การเปรียบเทียบประสิทธิภาพการพยากรณ์และการคัดเลือกตัวแปรของวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง สำหรับผลกระทบขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์

นางสาวทิฆัมพร สาระกอ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

COMPARING THE PREDICTION ACCURACY AND SUBSET SELECTION PERFORMANCES OF STEPWISE, LASSO, ELASTIC NET AND ADAPTIVE LASSO FOR SMALL AND SPARSE SIGNALS.

Miss Tikumporn Sarakor

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

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพการพยากรณ์และการคัดเลือกตัวแปรของวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง สำหรับผลกระทบขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์
โดย	นางสาวทิฆัมพร สาระกอ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. นัท กุลวานิช

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

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

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร. นัท กุลวานิช)

.....กรรมการ
(อาจารย์ ดร. วิฐรา พึ่งพาพงศ์)

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

ทิฆัมพร สาระกอ : การเปรียบเทียบประสิทธิภาพการพยากรณ์และการคัดเลือกตัวแปรของวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุงสำหรับผลกระทบขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์. (COMPARING THE PREDICTION ACCURACY AND SUBSET SELECTION PERFORMANCES OF STEPWISE, LASSO, ELASTIC NET AND ADAPTIVE LASSO FOR SMALL AND SPARSE SIGNALS.) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. นัท กุลวานิช, 79 หน้า.

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยที่มีลักษณะข้อมูลขนาดเล็กและมีค่าสัมประสิทธิ์การถดถอยบางตัวเป็นศูนย์ ด้วยวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง โดยใช้ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงบวก ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ และค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยเป็นเครื่องมือในการวัดประสิทธิภาพของการคัดเลือกตัวแปร โดยที่การคัดเลือกตัวแปรวิธีใดที่ให้ค่าของเกณฑ์ทั้ง 3 ต่ำสุดโดยสอดคล้องกันจะถือว่าการคัดเลือกตัวแปรวิธีนั้นเป็นวิธีที่มีประสิทธิภาพและเหมาะสมกับข้อมูลที่จำลองขึ้นมามากที่สุด

ผลการวิจัยพบว่า การคัดเลือกตัวแปรด้วยวิธีแลสโซ่ปรับปรุงนั้นให้ประสิทธิภาพดีที่สุดในหลายสถานการณ์ แต่สำหรับกรณีที่ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เมื่อเทียบกับจำนวนตัวแปรอิสระอยู่ในระดับสูง การคัดเลือกตัวแปรด้วยวิธีอีลาสติคเน็ตจะให้ประสิทธิภาพที่ดีกว่าการคัดเลือกตัวแปรด้วยวิธีแลสโซ่ปรับปรุง

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5581543626 : MAJOR STATISTICS
KEYWORDS: SUBSET SELECTION / SPARSE COEFFICIENTS / SMALL SIGNALS

TIKUMPORN SARAKOR: COMPARING THE PREDICTION ACCURACY AND SUBSET SELECTION PERFORMANCES OF STEPWISE, LASSO, ELASTIC NET AND ADAPTIVE LASSO FOR SMALL AND SPARSE SIGNALS.. ADVISOR: NAT KULVANICH, Ph.D., 79 pp.

This study aimed to compare the performances of the subset selection methods: Stepwise, Lasso, Elastic Net and Adaptive Lasso for small and sparse signals. The criteria for the performance measuring are False Positive, False Negative and Mean Absolute Error. The Variable selection method that provides a value of 3 minimum criteria will be considered as the best method and fit with data are simulated most.

The results showed the Adaptive Lasso offers the best performance in many situations. In case of the percentage of the sparse coefficients low, Elastic Net provides better performance than Adaptive Lasso

Department: Statistics
Field of Study: Statistics
Academic Year: 2013

Student's Signature .....
Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จลุล่วงได้ด้วยความอนุเคราะห์และการดูแลเอาใจใส่ของ อาจารย์ ดร. นัท กุลวานิช อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่ให้คำแนะนำ พร้อมทั้งคำปรึกษาอีกมากมายในการปรับปรุงแก้ไขและนำทางให้กับวิทยานิพนธ์ฉบับนี้ด้วยดีเสมอมา

ขอกราบขอบพระคุณ อาจารย์ ดร. วิฐรา พึ่งพาพงศ์ ที่สละเวลาให้คำแนะนำที่ดีและมีประโยชน์ต่อการปรับปรุงงาน จนกระทั่งวิทยานิพนธ์เล่มนี้สำเร็จสมบูรณ์

ขอกราบขอบพระคุณรองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา อาจารย์ ดร. วิฐรา พึ่งพาพงศ์ และอาจารย์ ดร. อรุณี กำลัง ประธานกรรมการและกรรมการสอบสอบวิทยานิพนธ์ที่กรุณาให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น และขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติที่ให้โอกาสทางการศึกษาและถ่ายทอดความรู้ให้แก่ผู้วิจัยจนกระทั่งสำเร็จการศึกษา

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณคุณพ่อและคุณแม่ที่ช่วยสนับสนุน ส่งเสริม เป็นกำลังใจซึ่งเป็นแรงผลักดันที่ดีที่สุดให้กับผู้วิจัยในการศึกษาเล่าเรียนจนสำเร็จการศึกษา และขอขอบคุณเพื่อนๆ ทุกคน ที่ให้ความช่วยเหลือ และเป็นกำลังใจให้ผู้วิจัยตลอดมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ตัวแบบการถดถอยเชิงเส้น (Linear Regression Model) เป็นหนึ่งในวิธีที่ใช้พยากรณ์เพื่อค้นหาความสัมพันธ์เชิงเส้นตรงระหว่างค่าพยากรณ์หรือตัวแปรตอบสนอง (Response) กับปัจจัยอื่นๆ หรือตัวพยากรณ์ (Predictor) ใดๆที่ส่งผลต่อค่าพยากรณ์ ซึ่งตัวแบบการถดถอยเชิงเส้นนี้มีการใช้งานอย่างแพร่หลาย เนื่องจากการใช้งานที่ค่อนข้างง่าย สะดวก และสามารถให้คำอธิบายที่เพียงพอในการตีความ ซึ่งบางครั้งตัวแบบเชิงเส้นนี้ยังสามารถทำงานได้ดีกว่าตัวแบบที่ไม่เป็นเชิงเส้น (Nonlinear Model) อีกด้วย วิธีการหนึ่งที่นิยมใช้ในการประมาณค่าอิทธิพลของพารามิเตอร์หรือค่าสัมประสิทธิ์การถดถอย ได้แก่ วิธีกำลังสองน้อยสุด (Ordinary Least Squares Method, OLS) ซึ่งวิธีนี้จะให้ค่าผลรวมกำลังสองของความคลาดเคลื่อน (Residual Sum of Squares) น้อยสุด แต่มีเหตุผล 2 ประการที่ทำให้การวิเคราะห์ข้อมูลด้วยตัวประมาณแบบวิธีกำลังสองน้อยสุด ได้ผลไม่ดีเท่าที่ควร ประการแรกคือ ความแม่นยำการทำนาย (Prediction Accuracy) ตัวประมาณแบบวิธีกำลังสองน้อยสุดมักให้ค่าความเอนเอียงต่ำแต่ค่าความแปรปรวนสูง ประการที่สอง คือ การตีความ (Interpretation) กรณีเมื่อมีตัวแปรพยากรณ์จำนวนมาก เรามักสนใจตัวแปรพยากรณ์บางส่วนที่ส่งผลกระทบอย่างรุนแรงกับตัวแปรตอบสนองเท่านั้น นอกจากนี้หากตัวแปรอิสระเหล่านั้นเกิดความสัมพันธ์ร่วมเชิงพหุ (Multicollinearity) ซึ่งกันและกัน จะทำให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุดไม่สามารถหาค่าได้

เพื่อเป็นการปรับปรุงวิธีกำลังสองน้อยสุด มี 2 เทคนิคพื้นฐานที่สามารถพัฒนาประสิทธิภาพให้กับตัวประมาณกำลังสองน้อยสุดได้ นั่นคือ การคัดเลือกตัวแปร (Subset Selection) และการวิเคราะห์การถดถอยริดจ์ (Ridge Regression) วิธีการคัดเลือกตัวแปร (Efroymson 1966) เป็นวิธีที่ทำการคัดเลือกตัวแปรเพียงบางส่วนที่มีผลกระทบต่อตัวแปรตอบสนองนำเข้าสมการ ส่วนตัวแปรพยากรณ์ตัวอื่นที่ไม่ส่งผลกระทบจะต้องไม่ถูกนำเข้าสมการพยากรณ์ ดังนั้นวิธีการคัดเลือกตัวแปรพยากรณ์จึงมีความจำเป็นเพื่อให้ได้สมการพยากรณ์ที่ดีที่สุดและประหยัดที่สุด แต่วิธีการคัดเลือกตัวแปรกลับมีข้อเสีย ถึงแม้ว่าจะให้สมการที่สามารถตีความได้ แต่ก็เป็นไปได้ที่จะมีตัวแปรมากหรือน้อยเกินไป เนื่องจากเป็นกระบวนการแบบไม่ต่อเนื่อง (Discrete) ที่ซึ่งตัวแปรอาจถูกเก็บเอาไว้หรือถูกตัดทิ้งไปโดยไม่จำเป็น และการเปลี่ยนแปลงข้อมูลเพียงเล็กน้อยนั้นอาจส่งผลให้สมการมีค่าเปลี่ยนแปลงไปจากสมการที่เลือกไว้แล้วได้ ซึ่งส่งผลให้ความแม่นยำในการทำนายลดลง ส่วนการ

วิเคราะห์การถดถอยริดจ์นั้นเป็นกระบวนการต่อเนื่อง (Continuous) ที่สามารถลดทอนค่าสัมประสิทธิ์และเพิ่มเสถียรภาพให้กับสมการ แต่การถดถอยริดจ์นั้นกลับไม่ได้ทำการคัดเลือกตัวแปร จึงไม่สามารถให้สมการที่สะดวกต่อการตีความ

จากข้อเสียดังกล่าวจึงได้มีผู้คิดค้นวิธีที่สามารถรักษาไว้ซึ่งข้อดีของทั้งวิธีการคัดเลือกตัวแปรและการวิเคราะห์การถดถอยริดจ์ นั่นคือ วิธีแลสโซ่ (LASSO) (Tibshirani 1996) เป็นการปรับค่าให้วิธีกำลังสองน้อยสุดโดยกำหนด L_1-penalty ในการหาค่าสัมประสิทธิ์การถดถอย แลสโซ่จะทำการลดค่าแบบต่อเนื่อง (Continuous shrinkage) และทำการคัดเลือกตัวแปรไปพร้อมๆกัน แม้ว่าแลสโซ่จะสามารถใช้ประโยชน์ได้หลากหลายสถานการณ์ แต่ Zou และ Hastie พบว่า ในกรณีที่จำนวนตัวแปรอิสระมีมากกว่าจำนวนค่าสังเกต ส่วนใหญ่แลสโซ่จะเลือกจำนวนตัวแปรมากที่สุดก่อนสิ้นสุดกระบวนการ เนื่องจากเป็นธรรมชาติของกระบวนการแก้ปัญหาแบบ Convex นอกจากนี้แลสโซ่มีเพียงขอบเขตของ L_1-norm ของค่าสัมประสิทธิ์ที่มีค่าน้อยกว่าค่าคงที่ค่าหนึ่งเท่านั้น และในกรณีที่จำนวนตัวแปรอิสระมีน้อยกว่าจำนวนค่าสังเกต ถ้าแต่ละตัวแปรพยากรณ์มีความสัมพันธ์กันสูง จะสังเกตได้ว่าประสิทธิภาพการทำนายของแลสโซ่นั้นจะต่ำกว่าการวิเคราะห์การถดถอยแบบริดจ์ Zou and Hastie (2005) จึงนำเสนอวิธีอีลาสติคเน็ต (Elastic Net) ขึ้น โดยที่วิธีนี้สามารถลดค่าแบบต่อเนื่องและทำการคัดเลือกตัวแปรไปพร้อมๆกันได้เช่นเดียวกับแลสโซ่ และปรับปรุงในส่วนที่แลสโซ่ทำงานได้ไม่ดีอีกด้วย ต่อมา Zou พบว่ามีเงื่อนไขบางประการที่ทำให้ Lasso ไม่คงเส้นคงวา จึงนำเสนอพจน์ค่าปรับด้วยการปรับค่าน้ำหนักให้กับ L_1-penalty ใน Lasso ซึ่งวิธีนี้สามารถใช้ได้กับสถานการณ์ที่ทำให้ Lasso ไม่คงเส้นคงวาได้ และการเลือกตัวแปรแบบ Adaptive Lasso (Zou 2006) สามารถทำงานได้ดีเมื่อตัวแบบที่แท้จริงได้ถูกรวมไว้ในตัวแบบเริ่มต้น

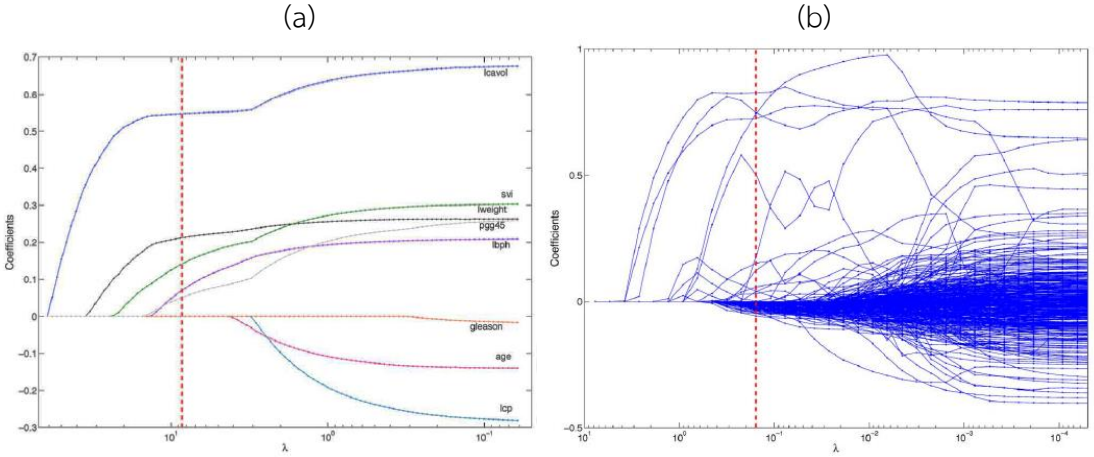
วิธีการดังกล่าวข้างต้นมีการนำไปใช้ในงานที่เกี่ยวข้องกับการศึกษาทางด้านชีวสารสนเทศ (Bioinformatics) เช่น ศึกษาระดับการแสดงออกของยีน การวินิจฉัยโรค การทำนายการเกิดโรค เป็นต้น ซึ่งข้อมูลประเภทนี้จะมีปัจจัยที่ก่อให้เกิดลักษณะที่สนใจเป็นจำนวนมาก ในขณะที่ตัวอย่างมีจำนวนน้อย ส่งผลให้ค่าสัมประสิทธิ์การถดถอยมีค่าเข้าใกล้ศูนย์มาก (Small Coefficient) ดังในผลการศึกษาการประมาณค่าสัมประสิทธิ์ด้วยวิธี SGLasso ในงานวิจัยของ (Ma, Song et al. 2007) ดังนั้นจึงเป็นเหตุผลหลักในการเลือกใช้ตัวแบบการถดถอยดังกล่าวในการวิเคราะห์ ซึ่งการวิเคราะห์ข้อมูลด้วยตัวแบบการถดถอยอื่นอาจทำให้เกิดการละทิ้งตัวแปรโดยไม่จำเป็น นอกจากนี้ Mateos (2010) ได้นำการวิเคราะห์แบบ Sparse Linear Regression ไปใช้กับข้อมูลทางด้าน Digital Signal Processing ซึ่งพัฒนาวิธี Distributed Lasso (DLasso) ขึ้นโดยทำการทดสอบกับข้อมูลจริง Prostate Cancer แล้วนำไปประยุกต์ใช้กับข้อมูลจำลองซึ่งเป็นข้อมูลการสื่อสารกันระหว่างกระบวนการโดยการประมาณค่าความหนาแน่นสเปกตรัมของคลื่นความถี่วิทยุในการประมวลผลเครือข่าย (Network Processing) ทางผู้วิจัยจึงสนใจศึกษาการจำลองข้อมูลในแบบต่างๆ เพื่อหาวิธีการคัดเลือกตัวแปรที่เหมาะสมกับตัวแบบการถดถอยเชิงเส้นที่ค่าสัมประสิทธิ์บางตัวเป็นศูนย์

(Sparse Linear Regression) และมีค่าสัญญาณขนาดเล็ก (Small Coefficient) เพื่อเปรียบเทียบประสิทธิภาพว่าวิธีการใดที่มีความเหมาะสมกับข้อมูลเหล่านี้มากที่สุด

1.2 ตัวอย่างลักษณะข้อมูล

งานวิจัยที่มีลักษณะของค่าสัมประสิทธิ์การถดถอยมีค่าบางตัวเป็นศูนย์และมีค่าเข้าใกล้ศูนย์นั้นสามารถเห็นได้จากงานวิจัยของ Ma, Song et al. (2007) ในการศึกษาข้อมูลจริง 2 ชุด ซึ่งมีลักษณะข้อมูลเป็นกรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง โดยชุดแรกเป็นชุดข้อมูลยีน Colon จำนวน 6,500 ยีน จากผู้ที่เป็นเนื้องอก 40 คน และคนปกติ 22 คน ส่วนชุดข้อมูลที่สองเป็นชุดข้อมูลยีน Nodal จำนวน 7,129 ยีน จากผู้ที่มีเนื้องอกที่เต้านมชนิดบวก 25 คน และชนิดลบ 24 คน โดยวิธีที่ใช้ในการเปรียบเทียบเพื่อจัดกลุ่มโครงสร้างยีนสำหรับการคัดเลือกยีนและการสร้างสมการทำนาย ได้แก่ Lasso, Group Lasso (GLasso) และ Supervised Group Lasso (SGLasso) ผลการวิจัยในแง่ของการคัดเลือกตัวแปรอิสระนั้นพบว่า จากชุดข้อมูลทั้งสองชุดข้างต้น วิธี Lasso และวิธี SGLasso จะให้ประสิทธิภาพที่ดีอย่างใกล้เคียงกัน ข้อแตกต่างคือวิธี SGLasso นั้นสามารถแบ่งกลุ่มตัวแปรอิสระที่ผ่านการคัดเลือกตัวแปรแล้วได้อีกด้วย ส่วนในแง่ของการประมาณค่าสัมประสิทธิ์ ตัวแปรอิสระที่ถูกคัดออกจากตัวแบบนั่นคือตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยเป็นศูนย์ และตัวแปรอิสระที่ถูกคัดเลือกเข้าตัวแบบนั่นคือตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เป็นศูนย์ ซึ่งผลของการประมาณค่าสัมประสิทธิ์ที่มีค่าไม่เป็นศูนย์ที่แสดงไว้ในงานวิจัยนี้จะเป็นค่าขนาดเล็กซึ่งมีค่าอยู่ระหว่าง -1 ถึง 1 ในทั้งสองชุดข้อมูล

