตัดสินใจเพื่อวิเคราะห์ข้อมูลและสถิติที่ได้ คือ ค่าความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

2.1 การวิเคราะห์การถดถอยพหุเชิงเส้น (Multiple Linear Regression Analysis) (สุพล ดุรงค์วัฒนา, 2015)

การวิเคราะห์การถดถอยพหุเชิงเส้นเป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตาม Y (Response Variable) เพียงตัวเดียวกับตัวแปรอิสระ X (Explanatory Variable) หลายตัวแปร $\left(X_1, X_2, \ldots, X_p\right)$ โดยที่ข้อมูลของตัวแปรตามเป็นเชิงปริมาณ และข้อมูลของตัวแปรอิสระเป็นเชิงปริมาณหรือคุณภาพก็ได้ จะสามารถเขียนตัวแบบได้ดังนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad ; i = 1, 2, \ldots, n \qquad (1)$$

โดยที่

X_{ij} คือ ตัวแปรอิสระ ทุก $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$

Y_i คือ ตัวแปรตาม ทุก $i = 1, 2, \ldots, n$

β_0, β_j คือ พารามิเตอร์ที่ไม่ทราบค่าจริง เรียกว่า สัมประสิทธิ์ความถดถอย (Coefficients of Regression) ทุก $j = 1, 2, \ldots, p$

ε_i คือ ค่าความคลาดเคลื่อน โดยที่ $\varepsilon \overset{iid}{\sim} N\left(0, \sigma^2 I_n\right)$

การประมาณค่าพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$

ในการประมาณค่าของพารามิเตอร์ จะใช้วิธีที่เรียกว่า วิธีกำลังสองของความคลาดเคลื่อนน้อยที่สุดแบบทั่วไป (Ordinary Least Squares Method : OLS) โดยที่ใช้ค่าสังเกตของตัวอย่างนั่นคือ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ เป็นค่าประมาณของ $\beta_0, \beta_1, \ldots, \beta_p$ ตามลำดับ และ y_i, x_{ij} เป็นข้อมูลค่าสังเกตที่เก็บรวบรวมได้จริง จะได้ตัวแบบการถดถอยดังนี้ $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_p x_{ip} + e_i$; $i = 1, 2, \ldots, n$

โดยจะกำหนดให้ $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_p x_{ip}$

$$y_i = \hat{y}_i + e_i$$

เรียก $\hat{y}_i$ ว่า ค่าพยากรณ์ (Predicted values) และเรียก e_i ว่า เศษเหลือ (Residuals)

ซึ่งการหาค่าประมาณ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ โดยวิธีกำลังสองของเศษเหลือน้อยที่สุด นั่นคือ $\sum_{i=1}^{n} e_i^2 = \tilde{e}^T \tilde{e}$ น้อยสุด ซึ่งสามารถทำได้ดังนี้ $\frac{\partial}{\partial b_j}\left(\sum_{i=1}^{n} e_i{}^2\right) = \frac{\partial}{\partial b_j}\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right) = 0$; $j = 1, 2, \ldots, \mathrm{p}$

จะได้ว่า

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \hat{\beta}_p \sum_{i=1}^{n} x_{i1} x_{ip} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{ip} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} x_{ip} + \ldots + \hat{\beta}_p \sum_{i=1}^{n} x_{ip}^2 = \sum_{i=1}^{n} x_{ip} y_i$$

เมื่อแก้สมการทั้งหมดก็จะสามารถหาค่า $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ ได้หรือสามารถเขียนให้อยู่ในรูปของเมทริกซ์ได้ดังนี้ $\tilde{\beta} = \left(x^T x\right)^{-1} x^T \tilde{y}$

การทดสอบสมมติฐาน

เป็นการทดสอบสมมติฐานทางสถิติของค่าสัมประสิทธิ์แต่ละตัวที่ได้จากการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ ว่ามีความแตกต่างจากศูนย์อย่างมีนัยสำคัญทางสถิติหรือไม่

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_a &: \beta_j \neq 0 \end{aligned} \quad ; \quad j = 1, 2, \ldots, p$$

ตัวสถิติทดสอบสำหรับสมมติฐานข้างต้นสามารถคำนวณได้ดังนี้

$$t = \frac{\hat{\beta}_j - 0}{SE\left(\hat{\beta}_j\right)} \sim t_{n-p-1}$$

ซึ่งเป็นตัวสถิติทดสอบแบบ t ที่มีองศาความเป็นอิสระเท่ากับ n-p-1

ซึ่งในที่นี้เราใช้งานวิธีการวิเคราะห์การถดถอยแบบตัวแปรเดียว (Univariate Regression : UR) เนื่องจากว่าจำนวนตัวแปรมีมากกว่าขนาดตัวอย่างจึงไม่สามารถใช้งานด้วยวิธีหลายตัวแปรได้ ดังนั้นตัวแบบจะอยู่ในรูปแบบดังนี้

$$Y_i = \beta_0 + \beta_j X_{ij} + \varepsilon_i \ ; \ i = 1, 2, \ldots, n \qquad (1.1)$$

วิธีการคัดกรองโดยการวิเคราะห์การถดถอยแบบตัวแปรเดียวนั้น จะทำการเก็บตัวแปรอิสระที่ผ่านการคัดกรองให้อยู่ในเซต T_i โดยใช้ค่า p – value (P) ซึ่งในงานวิจัยนี้ใช้การควบคุม False Discovery Rate (FDR) (Benjamini & Hochberg, 1995) ในการทดสอบสมมติฐาน มีขั้นตอนดังนี้

1. เรียงลำดับค่า p – value จากค่าน้อยไปหาค่ามาก $P_{j(1)} \le P_{j(2)} \le \ldots \le P_{j(m)}$; $j = 1, 2, \ldots, p$

2. ทดสอบสมมติฐาน ถ้า $P_{j(k)} \le k \times \frac{0.05}{p}$; $k = 1, 2, \ldots, \mathrm{m}$ เราจะปฏิเสธการทดสอบ $H_{0\,j(k)}$; $k = 1, 2, \ldots, \mathrm{m}$ นั้นคือ ตัวแปรอิสระตัวที่ j เป็นตัวที่ผ่านการคัดกรอง

3. จัดเก็บตัวแปรอิสระตัวที่ j จากข้อ 2. ไว้ในเซตของ T_i โดยจะเก็บทั้งหมด n ตัว โดยตัวที่เหลือนั้นถึงแม้จะมีค่า $P_{j(k)} \le k \times \frac{0.05}{p}$ ก็จะถือว่าไม่ผ่านการคัดกรอง

2.2 วิธีลาสโซ (Least Absolute Shrinkage And Selection Operator : LASSO)

(Tibshirani, 2011)

Tibshirani (1996) ได้เสนอวิธี Lasso ในบทความเรื่อง Regression Shrinkage and Selection via the Lasso ซึ่งเป็นวิธีที่สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่า β ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง (n<p) ได้ในคราวเดียวกัน โดยวิธีนี้จะทำให้ค่าสัมประสิทธิ์ β ส่วนใหญ่เป็นศูนย์และค่าสัมประสิทธิ์ β บางส่วนไม่เท่ากับศูนย์ ซึ่งสามารถหาค่าประมาณของสัมประสิทธิ์ β ได้ดังนี้

$$\hat{\beta} = \arg\min \left[\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 \right] + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \qquad (2)$$

โดยที่

X_{ij} คือ ตัวแปรอิสระ ทุก $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$

Y_i คือ ตัวแปรตาม ทุก $i = 1, 2, \ldots, n$

λ คือ tuning parameter มีค่ามากกว่าหรือเท่ากับ 0

วิธีการคัดกรองโดยวิธีลาสโซนั้น จะทำการเก็บตัวแปรอิสระที่ผ่านการคัดกรองให้อยู่ในเซต T_i โดยใช้ค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ ที่ได้จากสมการที่ (2) ซึ่งมีขั้นตอนดังนี้

1. นำค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ มาใส่ค่าสัมบูรณ์ จากนั้นเรียงลำดับจากค่ามากไปหาค่าน้อย

2. จัดเก็บตัวแปรอิสระจากข้อ 1 ไว้ในเซตของ T_i โดยจะเก็บเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ ไม่เท่ากับศูนย์

2.3 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ (Sure Independence Screening : SIS) (Fan & Lv, 2008; Li, Zhong, & Zhu, 2012)

Jianqing Fan และ Jinchi Lv (2007) ได้เสนอวิธี SIS ในบทความเรื่อง Sure Independence Screening for Ultra-High Dimensional Feature Space ซึ่งเป็นวิธีในการลดมิติของข้อมูลที่มีมิติสูง โดยจะทำการคัดเลือกตัวแปรอิสระที่มีความสำคัญกับตัวแปรตาม ซึ่งจะดูจากค่าความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม ดังนี้

$$\omega_j^{SIS} = corr(\mathrm{X}_j, \mathrm{Y}) = \frac{\mathrm{cov}\left(\mathrm{X}_j, \mathrm{Y}\right)}{\sqrt{\mathrm{var}\left(X_j\right)}\sqrt{\mathrm{var}\left(Y\right)}} \qquad (3)$$

โดยที่

X_j คือ ตัวแปรอิสระ ทุก $j = 1, 2, \ldots, p$

Y คือ ตัวแปรตาม

ω_j^{SIS} คือ ค่าความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม ทุก $j = 1, 2, \ldots, p$

จากนั้นทำการลดจำนวนตัวแปรอิสระให้มีขนาด $d = \lfloor \gamma n \rfloor < n$ โดยที่ $\gamma \in (0,1)$ และ $\lfloor \gamma n \rfloor$ คือ ส่วนที่เป็นจำนวนเต็มของ γn

วิธีการคัดกรองโดยวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญนั้น จะทำการเก็บตัวแปรอิสระที่ผ่านการคัดกรองให้อยู่ในเซต T_i โดยใช้ค่าความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม ω_j^{SIS} ที่ได้จากสมการที่ (3) ซึ่งมีขั้นตอนดังนี้

1. นำค่าความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม ω_j^{SIS} มาใส่ค่าสัมบูรณ์ จากนั้นเรียงลำดับจากค่ามากไปหาค่าน้อย

2. จัดเก็บตัวแปรอิสระจากข้อ 1 ไว้ในเซตของ T_i โดยจะเก็บเฉพาะตัวแปรอิสระมาจำนวน d ตัวแรก

<u>หมายเหตุ</u> ในงานวิจัยนี้กำหนดจำนวนตัวแปรอิสระที่ผ่านการกรองมาจำนวน d ตัวแรกเป็น

$$d1 = \frac{n}{\ln(n)}\ ,\ d2 = 2 \times \left(\frac{n}{\ln(n)}\right) \text{ และ } d3 = 3 \times \left(\frac{n}{\ln(n)}\right)$$

2.4 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง (Distance Correlation Sure Independence Screening : DC-SIS) (Li et al., 2012)

Runze Li , Wei Zhong และ Liping Zhu (2012) ได้เสนอวิธี DC-SIS ในบทความเรื่อง Feature Screening via Distance Correlation Learning ซึ่งเป็นวิธีในการลดมิติของข้อมูลที่มีมิติสูง และเป็นวิธีที่มีความคล้ายคลึงกับวิธี SIS แต่จะสามารถใช้กับข้อมูลที่เป็นแบบเชิงเส้น (Linear) หรือไม่ใช่เชิงเส้น (Nonlinear) ก็ได้ และสามารถใช้กับตัวแปรตามที่มีตัวแปรเดียว (Univariate Response) หรือหลายตัวแปร (Multivariate Responses) ก็ได้ โดยจะทำการคัดเลือกตัวแปรอิสระที่มีความสำคัญกับตัวแปรตาม ซึ่งจะดูจากค่า Distance Correlation ระหว่างตัวแปรอิสระกับตัวแปรตาม ดังนี้

$$\omega_j^{DC-SIS} = dcorr^2\left(X_j, \mathrm{Y}\right) = \frac{r\times\arcsin\left(r\right)+\sqrt{1-r^2}-r\times\arcsin\left(\frac{r}{2}\right)-\sqrt{4-r^2}+1}{1+\frac{\pi}{3}-\sqrt{3}} \qquad (4)$$

โดยที่

X_j คือ ตัวแปรอิสระ ทุก $j=1,\ldots,p$

Y คือ ตัวแปรตาม

r คือ ค่า Pearson correlation ระหว่างตัวแปรอิสระกับตัวแปรตาม

ω_j^{DC-SIS} คือ ค่า Distance Correlation ระหว่างตัวแปรอิสระกับตัวแปรตาม ทุก $j=1,\ldots,p$

จากนั้นทำการลดจำนวนตัวแปรอิสระให้มีขนาด $d=\left[\gamma n\right]<n$ โดยที่ $\gamma\in\left(0,1\right)$ และ $\left[\gamma n\right]$ คือส่วนที่เป็นจำนวนเต็มของ γn

วิธีการคัดกรองโดยวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่างนั้น จะทำการเก็บตัวแปรอิสระที่ผ่านการคัดกรองให้อยู่ในเซต T_i โดยใช้ค่าความสัมพันธ์ของระยะห่างระหว่างตัวแปรอิสระกับตัวแปรตาม ω_j^{DC-SIS} ที่ได้จากสมการที่ (4) ซึ่งมีขั้นตอนดังนี้

1. นำค่าความสัมพันธ์ของระยะห่างระหว่างตัวแปรอิสระกับตัวแปรตาม ω_j^{DC-SIS} มาใส่ค่าสัมบูรณ์จากนั้นเรียงลำดับจากค่ามากไปหาค่าน้อย
2. จัดเก็บตัวแปรอิสระจากข้อ 1 ไว้ในเซตของ T_i โดยจะเก็บเฉพาะตัวแปรอิสระมาจำนวน d ตัวแรก

<u>หมายเหตุ</u> ในงานวิจัยนี้กำหนดจำนวนตัวแปรอิสระที่ผ่านการกรองมาจำนวน d ตัวแรกเป็น

$d1 = \frac{n}{\ln(n)}$, $d2 = 2\times\left(\frac{n}{\ln(n)}\right)$ และ $d3 = 3\times\left(\frac{n}{\ln(n)}\right)$

2.5 วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ (Iteratively Thresholded Ridge Regression Screener : ITRRS) (Fan & Lv, 2008)

Jianqing Fan และ Jinchi Lv (2007) ได้เสนอวิธี ITRRS ในบทความเรื่อง Sure Independence Screening for Ultra-High Dimensional Feature Space เป็นวิธีในการลดจำนวนตัวแปรอิสระที่ทำการขยายออกมาจากวิธี SIS ซึ่งมีความแตกต่างคือ จะทำการลดตัวแปรอิสระลงให้เหลือ $[\gamma p]$ จนกระทั่งมีจำนวนน้อยกว่าขนาดตัวอย่าง n โดยจะดูจากค่าสัมประสิทธิ์การถอถอยริดจ์ระหว่างตัวแปรอิสระกับตัวแปรตาม

$$\omega_j^{ITRRS} = \arg\min\left[\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2\right] + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (5)$$

โดยที่

X_{ij} คือ ตัวแปรอิสระ ทุก $i = 1,2,\ldots,n$, $j = 1,2,\ldots,p$

Y_i คือ ตัวแปรตาม ทุก $i = 1,2,\ldots,n$

λ คือ tuning parameter มีค่ามากกว่าหรือเท่ากับ 0

ω_j^{ITRRS} คือ ค่าสัมประสิทธิ์ของ Ridge Regression ทุก $j = 1,2,\ldots,p$

โดยเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม $[\gamma p]$ ลำดับแรก โดยที่ $\gamma \in (0,1)$ และ $[\gamma p]$ คือ ส่วนที่เป็นจำนวนเต็มของ γp

วิธีการคัดกรองโดยวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำนั้น จะทำการเก็บตัวแปรอิสระที่ผ่านการคัดกรองให้อยู่ในเซต T_i โดยใช้ค่าสัมประสิทธิ์ของ Ridge Regression ω_j^{ITRRS} ที่ได้จากสมการที่ (5) ซึ่งมีขั้นตอนดังนี้

1. นำค่าสัมประสิทธิ์ของ Ridge Regression ω_j^{ITRRS} มาใส่ค่าสัมบูรณ์ จากนั้นเรียงลำดับจากค่ามากไปหาค่าน้อย

2. จัดเก็บตัวแปรอิสระจากข้อ 1 ไว้ในเซตของ T_i จากการวนซ้ำจนได้จำนวนตัวแปรอิสระ $d < n$

<u>หมายเหตุ</u> ในงานวิจัยนี้กำหนดให้จำนวนตัวแปรอิสระที่ลดลงในแต่ละครั้งของการวนรอบเป็น $d = 0.8\times p$

2.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการคัดกรองตัวแปรวิธีใดเหมาะสมในการคัดกรองตัวแปรของข้อมูลที่มีมิติสูงมากที่สุด โดยจะพิจารณาจากความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

กำหนดให้ $s_i = \{\mathrm{j} : \beta_j \neq 0\}$ เป็นเซตของดัชนีสัมประสิทธิ์การถดถอยที่แท้จริง β_j ไม่เท่ากับ 0 ในการจำลองข้อมูลครั้งที่ i (สมมติว่าค่าในเซต s_i เรียงลำดับจากน้อยไปมาก)

$T_i = \{\mathrm{j}:$ ตัวแปรอิสระตัวที่ j ที่ผ่านการคัดกรองตัวแปร โดยเรียงลำดับตามความสัมพันธ์ของตัวแปรอิสระจากค่าที่มีความสัมพันธ์มากไปหาค่าที่มีความสัมพันธ์น้อย ซึ่งจะมีความแตกต่างกันในแต่ละวิธี $\}$

1. ค่าความถูกต้องในการคัดกรองตัวแปร

คือ การพิจารณาเซตของดัชนีสัมประสิทธิ์การถดถอยที่ (s_i) เป็นซับเซตของเซตที่ได้มาจากวิธีคัดกรองตัวแปร (T_i) นั่นก็คือ $u_i = \begin{cases} 1_{\{s_i \subset T_i\}} \\ 0_{\{s_i \not\subset T_i\}} \end{cases}$ ในการจำลองครั้งที่ i ซึ่งจากการจำลองข้อมูลทั้งหมด 100 ครั้ง เพราะฉะนั้นค่าความถูกต้องในการคัดกรองตัวแปร คือ $U = \sum_{i=1}^{100} u_i$ โดยที่ดูค่าความถูกต้องในการคัดกรองตัวแปรที่มีค่ามากที่สุด

2. ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

ในการหาค่าค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ จะสามารถคำนวณได้ก็ต่อเมื่อ $s_i \subset T_i$ เท่านั้น ดังนั้นจึงกำหนดให้ค่าที่ได้จากการนับตำแหน่งของตัวแปรอิสระของเซต T_i ตั้งแต่ตัวแรกจนกว่าจะมีตัวแปรอิสระในเซต s_i ครบทุกตัวจึงหยุด โดยที่ค่าของตำแหน่งที่นับจนครบนั้นมีค่าเป็น D_i ในการจำลองครั้งที่ i ซึ่งจากการจำลองข้อมูลทั้งหมด 100 ครั้ง จึงดูค่าเฉลี่ยและค่าเบี่ยงเบงมาตรฐานที่มีค่าน้อยที่สุด ซึ่งหาได้จาก $M = \frac{\sum D_i}{U}$ และ $S.D. = \sqrt{\frac{\sum (D_i - M)^2}{U - 1}}$ ตามลำดับ

