การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง

โดย

นางสาวชุติกาญจน์ ชูสวัสดิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชาคณิตศาสตร์และสถิติ
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2560

การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง

โดย

นางสาวชุติกาญจน์ ชูสวัสดิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชาคณิตศาสตร์และสถิติ
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2560

PERFORMANCE COMPARISON OF PENALIZED REGRESSION METHODS IN POISSON REGRESSION UNDER HIGH-DIMENSIONAL SPARSE DATA WITH MULTICOLLINEARITY

BY

MISS CHUTIKARN CHOOSAWAT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (APPLIED STATISTICS)
DEPARTMENT OF MATHEMATICS AND STATISTICS
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2017

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นางสาวชุติกาญจน์ ชูสวัสดิ์

เรื่อง

การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)

เมื่อ วันที่ 5 มิถุนายน พ.ศ. 2561

ประธานกรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.แสงหล้า ชัยมงคล)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.สุปราณี ลิสวัสดิ์)

กรรมการสอบวิทยานิพนธ์ _____
(รองศาสตราจารย์ ดร.วราฤทธิ์ พานิชกิจโกศลกุล)

กรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.พัทธ์ชนก ศรีสุรเดชชัย)

กรรมการสอบวิทยานิพนธ์ _____
(รองศาสตราจารย์ ดร.เสาวณิต สุขภารังษี)

คณบดี _____
(รองศาสตราจารย์ ดร.สมชาย ชคตระการ)

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ ในตัวแบบการถดถอยปัวซงภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง
ชื่อผู้เขียน	นางสาวชุติกาญจน์ ชูสวัสดิ์
ชื่อปริญญา	วิทยาศาตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.สุปราณี ลิสวัสดิ์
ปีการศึกษา	2560

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์การถดถอยปัวซงด้วยวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา หรือกล่าวอีกอย่างหนึ่งว่า กรณีที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง และมีตัวแปรอิสระจำนวนน้อยที่ควรอยู่ในตัวแบบ นอกจากนี้ ตัวแปรอิสระมีความสัมพันธ์กันค่อนข้างสูง นั่นคือ $r = 0.5, 0.6, 0.7, 0.8$ และ 0.9 โดยศึกษารูปแบบความสัมพันธ์ของตัวแปรอิสระ 3 รูปแบบ มีดังนี้ รูปแบบความสัมพันธ์แบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz และรูปแบบความสัมพันธ์แบบ Hub Toeplitz โดยกำหนดให้พิจารณาตัวแปรอิสระเป็น 2 ลักษณะ คือ ตัวแปรอิสระ 1 กลุ่มและตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม หลังจากนั้นทำการจำลองข้อมูลด้วยโปรแกรม R ภายใต้สถานการณ์ต่างๆ และทำซ้ำ 1,000 ครั้ง และประยุกต์ใช้กับข้อมูลจริง

เมื่อพิจารณาประสิทธิภาพในการพยากรณ์ของการวิเคราะห์การถดถอยแบบพีนอลไลซ์ทั้ง 3 วิธี พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ในทุกกรณี และเมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มสูง จาก $r = 0.5$ ไปจนถึง $r = 0.9$ จะได้ว่า วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ทั้ง 3 วิธี มีแนวโน้มที่จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำลง และจากประสิทธิภาพในการแก้ไขปัญหาภาวะร่วมเชิงเส้นของตัวแปรอิสระ ทำให้ได้ว่า ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด เมื่อตัวแปร

อิสระมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz มีค่าต่ำที่สุด รองลงมา คือ รูปแบบความสัมพันธ์แบบ Constant สุดท้าย คือ รูปแบบความสัมพันธ์แบบ Toeplitz นอกจากนี้ ในกรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่า กรณีที่ตัวแปรอิสระมีเพียง 1 กลุ่ม

และเมื่อพิจารณาความผิดพลาดในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จากการศึกษาพบว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุง มีโอกาสในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆที่ตัวแปรอิสระนั้นไม่ควรอยู่ในตัวแบบ โดยวัดอัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) มากกว่า วิธีการวิเคราะห์แบบแลซโซ แต่มีโอกาสที่จะไม่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆ ที่ตัวแปรอิสระนั้นควรอยู่ในตัวแบบ โดยวัดจากอัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) และผลการวิจัยในตัวแปรอิสระ 1 กลุ่ม และ 2 กลุ่ม ให้ผลไปในทิศทางเดียวกัน