และในงานวิจัยของ Mateos (2010) ได้ปรับปรุงการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Lasso เพื่อนำไปประยุกต์ใช้ในการระบุตำแหน่งที่มาของสัญญาณ หรือการประมาณสนามพลังเมื่อข้อมูลสอน (Training Data) มีการกระจายข้ามตำแหน่ง โดยศึกษาทดลองกับข้อมูลจริง คือ Prostate Cancer ซึ่งมี 8 ตัวแปรอิสระ จากคนไข้ 67 คน แบ่งออกเป็น 7 กลุ่ม โดยข้อมูลในแต่ละกลุ่มจะแทนที่แต่ละตำแหน่ง ภายในกลุ่มข้อมูลนั้นจะเปรียบเสมือนมีการสื่อสารกันระหว่างตำแหน่งเกิดขึ้น ผลการทดสอบพบว่าวิธี Distributed Lasso นั้นสามารถคัดเลือกตัวแปรอิสระและประมาณค่าสัมประสิทธิ์การถดถอยที่มีค่าเข้าใกล้ศูนย์ได้ หลังจากนั้นจึงนำไปใช้กับข้อมูลจำลองของ Spectrum ซึ่งมี 968 ตัวแปรอิสระ 400 ตัวอย่างดังภาพที่ 1.2.1



ภาพที่ 1.2.1 แสดงแนวเส้นการประมาณค่าสัมประสิทธิ์แต่ละตัวแปรด้วยวิธี Distributed Lasso ของ (a) ข้อมูล Prostate Cancer (b) ข้อมูล Spectrum โดยเส้นประในแนวตั้งแสดงค่า λ ที่ให้ค่าประมาณของค่าคลาดเคลื่อนการทำนายน้อยที่สุด

ผลการวิจัยพบว่า ที่ค่า λ ที่ให้ค่าคลาดเคลื่อนการทำนายน้อยที่สุด ค่าสัมประสิทธิ์ที่มีค่าเป็นศูนย์จะถูกคัดออกจากโมเดล ส่วนค่าสัมประสิทธิ์ที่มีค่าไม่เป็นศูนย์จะมีค่าอยู่ระหว่าง -1 ถึง 1 ดังในภาพที่ 1.2.1

จากตัวอย่างงานวิจัยข้างต้น ผู้วิจัยจึงเลือกลักษณะข้อมูลที่จะทำการจำลองโดยใช้ทั้งกรณีที่จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง และกรณีจำนวนตัวแปรอิสระมากกว่าตัวอย่าง โดยที่ค่าสัมประสิทธิ์ที่แท้จริงนั้นมีค่าขนาดเล็กที่มีค่าอยู่ระหว่าง -1 ถึง 1 และค่าสัมประสิทธิ์บางตัวเป็นศูนย์ เพื่อทำการเปรียบเทียบประสิทธิภาพของวิธีคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ทั้ง 4 วิธี เพื่อพิจารณาว่าการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์วิธีใดจะมีประสิทธิภาพที่ดีสุดสำหรับข้อมูลที่มีค่าสัมประสิทธิ์ขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์

1.3 วัตถุประสงค์ของการศึกษา

เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยที่มีลักษณะข้อมูลขนาดเล็กและมีค่าสัมประสิทธิ์การถดถอยบางตัวเป็นศูนย์ ด้วยวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง

1.4 ขอบเขตของการศึกษา

ในการวิจัยครั้งนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1.4.1 รูปแบบการความสัมพันธ์การถดถอยเชิงเส้น คือ

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

โดยที่ $\boldsymbol{y} = (y_1, ..., y_n)^T$ เป็นเวกเตอร์ของตัวแปรตอบสนองขนาด n

$\beta = (\beta_1, ..., \beta_p)^T$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p

$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ เป็นเมตริกซ์ตัวแปรพยากรณ์ขนาด $n \times p$ ที่มีการแจกแจงแบบปกติที่ค่าเฉลี่ยเป็นศูนย์ มีเมตริกซ์ความแปรปรวนร่วมเป็น $\boldsymbol{\Sigma} = \text{cov}[\boldsymbol{x}_i, \boldsymbol{x}_j]$

$\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n

เมื่อ n เป็นขนาดตัวอย่าง

p เป็นจำนวนตัวแปรอิสระ

1.4.2 ความสัมพันธ์ระหว่างตัวแปรอิสระ $\boldsymbol{x}_i$ และ $\boldsymbol{x}_j$ เป็น $\gamma^{|i-j|}$ เมื่อ $i = 1, 2, ..., n$, $j = 1, 2, ..., p$ และ $i \neq j$

1.4.3 ค่าความคลาดเคลื่อนมีการแจกแจงแบบปกติที่ค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1

1.4.4 ขนาดตัวอย่าง 2 ระดับ ได้แก่ 20 และ 100

1.4.5 จำนวนตัวแปรอิสระแบ่งออกเป็น 2 กรณี ได้แก่ กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง ($p < n$) และกรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง ($p > n$) ซึ่งจะคิดเป็นสัดส่วนกับขนาดตัวอย่าง

1.4.6 ค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์ (Nonzero Coefficient) จะถูกกำหนดให้มีค่าเข้าใกล้ศูนย์ (Small Coefficient) โดยมีค่าอยู่ระหว่าง -1 ถึง 1 ซึ่งมีจำนวนเป็นร้อยละ 10, 40 และ 70 ของจำนวนตัวแปร ส่วนจำนวนสัมประสิทธิ์การถดถอยที่เหลือจะถูกกำหนดให้มีค่าเป็น 0

1.4.8 ทำการจำลองข้อมูลตามสถานการณ์ที่แตกต่างตามกำหนด โดยจำลองแต่ละสถานการณ์เป็นจำนวน 500 รอบ

1.5 วิธีดำเนินการศึกษา

1.5.1 ศึกษาค้นคว้าเอกสารและข้อมูลที่เกี่ยวข้องกับงานวิจัย

1.5.2 กำหนดเงื่อนไขและขอบเขตของการวิจัย ได้แก่ ความสัมพันธ์ระหว่างตัวแปรอิสระ (γ) ขนาดตัวอย่าง (n) จำนวนตัวแปรอิสระ (p) และค่าสัมประสิทธิ์การถดถอยที่แท้จริง (β)

1.5.3 จำลองข้อมูลของตัวแปรพยากรณ์ $(\boldsymbol{x})$ และค่าคลาดเคลื่อน (ε) ตามการแจกแจงและขอบเขตที่ต้องการศึกษา

1.5.4 คำนวณค่าตอบสนอง $(\boldsymbol{y})$ ตามสมการ

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

1.5.5 ค่า $\boldsymbol{x}$ และ $\boldsymbol{y}$ ที่ได้จากการคำนวณนำมาประมาณค่าสัมประสิทธิ์การถดถอย $(\hat{\beta})$ โดยใช้วิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง

1.5.6 ทำการวิเคราะห์ข้อมูลและเปรียบเทียบประสิทธิภาพการทำนาย โดยใช้ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงบวก (False Positive) ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (Mean Absolute Error)

1.5.7 สรุปผลที่ได้จากการศึกษา

1.6 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้วิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์ด้วยตัวปรับค่าที่เหมาะสมสำหรับข้อมูลที่มีค่าสัมประสิทธิ์การถดถอยที่มีขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์ และเพื่อเปรียบเทียบประสิทธิภาพของตัวแบบพยากรณ์ที่ได้จากแต่ละวิธี

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ในการศึกษานี้มีทฤษฎีและสถิติที่เกี่ยวข้อง คือ การคัดเลือกตัวแปร (Subset Selection) และเกณฑ์ที่ใช้ในการวัดประสิทธิภาพ โดยการคัดเลือกตัวแปรที่ใช้ในการศึกษานี้มี 4 วิธี ได้แก่ วิธีการคัดเลือกแบบขั้นตอน (Stepwise Selection) วิธีแลสโซ่ (Lasso) วิธีอีลาสติคเน็ต (Elastic net) วิธีแลสโซ่ปรับปรุง (Adaptive Lasso) และเกณฑ์ที่ใช้ในการวัดประสิทธิภาพ ได้แก่ ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive) ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (Mean Absolute Error)

2.1 การคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์การถดถอย

2.1.1 วิธีการคัดเลือกแบบขั้นตอน (Stepwise Selection)

การคัดเลือกตัวแปรแบบนี้เป็นการผสมผสานระหว่างวิธีการคัดเลือกตัวแปรพยากรณ์ทั้งแบบก้าวหน้าและแบบถอยหลังเข้าด้วยกัน ในขั้นแรกจะเลือกตัวแปรพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตามสูงที่สุดและมีนัยสำคัญทางสถิติ และมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรพยากรณ์ด้วยกันมีค่าน้อยและไม่มีนัยสำคัญทางสถิติเข้าสมการก่อน จากนั้นจะทดสอบตัวแปรพยากรณ์ที่ไม่ได้อยู่ในสมการว่ามีตัวแปรพยากรณ์ใดบ้างมีสิทธิ์เข้ามาอยู่ในการสมการด้วยวิธีการคัดเลือกแบบก้าวหน้า (Forward Selection) และขณะเดียวกันก็จะทดสอบตัวแปรพยากรณ์ที่อยู่ในสมการด้วยว่า ตัวแปรพยากรณ์ที่อยู่ในสมการตัวใดมีโอกาสที่จะถูกขจัดออกจากสมการด้วยวิธีการคัดเลือกแบบถอยหลัง (Backward Selection) โดยจะกระทำการคัดเลือกผสมทั้งสองวิธีนี้ในทุกขั้นตอนจนกระทั่งไม่มีตัวแปรใดที่ถูกคัดออกจากสมการและไม่มีตัวแปรใดที่จะถูกนำเข้าสมการ โดยจะเลือกสมการที่เล็กที่สุดที่ให้ค่าคาดหวังของความคลาดเคลื่อนการทำนายน้อยที่สุด

วิธีการคัดเลือกแบบก้าวหน้า (Forward Selection) นี้จะไม่มีตัวแปรพยากรณ์ใดๆอยู่ในสมการเริ่มต้น ในขั้นถัดไปจะทำการเพิ่มตัวแปรพยากรณ์เข้าไปทีละตัว จนกระทั่งตัวแปรทั้งหมดอยู่ในสมการ หรือผลตามเกณฑ์ที่กำหนด ตัวแปรใดที่ให้ค่า F-Ratio สูงที่สุด จะถูกนำเข้าสมการถ้าค่า F-

Ratio ไม่น้อยกว่าค่าที่กำหนด นั่นคือ ค่า F-Ratio ของตัวแปรที่ i (F_i) จะถูกนำเข้าในเทอม p ของสมการ ตามเกณฑ์ดังนี้

$$F_i = \max_i \left(\frac{RSS_p - RSS_{p+i}}{\hat{\sigma}^2_{p+i}} \right) > F_{in}$$

โดยที่ $p+i$ คือ จำนวนตัวแปรพยากรณ์ที่ใช้คำนวณเมื่อตัวแปรพยากรณ์ที่ i ถูกเลือกเข้าสู่สมการในเทอม p

วิธีการคัดเลือกแบบถอยหลัง (Backward Selection) จะเริ่มต้นด้วยกับสมการที่รวมทุกตัวแปรพยากรณ์ และจะจำกัดออกทีละตัวในขั้นตอนถัดไป ตัวแปรใดที่ให้ค่า F-Ratio ต่ำที่สุด จะถูกนำออกจากสมการถ้าค่า F-Ratio ไม่เกินค่าที่กำหนด นั่นคือ ค่า F-Ratio ของตัวแปรที่ i (F_i) จะถูกนำออกจากเทอม p ของสมการ ตามเกณฑ์ดังนี้

$$F_i = \min_i \left(\frac{RSS_{p-i} - RSS_p}{\hat{\sigma}^2_p} \right) < F_{out}$$

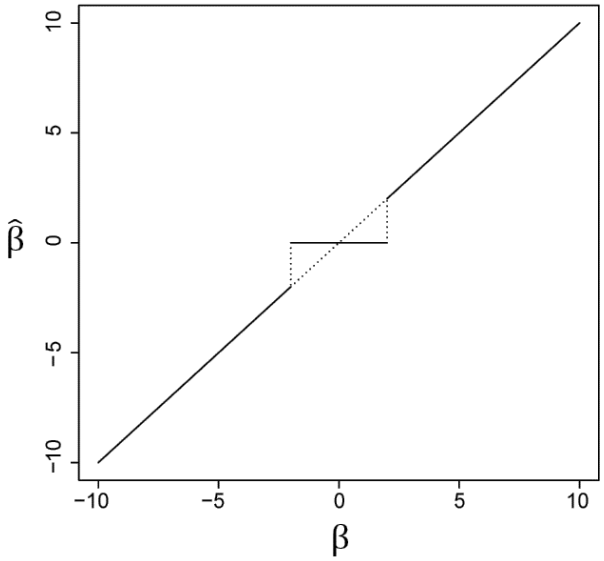
โดยที่ RSS_{p-i} คือ ค่าผลรวมกำลังสองของความคลาดเคลื่อน เมื่อตัวแปรพยากรณ์ที่ i ถูกกำจัดออกจากสมการปัจจุบัน

ค่าที่เหมาะสมสำหรับ F_{in} และ F_{out} เพื่อใช้ในการหยุดการทำงานของวิธีการคัดเลือกแบบขั้นตอนนั้นจะใช้เกณฑ์ดังต่อไปนี้

$$F_{in} = F\left(\alpha, 1, n-p-1\right)$$

และ

$$F_{out} = F\left(\alpha, 1, n-p\right)$$



ภาพที่ 2.1.1 แสดงคุณลักษณะของการประมาณค่าแบบ Subset Selection **(——)** โดยเทียบกับคุณลักษณะการประมาณค่าแบบ OLS (**.....**)

จากภาพที่ 2.1.1 แสดงให้เห็นว่าการประมาณค่าแบบ Subset Selection นั้นจะให้ค่าประมาณสัมประสิทธิ์เทียบเท่าการประมาณค่าแบบ OLS แต่เมื่อค่าสัมประสิทธิ์ที่แท้จริงนั้นมีค่าลู่เข้าสู่ค่าคงที่ค่าหนึ่ง Subset Selection จะปรับให้ค่าประมาณของสัมประสิทธิ์ตัวนั้นมีค่าเป็นศูนย์

2.1.2 วิธีแลสโซ่ (Least Absolute Shrinkage and Selection Operator)

เป็นวิธีที่สามารถลดค่าให้กับสัมประสิทธิ์การถดถอยบางตัวและทำให้สัมประสิทธิ์การถดถอยตัวอื่นๆ เป็นศูนย์ได้ และยังเป็นวิธีที่สามารถรักษาไว้ซึ่งคุณลักษณะที่ดีของการคัดเลือกตัวแปร (Subset Selection) นั่นคือ คุณสมบัติของการให้ตัวแบบที่ง่ายต่อการตีความ และอีกหนึ่งคุณลักษณะที่ดีของการถดถอยริดจ์ (Ridge Regression) นั่นคือ การลดค่าสัมประสิทธิ์และเพิ่มเสถียรภาพ โดยทำให้ค่าผลรวมความคลาดเคลื่อนกำลังสองมีค่าน้อยสุดโดยกำหนด ℓ_1- Penalty ให้สัมประสิทธิ์การถดถอย

กำหนดให้มีข้อมูลเริ่มต้น $\left(\boldsymbol{x}^i, y_i\right), i = 1,2,...,n$ โดย $\boldsymbol{x}^i = \left(x_{i1},...,x_{ip}\right)$ เป็นตัวแปรพยากรณ์ y_i เป็นตัวแปรตอบสนอง และให้ $\hat{\beta} = \left(\hat{\beta}_1,...,\hat{\beta}_p\right)^T$ ซึ่งการประมาณค่าสัมประสิทธิ์การถดถอยแลสโซ่สามารถคำนวณได้ดังนี้

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \qquad \text{โดยที่} \qquad \sum_{j=1}^{p} |\beta_j| \le t$$

เมื่อ $\sum_{i=1}^{p} |\beta_j|$ เป็นพจน์ค่าปรับ (Penalty term) วิธีแลสโซ่

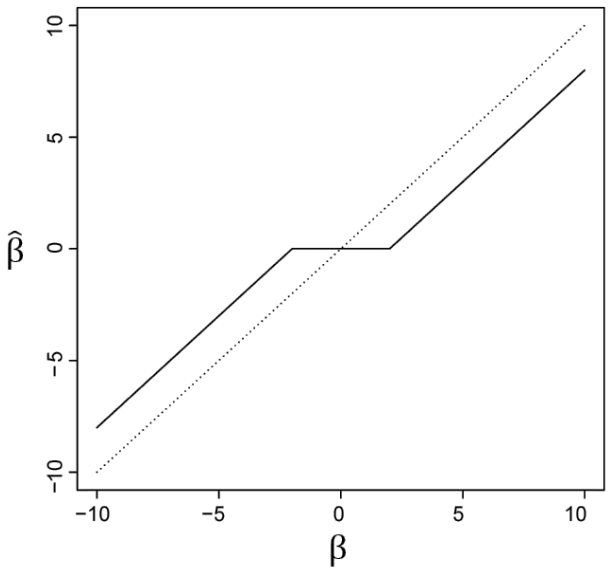
t เป็นพารามิเตอร์ปรับค่า (Tuning Parameter)

โดยธรรมชาติของเงื่อนไขข้างต้น เมื่อทำให้ t มีค่าเล็กเพียงพอแล้ว จะสามารถทำให้สัมประสิทธิ์การถดถอยบางตัวนั้นมีค่าเป็นศูนย์ได้ ควรเลือกใช้ t ที่ให้ค่าคลาดเคลื่อนการทำนายน้อยที่สุด

วิธีที่ใช้ในการประมาณพารามิเตอร์ t ในการศึกษานี้ คือ Ten-fold Cross-validation โดยตัวประมาณแลสโซ่จะถูกจัดให้อยู่ในรูปของ Normalized Parameter $s = t/\sum \hat{\beta}_j^0$ โดยที่ $\hat{\beta}_j^0$ เป็น Full Least Squares Estimates ซึ่งค่าพารามิเตอร์ s นี้จะมีค่าอยู่ระหว่าง 0 – 1 ค่าประมาณของ s ที่ดีที่สุดคือสามารถให้ค่าประมาณของค่าคลาดเคลื่อนการทำนาย (PE) น้อยที่สุด

ถึงแม้ว่าวิธีแลสโซ่นั้นจะสามารถทำงานได้ในหลากหลายสถานการณ์ แต่มีข้อจำกัดบางสถานการณ์เช่นกัน ในงานวิจัยของ Zou และ Hastie (2005) ได้กล่าวถึงข้อจำกัดของวิธีแลสโซ่ ไว้ดังนี้

1. กรณีจำนวนตัวแปรมากกว่าขนาดตัวอย่าง $(p > n)$ วิธีแลสโซ่จะเลือกจำนวนตัวแปรที่มากที่สุด ก่อนจะได้สมการที่เหมาะสม
2. กรณีจำนวนตัวแปรน้อยกว่าขนาดตัวอย่าง $(p < n)$ ถ้าตัวแปรอิสระมีค่าความสัมพันธ์ระหว่างตัวแปรสูง เมื่อพิจารณาจากประสิทธิภาพการทำนายแล้ว การถดถอยริดจ์จะทำงานได้ดีกว่าวิธีแลสโซ่



ภาพที่ 2.1.2 แสดงคุณลักษณะของการประมาณค่าแบบ Lasso (——) โดยเทียบกับคุณลักษณะการประมาณค่าแบบ OLS (.....)