บทที่ 3
วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการกรองตัวแปรอิสระสำหรับข้อมูลที่มีมิติสูง โดยใช้วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ โดยมีการจำลองข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) จานนั้นทำการกรองตัวแปรอิสระจากวิธีข้างต้นจนได้ตัวแปรอิสระที่มีจำนวนน้อยกว่าขนาดตัวอย่าง ซึ่งสามารถพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ โดยทำการวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชัน 3.3.3 ภายใต้ของเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการวิจัย

1. ศึกษาภายใต้อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ (n : p) ที่ 200:1000 , 200:2000 และ 200:4000

2. ศึกษาภายใต้ความสัมพันธ์ของ Correlation ของตัวแปรอิสระ 2 ระดับ คือ

ระดับที่ 1 : $\rho = 0.5$

ระดับที่ 2 : $\rho = 0.9$

กล่าวคือ สำหรับ $i = 1, 2, \ldots, n \quad X_i \sim N(0, \Sigma)$

โดยที่ $\Sigma = \begin{pmatrix} \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \cdots & 0 & 0 \\ \vdots & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \vdots & \vdots \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} \end{pmatrix}_{p \times p}$

3. ศึกษาภายใต้ค่าสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 โดยกำหนดให้มีค่าเป็น 1 กับ -1 ด้วยความน่าจะเป็นเท่ากัน

4. กำหนดให้ $s = \{j : \beta_j \neq 0\}$ เป็นเซตของดัชนีสัมประสิทธิ์การถดถอยที่แท้จริง β_j ไม่เท่ากับ 0 (สมมติว่าค่าในเซต s_i เรียงลำดับจากน้อยไปมาก) ศึกษาภายใต้จำนวนสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 ดังนี้

กรณีที่ p=1,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 6$

กรณีที่ p=2,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 10$

กรณีที่ p=4,000 ศึกษาภายใต้เงื่อนไข $|s_i| = 20$

5. ศึกษาภายใต้ค่าความแปรปรวนของค่าความคลาดเคลื่อน σ^2 โดยที่กำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio (SNR)) = 2 , 8

$$SNR = \sqrt{\frac{\beta^T X^T X \beta}{n\sigma^2}}$$

6. ศึกษาภายใต้ตัวแบบเชิงเส้น : ความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามเป็นแบบเชิงเส้น

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i \quad ; \varepsilon \overset{iid}{\sim} N\left(0, \sigma^2 I_n\right)$$

7. ศึกษาภายใต้การจำลองข้อมูลในแต่กรณีจำนวน 100 ครั้ง

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง

2. กำหนดการจำลองข้อมูล

    2.1 กำหนดค่าเริ่มต้นโดยการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และจำนวนพารามิเตอร์ p ตัว โดยใช้อัตราส่วน n:p ดังนี้ 200:1000 , 200:2000 และ 200:4000

    2.2 กำหนดให้จำนวนสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 ดังนี้ 6 , 10 และ 20

    2.3 กำหนดให้ค่าสัมประสิทธิ์ β_j ที่ไม่เท่ากับ 0 มีค่าเป็น 1 กับ -1 ด้วยความน่าจะเป็นที่เท่ากัน

    2.4 จำลองข้อมูลภายใต้รูปแบบความสัมพันธ์แบบเชิงเส้น

3. จำลองข้อมูลจากค่าเริ่มต้นที่กำหนด

    3.1 จำลองตัวแปรอิสระ X ให้มีการแจกแจงแบบปกติ $X_i \sim N(0,\Sigma)$

โดยที่ $\Sigma = \begin{pmatrix} \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \cdots & 0 & 0 \\ \vdots & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} & \vdots & \vdots \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & 0 & \overbrace{\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}}^{100} \end{pmatrix}_{p\times p}$

    3.2 จำลองค่าความคลาดเคลื่อน ε ให้มีการแจกแจงแบบปกติ $\varepsilon \overset{iid}{\sim} N(0,\sigma^2 I_n)$

    3.3 จำลองข้อมูลตัวแปรตามภายใต้รูปแบบความสัมพันธ์เชิงเส้น

4. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธีการดังต่อไปนี้ ในขั้นตอนการคัดกรองตัวแปรอิสระ

4.1 วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น

4.2 วิธีลาสโซ

4.3 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ

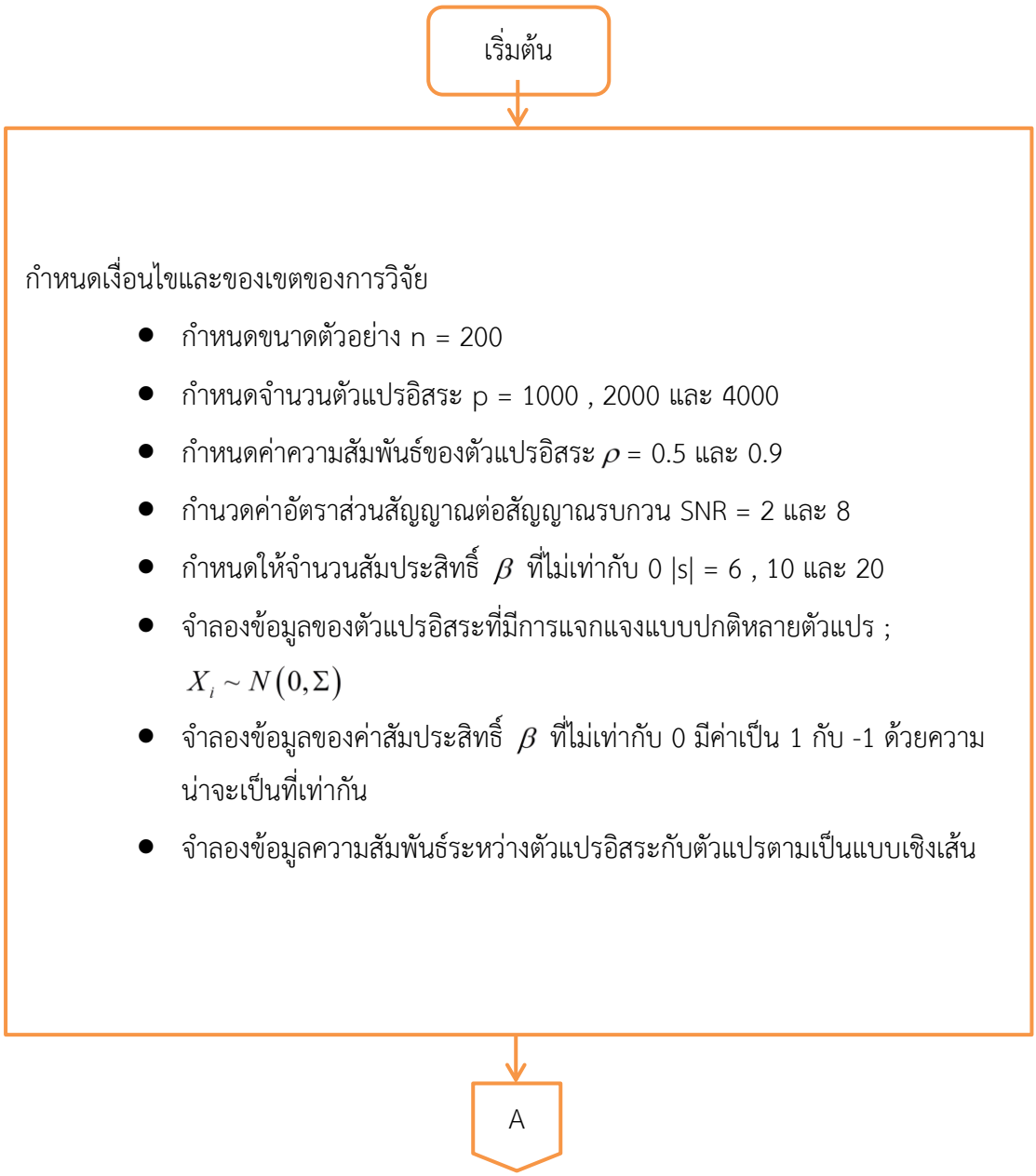
4.4 วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง

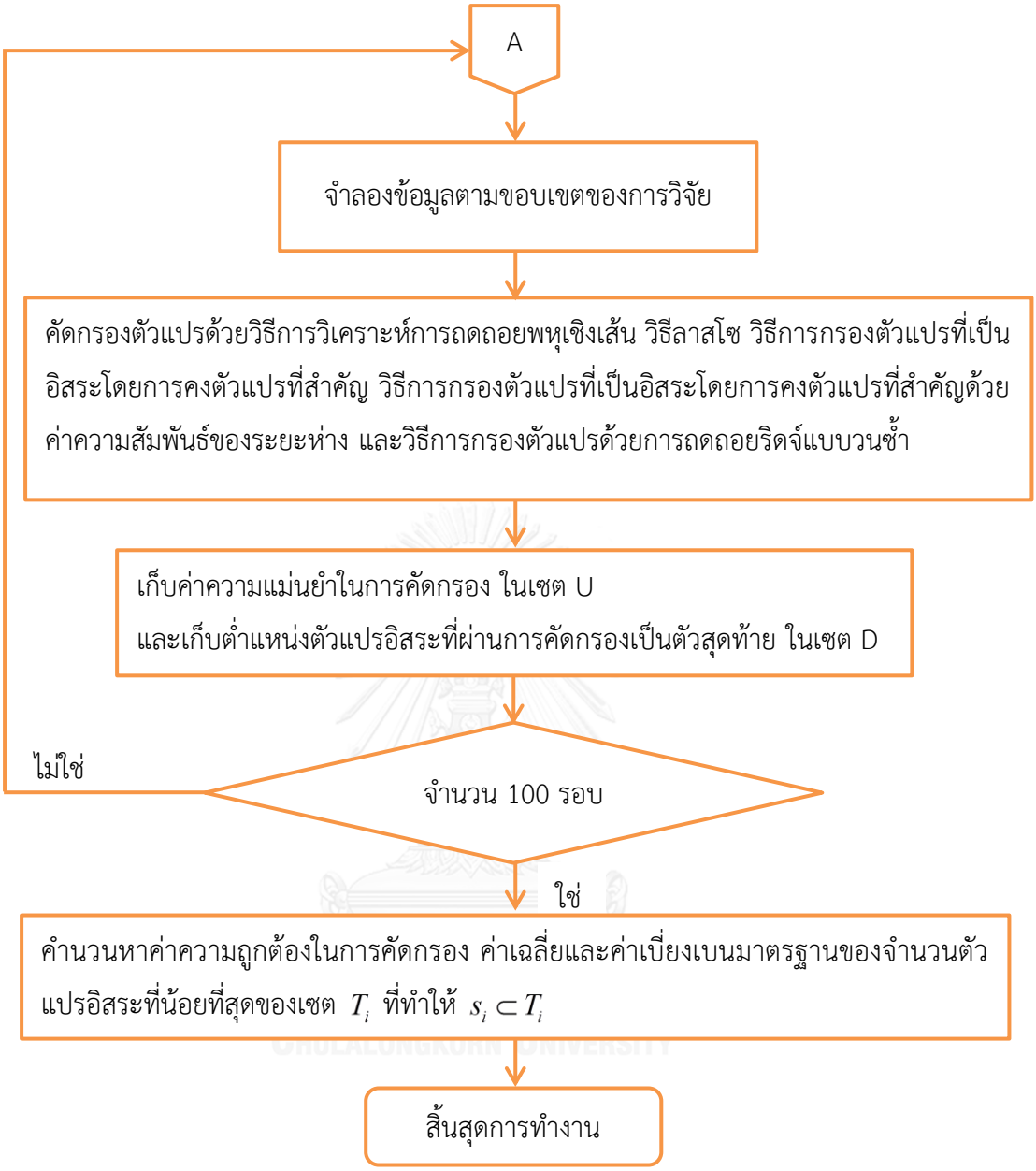
4.5 วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ

5. นำผลที่ได้จากข้อที่ 4 มาหาค่าความถูกต้องในการคัดกรองตัวแปร ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

6. วิเคราะห์และสรุปผลจากทั้ง 5 วิธี

3.3 ขั้นตอนการทำงานของโปรแกรม




A
จำลองข้อมูลตามขอบเขตของการวิจัย
คัดกรองตัวแปรด้วยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ
เก็บค่าความแม่นยำในการคัดกรอง ในเซต U
และเก็บตำแหน่งตัวแปรอิสระที่ผ่านการคัดกรองเป็นตัวสุดท้าย ในเซต D
ไม่ใช่
จำนวน 100 รอบ
ใช่
คำนวนหาค่าความถูกต้องในการคัดกรอง ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$
สิ้นสุดการทำงาน

บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบวิธีในการคัดกรองตัวแปรอิสระจากข้อมูลที่มีมิติสูง ประกอบไปด้วยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ โดยการจำลองข้อมูลซึ่งพิจารณาแยกตามขนาดตัวแปรอิสระ p = 1000 , 2000 และ 4000 รวมถึงค่าความสัมพันธ์ภายในตัวแปรอิสระและค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน SNR เกณฑ์ที่ใช้ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าความถูกต้องในการคัดกรองตัวแปรซึ่งเป็นการดูว่าตัวแปรอิสระที่ผ่านการกรองมานั้นเป็นตัวแปรอิสระที่แท้จริง ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ซึ่งเป็นการดูถึงความสัมพันธ์ของตัวแปรอิสระกับตัวแปรตามที่ผ่านการคัดกรอง

อักษรย่อและสัญลักษณ์ต่างๆ ที่ปรากฏในการำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆ แทนความหมายดังนี้

n แทน จำนวนขนาดตัวอย่าง

p แทน จำนวนขนาดตัวอย่าง

r แทน ค่าความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

s_i แทน เซตของดัชนีของสัมประสิทธิ์การถดถอยที่แท้จริง

T_i แทน เซตของดัชนีของตัวแปรอิสระที่ผ่านการคัดกรองตัวแปร

U แทน ค่าความแม่นยำในการคัดกรองตัวแปร

M แทน ค่าเบี่ยงเบนมาตรฐานของตำแหน่งตัวแปรอิสระที่ผ่านการคัดกรองเป็นตัวสุดท้าย

S.D. แทน ค่าเบี่ยงเบนมาตรฐานของตำแหน่งตัวแปรอิสระที่ผ่านการคัดกรองเป็นตัวสุดท้าย

UR แทน วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น

LASSO แทน วิธีลาสโซ

SIS แทน วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ

DC-SIS แทน วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง

ITRRS แทน วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 2 ส่วน คือ ในส่วนที่ 1 จะเปรียบเทียบความถูกต้องในการคัดกรองตัวแปรจากทั้ง 5 วิธี ส่วนที่ 2 จะเปรียบเทียบค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

โดยผลการวิจัยจะแบ่งออกเป็น 2 ส่วนดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร จากการคัดกรองตัวแปร ทั้ง 5 วิธี

เมื่อพิจารณาในกรณี

1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ n:p ที่ 200:1000 , 200:2000 และ200:4000

1.2 เมื่อกำหนดให้ความสัมพันธ์ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.5$ และ $\rho = 0.9$

1.3 เมื่อกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) SNR = 2 , 8

<u>ส่วนที่ 2</u> ผลเปรียบเทียบค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ จากการคัดกรองตัวแปร ทั้ง 5 วิธี

เมื่อพิจารณาในกรณี

1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ n:p ที่ 200:1000 , 200:2000 และ 200:4000

1.2 เมื่อกำหนดให้ความสัมพันธ์ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.5$ และ $\rho = 0.9$

1.3 เมื่อกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio)

SNR = 2 , 8

4.1 ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร จากการคัดกรองตัวแปร ทั้ง 5 วิธี

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิ์ภาพในการคัดกรองตัวแปรจากวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ ภายใต้ปัจจัย ดังต่อไปนี้

1. เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ n:p ที่ 200:1000 , 200:2000 และ 200:4000

2. เมื่อกำหนดให้ความสัมพันธ์ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.5$ และ $\rho = 0.9$

3. เมื่อกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio)

SNR = 2 , 8

โดยแสดงผลในตารางที่ 4.1 โดยแต่ละตารางมีลายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
U	4.1	- n : p - ความสัมพันธ์ของตัวแปรอิสระ - อัตราส่วนสัญญาณต่อสัญญาณรบกวน	1. UR 2. LASSO 3. SIS 4. DC-SIS 5. ITRRS

ตารางที่ 4.1 แสดงค่าความถูกต้องในการคัดกรองตัวแปรของแต่ละสถานการณ์ จากข้อมูลจำลอง 100 แบบจำลอง สำหรับตัวแบบเป็นแบบเชิงเส้น

วิธีการคัดกรองตัวแปร	ความแม่นยำในการคัดกรองตัวแปร (U)											
	ความสัมพันธ์ของตัวแปรอิสระ											
	SNR = 2						SNR = 8					
	$\rho = 0.5$			$\rho = 0.9$			$\rho = 0.5$			$\rho = 0.9$		
	200 : 1000	200 : 2000	200 : 4000	200 : 1000	200 : 2000	200 : 4000	200 : 1000	200 : 2000	200 : 4000	200 : 1000	200 : 2000	200 : 4000
UR	51	6	0	2	0	0	51	7	0	3	0	0
LASSO	100	100	2	73	9	0	100	100	100	100	100	64
SIS(d1)	13	0	0	0	0	0	19	0	0	0	0	0
SIS(d2)	21	0	0	0	0	0	24	2	0	0	0	0
SIS(d3)	31	4	0	0	0	0	36	2	0	0	0	0
DC-SIS(d1)	13	0	0	0	0	0	19	0	0	0	0	0
DC-SIS(d2)	21	0	0	0	0	0	24	2	0	0	0	0
DC-SIS(d3)	31	4	0	0	0	0	36	2	0	0	0	0
ITRRS	3	0	0	0	0	0	3	0	0	0	0	0

<u>หมายเหตุ</u> วิธี SIS และวิธี DC-SIS จะต้องกำหนดจำนวนตัวแปรอิสระที่ผ่านการกรองซึ่งจะต้องมีจำนวนน้อยว่าจำนวนขนาดตัวอย่าง (d<n) ในที่นี้เรากำหนดให้ $d1 = \frac{n}{\ln(n)}$, $d2 = 2 \times \left(\frac{n}{\ln(n)} \right)$ และ $d3 = 3 \times \left(\frac{n}{\ln(n)} \right)$

จากตารางที่ 4.1 ซึ่งแสดงผลของค่าความถูกต้องในการคัดกรองตัวแปร โดยนับจากข้อมูลจำลอง 100 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ พบว่า

1. ที่จำนวนตัวแปรอิสระ p เท่ากับ 1000

1.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่สามารถคัดกรองตัวแปรได้ถูกต้องมากที่สุด โดยสามารถคัดกรองข้อมูลได้ถูกต้องครบทั้ง 100 แบบจำลอง รองลงมาจะเป็น วิธี UR ซึ่งสามารถคัดกรองได้ถูกต้อง 51 แบบจำลอง โดยที่วิธี SIS และ วิธี DC-SIS เป็นวิธีที่มีความสามารถในการคัดกรองที่เท่ากัน และวิธี ITRRS สามารถคัดกรองข้อมูลได้น้อยที่สุดเพียง 3 แบบจำลอง จากค่า SNR จะพบว่าเมื่อค่า SNR เพิ่มขึ้นจะส่งผลให้ค่าความแม่นยำมีจำนวนเพิ่มขึ้นเล็กน้อย