คำสำคัญ: วิธีการวิเคราะห์การถดถอยแบบริดจ์, วิธีการวิเคราะห์การถดถอยแบบแลซโซ, วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง, รูปแบบความสัมพันธ์แบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz, รูปแบบความสัมพันธ์แบบ Hub Toeplitz, ปัญหาภาวะร่วมเชิงเส้น

Thesis Title	Performance Comparison of Penalized Regression Methods in Poisson Regression under High-Dimensional Sparse Data with Multicollinearity
Author	Miss Chutikarn Choosawat
Degree	Master of Science (Applied Statistics)
Department/Faculty/University	Mathematics and Statistics Faculty of Science and Technology Thammasat University
Thesis Advisor	Assistant Professor Supranee Lisawdi, Ph.D.
Academic Years	2017

ABSTRACT

The purpose of this study was to compare the performance of Poisson regression analysis among three methods of Penalized regression are as follow: Ridge regression, LASSO and Adaptive LASSO which under High dimensional sparse data or the number of independent variables more than sample size and a few of independent variables in the model. In addition, the independent variables which are highly correlated. ($r = 0.5, 0.6, 0.7, 0.8, 0.9$) We consider three correlation model of independent variables are as a follow: Constant correlation model, Toeplitz correlation model and Hub Toeplitz correlation model. We consider two types of independent variables, that is, one group of independent variables and two groups of independent variables. After performing 1,000 replications of simulation using R software under many situations and applied in real data.

When we consider the performance of predictive of three methods of Penalized regression, we found Adaptive LASSO gave the lowest median of predictive mean square error (mPMSE) in any cases. When the correlation is increased from $r = 0.5$ to $r = 0.9$, three methods of Penalized regression gave median of predictive mean square error (mPMSE) is lower. The performance to solve multicollinearity

problem. When independent variables had Hub Toeplitz correlation model, median of predictive mean square error (mPMSE) is the lowest. The second is the Constant correlation model and the last is Toeplitz correlation model. Furthermore, the median of predictive mean square error (mPMSE), two groups of independent variables gave lower than on group of independent variables.

For the case of the incorrect selection of independent variables into the model, it was found that the Adaptive LASSO has higher probability of incorrect selection by measuring the False Negative Rate (FNR) than LASSO. But Adaptive LASSO has lower probability of incorrect selection by measuring False Positive Rate (FPR) than LASSO. The results of one and two groups of independent variables were the same.

Keywords: Ridge regression, LASSO, Adaptive LASSO, Constant correlation model, Toeplitz correlation model, Hub Toeplitz correlation model, multicollinerity

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี ด้วยความช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.สุปราณี ลิสวัสดิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งกรุณาสละเวลาให้คำปรึกษาในทุกด้าน และมอบความรู้ พร้อมทั้งติดตามความคืบหน้าในทุกขั้นตอนการทำวิทยานิพนธ์ และตรวจสอบข้อบกพร่อง และให้คำแนะนำ เพื่อใช้ในการปรับปรุงแก้ไข ซึ่งเป็นประโยชน์ต่อการจัดทำวิทยานิพนธ์ตลอดระยะเวลาการจัดทำวิทยานิพนธ์จนสำเร็จ

ขอขอบพระคุณท่านคณะกรรมการสอบวิทยานิพนธ์ อันได้แก่ ผู้ช่วยศาสตราจารย์ ดร.แสงหล้า ชัยมงคล รองศาสตราจารย์ ดร.เสาวณิต สุขภารังษี รองศาสตราจารย์ ดร.วราฤทธิ์ พานิช-กิจโกศลกุล และผู้ช่วยศาสตราจารย์ ดร.พัทธ์ชนก ศรีสุรเดชชัย ที่ให้แนวคิดและความรู้เพิ่มเติมอันเป็นประโยชน์ต่องานวิจัย พร้อมทั้งตรวจสอบข้อผิดพลาดให้มีความสมบูรณ์เพิ่มมากขึ้น ซึ่งเป็นประโยชน์ต่อการทำวิจัยเป็นอย่างมาก

ขอขอบพระคุณ คณาจารย์สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่เป็นผู้เริ่มให้ความรู้ทางสถิติแก่ผู้วิจัย เพื่อมาประยุกต์ใช้ในวิทยานิพนธ์ รวมท้งเจ้าหน้าที่ภาควิชาคณิตศาสตร์และสถิติ ที่ให้ความเชื่อ และอำนวยความสะดวกในการจัดทำวิทยานิพนธ์ครั้งนี้เป็นอย่างดี