จากภาพที่ 2.1.2 แสดงให้เห็นว่าการประมาณค่าแบบ Lasso นั้นจะให้ค่าประมาณสัมประสิทธิ์ต่ำกว่าการประมาณค่าแบบ OLS แต่เมื่อค่าสัมประสิทธิ์ที่แท้จริงนั้นมีค่าลู่เข้าสู่ค่าคงที่ค่าหนึ่ง Lasso จะปรับให้ค่าประมาณของสัมประสิทธิ์ตัวนั้นมีค่าเป็นศูนย์

2.1.3 วิธีอีลาสติคเน็ต (Elastic net)

เป็นวิธีที่สามารถลดค่าสัมประสิทธิ์การถดถอยและทำให้สัมประสิทธิ์การถดถอยตัวอื่นเป็นศูนย์ได้เฉกเช่นเดียวกับวิธีแลสโซ่และสามารถนำไปใช้กับข้อมูลของกลุ่มตัวแปรอิสระที่มีความสัมพันธ์กันสูงได้ นอกจากนี้ยังให้สมการที่ค่าสัมประสิทธิ์การถดถอยบางตัวมีค่าเป็นศูนย์ด้วยกับการทำนายที่มีความแม่นยำ วิธีนี้เหมาะสมกับข้อมูลที่จำนวนตัวแปรอิสระนั้นมีจำนวนมากกว่าขนาดตัวอย่างมากๆ ซึ่งวิธีแลสโซ่นั้นให้ประสิทธิภาพในสถานการณ์นี้ได้ไม่ดีนัก ซึ่งการประมาณค่าสัมประสิทธิ์การถดถอยอีลาสติคเน็ตสามารถคำนวณได้ดังนี้

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \qquad \text{โดยที่} \quad \sum_{j=1}^{p} \left(\alpha \left|\beta_j\right|^2 + (1-\alpha)\left|\beta_j\right| \right) \le t$$

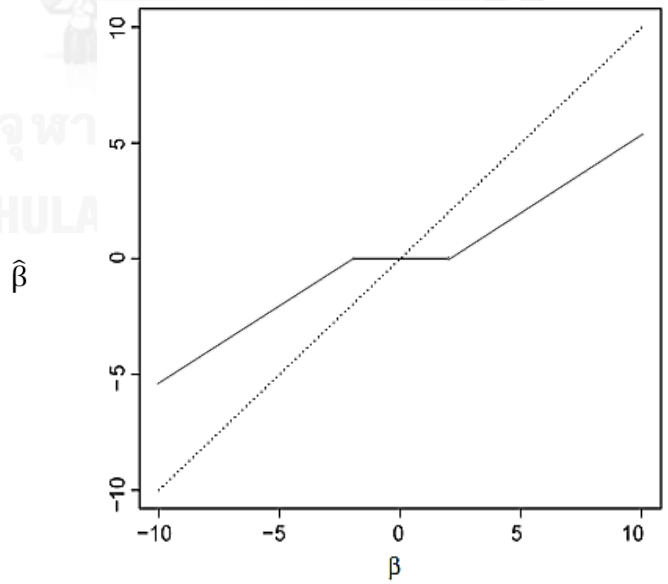
เมื่อ $\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right|^2+(1-\alpha)\left|\beta_j\right|\right)$ เป็นพจน์ค่าปรับ (Penalty term) วิธีอีลาสติคเน็ต

t เป็นพารามิเตอร์ปรับค่า (Tuning Parameter)

เงื่อนไขข้างต้น หาก $\alpha=0$ การประมาณค่าสัมประสิทธิ์ข้างต้นจะกลายเป็นการประมาณค่าสัมประสิทธิ์ด้วยการถดถอยแลสโซ่ และหาก $\alpha=1$ การประมาณค่าสัมประสิทธิ์ข้างต้นจะกลายเป็นการประมาณค่าสัมประสิทธิ์ด้วยการถดถอยริดจ์ทันที ดังนั้นค่าของ α สำหรับพจน์ค่าปรับของอีลาสติคเน็ตจะมีค่าระหว่าง 0 ถึง 1

การประมาณพารามิเตอร์ t สำหรับวิธีอีลาสติคเน็ตนั้นสามารถใช้วิธี Ten-fold Cross-validation ที่ให้ค่าประมาณของค่าคลาดเคลื่อนการทำนายน้อยที่สุดได้เฉกเช่นเดียวกับวิธีแลสโซ่

ในพจน์แรกของพจน์ค่าปรับสามารถเปลี่ยนคุณลักษณะที่มีความสัมพันธ์กันสูงให้มีความสัมพันธ์กันในระดับเฉลี่ยได้ ในขณะที่พจน์ที่สองจะสามารถทำให้เกิดผลลัพธ์ของสัมประสิทธิ์การถดถอยบางตัวนั้นมีค่าเป็นศูนย์ได้ นอกจากนี้วิธีอีลาสติคเน็ตยังสามารถวิเคราะห์ผลกระทบกลุ่ม เมื่อตัวแปรอิสระนั้นมีความสัมพันธ์กันสูงมาก ซึ่งส่งผลให้หากถูกคัดออกหรือนำเข้าสู่สมการจะมีแนวโน้มในการถูกกระทำกันเป็นกลุ่ม วิธีอีลาสติคเน็ตนี้มีประโยชน์อย่างมากเมื่อจำนวนตัวแปรอิสระนั้นมีมากกว่าขนาดตัวอย่างมากๆ



ภาพที่ 2.1.3 แสดงคุณลักษณะของการประมาณค่าแบบ Elastic Net (——) โดยเทียบกับคุณลักษณะการประมาณค่าแบบ OLS (·····)

จากภาพที่ 2.1.3 แสดงให้เห็นว่าการประมาณค่าแบบ Elastic Net นั้นจะให้ค่าประมาณสัมประสิทธิ์ต่ำกว่าการประมาณค่าแบบ OLS โดยเฉพาะเมื่อค่าสัมประสิทธิ์ที่แท้จริงมีค่ามากขึ้น ค่าประมาณสัมประสิทธิ์จะมีค่าลดลงมากยิ่งขึ้น แต่เมื่อค่าสัมประสิทธิ์ที่แท้จริงนั้นมีค่าลู่เข้าสู่ค่าคงที่ค่าหนึ่ง Elastic Net จะปรับให้ค่าประมาณของสัมประสิทธิ์ตัวนั้นมีค่าเป็นศูนย์

2.1.4 วิธีแลสโซ่ปรับปรุง (Adaptive LASSO)

วิธีนี้เป็นการเพิ่มค่าน้ำหนักให้กับพจน์ค่าปรับวิธีแลสโซ่ ซึ่งจะให้ตัวประมาณค่าสัมประสิทธิ์การถดถอยมีความคงเส้นคงวาโดยที่ยังคงไว้ซึ่งคุณสมบัติของการคัดเลือกตัวแปรแบบวิธีแลสโซ่ การประมาณค่าสัมประสิทธิ์การถดถอยแลสโซ่ปรับปรุงสามารถคำนวณได้ดังนี้

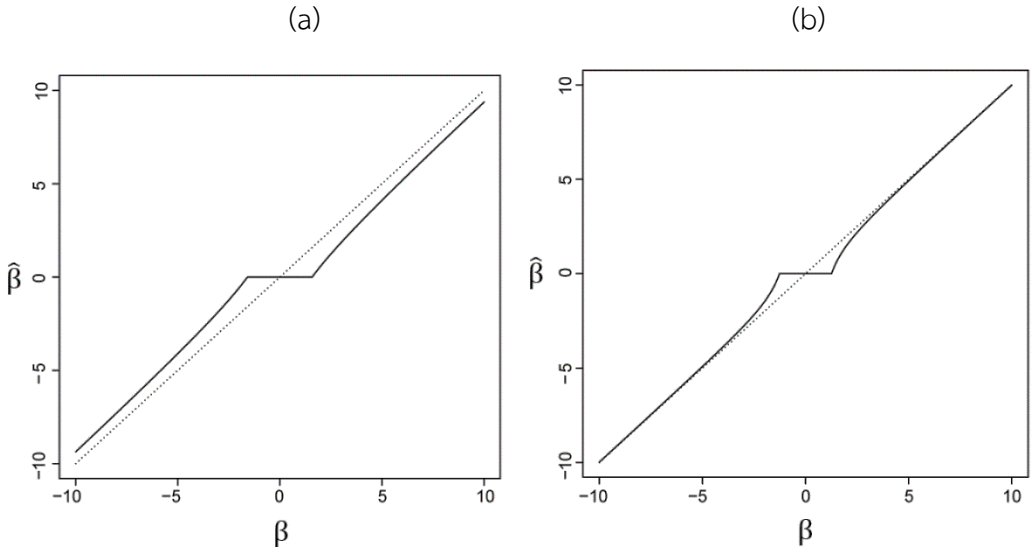
$$\hat{\beta}_{Adap.lasso} = \arg\min_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\right\} \quad \text{โดยที่} \quad \sum_{j=1}^{p} w_j|\beta_j| \le t \quad , \quad w_j = \frac{1}{|\hat{\beta}_j|^{\nu}}$$

เมื่อ $\sum_{j=1}^{p} w_j|\beta_j|$ เป็นพจน์ค่าปรับ (Penalty term) วิธีแลสโซ่ปรับปรุง

t เป็นพารามิเตอร์ปรับค่า (Tuning Parameter)

$\hat{\beta}_j$ เป็นค่าประมาณสัมประสิทธิ์ที่ได้จากวิธีกำลังสองน้อยสุด หากไม่สามารถหาค่าได้จากวิธีดังกล่าว สามารถใช้ค่าประมาณสัมประสิทธิ์ที่ได้จากการถดถอยริดจ์แทนได้

ν เป็นพารามิเตอร์ที่มีค่ามากกว่า 0



ภาพที่ 2.1.4 แสดงคุณลักษณะของการประมาณค่าแบบ Adaptive Lasso **(——)** โดยเทียบกับคุณลักษณะการประมาณค่าแบบ OLS (**·····**) เมื่อ (a) ν = 0.5 (b) ν = 2

จากภาพที่ 2.1.4 แสดงให้เห็นว่าการประมาณค่าแบบ Adaptive Lasso นั้นจะให้ค่าประมาณสัมประสิทธิ์ต่ำกว่าการประมาณค่าแบบ OLS ไม่มากนัก ที่ ν มีค่าน้อยจะลดค่าประมาณสัมประสิทธิ์ได้ดีกว่าที่ ν มีค่ามาก แต่เมื่อค่าสัมประสิทธิ์ที่แท้จริงนั้นมีค่าลู่เข้าสู่ค่าคงที่ค่าหนึ่ง Adaptive Lasso จะปรับให้ค่าประมาณของสัมประสิทธิ์ตัวนั้นมีค่าเป็นศูนย์

2.2 เกณฑ์ที่ใช้ในการตัดสิน

แนวทางในการพิจารณาว่าวิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์วิธีการใดมีความเหมาะสมสำหรับข้อมูลที่มีค่าสัมประสิทธิ์การถดถอยที่มีขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์ ได้แก่ ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive) ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (Mean Absolute Error) โดยวิธีการคัดเลือกตัวแปรและการประมาณค่าวิธีใดที่ให้ค่าเหล่านี้ต่ำโดยสอดคล้องกัน จะถือว่ากรณีนั้นมีความเหมาะสมกับข้อมูลที่มีค่าสัมประสิทธิ์การถดถอยที่มีขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์

2.2.1 ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive)

เป็นค่าของความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่ได้จากการคัดเลือกตัวแปรอิสระแบบต่างๆ โดยค่าสัมประสิทธิ์ที่ประมาณได้นั้นให้ค่าไม่เท่ากับศูนย์ในขณะที่ค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่แท้จริงมีค่าเป็นศูนย์ ซึ่งสามารถคำนวณได้จากสูตร

$$FP = \sum_{j=1}^{p} 1_{\{\beta_j = 0 \text{ and } \hat{\beta}_j \neq 0\}}$$

เมื่อ p เป็นจำนวนตัวแปรอิสระ

ความหมาย คือ ตัวแปรอิสระที่ j นั้นแท้จริงไม่มีผลกระทบต่อสมการการถดถอยเชิงเส้น แต่วิธีการคัดเลือกตัวแปรเหล่านั้นกลับให้ค่าที่ผิดพลาดเป็น ตัวแปรอิสระที่ j นั้นมีผลกระทบต่อสมการการถดถอยเชิงเส้น ซึ่งถือเป็นข้อผิดพลาดที่ไม่สมควรจะเกิดขึ้น ดังนั้นการคัดเลือกตัวแปรอิสระที่เหมาะสมควรให้ค่าความผิดพลาดในการตรวจจับเชิงบวกมีค่าน้อย จึงจะแสดงว่าการคัดเลือกตัวแปรอิสระวิธีนั้นๆ มีประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่ดี

2.2.2 ค่าความผิดพลาดในการตรวจจับเชิงลบ (False Negative)

เป็นค่าของความผิดพลาดในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่ได้จากการคัดเลือกตัวแปรอิสระแบบต่างๆ โดยค่าสัมประสิทธิ์ที่ประมาณได้นั้นให้ค่าเป็นศูนย์ในขณะที่ค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่แท้จริงมีค่าไม่เท่ากับศูนย์ ซึ่งสามารถคำนวณได้จากสูตร

$$FN = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0 \text{ and } \hat{\beta}_j = 0\}}$$

เมื่อ p เป็นจำนวนตัวแปรอิสระ

ความหมาย คือ ตัวแปรอิสระที่ j นั้นแท้จริงมีผลกระทบต่อสมการการถดถอยเชิงเส้น แต่วิธีการคัดเลือกตัวแปรเหล่านั้นกลับให้ค่าที่ผิดพลาดเป็น ตัวแปรอิสระที่ j นั้นไม่มีผลกระทบต่อสมการการถดถอยเชิงเส้น ซึ่งถือเป็นข้อผิดพลาดที่ไม่สมควรจะเกิดขึ้น ดังนั้นการคัดเลือกตัวแปรอิสระที่เหมาะสมจึงควรให้ค่าความผิดพลาดในการตรวจจับเชิงลบนี้มีค่าน้อย จึงจะแสดงว่าการคัดเลือกตัวแปรอิสระวิธีนั้นๆ มีประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นที่ดี

2.2.3 ค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (Mean Absolute Error)

เป็นการตรวจสอบความแตกต่างระหว่างค่าสัมประสิทธิ์การถดถอยเชิงเส้นแท้จริงกับค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นที่ได้จากการคัดเลือกตัวแปรแบบต่างๆ โดยการวัดขนาดของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยที่แท้จริงกับค่าสัมประสิทธิ์การถดถอยที่ได้จากการประมาณ หากค่าคลาดเคลื่อนสมบูรณ์โดยเฉลี่ยนี้มีค่าน้อย แสดงว่าวิธีการคัดเลือกตัวแปรนั้นๆ สามารถประมาณค่าสัมประสิทธิ์ได้ใกล้เคียงกับค่าจริง ซึ่งสามารถคำนวณได้จากสูตร

$$MAE = \frac{\sum_{j=1}^{p}\left|\hat{\beta}_j - \beta_j\right|}{p}$$

เมื่อ p เป็นจำนวนตัวแปรอิสระ

นอกจากนี้ยังมีการพิจารณาประสิทธิภาพโดยจำแนกตามวัตถุประสงค์ของเกณฑ์วัดประสิทธิภาพด้วย ซึ่งได้แก่ การพิจารณาในแง่ของความสัมพันธ์ระหว่างตัวแปร โดยจะพิจารณาจากค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงบวก และค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ แต่ในงานวิจัยนี้จะสนใจที่เกณฑ์ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ เนื่องจากต้องการพิสูจน์ว่าวิธีการคัดเลือกตัวแปรที่เหมาะสมกับข้อมูลตามขอบเขตที่กำหนดที่สุดนั้นจะยังคงรักษาตัวแปรอิสระที่สัมประสิทธิ์การถดถอยนั้นมีค่าขนาดเล็กหรือค่าเข้าใกล้ศูนย์ให้คงอยู่ในสมการการถดถอยได้ และการพิจารณาในอีกแง่หนึ่งคือความแม่นยำการทำนายซึ่งจะพิจารณาจากค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย

บทที่ 3

วิธีการดำเนินการศึกษา

ในงานวิจัยครั้งนี้เป็นการศึกษาการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ในกรณีที่ค่าสัญญาณมีค่าเข้าใกล้ศูนย์มากๆ เพื่อหาวิธีการคัดเลือกตัวแปรที่เหมาะสมกับตัวแบบเชิงเส้นที่มีค่าสัมประสิทธิ์บางตัวเป็นศูนย์และมีค่าสัญญาณขนาดเล็ก โดยการวิเคราะห์ข้อมูลทั้งหมดทำบนโปรแกรม R เวอร์ชัน 3.0.3 โดยวิเคราะห์ครอบคลุมตามขอบเขตของการวิจัย ในบทนี้จะกล่าวถึงขอบเขตของการศึกษา และขั้นตอนในการดำเนินการศึกษา ตามลำดับ

3.1 ขอบเขตของการศึกษา

ในการวิจัยครั้งนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

3.1.1 รูปแบบการความสัมพันธ์การถดถอยเชิงเส้น คือ

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$$

โดยที่ $\boldsymbol{y} = (y_1, ..., y_n)^T$ เป็นเวกเตอร์ของตัวแปรตอบสนองขนาด n

$\beta = (\beta_1, ..., \beta_p)^T$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p