1.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่สามารถคัดกรองตัวแปรได้ถูกต้องมากที่สุด โดยสามารถคัดกรองข้อมูลได้ถูกต้อง 73 แบบจำลอง สำหรับค่า SNR = 2 และสามารถคัดกรองได้ถูกต้องมากขึ้นเป็น 100 แบบจำลอง สำหรับค่าSNR = 8 ส่วนวิธี UR เป็นอีกวิธีที่สามารถคัดกรองได้โดยสามารถคัดกรองได้เพียง 2 และ 3 แบบจำลอง สำหรับค่า SNR = 2 และ 8 ตามลำดับ ส่วนวิธีอื่น ๆ นั้นไม่สามารถคัดกรองตัวแปรอิสระออกมาได้ถูกต้องเลย

2. ที่จำนวนตัวแปรอิสระ p เท่ากับ 2000

2.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่สามารถคัดกรองตัวแปรได้ถูกต้องมากที่สุด โดยสามารถคัดกรองข้อมูลได้ถูกต้องครบทั้ง 100 แบบจำลอง ส่วนวิธี UR สามารถคัดกรองได้โดยสามารถคัดกรองได้เพียง 6 และ 7 แบบจำลอง สำหรับค่า SNR = 2 และ 8 ตามลำดับ โดยที่วิธี SIS กับวิธี DC-SIS นั้น

สามารถคัดกรองได้ในกรณีที่ $d3 = 3\times\left(\frac{n}{\log(n)}\right)$ ที่ 4 แบบจำลอง และกรณี $d2 = 2\times\left(\frac{n}{\log(n)}\right)$ กับ $d3 = 3\times\left(\frac{n}{\log(n)}\right)$ ที่ 2 แบบจำลอง สำหรับค่า SNR = 2 และ 8 ตามลำดับ ส่วนวิธี ITRRS นั้นไม่สามารถคัดกรองตัวแปรอิสระออกมาได้ถูกต้องเลย

2.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถคัดกรองตัวแปรได้ถูกต้อง โดยสามารถคัดกรองข้อมูลได้ถูกต้อง 9 แบบจำลอง สำหรับค่า SNR = 2 และสามารถคัดกรองได้ถูกต้องมากขึ้นเป็น 100 แบบจำลอง สำหรับค่า SNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถคัดกรองตัวแปรอิสระออกมาได้ถูกต้องเลย

3. ที่จำนวนตัวแปรอิสระ p เท่ากับ 4000

3.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถคัดกรองตัวแปรได้ถูกต้อง โดยสามารถคัดกรองข้อมูลได้ถูกต้อง 2 แบบจำลอง สำหรับค่า SNR = 2 และสามารถคัดกรองได้ถูกต้องมากขึ้นเป็น 100 แบบจำลอง สำหรับค่า SNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถคัดกรองตัวแปรอิสระออกมาได้ถูกต้องเลย

3.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถคัดกรองตัวแปรได้ถูกต้อง โดยสามารถคัดกรองได้ 64 แบบจำลอง สำหรับค่า SNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถคัดกรองตัวแปรอิสระออกมาได้ถูกต้องเลย

4.2 ผลเปรียบเทียบค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ทั้ง 5 วิธี

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบประสิทธิ์ภาพในการคัดกรองตัวแปรจากวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ ภายใต้ปัจจัย ดังต่อไปนี้

1. เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ n:p ที่ 200:1000 , 200:2000 และ 200:4000

2. เมื่อกำหนดให้ความสัมพันธ์ของตัวแปรอิสระ 2 ระดับ คือ $\rho = 0.5$ และ $\rho = 0.9$

3. เมื่อกำหนดให้อัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) SNR = 2 , 8

โดยแสดงผลในตารางที่ 4.2 โดยแต่ละตารางมีลายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
M S.D	4.2	- n : p - ความสัมพันธ์ของตัวแปรอิสระ - อัตราส่วนสัญญาณต่อสัญญาณรบกวน	1. UR 2. LASSO 3. SIS 4. DC-SIS 5. ITRRS

ตารางที่ 4.2 แสดงค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ของแต่ละสถานการณ์ จากข้อมูลจำลอง 100 แบบจำลอง สำหรับตัวแบบเป็นแบบเชิงเส้น

วิธีการคัดกรองตัวแปร	ความสัมพันธ์ของตัวแปรอิสระ											
	SNR = 2											
	$\rho = 0.5$						$\rho = 0.9$					
	s = 6		s = 10		s = 20		s = 6		s = 10		s = 20	
	200 : 1000		200 : 2000		200 : 4000		200 : 1000		200 : 2000		200 : 4000	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
UR	94.2	60.93	112.5	45.79	-	-	151.5	17.68	-	-	-	-
LASSO	6	0	12.17	5.96	77	15.56	13.68	12.85	51.22	32	-	-
SIS(d1)	19.15	8.47	-	-	-	-	-	-	-	-	-	-
SIS(d2)	31.86	18.76	-	-	-	-	-	-	-	-	-	-
SIS(d3)	53.32	35.78	83.5	6.56	-	-	-	-	-	-	-	-
DC-SIS(d1)	19.15	8.47	-	-	-	-	-	-	-	-	-	-
DC-SIS(d2)	31.86	18.76	-	-	-	-	-	-	-	-	-	-
DC-SIS(d3)	53.32	35.78	83.5	6.56	-	-	-	-	-	-	-	-
ITRRS	168	0	-	-	-	-	-	-	-	-	-	-

วิธีการคัดกรองตัวแปร	ความสัมพันธ์ของตัวแปรอิสระ											
	SNR = 8											
	$\rho = 0.5$						$\rho = 0.9$					
	s = 6		s = 10		s = 20		s = 6		s = 10		s = 20	
	200 : 1000		200 : 2000		200 : 4000		200 : 1000		200 : 2000		200 : 4000	
	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.	M	S.D.
UR	80.92	61.91	139.57	57.92	-	-	184.67	13.61	-	-	-	-
LASSO	6	0	10	0	20	0	6	0	10	0	41.23	32.6
SIS(d1)	17.47	9.74	-	-	-	-	-	-	-	-	-	-
SIS(d2)	24.21	16.48	61	19.8	-	-	-	-	-	-	-	-
SIS(d3)	48.47	37.94	61	19.8	-	-	-	-	-	-	-	-
DC-SIS(d1)	17.47	9.74	-	-	-	-	-	-	-	-	-	-
DC-SIS(d2)	24.21	16.48	61	19.8	-	-	-	-	-	-	-	-
DC-SIS(d3)	48.47	37.94	61	19.8	-	-	-	-	-	-	-	-
ITRRS	168	0	-	-	-	-	-	-	-	-	-	-

<u>หมายเหตุ</u> วิธี SIS และวิธี DC-SIS จะต้องกำหนดจำนวนตัวแปรอิสระที่ผ่านการกรองซึ่งจะต้องมีจำนวนน้อยว่าจำนวนขนาดตัวอย่าง (d<n) ในที่นี้เรากำหนดให้ $d1 = \frac{n}{\ln(n)}$, $d2 = 2 \times \left(\frac{n}{\ln(n)} \right)$ และ $d3 = 3 \times \left(\frac{n}{\ln(n)} \right)$

จากตารางที่ 4.2. ซึ่งแสดงผลของค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ โดยนับจากข้อมูลจำลอง 100 ชุดข้อมูล ระหว่างการคัดกรองตัวแปรด้วยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ พบว่า

1. ที่จำนวนตัวแปรอิสระ p เท่ากับ 1000

1.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่ให้ค่าเฉลี่ยน้อยที่สุดโดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น 6 และค่าเบี่ยงเบนมาตรฐานเป็น 0 ต่อมาจะเป็นวิธี SIS และ วิธี DC-SIS ที่ให้ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานที่เท่ากัน ต่อมาจะเป็นวิธี UR ซึ่งให้ค่าเฉลี่ยที่สูงขึ้นเป็น 80.92 และค่าเบี่ยงเบนมาตรเป็น 61.91 ส่วนวิธี ITRRS ให้ค่าเฉลี่ยที่ 168 และค่าเบี่ยงเบนมาตรฐานเป็น 0 โดยที่เมื่อค่า SNR เพิ่มขึ้นจะส่งผลให้ค่าเฉลี่ยมีค่าลดลงแต่ค่าเบี่ยงเบนมาตรฐานนั้นไม่เห็นการเปลี่ยนแปลงอย่างชัดเจน

1.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่ให้ค่าเฉลี่ยน้อยที่สุดโดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น13.68 และค่าเบี่ยงเบนมาตรฐานเป็น 12.85 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น 6 และค่าเบี่ยงเบนมาตรฐานเป็น 0 สำหรับค่าSNR = 8 ส่วนวิธี UR เป็นอีกวิธีที่สามารถหาได้โดยมีค่าเฉลี่ยเป็น 151.5 และค่าเบี่ยงเบนมาตรฐานเป็น 17.68 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น 184.67 และค่าเบี่ยงเบนมาตรฐานเป็น 13.61 สำหรับค่าSNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้