สุดท้ายนี้ขอขอบพระคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่ให้โอกาสในการศึกษาหาความรู้ทางด้านสถิติ โดยมอบทุนการศึกษาตลอดการศึกษาในระดับปริญญาโทแก่ผู้วิจัย

นางสาวชุติกาญจน์ ชูสวัสดิ์

มิถุนายน 2561

สารบัญ

สารบัญตาราง

สารบัญภาพ

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ	คำเต็ม/คำจำกัดความ
n	ขนาดตัวอย่าง
p	จำนวนตัวแปรอิสระ
β_j	ค่าพารามิเตอร์สัมประสิทธิ์การถดถอยตัวที่ j
$\boldsymbol{\beta}$	เวกเตอร์ค่าพารามิเตอร์สัมประสิทธิ์การถดถอย
$\hat{\boldsymbol{\beta}}_{ols}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด ในตัวแบบถดถอยเชิงเส้น
$\hat{\boldsymbol{\beta}}_{ML}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด ในตัวแบบถดถอยปัวซง
$\hat{\boldsymbol{\beta}}^{*}_{ridge}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ ในตัวแบบถดถอยเชิงเส้น
$\hat{\boldsymbol{\beta}}_{ridge}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ ในตัวแบบถดถอยปัวซง
$\hat{\boldsymbol{\beta}}^{*}_{lasso}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบแลซโซ ในตัวแบบถดถอยเชิงเส้น
$\hat{\boldsymbol{\beta}}_{lasso}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบแลซโซ ในตัวแบบถดถอยปัวซง
$\hat{\boldsymbol{\beta}}^{*}_{adaplasso}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แลซโซแบบปรับปรุง ในตัวแบบถดถอยเชิงเส้น

$\hat{\boldsymbol{\beta}}_{adaplasso}$	เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แลซโซแบบปรับปรุง ในตัวแบบถดถอยปัวซง
r	ค่าสหสัมพันธ์บางส่วนของตัวแปรอิสระ 2 ตัวใดๆ
k	จำนวนกลุ่มการจำแนก/จำนวนตัวแปรอิสระ $(k>0)$
Σ_k	เมทริกซ์สหสัมพันธ์ขนาด $(k\times k)$
g_k	ขนาดของจำนวนกลุ่มการจำแนก/ขนาดตัวอย่าง
ρ_k	ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม โดยมีค่าอยู่ระหว่าง $0 \le \rho_k \le 1$
$\rho_{\max}$	ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม ที่มีค่ามากที่สุด $\max\{\rho_1,...,\rho_k\}$
$\rho_{\min}$	ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม ที่มีค่าน้อยที่สุด $\min\{\rho_1,...,\rho_k\}$
τ_k	ช่วงห่างในแถว/คอลัมน์แรกภายในvกลุ่ม
α_k	ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่มในรูปแบบความสัมพันธ์แบบ Hub Toeplitz
T	ข้อมูลทั้งหมด
v	จำนวนข้อมูลที่ถูกแบ่งออกเป็นส่วนๆ
$T-T_v$	ข้อมูลชุดทดลอง (Training Set)
T_v	ข้อมูลชุดทดสอบ (Test Set)
λ	พารามิเตอร์ปรับแต่ง (Tuning Parameter) $(\lambda>0)$

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

การนำสถิติศาสตร์มาใช้ในงานวิจัยและการวิเคราะห์ข้อมูลเป็นสิ่งสำคัญ คือ การเลือกวิธีการที่มีความเหมาะสมกับลักษณะข้อมูลและวัตถุประสงค์ของการวิจัย โดยในหลายๆงานวิจัยมีวัตถุประสงค์เพื่อศึกษาลักษณะความสัมพันธ์ระหว่างตัวแปรที่ต้องการศึกษากับตัวแปรต่างๆ ที่เป็นปัจจัยที่เกี่ยวข้องกับตัวแปรที่ต้องการศึกษา โดยเรียกตัวแปรที่ต้องการศึกษาว่า ตัวแปรตอบสนอง (Response Variable) และเรียกตัวแปรต่างๆที่เกี่ยวข้องกับการศึกษาว่าตัวแปรอิสระ (Independent Variable)