$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ เป็นเมตริกซ์ตัวแปรพยากรณ์ขนาด $n \times p$ ที่มีการแจกแจงแบบปกติที่ค่าเฉลี่ยเป็นศูนย์ มีเมตริกซ์ความแปรปรวนร่วมเป็น $\boldsymbol{\Sigma} = \mathrm{cov}[\boldsymbol{x}_i, \boldsymbol{x}_j]$

$\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n

เมื่อ n เป็นขนาดตัวอย่าง

p เป็นจำนวนตัวแปรอิสระ

3.1.2 ค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ $\boldsymbol{x}_i$ และ $\boldsymbol{x}_j$ (ρ) เป็น $\gamma^{|i-j|}$ เมื่อ $i = 1, 2, ..., p$, $j = 1, 2, ..., p$ และ $i \neq j$ โดยกำหนดให้ γ มี 3 ระดับ คือ 0.1, 0.5 และ 0.9

3.1.3 ขนาดตัวอย่าง 2 ระดับ ได้แก่ 20 และ 100

3.1.4 จำนวนตัวแปรอิสระจะกำหนดตามอัตราส่วนของจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง ซึ่งแบ่งออกเป็น 2 กรณี ดังนี้

กรณีที่ 1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ กำหนดให้อัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง $(p : n)$ มี 3 ระดับ ได้แก่

ตารางที่ 3.1.4.1 ขอบเขตข้อมูลเริ่มต้นกรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง โดยค่าอัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่างเป็น 1:5 1:2 และ 4:5

อัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง	$p : n$
1:5	4:20
	20:100
1:2	10:20
	50:100
4:5	16:20
	80:100

กรณีที่ 2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ กำหนดให้อัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง $(p : n)$ มี 3 ระดับ ได้แก่

ตารางที่ 3.1.4.2 ขอบเขตข้อมูลเริ่มต้นกรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง โดยค่าอัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่างเป็น 2:1 5:1 และ 10:1

อัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง	$p : n$
2:1	40:20
	200:100
5:1	100:20
	500:100
10:1	200:20
	1000:100

3.1.5 ค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์ (Nonzero Coefficient) จะถูกกำหนดให้มีค่าเข้าใกล้ศูนย์ (Small Coefficient) โดยมีค่าอยู่ระหว่าง -1 ถึง 1 ซึ่งมีจำนวนเป็นร้อยละ 10, 40 และ 70 ของจำนวนตัวแปร ส่วนจำนวนสัมประสิทธิ์การถดถอยที่เหลือจะถูกกำหนดให้มีค่าเป็น 0

ตารางที่ 3.1.5.1 ขอบเขตข้อมูลเริ่มต้นกรณีค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเป็นศูนย์และค่าไม่เป็นศูนย์ เมื่อจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง กรณีขนาดตัวอย่างเท่ากับ 100

n	p	ร้อยละของ β ที่ไม่เท่ากับ 0	จำนวนของ β ที่ไม่เท่ากับ 0	จำนวนของ β ที่เท่ากับ 0
100	20	10	2	18
		40	8	12
		70	14	6
100	50	10	5	45
		40	20	30
		70	35	15
100	80	10	8	72
		40	32	48
		70	56	24

3.1.6 ค่าความคลาดเคลื่อนมีการแจกแจงแบบปกติที่ค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 1 นั่นคือ $\varepsilon \sim N(0,1)$

3.1.7 วิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์สำหรับกรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ ได้แก่ วิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง ส่วนกรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ ได้แก่ วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง

3.1.8 จำลองข้อมูลตามสถานการณ์ที่แตกต่างกันตามกำหนด โดยจำลองแต่ละสถานการณ์เป็นจำนวน 500 รอบ

3.2 ขั้นตอนการดำเนินการศึกษา

3.2.1 ศึกษาค้นคว้าเอกสารและข้อมูลที่เกี่ยวข้องกับงานวิจัย

3.2.2 กำหนดเงื่อนไขและขอบเขตของการวิจัย

- ค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ (ρ)
- ขนาดตัวอย่าง (n)
- จำนวนตัวแปรอิสระ (p)
- ค่าสัมประสิทธิ์การถดถอยที่แท้จริง (β)

3.2.3 จำลองข้อมูลของตัวแปรพยากรณ์ (x) ตามการแจกแจงและขอบเขตที่ต้องการศึกษา

3.2.4 คำนวณค่าตอบสนอง (y) ดังสมการ

$$y = X\beta + \varepsilon$$

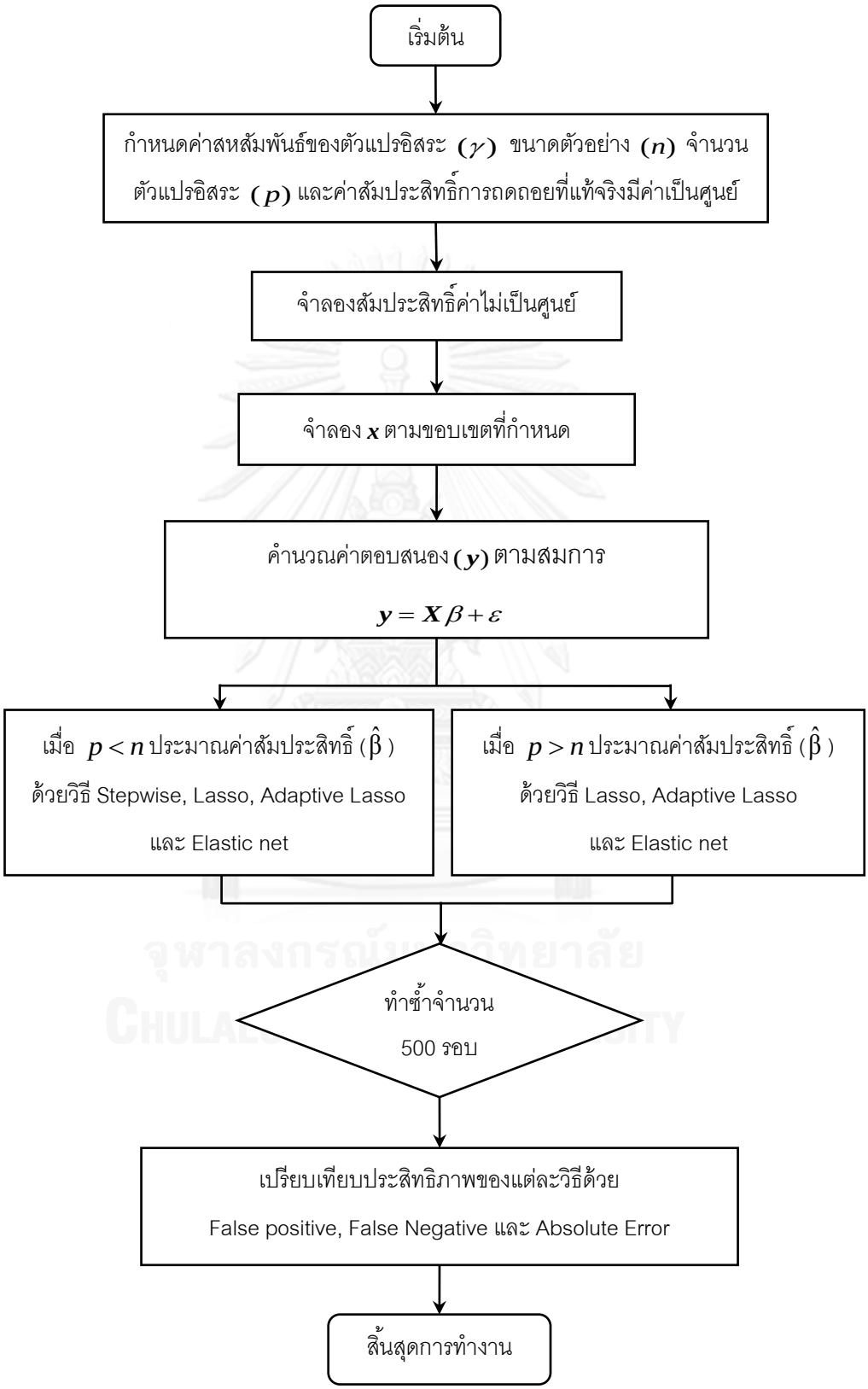
3.2.5 ค่า x และ y ที่ได้จากการคำนวณ จะนำมาประมาณค่าสัมประสิทธิ์การถดถอย $(\hat{\beta})$ โดยใช้วิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง

3.2.6 คำนวณค่าเฉลี่ยของประสิทธิภาพการทำนาย โดยใช้ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงบวก (False Positive) ค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ (False Negative) และค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (Mean Absolute Error)

3.2.7 ทำการวิเคราะห์ข้อมูลและสรุปผลที่ได้จากการศึกษา

3.3 ขั้นตอนการทำงานของโปรแกรม

ภาพที่ 3.1 แผนภาพการเขียนโปรแกรม



บทที่ 4

ผลการวิเคราะห์ข้อมูล

การศึกษาในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการคัดเลือกตัวแปรอิสระและประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบเชิงเส้นทั้ง 4 วิธี ได้แก่ วิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง สำหรับข้อมูลที่มีผลกระทบขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์ ซึ่งเกณฑ์ที่ใช้ในการพิจารณา ได้แก่ ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) และค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) หากวิธีใดให้ค่าเกณฑ์ดังกล่าวต่ำสุดโดยสอดคล้องกัน ถือว่าเป็นวิธีที่มีประสิทธิภาพและเหมาะกับข้อมูลตามขอบเขตที่กำหนด

ในการนำเสนอผลการวิจัยจะแสดงในรูปของตารางและกราฟในการเปรียบเทียบ โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่างๆ ดังนี้

n	แทน	ขนาดตัวอย่าง
p	แทน	จำนวนตัวแปรอิสระ
$p:n$	แทน	อัตราส่วนจำนวนตัวแปรอิสระต่อขนาดตัวอย่าง
γ	แทน	ค่าความสัมพันธ์ระหว่างตัวแปรอิสระ
%nonzero	แทน	ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เมื่อเทียบกับจำนวนตัวแปรอิสระ
ST	แทน	วิธีเพิ่มลดตัวแปรแบบขั้นตอน (Stepwise)
LA	แทน	วิธีแลสโซ่ (LASSO)
EN	แทน	วิธีอีลาสติคเน็ต (Elastic Net)
AD	แทน	วิธีแลสโซ่ปรับปรุง (Adaptive Lasso)
FP	แทน	ค่าความผิดพลาดในการตรวจจับเชิงบวก
FN	แทน	ค่าความผิดพลาดในการตรวจจับเชิงลบ

MAE แทน ค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย

เมื่อพิจารณาตามขอบเขตการศึกษาแล้ว จะมีทั้งสิ้น 108 กรณีศึกษา ดังนี้

ขอบเขตการศึกษากรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

n	p	γ = 0.1	γ = 0.5	γ = 0.9
20	4	%nonzero = 10, 40, 70		
	10			
	16			
100	20			
	50			
	80			

ขอบเขตการศึกษากรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

n	p	γ = 0.1	γ = 0.5	γ = 0.9
20	40	%nonzero = 10, 40, 70		
	100			
	200			
100	20			
	50			
	80			

การนำเสนอผลการศึกษาจะแบ่งออกเป็น 3 ส่วน โดยแบ่งตามเกณฑ์การวัดประสิทธิภาพทั้ง 3 เกณฑ์ มีรายละเอียดในแต่ละหัวข้อดังนี้

4.1 การเปรียบเทียบค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.1.1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

4.1.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

4.2 การเปรียบเทียบค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) ที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.2.1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

4.2.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

4.3 การเปรียบเทียบค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) ที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.3.1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

4.3.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

4.1 การเปรียบเทียบค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวกที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.1.1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

ตารางที่ 4.1.1.1 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
4	10	0.7160	1.2860	1.8140	**0.4000**	0.8100	1.2680	1.8860	**0.4880**	0.9800	1.4020	1.7040	**0.6100**
		(0.8079)	(1.2261)	(0.9061)	(0.7274)	(0.8941)	(1.2079)	(0.9310)	(0.8119)	(0.8816)	(1.1896)	(0.8775)	(0.7869)
	40	0.4420	0.9560	1.3100	**0.3640**	0.5260	0.9660	1.2900	**0.3720**	0.5860	1.9740	1.1820	**0.4300**
		(0.6127)	(0.8852)	(0.7176)	(0.6575)	(0.6680)	(0.8616)	(0.7204)	(0.6344)	(0.6476)	(0.8165)	(0.7084)	(0.6114)
	70	0.2100	0.5620	0.6980	0.2080	0.2520	0.5200	0.6720	**0.2100**	0.3000	0.5440	0.6400	**0.2380**
		(0.4077)	(0.4966)	(0.4596)	(0.4063)	(0.4346)	(0.5001)	(0.4700)	(0.4077)	(0.4587)	(0.4986)	(0.4805)	(0.4263)
10	10	2.9480	2.6240	4.4520	**0.6260**	3.2340	2.7580	4.2860	**0.7200**	3.3800	2.8980	3.9320	**0.9080**
		(1.8396)	(2.8697)	(1.9191)	(1.3060)	(1.8071)	(2.8326)	(2.0290)	(1.3480)	(1.8694)	(2.7045)	(2.2328)	(1.3443)
	40	1.9480	2.2320	3.7870	**0.8480**	2.2900	2.6800	3.6740	**1.0200**	2.3500	2.3400	2.99600	**0.8060**
		(1.2857)	(1.9891)	(1.5014)	(1.5464)	(1.4304)	(2.0780)	(1.6288)	(1.4994)	(1.3532)	(1.8448)	(1.5941)	(1.1606)
	70	1.1150	1.7880	2.0540	**0.6420**	0.9760	1.3560	1.8620	**0.6740**	1.1660	1.2160	1.6580	**0.6080**
		(0.8396)	(1.0404)	(0.9054)	(0.9646)	(0.8419)	(1.1313)	(0.9256)	(0.8520)	(0.8900)	(1.0448)	(0.9894)	(0.7610)
16	10	5.3540	2.5580	6.7020	**1.1340**	5.4500	2.7220	6.1060	**0.9800**	5.4960	3.3800	5.64800	**1.2060**
		(1.8438)	(3.2735)	(3.1858)	(2.2284)	(1.7475)	(3.3716)	(3.0208)	(1.9584)	(1.6946)	(3.4865)	(3.1787)	(1.7566)
	40	3.3460	2.0840	5.2440	**1.3980**	3.5080	2.3240	5.2140	**1.4180**	3.4280	2.5260	4.4420	**1.3160**
		(1.4624)	(2.3426)	(2.3063)	(1.9481)	(1.4021)	(2.4826)	(2.2464)	(1.9675)	(1.4088)	(2.1386)	(2.3241)	(1.6497)
	70	1.5780	1.1040	2.5880	**0.8720**	1.5420	1.1520	2.6020	**0.9100**	1.5260	1.1740	2.1700	**0.7380**
		(0.9475)	(1.2265)	(1.1820)	(1.1551)	(0.9789)	(1.2649)	(1.1791)	(1.1526)	(0.9608)	(1.1944)	(1.2182)	(0.9876)

จากตารางที่ 4.1.1.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 16 นั้นให้แนวโน้มค่าเฉลี่ย FP ที่ไม่ชัดเจน
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 4 และ 10 ในทุกระดับของ %nonzero นั้นให้แนวโน้มค่าเฉลี่ย FP ที่ไม่ชัดเจน
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี p = 4, %nonzero = 10 กรณี p = 10, %nonzero = 40 และกรณี p = 16, %nonzero = 70 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.5 และ 0.9 มีค่าสูงสุดและต่ำสุด
   - วิธี Adaptive Lasso ที่ระดับ p = 4 นั้นให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้น ส่วนกรณี p = 10 และ 16 นั้น ให้แนวโน้มค่าเฉลี่ย FP ที่ไม่ชัดเจน
2. เมื่อร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise วิธี Lasso และวิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงในทุกกรณี
   - วิธี Adaptive Lasso ในทุกระดับของ γ ที่ระดับ p = 4 นั้นให้แนวโน้มค่าเฉลี่ย FP ลดลง ที่ระดับ p = 10 นั้นให้แนวโน้มค่าเฉลี่ย FP ไม่ชัดเจน และที่ระดับ p = 16 นั้นให้ค่าเฉลี่ย FP ที่ระดับ %nonzero = 40 และ 70 มีค่าสูงสุดและต่ำสุดตามลำดับ
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise วิธี Elastic Net และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นในทุกกรณี
   - วิธี Lasso ให้ค่าเฉลี่ย FP ที่ระดับ p = 4 และ 10 มีค่าต่ำสุดและสูงสุดโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.9, %nonzero = 10 และ 40 นั้นให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้น
4. ค่าเฉลี่ย FP ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso ให้ค่าเฉลี่ย FP ต่ำที่สุด

ตารางที่ 4.1.1.2 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
20	10	1.7300 (1.2917)	6.5240 (6.1775)	9.0220 (3.1127)	0.8240 (1.6144)	1.8760 (1.3327)	5.8600 (5.7868)	8.5100 (3.0007)	0.8300 (1.4565)	2.1440 (1.3546)	6.1360 (5.2727)	7.0840 (3.1190)	1.2940 (1.7267)
	40	1.1200 (0.9877)	6.5520 (3.8132)	8.4960 (1.9395)	1.7620 (2.0962)	1.2160 (1.0824)	6.2240 (3.7787)	8.2580 (1.9096)	2.0080 (2.0553)	1.4800 (1.1766)	5.7980 (3.7642)	7.3000 (2.3836)	2.23800 (2.1077)
	70	0.5800 (0.6874)	4.1640 (1.8783)	5.0460 (1.0049)	1.7160 (1.4573)	0.6020 (0.7407)	4.2240 (1.8112)	5.0020 (1.0660)	1.9200 (1.4987)	0.7400 (0.8377)	3.4400 (2.0422)	4.2600 (1.3084)	1.7200 (1.4933)
50	10	2.5540 (1.4858)	12.4380 (10.9470)	21.5100 (5.1671)	1.8180 (2.4700)	2.8680 (1.5425)	13.9180 (11.1422)	20.3880 (5.6669)	2.2220 (3.0389)	3.3320 (1.5744)	13.1160 (10.3210)	16.8980 (6.3029)	3.1220 (3.0968)
	40	1.8000 (1.2550)	16.1080 (7.0033)	21.8160 (3.2790)	4.6900 (3.4869)	1.9100 (1.3332)	16.0820 (6.9892)	21.1120 (3.4045)	5.2600 (3.6164)	2.0980 (1.2930)	12.8920 (6.8675)	17.7220 (4.2948)	5.4760 (3.8082)
	70	0.9180 (0.9236)	10.5040 (3.3747)	12.5580 (1.6529)	4.9300 (2.9289)	0.8720 (0.8973)	10.4520 (3.4157)	12.3120 (1.7495)	5.3220 (2.8359)	1.0680 (0.9764)	7.7560 (3.8149)	10.2560 (2.3515)	4.2000 (2.5817)
80	10	4.7700 (1.6069)	13.8100 (13.2380)	32.9600 (7.7285)	2.8440 (3.1988)	4.9120 (1.6154)	13.4200 (13.1379)	30.9280 (7.5637)	3.0580 (3.7216)	5.0420 (1.5929)	15.2720 (13.2729)	25.2280 (8.1038)	4.7640 (4.3287)
	40	3.2100 (1.5423)	17.3840 (8.9604)	33.3460 (4.4652)	8.4240 (5.5090)	3.2840 (1.5046)	17.0900 (8.6658)	32.1860 (4.7312)	9.6380 (5.6529)	3.4640 (1.5523)	14.8360 (8.4581)	26.4920 (6.3997)	9.0240 (5.4231)
	70	1.7060 (1.2126)	12.6520 (4.7556)	19.5420 (2.5332)	9.2600 (4.2095)	1.7340 (1.2419)	11.9340 (4.6035)	18.8340 (2.7939)	9.1120 (4.1369)	1.8140 (1.2420)	8.8640 (4.4218)	15.3060 (3.6501)	6.5120 (3.8019)