2. ที่จำนวนตัวแปรอิสระ p เท่ากับ 2000

2.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีที่ให้ค่าเฉลี่ยน้อยที่สุดโดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น 12.17 และค่าเบี่ยงเบนมาตรฐานเป็น 5.96 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น 10 และค่าเบี่ยงเบนมาตรฐานเป็น 0 สำหรับค่าSNR = 8 ต่อมาเป็นวิธี SIS และ วิธี DC-SIS ที่ให้ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานที่เท่ากันโดยสามารถหาได้ในกรณีที่ $d3 = 3\times\left(\frac{n}{\log(n)}\right)$ ซึ่งมีค่าเฉลี่ยเป็น 83.5 และค่าเบี่ยงเบนมาตรฐานเป็น 6.56 สำหรับค่า SNR = 2 และกรณี $d2 = 2\times\left(\frac{n}{\log(n)}\right)$ กับ $d3 = 3\times\left(\frac{n}{\log(n)}\right)$ มีค่าเฉลี่ยเป็น 61 และค่าเบี่ยงเบนมาตรฐานเป็น 19.8 สำหรับค่า SNR = 8 ส่วนวิธี UR เป็นอีกวิธีที่สามารถหาได้โดยมีค่าเฉลี่ยเป็น 112.5 และค่าเบี่ยงเบนมาตรฐานเป็น 45.79 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น 139.57 และค่าเบี่ยงเบนมาตรฐานเป็น 57.92 สำหรับค่าSNR = 8 ส่วนวิธี ITRRS นั้นไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้

2.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถให้ค่าเฉลี่ยได้โดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น 51.22 และค่าเบี่ยงเบนมาตรฐานเป็น 32 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น10 และค่าเบี่ยงเบนมาตรฐานเป็น 0 สำหรับค่าSNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้

3. ที่จำนวนตัวแปรอิสระ p เท่ากับ 4000

3.1 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถให้ค่าเฉลี่ยได้โดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น 77 และค่าเบี่ยงเบนมาตรฐานเป็น 15.56 สำหรับค่า SNR = 2 และมีค่าเฉลี่ยเป็น 20 และค่าเบี่ยงเบนมาตรฐานเป็น 0 สำหรับค่าSNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้

3.2 เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.9$

จากการคัดกรองตัวแปรทั้ง 5 วิธีข้างต้น พบว่าวิธี LASSO เป็นวิธีเดียวที่สามารถให้ค่าเฉลี่ยได้โดยที่ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ มีค่าเฉลี่ยเป็น 41.23 และค่าเบี่ยงเบนมาตรฐานเป็น 32.6 สำหรับค่า SNR = 8 ส่วนวิธีอื่น ๆ นั้นไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรระหว่างวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ โดยพิจารณาตามจำนวนตัวแปรอิสระเป็น 1000 , 2000 และ 4000 รวมถึงความสัมพันธ์ภายในตัวแปรอิสระเป็น 0.5 และ 0.9 โดยมีเกณฑ์ที่ใช้ในการพิจารณาประสิทธิภาพของแต่ละวิธีจากค่าความถูกต้องในการคัดกรองตัวแปรค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ โดยสรุปผลวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 แบ่งผลการวิจัยออกเป็น 2 ส่วน โดยพิจารณาตามขนาดตัวอย่าง ดังนี้

ส่วนที่ 1 ผลการเปรียบเทียบความถูกต้องในการคัดกรองตัวแปร จากการคัดกรองตัวแปร ทั้ง 5 วิธี

ตารางที่ 5.1.1 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุดสำหรับตัวแบบเชิงเส้น เมื่อพิจารณาความถูกต้องในการคัดกรองตัวแปร โดยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 200 โดยจำแนกตามอัตราส่วนขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) , ความสัมพันธ์ของตัวแปรอิสระ และ อัตราส่วนสัญญาณต่อสัญญาณรบกวน

n:p	SNR = 2		SNR = 8	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
200:1000	LASSO	LASSO	LASSO	LASSO
200:2000	LASSO	LASSO	LASSO	LASSO
200:4000	LASSO	-	LASSO	LASSO

จากตารางที่ 5.1.1 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่างเท่ากับ 200 พิจารณาจากความถูกต้องในการคัดกรองตัวแปร โดยค่าความถูกต้องในการคัดกรองตัวแปร จะได้วิธีที่เหมาะสมในการคัดกรองตัว

แปรคล้ายกัน นั้นคือการคัดกรองตัวแปรด้วยวิธี LASSO จะเหมาะสมมากที่สุด โดยสามารถทำงานได้ดีในทุกระดับความสัมพันธ์ของตัวแปรอิสระและทุกประเภทของข้อมูล ยกเว้นกรณีที่จำนวนตัวแปรอิสระเป็น 4000 ค่า SNR = 2 และ $\rho = 0.9$ จะไม่มีวิธีที่ดีที่สุด

ส่วนที่ 2 ผลเปรียบเทียบค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ทั้ง 5 วิธี

ตารางที่ 5.1.2 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุดสำหรับตัวแบบเชิงเส้น เมื่อพิจารณาจากค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ โดยวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 200 โดยจำแนกตามอัตราส่วนขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) , ความสัมพันธ์ของตัวแปรอิสระ และ อัตราส่วนสัญญาณต่อสัญญาณรบกวน

n:p	SNR = 2		SNR = 8	
	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$	$\rho = 0.9$
200:1000	LASSO	LASSO	LASSO	LASSO
200:2000	LASSO	LASSO	LASSO	LASSO
200:4000	LASSO	-	LASSO	LASSO

จากตารางที่ 5.1.2 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่างเท่ากับ 200 พิจารณาจากความถูกต้องในการคัดกรองตัวแปร โดยค่าความถูกต้องในการคัดกรองตัวแปร จะได้วิธีที่เหมาะสมในการคัดกรองตัวแปรคล้ายกัน นั้นคือการคัดกรองตัวแปรด้วยวิธี LASSO จะเหมาะสมมากที่สุด โดยสามารถทำงานได้ดีในทุกระดับความสัมพันธ์ของตัวแปรอิสระและทุกประเภทของข้อมูล ยกเว้นกรณีที่จำนวนตัวแปรอิสระเป็น 4000 ค่า SNR = 2 และ $\rho = 0.9$ จะไม่มีวิธีที่ดีที่สุด

5.1.2 ผลจากความแตกต่างระหว่างจำนวนตัวแปรอิสระ

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าความถูกต้องในการคัดกรองตัวแปร (U) ของวิธีการคัดกรองทั้ง 5 วิธี วิธี LASSO เป็นวิธีที่สามารถคัดกรองตัวแปรอิสระได้ถูกต้องมากที่สุดแต่จะคัดกรองได้ถูกต้องน้อยลงเมื่อจำนวนตัวแปรอิสระเพิ่งขึ้น โดยที่วิธี UR เป็นวิธีที่มีความสามารถรองลงมาแต่จะทำได้บ้างในกรณีที่มีจำนวนตัวแปรอิสระเท่ากับ 1000 ต่อมาเป็นวิธี SIS และวิธี DC-SIS นั้นสามารถคัดกรองได้เหมือนกันโดยจะสามารถคัดกรองได้บ้างในกรณีที่มีจำนวนตัวแปรอิสระเท่ากับ 1000 และวิธีสุดท้ายคือวิธี ITRRS เป็นวิธีที่คัดกรองได้ถูกต้องน้อยที่สุดเพียง 3 ครั้ง จากการจำลองข้อมูลทั้งหมด 100 ครั้ง ซึ่งจะสามารถคัดกรองได้บ้างในกรณีที่มีจำนวนตัวแปรอิสระเท่ากับ 1000 เท่านั้น โดยที่ทั้ง 4 วิธี คือ วิธี UR วิธี SIS วิธี DC-SIS และวิธี ITRRS จะไม่สามารถคัดกรองตัวแปรอิสระได้ถูกต้อง เมื่อจำนวนตัวแปรอิสระมีมากขึ้นรวมถึงค่าความสัมพันธ์ของตัวแปรอิสระด้วย และประสิทธิภาพในการหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ นั้น วิธี LASSO เป็นวิธีที่ให้ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานต่ำที่สุด โดยที่วิธี SIS และ วิธี DC-SIS เป็นวิธีที่มีความสามารถรองลงมาแต่จะทำได้ดีในกรณีที่มีจำนวนตัวแปรอิสระเท่ากับ 1000 ต่อมาเป็นวิธี UR ที่จะให้ค่าเฉลี่ยสูงขึ้น และวิธี ITRRS จะให้ค่าเฉลี่ยที่สูงที่สุดเป็น 168 แต่ค่าเบี่ยงเบนมาตรฐานเป็น 0 โดยจะมีส่วนที่ไม่สามารถหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานได้ เนื่องจากในการหาค่าความถูกต้องนั้นไม่สามารถหาได้

5.1.3 ผลจากความแตกต่างระหว่างความสัมพันธ์ภายในตัวแปรอิสระ

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าความถูกต้องในการคัดกรองตัวแปร (U) ของวิธีการคัดกรองทั้ง 5 วิธี เมื่อขนาดของความสัมพันธ์เป็น 0.5 จะสามารถคัดกรองได้ถูกต้องมากกว่าเมื่อขนาดของความสัมพันธ์เป็น 0.9 และประสิทธิภาพในการหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ของวิธีการคัดกรองทั้ง 5 วิธี เมื่อขนาดของความสัมพันธ์เป็น 0.5 จะให้ค่าเฉลี่ยน้อยกว่า เมื่อขนาดของความสัมพันธ์เป็น 0.9

5.1.4 ผลจากความแตกต่างระหว่างค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวน

จากผลที่ได้จะพบว่าประสิทธิภาพในการหาค่าความถูกต้องในการคัดกรองตัวแปร (U) ของวิธีการคัดกรองทั้ง 5 วิธี เมื่อค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวนเป็น 8 จะสามารถคัดกรองได้ถูกต้องมากกว่า เมื่อค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวนเป็น 2 และประสิทธิภาพในการหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$

ของวิธีการคัดกรองทั้ง 5 วิธี เมื่อขนาดของความสัมพันธ์เป็น 8 จะให้ค่าเฉลี่ยน้อยกว่า เมื่อขนาดของความสัมพันธ์เป็น 2