ในกรณีที่ตัวแปรตอบสนองเป็นตัวแปรสุ่มชนิดต่อเนื่อง วิธีการหนึ่งที่ใช้กันอย่างแพร่หลาย คือ การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression Analysis) แต่ถ้าตัวแปรตอบสนองเป็นข้อมูลจำนวนนับ (Count Data) ตัวแบบพื้นฐานที่นิยมนำมาใช้กันอย่างแพร่หลาย คือ ตัวแบบการถดถอยปัวซง (Poisson Regression Model)

การวิเคราะห์การถดถอยสามารถแบ่งออกเป็น 2 ชนิด ถ้ามีตัวแปรอิสระในตัวแบบเพียงตัวเดียว เรียกว่า การถดถอยอย่างง่าย (Simple Regression) แต่ถ้ามีตัวแปรอิสระมากกว่าสองตัวขึ้นไป จะเรียกว่า การถดถอยพหุคูณ (Multiple Regression)

การวิเคราะห์การถดถอยพหุคูณ (Multiple Regression Analysis) ใช้ศึกษาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตาม โดยที่ตัวแปรอิสระมีตั้งแต่ 2 ตัวขึ้นไป ซึ่งการวิเคราะห์การถดถอยเชิงเส้นพหุคูณที่นิยมใช้กัน คือ วิธีกำลังสองน้อยที่สุด (Ordinary Least Square : OLS) เป็นวิธีมาตรฐานในการประมาณค่าสัมประสิทธิ์ถดถอย โดยวิธีดังกล่าวมีคุณสมบัติไม่เอนเอียง ดังนั้น วิธีการยกกำลังสองน้อยที่สุด จึงมีคุณสมบัติเรียกสั้นๆ ว่า BLUE (Best Linear Unbiased Estimator) นอกจากนี้วิธีภาวะความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation : MLE) เป็นอีกวิธีหนึ่งที่มีคุณสมบัติใกล้เคียงกับวิธียกกำลังสองน้อยที่สุดด้วย

งานวิจัยนี้สนใจในกรณีที่มีจำนวนตัวแปรอิสระเป็นจำนวนมาก อาจทำให้เกิดปัญหาความสัมพันธ์กันเองของตัวแปรอิสระ ซึ่งเป็นปัญหาที่เห็นโดยทั่วไป เมื่อจำนวนตัวแปรอิสระมีจำนวนมาก เรียกว่า ปัญหามัลติคอลิเนียร์ลิตี้ หรือภาวะร่วมเชิงเส้น (Multicollinearity) ตัวอย่างเช่น ผู้วิจัย

ศึกษาตัวแปรทางด้านเศรษฐศาสตร์ โดยศึกษาการเปลี่ยนแปลงทางด้านเศรษฐกิจ โดยมีผลกระทบมาจากหลายๆตัวแปร โดยที่ตัวแปรเหล่านั้นส่งผลกระทบต่อเศรษฐกิจในลักษณะใกล้เคียงกัน ซึ่งรูปแบบความสัมพันธ์ของตัวแปรอิสระก็มีอยู่ด้วยกันหลายวิธี ในปี 2013 Johanna Hardin, Stephan Ramon Garcia และ David Golan ได้นำเสนอสร้างเมทริกซ์สหสัมพันธ์ (Correlation Matrix) 3 รูปแบบ คือ Constant correlation model, Toeplitz correlation model และ Hub Toeplitz correlation model เพื่อสร้างสถานการณ์ของตัวแปรอิสระให้แตกต่างกัน โดยรูปแบบความสัมพันธ์เหล่านี้ ส่วนมากจะใช้ในการวิเคราะห์การจำแนก และการจัดหมวดหมู่

ในบางครั้งการที่มีตัวแปรอิสระมากเกินไป แต่ขนาดของข้อมูลที่เราศึกษาหรือขนาดตัวอย่างมีน้อย ทำให้เกิดความไม่เพียงพอต่อการวิเคราะห์ข้อมูล จึงก่อให้เกิดปัญหาในการวิเคราะห์เช่นกัน โดยเรียกลักษณะข้อมูลนี้ว่า ข้อมูลที่มีมิติสูง (High-Dimensional) เช่น ในทางการแพทย์ ต้องการศึกษาเชื้อไวรัสในหนู 10 ตัว โดยเชื้อไวรัสที่ต้องการศึกษามีมากกว่า 100 ชนิด จะเห็นว่า เรามีขนาดตัวอย่างเพียง 10 ตัว แต่เชื้อไวรัสที่เราต้องการศึกษามีเป็นจำนวนมาก เป็นต้น นอกจากนี้ ในกรณีที่มีตัวแปรอิสระเป็นจำนวนมาก อาจจะมีตัวแปรอิสระบางตัวเท่านั้นที่ควรอยู่ในตัวแบบ เรียกว่า ตัวแปรที่มีผล (Active Variable) และบางตัวไม่ควรอยู่ในตัวแบบ (Inactive Variable) นั่นคือ ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์ และถ้าในตัวแบบมีค่าสัมประสิทธิ์การถดถอยส่วนใหญ่เป็นศูนย์ จะเรียกลักษณะตัวแบบนี้ว่า ตัวแบบบางเบา (Sparse Model)