จากตารางที่ 4.1.1.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 10, %nonzero = 70 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.5 และ 0.9 มีค่าต่ำสุดและสูงสุด ตามลำดับ
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี %nonzero = 10 ในทุกระดับของ p นั้นให้แนวโน้มค่าเฉลี่ย FP ที่ไม่ชัดเจน
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงในทุกกรณี
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี %nonzero = 70 ในทุกระดับของ p นั้นให้แนวโน้มค่าเฉลี่ย FP ที่ไม่ชัดเจน
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise ให้แนวโน้มค่าเฉลี่ย FP ลดลงในทุกกรณี
   - วิธี Lasso ให้ค่าเฉลี่ย FP ที่ระดับ %nonzero = 40 และ 70 สูงสุดและต่ำสุดตามลำดับ ยกเว้นกรณี γ = 0.9 นั้นให้แนวโน้มค่าเฉลี่ย FP ลดลง
   - วิธี Elastic Net ให้ค่าเฉลี่ย FP ที่ระดับ %nonzero = 40 และ 70 สูงสุดและต่ำสุดตามลำดับ ยกเว้นกรณี p = 4, γ = 0.1, 0.5 นั้นให้แนวโน้มค่าเฉลี่ย FP ลดลง
   - วิธี Adaptive Lasso ให้ค่าเฉลี่ย FP ที่ระดับ %nonzero = 10 และ 40 ต่ำสุดและสูงสุดตามลำดับ ยกเว้นกรณี p = 10, γ = 0.1, 0.5 และ p = 16, γ = 0.1 นั้นให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้น
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธีให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นในทุกกรณี
4. ที่ระดับ %nonzero = 10 การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso ให้ค่าเฉลี่ย FP ต่ำที่สุด และที่ระดับ %nonzero = 40 และ 70 การคัดเลือกตัวแปรด้วยวิธี Stepwise ให้ค่าเฉลี่ย FP ต่ำที่สุด

4.1.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

ตารางที่ 4.1.2.1 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	$\gamma = 0.1$			$\gamma = 0.5$			$\gamma = 0.9$		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
40	10	7.6640 (5.3734)	10.9060 (4.6936)	**1.2560** (1.9404)	6.9140 (5.6140)	9.7760 (4.2779)	**1.1700** (1.9283)	7.7340 (4.8302)	10.4520 (4.6501)	**2.0300** (2.4348)
	40	5.1800 (3.9680)	7.0540 (3.3616)	**1.1840** (1.7079)	4.9260 (3.4162)	7.1860 (3.3466)	**1.7180** (1.8979)	6.5700 (3.2482)	7.8260 (2.9073)	**1.9380** (1.6893)
	70	2.5700 (1.9912)	3.2880 (2.1229)	**0.5000** (0.8704)	2.7620 (2.1326)	3.5280 (2.1531)	**0.6900** (1.2284)	3.4600 (1.9039)	3.9660 (1.6138)	**1.2000** (1.2227)
100	10	6.3320 (5.3693)	12.1000 (4.3217)	**1.2340** (2.0194)	7.5540 (5.7020)	12.9760 (4.7021)	**1.3800** (1.8450)	10.5300 (5.0321)	15.5800 (5.6679)	**3.3760** (2.9082)
	40	4.3680 (4.1646)	7.4720 (3.0829)	**1.1240** (1.8162)	4.6640 (3.9497)	8.4520 (3.6024)	**1.2000** (1.7115)	7.5560 (3.2047)	11.3920 (3.5672)	**3.0480** (2.4450)
	70	2.3960 (2.2603)	4.1700 (2.0844)	**0.4480** (0.8447)	2.3260 (2.1783)	4.2360 (2.4031)	**0.5940** (1.1468)	3.5540 (1.7774)	5.7540 (2.4137)	**1.6540** (1.3587)
200	10	5.2820 (5.5960)	12.1200 (4.4930)	**0.6800** (1.5432)	6.1960 (5.7379)	13.3300 (5.3486)	**1.3740** (2.3035)	10.3860 (5.3348)	18.6600 (6.0339)	**3.3500** (3.3162)
	40	3.7340 (3.8895)	8.5560 (3.9858)	**1.3840** (1.9268)	4.5360 (4.2485)	8.8540 (4.0386)	**1.1100** (1.9169)	7.1440 (3.5902)	12.4820 (4.5847)	**2.7900** (2.7827)
	70	1.6140 (2.1268)	4.2340 (1.9738)	**0.6780** (1.0736)	2.0700 (2.2395)	4.1900 (2.3898)	**0.6000** (0.8040)	3.5440 (2.5042)	6.6540 (2.6605)	**1.3820** (1.3317)

จากตารางที่ 4.1.2.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี %nonzero = 40 ในทุกระดับของ p นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.5 และ 0.9 มีค่าสูงสุดและต่ำสุด
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso และวิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงในทุกกรณี
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.1, p = 200 และกรณี γ = 0.5, p = 40 นั้นให้ค่าเฉลี่ย FP ที่ระดับ %nonzero = 40 และ 70 มีค่าสูงสุดและต่ำสุด
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.9, %nonzero = 10, 40 นั้นให้ค่าเฉลี่ย FP ที่ระดับ p = 100 และ 200 มีค่าสูงสุดและต่ำสุด
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.5, %nonzero = 10 และกรณี γ = 0.9 ในทุกระดับของ %nonzero
4. ค่าเฉลี่ย FP ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso ให้ค่าเฉลี่ย FP ต่ำที่สุด รองลงมาเป็นวิธี Lasso และวิธี Elastic Net ตามลำดับ

ตารางที่ 4.1.2.2 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	$\gamma = 0.1$			$\gamma = 0.5$			$\gamma = 0.9$		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
40	10	50.6140 (18.4615)	59.8220 (9.9223)	7.8900 (5.0270)	47.0300 (20.2462)	57.0740 (10.3926)	7.4920 (4.8084)	49.2780 (18.9959)	49.8760 (10.3041)	9.4000 (4.6493)
	40	29.1660 (16.5577)	39.3000 (10.0438)	9.7560 (7.7941)	34.8740 (12.9441)	43.1160 (8.9712)	13.4320 (7.8910)	37.5800 (9.4411)	41.1960 (6.7878)	11.9600 (3.7844)
	70	13.1760 (8.2597)	19.0520 (6.8319)	4.1140 (4.9214)	16.5000 (7.1626)	20.5540 (5.7846)	6.7000 (4.4890)	19.9960 (5.8938)	20.4680 (4.1106)	6.6120 (2.5142)
100	10	40.5260 (23.9615)	65.1520 (16.7425)	12.5000 (10.8670)	47.0200 (24.8775)	67.2460 (16.9355)	12.8340 (10.6297)	61.3400 (15.1199)	80.3180 (11.8438)	23.0740 (6.5200)
	40	21.2560 (18.0462)	29.4620 (15.9021)	3.2020 (6.8372)	29.3160 (17.1798)	37.6660 (15.5636)	5.3920 (7.9594)	42.6880 (10.0100)	55.5820 (9.1178)	19.9480 (5.8040)
	70	11.6000 (9.9970)	13.3260 (6.8400)	1.0300 (2.2717)	13.5760 (8.4701)	16.9320 (7.8589)	2.9520 (4.5045)	20.5440 (5.8524)	26.9740 (5.3719)	10.2460 (3.8980)
200	10	29.5500 (26.9483)	48.2640 (21.9640)	4.0120 (8.7519)	37.6260 (27.4306)	52.3460 (20.8638)	6.4960 (10.687)	58.9280 (20.2867)	92.5280 (17.1532)	27.2960 (11.0218)
	40	19.1300 (19.4348)	27.9980 (14.9217)	1.3440 (1.9016)	23.9780 (20.4240)	33.3400 (15.9496)	3.2380 (5.9981)	44.1120 (11.1671)	64.1500 (11.7330)	20.2400 (8.6947)
	70	9.4000 (9.2299)	14.4320 (7.9738)	1.1800 (2.5201)	12.2380 (9.6272)	15.9640 (8.2522)	1.5520 (2.9996)	21.3760 (7.0834)	31.6840 (7.3180)	9.8800 (5.3092)

จากตารางที่ 4.1.2.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 200, %nonzero = 10 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.1 และ 0.5 มีค่าสูงสุดและต่ำสุดตามลำดับ
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 200, %nonzero = 10 นั้นให้แนวโน้มค่าเฉลี่ย FP ลดลง และที่ระดับของ %nonzero = 40, 70 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.1 และ 0.5 มีค่าต่ำสุดและสูงสุดตามลำดับ
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 200, %nonzero = 10 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.5 และ 0.9 มีค่าต่ำสุดและสูงสุดตามลำดับ และที่ระดับของ %nonzero = 40, 70 นั้นให้ค่าเฉลี่ย FP ที่ระดับ γ = 0.1 และ 0.5 มีค่าต่ำสุดและสูงสุดตามลำดับ
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso และวิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FP ลดลงในทุกกรณี
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี p = 200 ในทุกระดับของ γ ค่าเฉลี่ย FP ที่ระดับ %nonzero = 40 และ 70 มีค่าสูงสุดและต่ำสุด ตามลำดับ
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย FP ลดลงโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.9, %nonzero = 10 นั้นให้ค่าเฉลี่ย FP ที่ระดับ p = 200 และ 500 มีค่าต่ำสุดและสูงสุดตามลำดับ และที่ระดับของ %nonzero = 40, 70 นั้นให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้น
   - วิธี Elastic Net กรณี γ = 0.9 ในทุกระดับของ %nonzero ให้แนวโน้มค่าเฉลี่ย FP เพิ่มขึ้น ที่ระดับ %nonzero = 40, γ = 0.1, 0.5 และที่ระดับ %nonzero = 70, γ = 0.5 นั้นให้แนวโน้มค่าเฉลี่ย FP มีค่าลดลง
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FP ไม่ชัดเจน
4. ค่าเฉลี่ย FP ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso ให้ค่าเฉลี่ย FP ต่ำที่สุด รองลงมาเป็นวิธี Lasso และวิธี Elastic Net ตามลำดับ

4.2 การเปรียบเทียบค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.2.1 เมื่อจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

ตารางที่ 4.2.1.1 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
4	10	0.3280 (0.4700)	0.2820 (0.4504)	**0.1300** (0.3366)	0.4480 (0.4978)	0.3560 (0.4793)	0.2840 (0.4514)	**0.1520** (0.3594)	0.4760 (0.4999)	0.5220 (0.5000)	0.3500 (0.4774)	**0.2300** (0.4213)	0.5420 (0.4987)
	40	0.6420 (0.6684)	0.5320 (0.7308)	**0.2760** (0.4820)	0.8900 (0.7740)	0.7140 (0.6611)	0.6060 (0.7615)	**0.3040** (0.5061)	1.0060 (0.7610)	1.0340 (0.6676)	0.7460 (0.7712)	**0.5000** (0.5989)	1.1800 (0.6845)
	70	0.9800 (0.8444)	0.7360 (0.9717)	**0.3800** (0.6036)	1.3320 (1.0919)	1.1040 (0.8501)	0.7840 (0.9566)	**0.4580** (0.6551)	1.3860 (1.0970)	1.6480 (0.7933)	1.0660 (1.0333)	**0.7840** (0.7658)	1.7860 (0.8565)
10	10	0.3520 (0.4781)	0.3540 (0.4787)	**0.1900** (0.3927)	0.5660 (0.4961)	0.3620 (0.4811)	0.3800 (0.4859)	**0.2280** (0.4200)	0.5500 (0.4980)	0.5360 (0.4992)	0.4640 (0.4992)	**0.3320** (0.4714)	0.6940 (0.4613)
	40	1.4220 (0.9411)	1.1780 (1.2121)	**0.6900** (0.8290)	2.0540 (1.3002)	1.5580 (0.9319)	1.2940 (1.2739)	**0.7780** (0.8912)	2.2020 (1.2745)	2.2180 (0.9902)	1.9160 (1.3223)	**1.4820** (1.0918)	2.8680 (0.9862)
	70	2.6500 (1.2864)	1.8980 (1.8511)	**1.1160** (1.1770)	3.4500 (2.1496)	2.8940 (1.2238)	2.2660 (1.9367)	**1.3700** (1.3010)	3.8840 (2.0773)	3.8440 (1.3927)	3.3260 (2.0009)	**2.4300** (1.5171)	5.0300 (1.4959)
16	10	1.1220 (0.6841)	0.8840 (0.7561)	**0.4020** (0.5942)	1.1980 (0.7402)	1.1100 (0.6949)	0.9340 (0.7340)	**0.5140** (0.6250)	1.2360 (0.6879)	1.1980 (0.6897)	1.1420 (0.7609)	**0.8040** (0.7004)	1.4580 (0.6551)
	40	3.8380 (1.1792)	3.4740 (2.0371)	**1.4180** (1.3472)	4.2100 (2.1154)	3.8900 (1.1835)	3.5300 (2.0073)	**1.5660** (1.3818)	4.2180 (2.0225)	4.0980 (1.2663)	4.1440 (1.8618)	**2.7060** (1.5851)	5.1620 (1.4560)
	70	6.6680 (1.3286)	6.0740 (3.3519)	**2.3680** (2.0369)	6.8560 (3.5773)	6.8200 (1.4407)	6.4640 (3.2474)	**2.7300** (2.3021)	7.1280 (3.3355)	7.1340 (1.4478)	7.3420 (2.9151)	**4.4080** (2.4992)	8.6580 (2.3800)

จากตารางที่ 4.2.1.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น
   - วิธี Stepwise และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นโดยส่วนใหญ่
   - วิธี Lasso และวิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธีให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธีให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
4. ค่าเฉลี่ย FN ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Elastic Net ให้ค่าเฉลี่ย FN ต่ำที่สุด

ตารางที่ 4.2.1.2 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
20	10	1.2080	0.3120	**0.1620**	0.5220	1.2780	0.3400	**0.1740**	0.4840	1.4080	0.4960	**0.3700**	0.7740
		(0.6615)	(0.5686)	(0.3951)	(0.6560)	(0.6705)	(0.5771)	(0.3899)	(0.5987)	(0.6248)	(0.6862)	(0.5672)	(0.7294)
	40	4.8300	0.7100	**0.3700**	1.4860	4.9540	0.8720	**0.4260**	1.5940	5.5720	1.9320	**1.2060**	3.1000
		(1.1815)	(1.0753)	(0.6496)	(1.2494)	(1.1218)	(1.1723)	(0.6303)	(1.2839)	(1.1643)	(1.8027)	(1.2010)	(1.7216)
	70	8.5740	0.7140	**0.3280**	1.8180	8.7720	1.0060	**0.4720**	2.1560	9.9520	3.0820	**1.6940**	4.9020
		(1.2774)	(1.1946)	(0.5806)	(1.6279)	(1.2406)	(1.5002)	(0.7087)	(1.6471)	(1.2919)	(2.9456)	(1.4058)	(2.4699)
50	10	4.2640	0.8260	**0.5000**	1.2520	4.2560	0.8440	**0.5260**	1.3840	4.3920	1.4000	**1.0760**	2.1040
		(0.7947)	(0.9303)	(0.6441)	(1.0269)	(0.7795)	(0.9867)	(0.6887)	(1.1059)	(0.7179)	(1.1656)	(0.9445)	(1.1747)
	40	16.8660	1.7960	**0.9620**	3.7220	16.8800	2.2620	**1.2620**	4.1280	17.6080	5.7940	**3.5420**	8.7580
		(1.3942)	(1.8113)	(1.0482)	(2.1601)	(1.3495)	(2.2260)	(1.1934)	(2.2157)	(1.2121)	(3.6069)	(2.0210)	(3.0761)
	70	29.5100	2.1740	**1.0700**	5.1420	29.7340	2.7060	**1.3940**	5.9300	31.0100	9.6020	**5.2920**	14.5040
		(1.3949)	(2.2121)	(1.0580)	(2.7064)	(1.4210)	(2.6807)	(1.2352)	(2.9858)	(1.3877)	(6.4023)	(3.0208)	(5.2670)
80	10	7.1640	1.4960	**0.8100**	1.9940	7.1820	1.6680	**0.9160**	2.1200	7.3060	2.8540	**1.9240**	3.5480
		(0.8477)	(1.2842)	(0.8874)	(1.2858)	(0.8380)	(1.3585)	(0.9223)	(1.3210)	(0.7678)	(1.6352)	(1.2170)	(1.5007)
	40	28.9500	5.3140	**2.1180**	6.9220	29.0360	6.4820	**2.6980**	7.7020	29.4200	12.9800	**7.0380**	15.3520
		(1.3752)	(3.2749)	(1.5313)	(2.8831)	(1.3559)	(3.5821)	(1.6913)	(2.9791)	(1.3323)	(5.2596)	(2.9123)	(4.1297)
	70	50.5980	7.9460	**2.4240**	9.7160	50.7800	10.3680	**3.5120**	11.9000	51.4480	23.0080	**11.0960**	26.6080
		(1.4493)	(5.3864)	(1.7614)	(4.5993)	(1.4697)	(6.3287)	(2.1953)	(5.3747)	(1.4097)	(9.3776)	(4.8939)	(7.9315)