5.2 สรุปผลโดยรวม

จากผลการวิจัยในการเปรียบเทียบวิธีการคัดกรองตัวแปรอิสระจากวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีลาสโซ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ โดยใช้ข้อมูลจากการจำลองผลปรากฏว่าการคัดกรองตัวแปรด้วยวิธีลาสโซ สามารถคัดกรองตัวแปรได้มีประสิทธิภาพมากที่สุด รองลงมาคือวิธีการวิเคราะห์การถดถอยพหุเชิงเส้น วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญกับวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่างมีความสามารถเท่าเทียมกัน และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำเป็นวิธีที่มีความถูกต้องน้อยที่สุด ซึ่งเมื่อพิจารณาจากค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของจำนวนตัวแปรอิสระที่น้อยที่สุดของเซต T_i ที่ทำให้ $s_i \subset T_i$ ก็แสดงผลออกมาในทำนองเดียวกันคือ วิธีลาสโซ เป็นวิธีที่มีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานน้อยที่สุด โดยวิธีต่อมาคือวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญกับวิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่างมีความสามารถเท่าเทียมกัน ซึ่งจะแตกต่างจากค่าความถูกต้องที่วิธีการวิเคราะห์การถดถอยพหุเชิงจะเป็นลำดับที่ 2 โดยสำหรับค่าเฉลี่ยนั้นวิธีการวิเคราะห์การถดถอยพหุเชิงจะตามมาเป็นลำดับที่ 3 และวิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำเป็นวิธีที่มีความแม่นยำน้อยที่สุด

ดังนั้นจึงสามารถสรุปไว้ว่าวิธีการในการคัดกรองตัวแปรอิสระที่ดีที่สุดคือ วิธีลาสโซ อย่างไรก็ตามในการนำไปใช้งานจริงนั้นผู้นำไปใช้ควรพิจารณาลักษณะของข้อมูลต่างๆ ด้วย เนื่องจากผลสรุปที่ได้เป็นผลจากภายใต้ขอบเขตการวิจัยที่ทำการศึกษา

5.3 ข้อเสนอแนะ

จากงานวิจัยนี้ผู้สนใจอาจจะนำไปศึกษาต่อได้อีกในเรื่องของ

1. วิธีการคัดกรองตัวแปร ในงานวิจัยนี้เลือกมาศึกษาทั้งหมด 5 วิธี ซึ่งในความเป็นจริงยังมีวิธีอื่นๆ ที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการคัดกรองตัวแปรอื่นๆ มาร่วนพิจารณา

2. ขอบเขตในการวิจัย ในเรื่องของลักษณะข้อมูล , ขนาดตัวอย่าง , จำนวนตัวแปรอิสระ , ความสัมพันธ์ของตัวแปรอิสระ (Correlation) อาจจะมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้น

3. กรณีที่ Y ไม่มีการแจกแจงแบบปกติ

4. กรณีที่รูปแบบความสัมพันธ์ของตัวแปรอิสระเปลี่ยนไป

รายการอ้างอิง

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

Bühlmann, P., & Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics, 29*(3-4), 407-430.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(5), 849-911.

Li, R., Zhong, W., & Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association, 107*(499), 1129-1139.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(3), 273-282.

วิฐรา พึ่งพาพงศ์. (2015). บท วิเคราะห์ วิธี วิเคราะห์ การ ถดถอย เชิง เส้น สำหรับ ข้อมูล ที่ มี มิติ สูง. วารสาร วิทยาศาสตร์ และ เทคโนโลยี, *23*(2), 212-223.

สุพล ดุรงค์วัฒนา. (2015). Regresion Model: Analytics-based Approach. กรุงเทพ:แดเน็ตซ์ อินเตอร์คอร์ปอเรชั่น.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่าง กรณีที่มีขนาดตัวอย่างเท่ากับ 200 และจำนวนตัวแปรอิสระเท่ากับ 1000 เมื่อจำนวนสัมประสิทธิ์การถดถอยที่แท้จริง β_j ไม่เท่ากับ 0 เป็น 6 ที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0 และค่าอัตราส่วนสัญญาณต่อสัญญาณรบกวนเป็น 8 ซึ่งดำเนินการภายใต้ตัวแบบเชิงเส้น โดยมีการคัดกรองตัวแปรด้วยวิธี