จากตารางที่ 4.2.1.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นโดยส่วนใหญ่
   - วิธี Lasso และวิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธีให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย การคัดเลือกตัวแปรทั้ง 4 วิธีให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
4. ค่าเฉลี่ย FN ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Elastic Net ให้ค่าเฉลี่ย FN ต่ำที่สุด

4.2.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

ตารางที่ 4.2.2.1 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1			γ = 0.5			γ = 0.9		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
40	10	1.8480 (1.2368)	1.4900 (0.9999)	2.9160 (1.0926)	2.2120 (1.1573)	1.6860 (0.9535)	3.1540 (0.9252)	2.5000 (1.0493)	1.9740 (0.9894)	3.2860 (0.7795)
	40	10.1260 (3.3099)	7.9520 (2.3714)	13.6500 (2.6643)	10.3040 (3.3979)	8.2300 (2.5100)	13.3200 (2.3393)	10.2840 (2.2432)	8.8820 (2.3019)	13.6260 (1.6254)
	70	18.9980 (5.1512)	16.2600 (3.8181)	24.9440 (3.1264)	18.7280 (4.524)	15.6580 (3.6359)	24.6560 (3.2361)	18.0880 (3.6061)	15.8320 (3.4232)	23.6080 (2.4985)
100	10	7.6860 (1.9010)	6.1480 (1.5892)	8.9620 (1.3098)	7.6160 (1.8029)	6.3740 (1.5086)	9.1340 (1.2115)	7.4440 (1.6038)	6.5540 (1.7137)	9.0420 (1.1274)
	40	35.4600 (4.0262)	32.2820 (3.4818)	38.4120 (2.1932)	35.2700 (4.0772)	31.2020 (3.0017)	38.2120 (2.3368)	33.1940 (2.7659)	30.0600 (3.2716)	37.0760 (2.1285)
	70	62.3760 (5.8390)	57.8280 (4.5536)	67.7960 (2.8330)	62.5120 (5.3475)	56.9220 (4.8963)	67.5480 (2.8120)	59.0600 (3.8868)	52.9800 (3.9286)	64.8840 (2.9173)
200	10	18.2000 (1.7408)	16.3660 (1.7087)	19.6280 (0.7217)	18.0200 (1.9226)	16.3220 (1.9482)	19.2780 (1.2215)	17.6720 (1.4568)	15.6760 (1.7353)	19.1680 (0.9921)
	40	76.4580 (3.5884)	72.2980 (3.0626)	78.6940 (1.7215)	75.9500 (3.5516)	72.2960 (3.3584)	78.7660 (1.7702)	73.7920 (3.0987)	69.3140 (3.9407)	77.6320 (2.2503)
	70	135.3460 (5.5235)	127.7460 (4.6354)	137.7820 (2.5528)	134.1680 (5.3119)	127.9420 (4.1726)	138.0120 (2.3720)	130.1880 (4.3863)	121.6940 (4.9435)	135.3800 (3.3808)

จากตารางที่ 4.2.2.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย FN ลดลงโดยส่วนใหญ่ ยกเว้นกรณี p = 40, %nonzero = 10 การคัดเลือกตัวแปรทั้ง 3 วิธีนั้นให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้น
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
4. ค่าเฉลี่ย FN ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Elastic Net ให้ค่าเฉลี่ย FN ต่ำที่สุด รองลงมาเป็นวิธี Lasso และวิธี Adaptive Lasso ตามลำดับ

ตารางที่ 4.2.2.2 ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1			γ = 0.5			γ = 0.9		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
200	10	3.4780 (2.2539)	**2.9880** (1.6816)	5.7560 (2.2535)	4.3420 (2.5474)	**3.5500** (1.7137)	7.0960 (2.4745)	6.9780 (2.0124)	**5.9920** (2.0025)	10.8480 (2.2906)
	40	41.3820 (15.7831)	**31.5740** (7.5949)	55.9980 (12.4093)	37.0560 (10.5767)	**30.4480** (6.0022)	52.5280 (10.4770)	38.7120 (6.7109)	**35.1500** (4.8936)	57.2840 (4.2333)
	70	89.0980 (25.5588)	**71.9260** (14.7509)	118.4500 (19.1530)	79.7280 (18.1253)	**68.4700** (10.0699)	106.6120 (13.9283)	76.8540 (12.2486)	**72.8420** (7.0205)	109.7000 (5.5804)
500	10	31.4900 (7.6138)	**25.6440** (3.8964)	38.9600 (7.1208)	30.6860 (6.9266)	**25.5140** (4.5048)	38.4420 (6.6322)	30.5520 (4.1350)	**26.5200** (3.2333)	37.1900 (3.2464)
	40	174.9000 (17.6438)	**165.7300** (13.7421)	194.7520 (8.6461)	167.2420 (17.0928)	**158.1520** (14.0082)	191.9380 (10.4478)	157.9000 (8.5345)	**145.9000** (7.9283)	176.2200 (5.8094)
	70	312.1100 (29.1970)	**303.4420** (21.3779)	345.1640 (8.9834)	304.0560 (24.3851)	**292.9220** (19.4685)	338.4360 (13.9510)	286.5180 (13.7763)	**267.7820** (9.9174)	314.9760 (8.9459)
1000	10	89.0720 (8.5779)	**82.4340** (6.5262)	97.8320 (4.1562)	87.0480 (7.9429)	**82.5960** (5.5833)	96.8720 (4.5185)	83.1440 (5.9933)	**76.1280** (4.4365)	90.2640 (4.2489)
	40	380.6000 (18.2103)	**370.7340** (12.7365)	398.1000 (2.0719)	376.7620 (18.4479)	**367.0800** (12.7173)	395.6120 (6.4258)	360.9000 (10.4112)	**343.5780** (10.3458)	379.9780 (8.1408)
	70	671.0640 (27.1293)	**655.5860** (20.2919)	695.2560 (7.2060)	663.5140 (25.9454)	**651.3360** (21.3670)	694.4180 (7.3348)	641.3480 (17.1041)	**614.3380** (17.4020)	670.1000 (14.0177)

จากตารางที่ 4.2.2.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธีนั้นให้แนวโน้มค่าเฉลี่ย FN ลดลงโดยส่วนใหญ่
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย FN เพิ่มขึ้นในทุกกรณี
4. ค่าเฉลี่ย FN ในทุกกรณีพบว่า การคัดเลือกตัวแปรด้วยวิธี Elastic Net ให้ค่าเฉลี่ย FN ต่ำที่สุด รองลงมาเป็นวิธี Lasso และวิธี Adaptive Lasso ตามลำดับ

4.3 การเปรียบเทียบค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์ที่ได้จากการคัดเลือกตัวแปรอิสระด้วยวิธี Stepwise วิธี Lasso วิธี Elastic Net และวิธี Adaptive Lasso

4.3.1 จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$

ตารางที่ 4.3.1.1 ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์ (MAE) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
4	10	0.1356 (0.1082)	0.1286 (0.0960)	0.1348 (0.0981)	**0.0997** (0.0858)	0.1693 (0.1477)	0.1416 (0.1206)	0.1566 (0.1231)	**0.1134** (0.1012)	0.3549 (0.3341)	0.2586 (0.2819)	0.2211 (0.2479)	**0.1793** (0.2257)
	40	0.1562 (0.1013)	0.1683 (0.0912)	0.1636 (0.0888)	**0.1531** (0.0885)	0.1969 (0.1403)	0.1920 (0.1082)	0.1875 (0.1084)	**0.1842** (0.1191)	0.4083 (0.2985)	0.3123 (0.2348)	0.3041 (0.2436)	**0.2688** (0.1961)
	70	0.2010 (0.1007)	0.2134 (0.0997)	**0.1996** (0.0921)	0.2204 (0.1131)	0.2436 (0.1363)	0.2352 (0.1108)	**0.2258** (0.1125)	0.2421 (0.1186)	0.4849 (0.2944)	0.3888 (0.2295)	**0.3657** (0.2143)	0.3769 (0.2140)
10	10	0.1650 (0.1094)	0.0782 (0.0778)	0.0934 (0.0782)	**0.0499** (0.0573)	0.2162 (0.1388)	0.0927 (0.1004)	0.1025 (0.0895)	**0.0558** (0.0659)	0.5007 (0.3533)	0.1736 (0.2383)	0.1762 (0.2333)	**0.0957** (0.1507)
	40	0.2036 (0.0995)	0.1526 (0.0722)	0.1532 (0.0638)	**0.1426** (0.0646)	0.2777 (0.1382)	0.1956 (0.1104)	0.2002 (0.1116)	**0.1694** (0.0939)	0.5720 (0.3046)	0.2795 (0.2033)	0.2823 (0.2068)	**0.2283** (0.1398)
	70	0.2602 (0.1046)	0.2288 (0.0868)	**0.2210** (0.0831)	0.2389 (0.0894)	0.3156 (0.1290)	0.2528 (0.0895)	**0.2467** (0.0975)	0.2656 (0.0956)	0.6546 (0.3020)	0.3819 (0.2028)	0.3769 (0.2050)	**0.3686** (0.1336)
16	10	0.2809 (0.1764)	0.0720 (0.0833)	0.1073 (0.1058)	**0.0615** (0.0527)	0.3621 (0.4055)	0.0793 (0.0848)	0.1075 (0.1041)	**0.0624** (0.0598)	0.7673 (0.5107)	0.1450 (0.2222)	0.1730 (0.2539)	**0.0927** (0.1096)
	40	0.3547 (0.1721)	0.1879 (0.0999)	0.2045 (0.1185)	**0.1882** (0.0722)	0.4189 (0.2289)	0.2013 (0.0934)	0.2244 (0.1168)	**0.1988** (0.0796)	0.8199 (0.4742)	0.2836 (0.2834)	0.3171 (0.2579)	**0.2668** (0.1510)
	70	0.4407 (0.1704)	0.2977 (0.0920)	**0.2817** (0.1109)	0.3049 (0.0889)	0.5007 (0.3327)	0.3211 (0.1095)	**0.3137** (0.1306)	0.3310 (0.1015)	0.9338 (0.4851)	**0.4020** (0.2225)	0.4297 (0.2399)	0.4151 (0.1364)

จากตารางที่ 4.3.1.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นทุกกรณี
   - วิธี Lasso กรณี %nonzero = 10 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE ลดลง ส่วนกรณีอื่นนั้นให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน
   - วิธี Elastic Net กรณี %nonzero = 70 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้น ส่วนกรณีอื่นนั้นให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน
   - วิธี Adaptive Lasso นั้นให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน
4. ค่าเฉลี่ย MAE กรณี %nonzero = 10 และ 40 การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso ให้ค่าเฉลี่ย MAE ต่ำที่สุด แต่ในกรณี %nonzero = 70 วิธี Elastic Net ให้ค่าเฉลี่ย MAE ต่ำที่สุด ยกเว้นกรณี γ = 0.9, p = 10 และ 16 วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำสุด คือ วิธี Adaptive Lasso และวิธี Lasso ตามลำดับ

ตารางที่ 4.3.1.2 ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์ของค่าสัมประสิทธิ์ (MAE) กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง $(p < n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non zero	γ = 0.1				γ = 0.5				γ = 0.9			
		ST	LA	EN	AD	ST	LA	EN	AD	ST	LA	EN	AD
20	10	0.0477	0.0393	0.0389	**0.0167**	0.0551	0.0413	0.0417	**0.0177**	0.0897	0.0782	0.0639	**0.0362**
		(0.0249)	(0.0274)	(0.0188)	(0.0127)	(0.0266)	(0.0336)	(0.0215)	(0.0137)	(0.0451)	(0.0763)	(0.0428)	(0.0319)
	40	0.1330	0.0684	0.0692	**0.0502**	0.1450	0.0841	0.0818	**0.0604**	0.1980	0.1673	0.1601	**0.1377**
		(0.0395)	(0.0215)	(0.0188)	(0.0181)	(0.0392)	(0.0299)	(0.0246)	(0.0241)	(0.0592)	(0.0670)	(0.0575)	(0.0531)
	70	0.2265	0.0888	0.0846	**0.0770**	0.2378	0.1126	0.1051	**0.0973**	0.3058	0.2297	**0.2159**	0.2232
		(0.0475)	(0.0208)	(0.0174)	(0.0178)	(0.0491)	(0.0301)	(0.0233)	(0.0251)	(0.0629)	(0.0640)	(0.0564)	(0.0621)
50	10	0.0530	0.0323	0.0392	**0.0156**	0.0568	0.0387	0.0424	**0.0182**	0.0825	0.0697	0.0663	**0.0388**
		(0.0145)	(0.0202)	(0.0131)	(0.0075)	(0.0170)	(0.0260)	(0.0173)	(0.0110)	(0.0269)	(0.0574)	(0.0363)	(0.0228)
	40	0.1786	0.0757	0.0808	**0.0555**	0.1830	0.0903	0.0944	**0.0685**	0.2104	0.1676	0.1686	**0.1498**
		(0.0280)	(0.0185)	(0.0160)	(0.0143)	(0.0274)	(0.0246)	(0.0203)	(0.0184)	(0.0338)	(0.0552)	(0.0424)	(0.0385)
	70	0.3065	0.1047	0.1022	**0.0941**	0.3086	0.1276	0.1247	**0.1178**	0.3444	0.2477	**0.2360**	0.2500
		(0.0324)	(0.0191)	(0.0164)	(0.0168)	(0.0349)	(0.0248)	(0.0219)	(0.0223)	(0.0380)	(0.0475)	(0.0425)	(0.0452)
80	10	0.0610	0.0285	0.0401	**0.0154**	0.0658	0.0311	0.0422	**0.0174**	0.0953	0.0571	0.0629	**0.0383**
		(0.0128)	(0.0126)	(0.0130)	(0.0062)	(0.0143)	(0.0153)	(0.0149)	(0.0082)	(0.0255)	(0.0425)	(0.0288)	(0.0195)
	40	0.1956	0.0867	0.0944	**0.0682**	0.2007	0.1012	0.1086	**0.0828**	0.2327	0.1670	0.1785	**0.1635**
		(0.0231)	(0.0192)	(0.0176)	(0.0160)	(0.0240)	(0.0225)	(0.0216)	(0.0205)	(0.0323)	(0.0364)	(0.0425)	(0.0362)
	70	0.3303	0.1432	0.1376	**0.1321**	0.3357	0.1694	0.1610	**0.1600**	0.3696	0.2689	**0.2642**	0.2815
		(0.0272)	(0.0309)	(0.0236)	(0.0247)	(0.0279)	(0.0346)	(0.0280)	(0.0302)	(0.0351)	(0.0399)	(0.0438)	(0.0440)

จากตารางที่ 4.3.1.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 4 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Stepwise ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี %nonzero = 10, γ = 0.9 ที่ระดับ p = 50 และ 80 มีค่าต่ำสุดและสูงสุดตามลำดับ
   - วิธี Lasso กรณี %nonzero = 10 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE ลดลง ส่วนกรณี %nonzero = 40 และ 70 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้น
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี %nonzero = 10, γ = 0.5 และ 0.9 ให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน
   - วิธี Adaptive Lasso กรณี %nonzero = 10 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน ส่วนกรณี %nonzero = 40 และ 70 ในทุกระดับของ γ ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้น
4. วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุด คือวิธี Adaptive Lasso ยกเว้นกรณี γ = 0.9, %nonzero = 70 ในทุกระดับของ p พบว่า การคัดเลือกตัวแปรด้วยวิธี Elastic Net ให้ค่าเฉลี่ย MAE ต่ำที่สุด

4.3.2 จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$

ตารางที่ 4.3.2.1 ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์ของค่าสัมประสิทธิ์ (MAE) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 20 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1			γ = 0.5			γ = 0.9		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
40	10	0.0628 (0.0270)	0.0673 (0.0239)	**0.0469** (0.0174)	0.0678 (0.0371)	0.0689 (0.0302)	**0.0486** (0.0224)	0.0931 (0.0496)	0.0924 (0.0437)	**0.0672** (0.0364)
	40	0.1956 (0.0414)	**0.1906** (0.0400)	0.1947 (0.0359)	0.2040 (0.0421)	**0.2015** (0.0399)	0.2038 (0.0451)	0.2441 (0.0604)	0.2255 (0.0430)	**0.2233** (0.0439)
	70	0.3295 (0.0472)	**0.3192** (0.0443)	0.3450 (0.0474)	0.3420 (0.0458)	**0.3268** (0.0472)	0.3534 (0.0472)	0.3765 (0.0667)	**0.3445** (0.0519)	0.3726 (0.0563)
100	10	0.0547 (0.0131)	0.0579 (0.0128)	**0.0502** (0.0115)	0.0594 (0.0149)	0.0623 (0.0139)	**0.0520** (0.0115)	0.0746 (0.0190)	0.0724 (0.0171)	**0.0657** (0.0181)
	40	0.2082 (0.0254)	**0.2053** (0.0231)	0.2054 (0.0236)	0.2104 (0.0245)	0.2091 (0.0234)	**0.2079** (0.0228)	0.2291 (0.0310)	**0.2220** (0.0302)	0.2294 (0.0322)
	70	0.3580 (0.0304)	**0.3505** (0.0288)	0.3561 (0.0275)	0.3588 (0.0300)	**0.3530** (0.0295)	0.3590 (0.0319)	0.3725 (0.0312)	**0.3600** (0.0287)	0.3800 (0.0356)
200	10	0.0555 (0.0098)	0.0574 (0.0082)	**0.0526** (0.0080)	0.0571 (0.0100)	0.0591 (0.0095)	**0.0539** (0.0097)	0.0635 (0.0117)	0.0635 (0.0102)	**0.0594** (0.0121)
	40	**0.2069** (0.0162)	0.2080 1(0.0153)	0.2076 (0.0160)	0.2071 (0.0183)	0.2065 (0.0153)	**0.2031** (0.0174)	0.2218 (0.0218)	**0.2169** (0.0186)	0.2184 (0.0231)
	70	0.3587 (0.0218)	**0.3581** (0.0210)	0.3610 (0.0235)	0.3591 (0.0222)	**0.3570** (0.0193)	0.3590 (0.0206)	0.3694 (0.0258)	**0.3638** (0.0232)	0.3721 (0.0259)

จากตารางที่ 4.3.2.1 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
   - วิธี Elastic Net และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 200, %nonzero = 40 และ 70 ที่ระดับ γ = 0.5, 0.9 นั้นให้ค่าเฉลี่ย MAE มีค่าต่ำสุดและสูงสุดตามลำดับ
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย MAE ที่ไม่ชัดเจน
4. ที่ระดับ %nonzero = 10 ในทุกระดับของ p และ γ พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุด คือวิธี Adaptive Lasso ซึ่งวิธี Lasso และ Elastic Net ให้ค่าเฉลี่ย MAE ที่ไม่แตกต่างกันมากนัก ส่วนที่ระดับ %nonzero = 40 พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุดโดยส่วนใหญ่ คือวิธี Elastic net ซึ่งมีบางกรณีที่วิธี Lasso และ Adaptive Lasso นั้นให้ค่าต่ำสุด และที่ระดับ %nonzero = 70 ในทุกระดับของ p และ γ พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุด คือวิธี Elastic Net ซึ่งวิธี Lasso และ Elastic Net ให้ค่าเฉลี่ย MAE ที่ไม่แตกต่างกันมากนัก

ตารางที่ 4.3.2.2 ค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์ของค่าสัมประสิทธิ์ (MAE) กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง $(p > n)$ เมื่อกำหนดให้ขนาดตัวอย่างเท่ากับ 100 โดยแสดงค่าเบี่ยงเบนมาตรฐานไว้ในวงเล็บ

p	%non-zero	γ = 0.1			γ = 0.5			γ = 0.9		
		LA	EN	AD	LA	EN	AD	LA	EN	AD
200	10	0.0386 (0.0112)	0.0411 (0.0081)	**0.0201** (0.0058)	0.0409 (0.0122)	0.0437 (0.0094)	**0.0235** (0.0063)	0.0710 (0.0287)	0.0604 (0.0131)	**0.0429** (0.0114)
	40	0.1765 (0.0216)	**0.1698** (0.0180)	0.1776 (0.0203)	0.1797 (0.0222)	**0.1740** (0.0192)	0.1791 (0.0215)	0.2005 (0.0247)	**0.1898** (0.0184)	0.1951 (0.0198)
	70	0.3186 (0.0234)	**0.3074** (0.0201)	0.3407 (0.0201)	0.3197 (0.0226)	**0.3068** (0.0209)	0.3365 (0.0216)	0.3350 (0.0256)	**0.3161** (0.0216)	0.3514 (0.0246)
500	10	0.0511 (0.0071)	0.0535 (0.0064)	**0.0478** (0.0069)	0.0530 (0.0074)	0.0544 (0.0062)	**0.0482** (0.0070)	0.0619 (0.0075)	0.0622 (0.0066)	**0.0563** (0.0075)
	40	0.2053 (0.0125)	0.2037 (0.0112)	**0.2027** (0.0110)	0.2099 (0.0146)	0.2072 (0.0119)	**0.2057** (0.0131)	0.2213 (0.0137)	**0.2177** (0.0115)	0.2246 (0.0136)
	70	0.3551 (0.0155)	**0.3485** (0.0134)	0.3528 (0.0136)	0.3559 (0.0147)	**0.3502** (0.0141)	0.3576 (0.0154)	0.3689 (0.0177)	**0.3606** (0.0160)	0.3803 (0.0178)
1000	10	0.0540 (0.0057)	0.0549 (0.0047)	**0.0513** (0.0042)	0.0556 (0.0056)	0.0560 (0.0044)	**0.0523** (0.0049)	0.0608 (0.0059)	0.0626 (0.0048)	**0.0604** (0.0061)
	40	0.2082 (0.0113)	0.2072 (0.0088)	**0.2020** (0.0074)	0.2095 (0.0112)	0.2084 (0.0096)	**0.2041** (0.0086)	0.2244 (0.0089)	**0.2229** (0.0079)	0.2264 (0.0092)
	70	0.3568 (0.0121)	0.3554 (0.0098)	**0.3535** (0.0100)	0.3600 (0.0126)	0.3572 (0.0111)	**0.3554** (0.0121)	0.3729 (0.0121)	**0.3684** (0.0102)	0.3780 (0.0149)

จากตารางที่ 4.3.2.2 พบว่า

1. เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
   - วิธี Elastic Net และวิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี p = 200, %nonzero = 70 ที่ระดับ γ = 0.5, 0.9 นั้นให้ค่าเฉลี่ย MAE มีค่าต่ำสุดและสูงสุดตามลำดับ
2. เมื่อเปอร์เซ็นต์ของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เพิ่มขึ้น การคัดเลือกตัวแปรทั้ง 3 วิธี ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
3. เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น การคัดเลือกตัวแปรด้วย
   - วิธี Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่ ยกเว้นกรณี γ = 0.9, %nonzero = 10 ให้แนวโน้มค่าเฉลี่ย MAE ลดลง
   - วิธี Elastic Net ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นในทุกกรณี
   - วิธี Adaptive Lasso ให้แนวโน้มค่าเฉลี่ย MAE เพิ่มขึ้นโดยส่วนใหญ่
4. ที่ระดับ %nonzero = 10 ในทุกระดับของ p และ γ พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุด คือวิธี Adaptive Lasso ส่วนที่ระดับ %nonzero = 40 พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุดโดยส่วนใหญ่ คือวิธี Elastic Net ซึ่งมีเพียงกรณี p = 500,1000, γ = 0.1, 0.5 ที่วิธี Adaptive Lasso นั้นให้ค่าต่ำสุด และที่ระดับ %nonzero = 70 พบว่า วิธีการคัดเลือกตัวแปรที่ให้ค่าเฉลี่ย MAE ต่ำที่สุดโดยส่วนใหญ่ คือวิธี Elastic Net ซึ่งมีเพียงกรณี p = 1000, γ = 0.1, 0.5 ที่วิธี Adaptive Lasso นั้นให้ค่าต่ำสุด

บทที่ 5

สรุปผลการศึกษาและข้อเสนอแนะ

การศึกษางานวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการคัดเลือกตัวแปรอิสระและประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบเชิงเส้นทั้ง 4 วิธี ได้แก่ วิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง สำหรับข้อมูลที่มีผลกระทบขนาดเล็กและมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์ ซึ่งเกณฑ์ที่ใช้ในการพิจารณา ได้แก่ ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) และค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) โดยวิธีการคัดเลือกตัวแปรใดให้ค่าของเกณฑ์ทั้งสามต่ำสุดโดยสอดคล้องกัน จะถือว่าวิธีการคัดเลือกตัวแปรนั้นมีประสิทธิภาพที่ดีที่สุด นอกจากนี้ยังมีการพิจารณาโดยแยกตามวัตถุประสงค์ของเกณฑ์วัดประสิทธิภาพ ได้แก่ ในแง่ของความสัมพันธ์ระหว่างตัวแปร ซึ่งจะพิจารณาจากค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงบวก (FP) และค่าเฉลี่ยความผิดพลาดในการตรวจจับเชิงลบ (FN) และในแง่ของความแม่นยำการทำนายซึ่งจะพิจารณาจากค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ย (MAE) ซึ่งสามารถสรุปผลการศึกษาในกรณีต่าง ๆ ได้ดังนี้

5.1 สรุปผลการศึกษา

5.1.1 ผลการเปรียบเทียบประสิทธิภาพวิธีการคัดเลือกตัวแปรอิสระและประมาณค่าสัมประสิทธิ์

5.1.1.1 กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง

จากการจำลองข้อมูลตามขอบเขตที่กำหนด สามารถสรุปวิธีที่มีประสิทธิภาพในการคัดเลือกตัวแปรพยากรณ์และประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นในแต่ละสถานการณ์ต่างๆ ได้ดังนี้

ตารางที่ 5.1.1.1 สรุปผลวิธีการคัดเลือกตัวแปรอิสระที่ให้ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวกค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบและค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยของค่าสัมประสิทธิ์ต่ำสุด กรณี $p < n$ เมื่อ $n = 20$ (✕ หมายถึง ไม่สามารถสรุปผลการทดสอบได้)

p	%nonzero	γ	FP	FN	MAE	วิธีที่เหมาะสม
4	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
10	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	AD	AD
16	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	LA	✕

ตารางที่ 5.1.1.2 สรุปผลวิธีการคัดเลือกตัวแปรอิสระที่ให้ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวกค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบและค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยของค่าสัมประสิทธิ์ต่ำสุด กรณี $p<n$ เมื่อ $n=100$ (× หมายถึง ไม่สามารถสรุปผลการทดสอบได้)

p	%nonzero	γ	FP	FN	MAE	วิธีที่เหมาะสม
20	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	AD	×
	70	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	EN	EN
50	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	AD	×
	70	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	EN	EN
80	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	AD	×
	70	0.1	ST	EN	AD	×
		0.5	ST	EN	AD	×
		0.9	ST	EN	EN	EN

จากตารางที่ 5.1.1.1 และตารางที่ 5.1.1.2 จะพบว่า สำหรับกรณีที่จำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso นั้นจะให้ประสิทธิภาพในการคัดเลือกตัวแปรอิสระและการประมาณค่าสัมประสิทธิ์การถดถอยได้ดีกว่าในกรณีที่ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับต่ำ (%nonzero = 10) และการคัดเลือกตัวแปรด้วยวิธี Elastic Net นั้นให้ประสิทธิภาพที่ดีในกรณีที่ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับสูง (%nonzero = 70)

หากพิจารณาตามวัตถุประสงค์ของเกณฑ์การวัดประสิทธิภาพแล้ว ในแง่ของความสัมพันธ์ระหว่างตัวแปร จะพบว่าการคัดเลือกตัวแปรด้วยวิธี Elastic Net นั้นให้ประสิทธิภาพที่ดีที่สุดในทุกกรณี และหากพิจารณาในแง่ของความแม่นยำการทำนาย จะพบว่าการคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso นั้นจะให้ประสิทธิภาพที่ดีในกรณีที่ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับต่ำและกลาง (%nonzero = 10, 40) ส่วนในระดับของร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับสูง (%nonzero = 70) การคัดเลือกตัวแปรด้วยวิธี Elastic Net จะให้ผลได้ดีกว่าเมื่อขนาดตัวอย่างมีขนาดเล็ก แต่เมื่อขนาดตัวอย่างเพิ่มขึ้น การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso จะให้ผลของการทำนายมีความแม่นยำมากกว่า

5.1.1.2 กรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง

จากการจำลองข้อมูลตามขอบเขตที่กำหนด สามารถสรุปวิธีที่มีประสิทธิภาพในการคัดเลือกตัวแปรพยากรณ์และประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นในแต่ละสถานการณ์ต่างๆ ได้ดังนี้

ตารางที่ 5.1.2.1 สรุปผลวิธีการคัดเลือกตัวแปรอิสระที่ให้ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวกค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบและค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยของค่าสัมประสิทธิ์ต่ำสุด กรณี $p > n$ เมื่อ $n = 20$ (✕ หมายถึง ไม่สามารถสรุปผลการทดสอบได้)

p	%nonzero	γ	FP	FN	MAE	วิธีที่เหมาะสม
40	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	AD	AD
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
100	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	EN	EN
		0.5	AD	EN	AD	AD
		0.9	AD	EN	EN	EN
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
200	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	LA	✕
		0.5	AD	EN	AD	AD
		0.9	AD	EN	EN	EN
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN

ตารางที่ 5.1.2.2 สรุปผลวิธีการคัดเลือกตัวแปรอิสระที่ให้ค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงบวกค่าเฉลี่ยของค่าความผิดพลาดในการตรวจจับเชิงลบและค่าเฉลี่ยของค่าคลาดเคลื่อนสัมบูรณ์โดยเฉลี่ยของค่าสัมประสิทธิ์ต่ำสุด กรณี $p > n$ เมื่อ $n = 100$

p	%nonzero	γ	FP	FN	MAE	วิธีที่เหมาะสม
200	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
500	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	EN	EN
	70	0.1	AD	EN	EN	EN
		0.5	AD	EN	EN	EN
		0.9	AD	EN	EN	EN
1000	10	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	AD	AD
	40	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	EN	EN
	70	0.1	AD	EN	AD	AD
		0.5	AD	EN	AD	AD
		0.9	AD	EN	EN	EN

จากตารางที่ 5.1.2.1 และตารางที่ 5.1.2.2 สามารถสรุปได้ว่า สำหรับกรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง การคัดเลือกตัวแปรด้วยวิธี Elastic Net และวิธี Adaptive Lasso นั้นจะให้ประสิทธิภาพในการคัดเลือกตัวแปรอิสระและการประมาณค่าสัมประสิทธิ์การถดถอยได้ดีกว่าวิธี Stepwise และวิธี Lasso ซึ่งการคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso นั้นจะให้ประสิทธิภาพที่ดีกว่าวิธี Elastic Net เมื่อร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับต่ำและกลาง (%nonzero = 10, 40) และการคัดเลือกตัวแปรด้วยวิธี Elastic Net นั้นจะให้ประสิทธิภาพที่ดีกว่าวิธี Adaptive Lasso เมื่อร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเป็นศูนย์อยู่ในระดับสูง (%nonzero = 70)

หากพิจารณาตามวัตถุประสงค์ของเกณฑ์การวัดประสิทธิภาพแล้ว ในแง่ของความสัมพันธ์ระหว่างตัวแปร จะพบว่าการคัดเลือกตัวแปรด้วยวิธี Elastic Net นั้นให้ประสิทธิภาพที่ดีที่สุดในทุกกรณี และหากพิจารณาในแง่ของความแม่นยำการทำนาย จะพบว่าการคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso นั้นจะให้ประสิทธิภาพที่ดีในกรณีที่ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับต่ำและกลาง (%nonzero = 10, 40) ส่วนในระดับของร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับสูง (%nonzero = 70) การคัดเลือกตัวแปรด้วยวิธี Elastic Net จะให้ผลได้ดีกว่าเมื่อขนาดตัวอย่างมีขนาดเล็ก แต่เมื่อขนาดตัวอย่างเพิ่มขึ้น การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso จะให้ผลของการทำนายมีความแม่นยำมากกว่า

5.1.2 ผลกระทบจากระดับค่าความสัมพันธ์ระหว่างตัวแปรอิสระ (γ)

พิจารณาจากผลกระทบที่ได้รับจากค่าความสัมพันธ์ระหว่างตัวแปรอิสระ เมื่อจำนวนตัวแปรอิสระและร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์มีค่าคงที่พบว่า เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าเปลี่ยนแปลง จะไม่ส่งผลต่อประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระ ผลกระทบที่เห็นได้ชัดคือกรณีที่ค่าความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าสูงมาก การคัดเลือกตัวแปรด้วยวิธี Elastic Net จะมีประสิทธิภาพดีกว่า ทั้งนี้ขึ้นอยู่กับปัจจัยอื่นๆต้องมีค่าสูงด้วย

5.1.3 ผลกระทบจากระดับค่าร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์ (%nonzero)

พิจารณาจากผลกระทบที่ได้รับจากค่าร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์ เมื่อจำนวนตัวแปรอิสระและค่าความสัมพันธ์ระหว่างตัวแปรอิสระคงที่ พบว่า การคัดเลือกตัวแปรอิสระด้วยวิธี Adaptive Lasso จะมีประสิทธิภาพการทำงานที่ดีเมื่อค่าร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับต่ำถึงกลาง และเมื่อค่าร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับสูง การคัดเลือกตัวแปรอิสระด้วยวิธี Elastic Net จะให้ประสิทธิภาพที่ดีกว่า

5.1.4 ผลกระทบจากระดับจำนวนตัวแปรอิสระ (p)

พิจารณาจากผลกระทบที่ได้รับจากระดับจำนวนตัวแปรอิสระ เมื่อค่าความสัมพันธ์ระหว่างตัวแปรอิสระและร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์ คงที่ พบว่า กรณีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง เมื่อระดับจำนวนตัวแปร (p) เปลี่ยนแปลง จะไม่ส่งผลต่อประสิทธิภาพของวิธีการคัดเลือกตัวแปรอิสระ แต่สำหรับกรณีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง เมื่อร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์อยู่ในระดับกลางและสูง การคัดเลือกตัวแปรอิสระด้วยวิธี Elastic Net จะมีประสิทธิภาพการทำงานที่ดีเมื่อจำนวนตัวแปรอิสระต่อขนาดตัวอย่างอยู่ในระดับต่ำ แต่เมื่อจำนวนตัวแปรอิสระต่อขนาดตัวอย่างอยู่ในระดับกลางและระดับสูง การคัดเลือกตัวแปรอิสระด้วยวิธี Adaptive Lasso จะมีประสิทธิภาพการทำงานที่เหมาะสมมากกว่า

5.2 ข้อเสนอแนะ

เพื่อเป็นแนวทางให้ผู้ที่สนใจได้ศึกษาเพิ่มเติมในการศึกษาครั้งต่อๆไป อาจนำแนวทางในกรณีต่างๆ ไปศึกษาต่อ ดังนี้

1. การศึกษาในครั้งนี้ได้ทำการศึกษาที่ตัวแปรอิสระและค่าความคลาดเคลื่อนมีลักษณะการแจกแจงข้อมูลแบบปกติที่ค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเป็นหนึ่งเท่านั้น ดังนั้นในการศึกษาต่อไปอาจทำการศึกษาในกรณีที่ตัวแปรอิสระมาจากการแจกแจงแบบอื่น
2. การศึกษาในครั้งนี้ไม่ได้ทำการศึกษากลุ่มตัวแปรอิสระที่มีความสัมพันธ์กันสูง ดังนั้นในการศึกษาต่อไปอาจทำการศึกษาในกรณีที่มีกลุ่มของตัวแปรที่มีการปะปนอยู่กับ

ตัวแปรอื่นๆ เนื่องจากตัวแปรประเภทนี้เมื่อมีความสัมพันธ์กันสูงมาก จะมีแนวโน้มที่จะถูกนำเข้าหรือคัดออกจากตัวแบบไปพร้อมๆกัน

3. การศึกษาในครั้งนี้พบว่าในหลายๆ กรณี นั้นให้ผลที่ไม่สามารถเห็นแนวโน้มได้ชัดเจน การศึกษาถัดไปอาจเพิ่มกรณีที่ทำให้เห็นแนวโน้มได้มากขึ้น เช่น เพิ่มเป็น 5 ระดับ เป็นต้น

4. การศึกษาในครั้งนี้ได้ศึกษาวิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์การถดถอย 4 วิธี ซึ่งวิธียังมีอีกหลายวิธีในการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ ดังนั้นในการศึกษาครั้งต่อไป อาจใช้วิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์การถดถอยในแบบอื่นๆ ซึ่งอาจให้ประสิทธิภาพที่ดีกว่า

รายการอ้างอิง

Efroymson, M. A. (1966). "Stepwise Regression - A Backward and Forward Look." Presented at the Eastern Regional Meetings of the Inst. of Math. Statist., Florham Park, New Jerse.

Ma, S., et al. (2007). "Supervised group Lasso with applications to microarray data analysis." BMC Bioinformatics **8**(60).

Mateos, G. (2010). "Distributed Sparse Linear Regression." IEEE Transactions On Signals Processing **58**(10): 5262 -5276.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via The Lasso." Journal of the Royal Statistical Society **58**(1): 267 - 288.