- วิธีการวิเคราะห์การถดถอยพหุเชิงเส้น (UR)
- วิธีลาสโซ (LASSO)
- วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญ (SIS)
- วิธีการกรองตัวแปรที่เป็นอิสระโดยการคงตัวแปรที่สำคัญด้วยค่าความสัมพันธ์ของระยะห่าง (DC-SIS)
- วิธีการกรองตัวแปรด้วยการถดถอยริดจ์แบบวนซ้ำ (ITRRS)

```
library(mvtnorm)
library(glmnet)

n<-200;
p<-1000;
rho<-0.5;
s<-6;
SNR<-2;

d_1<-floor(n/log(n))
d_2<-floor(2*(n/log(n)))
d_3<-floor(3*(n/log(n)))

U_regre<-c()
U_lasso<-c()
U_sis<-c()
```

```
U_sis2<-c()
U_sis3<-c()
U_dcsis<-c()
U_dcsis2<-c()
U_dcsis3<-c()
U_itrrs<-c()
D_regre<-c()
D_lasso<-c()
D_sis<-c()
D_sis2<-c()
D_sis3<-c()
D_dcsis<-c()
D_dcsis2<-c()
D_dcsis3<-c()
D_itrrs<-c()

########## Mean and Sigma for X ##########

mean_X<-matrix(c(numeric(p)),nrow=p,ncol=1)
Sigma_X<-diag(p)
for(a in 1:(p/100)){
  Sigma_X[(a*100-99):(a*100),(a*100-99):(a*100)]<-rho
}
diag(Sigma_X)<-1

########## Simulation X and Y 100 Times ##########

for (b in 1:100){
```

```
########## Simulation X ##########

 X<-rmvnorm(n,mean_X,Sigma_X)

########## simulation Beta ##########

 Beta<-matrix(0,nrow=p,ncol=1)
 BetaValue<-c(-1,1)
 position_X<-sample(1:p,s)
 S<-sort(position_X,decreasing = FALSE)
 value<-sample(BetaValue,s,replace=TRUE)
  for(d in 1:s){
    Beta[S[d],]<-value[d]
  }

########## error ##########

 sigma_squared<-(t(Beta)%*%t(X)%*%X%*%Beta)/((SNR^2)*n)
 error<-matrix(rnorm(n,0,sqrt(sigma_squared)),nrow=n,ncol=1)

########## Y linear ##########

 Y<-X%*%Beta+error

                    ########## Variable Screening ##########
  ########## Univariate Regression (UR) ##########

  pvalue<-rep(NA,p)
```

```
 betaregre<-c()
 T_regre<-c()
 for (e in 1:p){
   model<-lm(Y~X[,e])
   summarymodel<-summary(model)
   pvalue[e]<-1-pf(summarymodel$fstatistic[1], summarymodel$fstatistic[2],
summarymodel$fstatistic[3])
 }

tmp_re<-order(pvalue)
 tmp2_re<-pvalue[tmp_re]<=(1:p)*0.05/p
 tmp3_re<-tmp_re[tmp2_re==TRUE]



 if (length(tmp3_re)>200){
   T_regre<-tmp3_re[c(1:200)]}else{
     T_regre<-tmp3_re
   }
u_regre<-0
 for(a_regre in 1:s){
   for(b_regre in 1:length(T_regre)){
     if(S[a_regre]==T_regre[b_regre]){
       u_regre<-u_regre+1}}
 }
 if(u_regre/s==1){
   U_re<-1}else{
     U_re<-0}
 U_regre<-c(U_regre,U_re)
```

```
if (U_re==1){
  d_regre<-c()
  for(a_regre2 in 1:s){
    for(b_regre2 in 1:length(T_regre)){
      if(S[a_regre2]==T_regre[b_regre2]){
        d_regre<-c(d_regre,b_regre2)}}

    D_re<-max(d_regre)}}else{
      D_re<-0}

D_regre<-c(D_regre,D_re)
D_regre1<-D_regre!=0
D_regre2<-D_regre[D_regre1==TRUE]

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

grid<-10^seq(10,-2,length=100)
lassomodel<-glmnet(X,Y,alpha=1,lambda=grid,intercept=TRUE)
cv.out<-cv.glmnet(X,Y,alpha=1)
bestlam<-cv.out$lambda.min
lassocoef<-predict.glmnet(lassomodel,type="coefficients",s=bestlam)

tmp_la<-order(abs(lassocoef),decreasing = TRUE)
tmp2_la<-lassocoef[tmp_la]!=0
T_lasso<-tmp_la[tmp2_la==TRUE]-1
```

```
u_lasso<-0
for(a_lasso in 1:s){
  for(b_lasso in 1:length(T_lasso)){
    if(S[a_lasso]==T_lasso[b_lasso]){
      u_lasso<-u_lasso+1}}
}
if(u_lasso/s==1){
  U_la<-1}else{
    U_la<-0}
U_lasso<-c(U_lasso,U_la)

if (U_la==1){
  d_lasso<-c()
  for(a_lasso2 in 1:s){
    for(b_lasso2 in 1:length(T_lasso)){
      if(S[a_lasso2]==T_lasso[b_lasso2]){
        d_lasso<-c(d_lasso,b_lasso2)}}

    D_la<-max(d_lasso)}}else{
      D_la<-0}

D_lasso<-c(D_lasso,D_la)
D_lasso1<-D_lasso!=0
D_lasso2<-D_lasso[D_lasso1==TRUE]

########## SIS ##########
Omega_SIS_V<-c()
```

```
for(f in 1:p){
  Corr_SIS<-cor(Y,X[,f])
  Omega_SIS_V<-c(Omega_SIS_V,Corr_SIS)
}

tmp_s<-order(abs(Omega_SIS_V),decreasing = TRUE)
T_sis<-tmp_s[c(1:d_1)]

u_sis<-0
for(a_sis in 1:s){
  for(b_sis in 1:length(T_sis)){
    if(S[a_sis]==T_sis[b_sis]){
      u_sis<-u_sis+1}}
}
if(u_sis/s==1){
  U_s<-1}else{
    U_s<-0}

U_sis<-c(U_sis,U_s)

if (U_s==1){
  d_sis<-c()
  for(a_sis2 in 1:s){
    for(b_sis2 in 1:length(T_sis)){
      if(S[a_sis2]==T_sis[b_sis2]){
        d_sis<-c(d_sis,b_sis2)}}

    D_s<-max(d_sis)}}else{
```

```
      D_s<-0}

D_sis<-c(D_sis,D_s)
D_sis_1<-D_sis!=0
D_sis_2<-D_sis[D_sis_1==TRUE]

T_sis2<-tmp_s[c(1:d_2)]

u_sis2<-0
for(a_sis_2 in 1:s){
  for(b_sis_2 in 1:length(T_sis2)){
    if(S[a_sis_2]==T_sis2[b_sis_2]){
      u_sis2<-u_sis2+1}}
}
if(u_sis2/s==1){
  U_s2<-1}else{
    U_s2<-0}

U_sis2<-c(U_sis2,U_s2)

if (U_s2==1){
  d_sis2<-c()
  for(a_sis2_2 in 1:s){
    for(b_sis2_2 in 1:length(T_sis2)){
      if(S[a_sis2_2]==T_sis2[b_sis2_2]){
        d_sis2<-c(d_sis2,b_sis2_2)}}

    D_s2<-max(d_sis2)}}else{
```

```
      D_s2<-0}

D_sis2<-c(D_sis2,D_s2)
D_sis2_1<-D_sis2!=0
D_sis2_2<-D_sis2[D_sis2_1==TRUE]

T_sis3<-tmp_s[c(1:d_3)]

u_sis3<-0
for(a_sis_3 in 1:s){
  for(b_sis_3 in 1:length(T_sis3)){
    if(S[a_sis_3]==T_sis3[b_sis_3]){
      u_sis3<-u_sis3+1}}
}
if(u_sis3/s==1){
  U_s3<-1}else{
    U_s3<-0}

U_sis3<-c(U_sis3,U_s3)

if (U_s3==1){
  d_sis3<-c()
  for(a_sis2_3 in 1:s){
    for(b_sis2_3 in 1:length(T_sis3)){
      if(S[a_sis2_3]==T_sis3[b_sis2_3]){
        d_sis3<-c(d_sis3,b_sis2_3)}}

    D_s3<-max(d_sis3)}}else{
```

```
      D_s3<-0}

D_sis3<-c(D_sis3,D_s3)
 D_sis3_1<-D_sis3!=0
D_sis3_2<-D_sis3[D_sis3_1==TRUE]

########## DC-SIS ##########
Omega_DCSIS_V<-c()
for(g in 1:p){
  r<-cor(Y,X[,g])
  Corr_DCSIS<-((r*asin(r))+(sqrt(1-(r^2)))-(r*asin(r/2))-(sqrt(4-(r^2)))+1)/(1+(pi/3)-sqrt(3))
  Omega_DCSIS_V<-c(Omega_DCSIS_V,Corr_DCSIS)
}

tmp_dc<-order(abs(Omega_DCSIS_V),decreasing = TRUE)
T_dcsis<-tmp_dc[c(1:d_1)]

u_dcsis<-0
for(a_dcsis in 1:s){
  for(b_dcsis in 1:length(T_dcsis)){
    if(S[a_dcsis]==T_dcsis[b_dcsis]){
      u_dcsis<-u_dcsis+1}}
}
if(u_dcsis/s==1){
  U_dc<-1}else{
    U_dc<-0}

U_dcsis<-c(U_dcsis,U_dc)
```

```
if (U_dc==1){
  d_dcsis<-c()
  for(a_dcsis2 in 1:s){
    for(b_dcsis2 in 1:length(T_dcsis)){
      if(S[a_dcsis2]==T_dcsis[b_dcsis2]){
        d_dcsis<-c(d_dcsis,b_dcsis2)}}

    D_dc<-max(d_dcsis)}}else{
      D_dc<-0}

D_dcsis<-c(D_dcsis,D_dc)
D_dcsis_1<-D_dcsis!=0
D_dcsis_2<-D_dcsis[D_dcsis_1==TRUE]

T_dcsis2<-tmp_dc[c(1:d_2)]

u_dcsis2<-0
for(a_dcsis_2 in 1:s){
  for(b_dcsis_2 in 1:length(T_dcsis2)){
    if(S[a_dcsis_2]==T_dcsis2[b_dcsis_2]){
      u_dcsis2<-u_dcsis2+1}}
}
if(u_dcsis2/s==1){
  U_dc2<-1}else{
  U_dc2<-0}

U_dcsis2<-c(U_dcsis2,U_dc2)
```

```
if (U_dc2==1){
  d_dcsis2<-c()
  for(a_dcsis2_2 in 1:s){
    for(b_dcsis2_2 in 1:length(T_dcsis2)){
      if(S[a_dcsis2_2]==T_dcsis2[b_dcsis2_2]){
        d_dcsis2<-c(d_dcsis2,b_dcsis2_2)}}

    D_dc2<-max(d_dcsis2)}}else{
      D_dc2<-0}

D_dcsis2<-c(D_dcsis2,D_dc2)
D_dcsis2_1<-D_dcsis2!=0
D_dcsis2_2<-D_dcsis2[D_dcsis2_1==TRUE]

T_dcsis3<-tmp_dc[c(1:d_3)]

u_dcsis3<-0
for(a_dcsis_3 in 1:s){
  for(b_dcsis_3 in 1:length(T_dcsis3)){
    if(S[a_dcsis_3]==T_dcsis3[b_dcsis_3]){
      u_dcsis3<-u_dcsis3+1}}
}
if(u_dcsis3/s==1){
  U_dc3<-1}else{
    U_dc3<-0}

U_dcsis3<-c(U_dcsis3,U_dc3)
```

```
  if (U_dc3==1){
    d_dcsis3<-c()
    for(a_dcsis2_3 in 1:s){
      for(b_dcsis2_3 in 1:length(T_dcsis3)){
        if(S[a_dcsis2_3]==T_dcsis3[b_dcsis2_3]){
          d_dcsis3<-c(d_dcsis3,b_dcsis2_3)}}

      D_dc3<-max(d_dcsis3)}}else{
        D_dc3<-0}

  D_dcsis3<-c(D_dcsis3,D_dc3)
  D_dcsis3_1<-D_dcsis3!=0
  D_dcsis3_2<-D_dcsis3[D_dcsis3_1==TRUE]

########## ITRRS ##########

  T_itrrs<-c(1:p)

  for(g in 1:14){
    if(length(T_itrrs)>n){
      grid<-10^seq(10,-2,length=100)
      ridgemodel<-glmnet(X[,sort(T_itrrs)],Y,alpha=0,lambda=grid,intercept=TRUE)
      cv.out<-cv.glmnet(X[,sort(T_itrrs)],Y,alpha=1)
      bestlam<-cv.out$lambda.min
      ridgecoef<-predict.glmnet(ridgemodel,type="coefficients",s=bestlam)
      Omega_ITRRS<-matrix(ridgecoef[-1],nrow=length(T_itrrs),ncol=1)
      M_itrrs<-
matrix(c(abs(Omega_ITRRS),sort(T_itrrs)),nrow=length(Omega_ITRRS),ncol=2)
```

```
    M_itrrs_2<-M_itrrs[order(Omega_ITRRS[,1],decreasing = TRUE),]
    T_itrrs<-M_itrrs_2[c(1:floor((length(T_itrrs)*0.8))),2]
    }
}

u_itrrs<-0
for(a_itrrs in 1:s){
  for(b_itrrs in 1:length(T_itrrs)){
    if(S[a_itrrs]==T_itrrs[b_itrrs]){
      u_itrrs<-u_itrrs+1}}
}

if(u_itrrs/s==1){
  U_it<-1}else{
    U_it<-0}
U_itrrs<-c(U_itrrs,U_it)


if (U_it==1){
  d_itrrs<-c()
  for(a_itrrs2 in 1:s){
    for(b_itrrs2 in 1:length(T_itrrs)){
      if(S[a_itrrs2]==T_itrrs[b_itrrs2]){
        d_itrrs<-c(d_itrrs,b_itrrs)}}

    D_it<-max(d_itrrs)}}else{
      D_it<-0}
```

```
  D_itrrs<-c(D_itrrs,D_it)
  D_itrrs1<-D_itrrs!=0
  D_itrrs2<-D_itrrs[D_itrrs1==TRUE]
}

##########  the accuracy of variable screening (U) ##########

sum(U_regre)
sum(U_lasso)
sum(U_sis)
sum(U_sis2)
sum(U_sis3)
sum(U_dcsis)
sum(U_dcsis2)
sum(U_dcsis3)
sum(U_itrrs)

##########  Mean (M) ##########

mean(D_regre2)
mean(D_lasso2)
mean(D_sis_2)
mean(D_sis2_2)
mean(D_sis3_2)
mean(D_dcsis_2)
mean(D_dcsis2_2)
mean(D_dcsis3_2)
mean(D_itrrs2)
```

########## Standard Deviation (S.D.) ##########

sd(D_regre2)

sd(D_lasso2)

sd(D_sis_2)

sd(D_sis2_2)

sd(D_sis3_2)

sd(D_dcsis_2)

sd(D_dcsis2_2)

sd(D_dcsis3_2)

sd(D_itrrs2)

ประวัติผู้เขียนวิทยานิพนธ์

นาย ทวีศักดิ์ เล็กตระกูลชัย เกิดวันเสาร์ที่ 13 มีนาคม พ.ศ. 2536 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2558