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

Zou, H. and T. Hastie (2005). "Regularization and Variable Selection Via The Elastic Net." Journal of the Royal Statistical Society **67**(2): 301 - 320.

บรรณานุกรม

Smith, S.W. “The Scientist and Engineer's Guide to Digital Signal Processing.” California Technical Publishing, San Diego, California, USA.

Hastie, T., et al. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.” Springer Science+Business Media, New York, USA.

ภาคผนวก

คำสั่งที่ใช้ในการวิเคราะห์ข้อมูลจากการจำลองด้วยโปรแกรม R

ตัวอย่างกรณีที่ตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง โดยที่ตัวแปรอิสระเท่ากับ 20 ขนาดตัวอย่างเท่ากับ 100 ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เมื่อเทียบกับจำนวนตัวแปรอิสระเท่ากับ 70 และค่าความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.1

```
library(mvtnorm)
library(lars)
library(elasticnet)
library(parcor)
library(glmnet)

########################################
################ Variable ###############
########################################

#################### p < n
n <- 100
p <- 20
zero <- 0.3
gamma <- 0.5
N <- 500

FPst  <- matrix(NA,N,1)                    #Matrix for FPSM
FPlas <- matrix(NA,N,1)
FPen  <- matrix(NA,N,1)
FPad  <- matrix(NA,N,1)
FNst  <- matrix(NA,N,1)                    #Matrix for FNSM
```

```
FNlas <- matrix(NA,N,1)
FNen  <- matrix(NA,N,1)
FNad  <- matrix(NA,N,1)
dis_st  <- matrix(NA,N,1)                    #Matrix for MAE
dis_las <- matrix(NA,N,1)
dis_en  <- matrix(NA,N,1)
dis_ad  <- matrix(NA,N,1)

for (i in 1:N){
#####################################
############### Generate X ##############
#####################################

mu_X   <- matrix(0, p, 1)                    #mean of X Matrix
corr_X <- matrix(, p, p)                     #empty matrix for correlation
for(k in 1:p){
  for(j in 1:p){
      corr_X[j,k] = gamma^abs(j-k)           #condition of correlation matrix
  }
}
X_Gen  <- rmvnorm(n, mu_X, corr_X)           #gen X with Multivariate norm fn


#####################################
############# Generate Error ############
#####################################
```

```
e_Gen <- matrix(rnorm(n,0,1))                    #gen standard Error ~ N(0,1)

#####################################
############### Gen Beta ###############
#####################################

sp_pos <- sample(1:p,floor(zero*p),replace=F)    #random the sparse position
b_Gen <- matrix(NA, p,1)                         #empty matrix
b_Gen[sp_pos] <- 0                               #set 0 at sparse position
sm_miss <- which(is.na(b_Gen))                   #find the NA position
sm_miss_val <- p - floor(zero*p)                 #count the NA position and
divide                                           ###to set beta small
b_Gen[sm_miss] <- runif(sm_miss_val, -1, 1)      #random for replace NA value
                                                 with small beta

#####################################
############### Simulate Y #############
#####################################

Y_sim <- X_Gen %*% b_Gen + e_Gen                 #linear model

#####################################
############ Adaptive LASSO ############
#####################################

b_ad <- matrix(NA,p+1,ncol=1)                    #empty matrix
ad_mod <- adalasso(X_Gen, Y_sim, use.Gram=F)
```

```
b_ad[1,1]<-ad_mod$intercept.adalasso                    #intercept
b_ad[2:(p+1),1] <- ad_mod$coefficients.adalasso         #coefficients

#######################################
############### Stepwise ###############
#######################################

step <- step(lm(Y_sim~
 X_Gen[, 1]+X_Gen[, 2]+X_Gen[, 3]+X_Gen[, 4]+X_Gen[, 5]
+X_Gen[, 6]+X_Gen[, 7]+X_Gen[, 8]+X_Gen[, 9]+X_Gen[,10]
+X_Gen[,11]+X_Gen[,12]+X_Gen[,13]+X_Gen[,14]+X_Gen[,15]
+X_Gen[,16]+X_Gen[,17]+X_Gen[,18]+X_Gen[,19]+X_Gen[,20]
), direction = "both")                                  #fit stepwise
lastbeta <- as.matrix(step$coefficients)
b_st<-matrix(0,p+1,1)
if (nrow(lastbeta) != 1){
for(m in 2:nrow(lastbeta)){
if(names(lastbeta[m,1])=="X_Gen[, 1]"){b_st[ 2,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 2]"){b_st[ 3,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 3]"){b_st[ 4,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 4]"){b_st[ 5,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 5]"){b_st[ 6,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 6]"){b_st[ 7,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 7]"){b_st[ 8,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 8]"){b_st[ 9,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[, 9]"){b_st[10,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,10]"){b_st[11,1]<-lastbeta[m,1]}
```

```
if(names(lastbeta[m,1])=="X_Gen[,11]"){b_st[12,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,12]"){b_st[13,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,13]"){b_st[14,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,14]"){b_st[15,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,15]"){b_st[16,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,16]"){b_st[17,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,17]"){b_st[18,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,18]"){b_st[19,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,19]"){b_st[20,1]<-lastbeta[m,1]}
if(names(lastbeta[m,1])=="X_Gen[,20]"){b_st[21,1]<-lastbeta[m,1]}
}}
else {b_st[1,1] <- lastbeta[1,1]}

#####################################
################ LASSO ###############
#####################################
las_mod <- lars(x=X_Gen,y=Y_sim,type="lasso",use.Gram=FALSE,normalize=
TRUE,intercept=TRUE)
cvlas   <- cv.lars(x=X_Gen,y=Y_sim,K=10,type='lasso',plot.it=FALSE)
sAtBest <- cvlas$index[which.min(cvlas$cv.error)]
tmp     <- predict.lars(las_mod, type="coefficients", s=sAtBest, mode="fraction")
blas    <- as.matrix(tmp$coefficients)
X_zero  <- as.data.frame(t(rep(0,p)))
beta0   <- predict.lars(las_mod,X_zero, s=0.5, mode="lambda", type="fit")
b_las   <- matrix(NA,p+1,ncol=1)                              #empty matrix
b_las[1,1] <- beta0$fit                                       #replace with intercept
b_las[2:(p+1),1] <- blas                                      #replace with coefficients
```

```
#####################################
################ ENET ################
#####################################

b_en   <- matrix(NA,p+1,1)
cv     <- cv.glmnet(x=X_Gen,y=Y_sim,family="gaussian",nfolds=10)
en_mod <- glmnet(x=X_Gen,y=Y_sim,family="gaussian",alpha=0.5,
lambda=cv$lambda.min,intercept=TRUE)
enb0   <- as.matrix(en_mod$a0)
enb    <- as.matrix(en_mod$beta)
b_en   <- rbind(t(enb0),enb)

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

bhat_st <- as.matrix(b_st[2:(p+1),1])          #beta of stepwise
bhat_las<- as.matrix(b_las[2:(p+1),1])         #beta of lasso
bhat_en <- as.matrix(b_en[2:(p+1),1])          #beta of enet
bhat_ad <- as.matrix(b_ad[2:(p+1),1])          #beta of adap

#####################################
################ FP ################
#####################################

##### bset == 0 & bhat != 0}

zuu=0
zvv=0
```

```
zww=0
zxx=0

for (l in 1:p){
if (bhat_st[l,1]!=0 & b_Gen[l,1]==0){zuu = zuu+1}
if (bhat_las[l,1]!=0& b_Gen[l,1]==0){zvv = zvv+1}
if (bhat_en[l,1]!=0 & b_Gen[l,1]==0){zww = zww+1}
if (bhat_ad[l,1]!=0 & b_Gen[l,1]==0){zxx = zxx+1}
}

FPst[i,1]  <- zuu
FPlas[i,1] <- zvv
FPen[i,1]  <- zww
FPad[i,1]  <- zxx

#######################################
################## FN #################
#######################################

##### bset != 0& bhat == 0

zqq=0
zrr=0
zss=0
ztt=0

for (l in 1:p){
```

```
if (b_Gen[l,1]!=0 & bhat_st[l,1]==0){zqq = zqq+1}
if (b_Gen[l,1]!=0 & bhat_las[l,1]==0){zrr = zrr+1}
if (b_Gen[l,1]!=0 & bhat_en[l,1]==0){zss = zss+1}
if (b_Gen[l,1]!=0 & bhat_ad[l,1]==0){ztt = ztt+1}
}

FNst[i,1]  <- zqq
FNlas[i,1] <- zrr
FNen[i,1]  <- zss
FNad[i,1]  <- ztt

#######################################
######### Distance between bhat&b #######
#######################################

d_st  <- abs(bhat_st  - b_Gen)
d_las <- abs(bhat_las - b_Gen)
d_en  <- abs(bhat_en  - b_Gen)
d_ad  <- abs(bhat_ad  - b_Gen)

dis_st[i,1]  <- colMeans(d_st)
dis_las[i,1] <- colMeans(d_las)
dis_en[i,1]  <- colMeans(d_en)
dis_ad[i,1]  <- colMeans(d_ad)
}
```

```
#####################################
################ Output ##############
#####################################

aveFP_st  <- colMeans(FPst)
aveFP_las <- colMeans(FPlas)
aveFP_en  <- colMeans(FPen)
aveFP_ad  <- colMeans(FPad)
sdFP_st   <- sd(FPst)
sdFP_las  <- sd(FPlas)
sdFP_en   <- sd(FPen)
sdFP_ad   <- sd(FPad)

aveFN_st  <- colMeans(FNst)
aveFN_las <- colMeans(FNlas)
aveFN_en  <- colMeans(FNen)
aveFN_ad  <- colMeans(FNad)
sdFN_st   <- sd(FNst)
sdFN_las  <- sd(FNlas)
sdFN_en   <- sd(FNen)
sdFN_ad   <- sd(FNad)

avedis_st  <- colMeans(dis_st)
avedis_las <- colMeans(dis_las)
avedis_en  <- colMeans(dis_en)
avedis_ad  <- colMeans(dis_ad)
sddis_st   <- sd(dis_st)
```

```
sddis_las  <- sd(dis_las)
sddis_en   <- sd(dis_en)
sddis_ad   <- sd(dis_ad)

Output <- cbind(n,p,zero,gamma,
aveFP_st , aveFP_las, aveFP_en , aveFP_ad ,
sdFP_st , sdFP_las, sdFP_en , sdFP_ad ,
aveFN_las , aveFN_st , aveFN_en , aveFN_ad ,
sdFN_las , sdFN_st , sdFN_en , sdFN_ad ,
avedis_las , avedis_st , avedis_en , avedis_ad ,
sddis_las , sddis_st , sddis_en , sddis_ad)
write.table(Output,file="D:/CU Stat/My Thesis/Thesis B/DATA/1
Output.csv",row.names=FALSE,col.name=FALSE,append=TRUE,na = "NA",sep=",")
```

ตัวอย่างกรณีที่ตัวแปรอิสระมากกว่าขนาดตัวอย่าง โดยที่ตัวแปรอิสระเท่ากับ 200 ขนาดตัวอย่างเท่ากับ 100 ร้อยละของจำนวนค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เป็นศูนย์เมื่อเทียบกับจำนวนตัวแปรอิสระเท่ากับ 70 และค่าความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.1

```
library(mvtnorm)
library(lars)
library(elasticnet)
library(parcor)
library(glmnet)

#######################################
############### Variable ##############
#######################################
```

```
#################### p > n

n <- 100
p <- 200
zero <- 0.3
gamma <- 0.1

N <- 500

FPlas <- matrix(NA,N,1)
FPen  <- matrix(NA,N,1)
FPad  <- matrix(NA,N,1)
FNlas <- matrix(NA,N,1)
FNen  <- matrix(NA,N,1)
FNad  <- matrix(NA,N,1)
dis_las <- matrix(NA,N,1)
dis_en  <- matrix(NA,N,1)
dis_ad  <- matrix(NA,N,1)

for (i in 1:N){

#####################################
################ Generate X ############
#####################################

mu_X   <- matrix(0, p, 1)                    #mean of X Matrix
corr_X <- matrix(, p, p)                     #empty matrix for correlation
```

```
for(k in 1:p){
  for(j in 1:p){
      corr_X[j,k] = gamma^abs(j-k)          #condition of correlation matrix
  }
}
X_Gen  <- rmvnorm(n, mu_X, corr_X)          #gen X with Multivariate normal
fn

#######################################
############ Generate Error ############
#######################################

e_Gen <- matrix(rnorm(n,0,1))               #gen standard Error ~ N(0,1)

#######################################
############## Gen Beta ##############
#######################################

sp_pos <- sample(1:p,floor(zero*p),replace=F)   #random the sparse position
b_Gen <- matrix(NA, p,1)                        #empty matrix
b_Gen[sp_pos] <- 0                              #set 0 at sparse position
sm_miss <- which(is.na(b_Gen))                  #find the NA position
sm_miss_val <- p - floor(zero*p)                #count the NA position and
                                                divide to set beta small
b_Gen[sm_miss] <- runif(sm_miss_val, -1, 1)     #random for replace NA value
                                                with small beta

#######################################
############## Simulate Y #############
```

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

Y_sim <- X_Gen %*% b_Gen + e_Gen                    #linear model

#####################################
############ Adaptive LASSO ############
#####################################

b_ad <- matrix(NA,p+1,ncol=1)                       #empty matrix
ad_mod <- adalasso(X_Gen, Y_sim, use.Gram=F)
b_ad[1,1]<-ad_mod$intercept.adalasso                #replace with intercept
b_ad[2:(p+1),1] <- ad_mod$coefficients.adalasso     #replace with coefficients


#####################################
############### LASSO ###############
#####################################

las_mod <- lars(x=X_Gen,y=Y_sim,type="lasso",use.Gram=FALSE,normalize=
TRUE,intercept=TRUE)
cvlas   <- cv.lars(x=X_Gen,y=Y_sim,K=10,type='lasso',plot.it=FALSE)
sAtBest <- cvlas$index[which.min(cvlas$cv.error)]
tmp     <- predict.lars(las_mod, type="coefficients", s=sAtBest, mode="fraction")
blas    <- as.matrix(tmp$coefficients)
X_zero  <- as.data.frame(t(rep(0,p)))
beta0   <- predict.lars(las_mod,X_zero, s=0.5, mode="lambda", type="fit")
b_las   <- matrix(NA,p+1,ncol=1)                    #empty matrix
```

```
b_las[1,1] <- beta0$fit                          #replace with intercept
b_las[2:(p+1),1] <- blas                         #replace with coefficients

#######################################
################ ENET ###############
#######################################

b_en   <- matrix(NA,p+1,1)
cv     <- cv.glmnet(x=X_Gen,y=Y_sim,family="gaussian",nfolds=10)
en_mod <- glmnet(x=X_Gen,y=Y_sim,family="gaussian",alpha=0.5,
lambda=cv$lambda.min,intercept=TRUE)
enb0   <- as.matrix(en_mod$a0)
enb    <- as.matrix(en_mod$beta)
b_en   <- rbind(t(enb0),enb)

#####################################
bhat_las<- as.matrix(b_las[2:(p+1),1])           #beta of lasso
bhat_en <- as.matrix(b_en[2:(p+1),1])            #beta of enet
bhat_ad <- as.matrix(b_ad[2:(p+1),1])            #beta of adap

#####################################
################ FP #################
#####################################

##### bset == 0 & bhat == small-{0}

zvv=0
```

```
zww=0
zxx=0

for (l in 1:p){
if (bhat_las[l,1]!=0& b_Gen[l,1]==0){zvv = zvv+1}
if (bhat_en[l,1]!=0 & b_Gen[l,1]==0){zww = zww+1}
if (bhat_ad[l,1]!=0 & b_Gen[l,1]==0){zxx = zxx+1}
}

FPlas[i,1] <- zvv
FPen[i,1]  <- zww
FPad[i,1]  <- zxx

########################################
################## FN #################
########################################

##### bset != } & bhat == 0
zrr=0
zss=0
ztt=0

for (l in 1:p){
if (b_Gen[l,1]!=0 & bhat_las[l,1]==0){zrr = zrr+1}
if (b_Gen[l,1]!=0 & bhat_en[l,1]==0){zss = zss+1}
if (b_Gen[l,1]!=0 & bhat_ad[l,1]==0){ztt = ztt+1}
}
```

```
FNlas[i,1] <- zrr
FNen[i,1]  <- zss
FNad[i,1]  <- ztt

#####################################
######## Distance between bhat&b #######
#####################################

d_las <- abs(bhat_las - b_Gen)
d_en  <- abs(bhat_en  - b_Gen)
d_ad  <- abs(bhat_ad  - b_Gen)

dis_las[i,1] <- colMeans(d_las)
dis_en[i,1]  <- colMeans(d_en)
dis_ad[i,1]  <- colMeans(d_ad)
}
#####################################
############### Output ##############
#####################################

aveFP_las <- colMeans(FPlas)
aveFP_en  <- colMeans(FPen)
aveFP_ad  <- colMeans(FPad)
sdFP_las  <- sd(FPlas)
sdFP_en   <- sd(FPen)
sdFP_ad   <- sd(FPad)
```

```
aveFN_las <- colMeans(FNlas)
aveFN_en  <- colMeans(FNen)
aveFN_ad  <- colMeans(FNad)
sdFN_las  <- sd(FNlas)
sdFN_en   <- sd(FNen)
sdFN_ad   <- sd(FNad)

avedis_las <- colMeans(dis_las)
avedis_en  <- colMeans(dis_en)
avedis_ad  <- colMeans(dis_ad)
sddis_las  <- sd(dis_las)
sddis_en   <- sd(dis_en)
sddis_ad   <- sd(dis_ad)

Output <- cbind(n,p,zero,gamma,
aveFP_las,sdFP_las,aveFN_las,sdFN_las,avedis_las,sddis_las,
aveFP_en ,sdFP_en ,aveFN_en ,sdFN_en ,avedis_en ,sddis_en,
aveFP_ad ,sdFP_ad ,aveFN_ad ,sdFN_ad ,avedis_ad ,sddis_ad)

write.table(Output,file="D:/CU Stat/My Thesis/Thesis B/DATA/2
Output.csv",row.names=FALSE,col.name=FALSE,append=TRUE,na = "NA",sep=",")
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวทิฆัมพร สาระกอ เกิดวันเสาร์ที่ 21 พฤศจิกายน พ.ศ. 2530 สำเร็จการศึกษาปริญญาวิศวกรรมศาสตร์บัณฑิต (วศ.บ.) สาขาวิชาวิศวกรรมอุตสาหการ ภาควิชาวิศวกรรมอุตสาหการ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ในปีการศึกษา 2552 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555