งานวิจัยนี้จึงได้ศึกษาการวิเคราะห์การถดถอยพหุคูณในตัวแบบปัวซง ในลักษณะข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูงอีกด้วย เมื่อเกิดปัญหาเหล่านี้ วิธีการวิเคราะห์การถดถอยเชิงเส้นพหุคูณด้วยวิธีกำลังสองน้อยที่สุดหรือวิธีภาวะน่าจะเป็นสูงสุด อาจจะไม่เป็นวิธีที่ดีนัก เพราะเมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง ส่งผลให้ตัวประมาณด้วยวิธีกำลังสองน้อยที่สุดมีความแปรปรวนสูงขึ้น และทำให้การประมาณค่าไม่ถูกต้อง และไม่มีประสิทธิภาพ นอกจากนี้การอธิบายผลของตัวแบบมีความยากและซับซ้อนมากขึ้น ดังนั้น วิธีกำลังสองน้อยที่สุด จึงเป็นวิธีที่ไม่เหมาะสมกับสถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์กันสูง หรือในข้อมูลที่มีมิติสูง วิธีการหนึ่งที่นิยมใช้ในการวิเคราะห์การถดถอยในข้อมูลที่มีลักษณะดังกล่าว คือ วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ (Penalized Regression) เพื่อหาค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอย $\left(\boldsymbol{\beta}\right)$ ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) ดังสมการ $\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\left(-l\left(\boldsymbol{\beta}\right)\right) + P_{\lambda}\left(\boldsymbol{\beta}\right)$ มีค่าน้อยที่สุดภายใต้เงื่อนไขที่แตกต่างกัน ที่เรียกว่า ฟังก์ชันพีนอลตี้ (Penalty Function) โดยฟังก์ชันนี้จะมีด้วยกันหลายรูปแบบ ซึ่งจะให้ค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอยแตกต่างกัน ในงานวิจัยนี้จะศึกษาวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ (Penalized Regression) 3 วิธี คือ วิธีการ

วิเคราะห์การถดถอยแบบริดจ์ (Ridge regression), วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแบบแลซโซ (Adaptive LASSO)

ในปี 1970 Hoerl และ Kennard ได้เสนอวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย เรียกว่า การถดถอยแบบริดจ์ (Ridge Regression) ในตัวแบบเชิงเส้น เพื่อแก้ปัญหาตัวแปรอิสระมีความสัมพันธ์กัน หรือเกิดภาวะร่วมเชิงเส้น โดยตัวประมาณที่ได้จากวิธีนี้จะช่วยลดความแปรปรวนและความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เมื่อเกิดปัญหาดังกล่าว แต่เป็นตัวประมาณที่เอนเอียง (Bias Estimator) และต่อมา Kristofer และ Ghazi (2011) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ สำหรับตัวแบบการถดถอยปัวซง เป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง สามารถคำนวณหาค่าสัมประสิทธิ์การถดถอย จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood จากวิธีประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุด แต่เนื่องจากวิธีของริดจ์ สามารถใช้ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยทุกตัวให้มีค่าเข้าใกล้ศูนย์ หรือทำให้มีขนาดเล็กลง (Shrink) และตัวประมาณที่ได้จะมีความเสถียร แต่วิธีนี้ยังขาดคุณสมบัติในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ

ในปี 1996 Tibshirani จึงได้นำเสนอวิธีการหนึ่งเพื่อแก้ไขคุณสมบัติของตัวประมาณด้วยวิธีริดจ์ในตัวแบบเชิงเส้นให้ดีขึ้น เรียกการประมาณค่านี้ว่า แลซโซ (LASSO: Least Absolute Shrinkage and Selection Operator) โดยวิธีนี้นอกจากจะประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย ยังสามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้อีกด้วย ตัวประมาณที่ได้จากวิธีแลซโซเป็นตัวประมาณที่เอนเอียง แต่สามารถลดความแปรปรวนได้เช่นเดียวกับตัวประมาณด้วยวิธีริดจ์ ต่อมาในปี 2007 Park และ Hastie ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซโซ สำหรับตัวแบบการถดถอยปัวซง แต่ตัวประมาณที่ได้จากวิธีของแลซโซ ถึงแม้ว่าจะสามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้นั้น แต่ถ้ากรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง หรือเกิดภาวะร่วมเชิงเส้น วิธีแลซโซ จะเลือกตัวแปรเพียงตัวเดียวจากตัวแปรที่มีความสัมพันธ์กันในกลุ่มนั้น โดยไม่คำนึงว่าตัวแปรอิสระนั้นมีความสัมพันธ์กันตัวแปรสนองมากที่สุดหรือไม่ ดังนั้น วิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีแลซโซ ยังมีข้อจำกัดบางประการในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบที่ไม่มีความคงเส้นคงวา (Consistency)

ในปี 2006 Zou ได้นำเสนอวิธีการหนึ่งเพื่อแก้ไขคุณสมบัติความไม่คงเส้นคงวาในการคัดเลือกตัวแปรเข้าสู่ตัวแบบของตัวประมาณด้วยวิธีแลซโซ สำหรับตัวแบบเชิงเส้น โดยมีการถ่วงน้ำหนัก เพื่อให้ความสำคัญกับตัวแปรแต่ละตัวในการคัดเลือกเข้าสู่ตัวแบบ ซึ่งการถ่วงน้ำหนักนี้ให้กับตัวประมาณแลซโซแบบเดิมนั้น มีชื่อเรียกใหม่ว่า Adaptive LASSO ในปี 2007 Park และ Hastie

ได้ศึกษาต่อจากงานของ Fan และ Li ในปี 2001 จากข้อจำกัดที่ว่า วิธี LASSO ยังขาดคุณสมบัติความคงเส้นคงวา ในตัวแบบการถดถอยปัวซง ส่งผลทำให้การประมาณค่าด้วยวิธีนี้มีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบมากขึ้น ทำให้ช่วยลดเอนเอียงในการประมาณค่าดีกว่าวิธีของแลซโซแบบเดิม

ดังนั้นในงานวิจัยนี้ ผู้วิจัยจึงได้ศึกษาการเปรียบเทียบประสิทธิภาพของตัวประมาณด้วยวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 3 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge regression), วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแบบแลซโซ (Adaptive LASSO) ในตัวแบบการถดถอยปัวซง กรณีที่ข้อมูลมีมิติสูง และตัวแปรอิสระเกิดภาวะร่วมเชิงเส้นสูง ในแต่ละรูปแบบความสัมพันธ์ โดยสนใจทั้งหมด 3 รูปแบบ คือ Constant model , Toeplitz model, และ Hub Toeplitz model และมีค่าสัมประสิทธิ์การถดถอยแบบบางเบา เรียกว่า ตัวแบบบางเบา (Sparse Model) หรือกล่าวอีกอย่างหนึ่งว่าค่าสัมประสิทธิ์การถดถอยส่วนน้อยไม่เป็นศูนย์ และส่วนมากเป็นศูนย์ อยู่ภายใต้สถานการณ์ที่แตกต่างกันหลายเงื่อนไข โดยเงื่อนไขที่กำหนดจะครอบคลุมคุณสมบัติของตัวประมาณในแต่ละวิธี

1.2 วัตถุประสงค์ของการศึกษา

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบการวิเคราะห์การถดถอยปัวซง (Poisson Regression) สำหรับข้อมูลที่มีมิติสูงแบบบางเบา ในกรณีที่ตัวแปรอิสระเกิดภาวะร่วมเชิงเส้น ทั้ง 3 รูปแบบ นั่นคือ Constant model , Toeplitz model, และ Hub Toeplitz โดยเปรียบเทียบประสิทธิภาพของเครื่องมือในการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 3 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge regression), วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแบบแลซโซ (Adaptive LASSO)

1.3 ขอบเขตของการศึกษา

ผู้วิจัยจำลองข้อมูลที่ใช้ในการทดลองภายใต้สถานการณ์ต่างๆ ดังนี้

1. ขนาดตัวอย่าง $n = 25$ และ 50
2. จำนวนตัวแปรอิสระ $p = 50,100$ และ 200
3. กำหนดสถานการณ์ของการทดลองสำหรับแต่ละค่า n และ p โดยกำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระ $\mathbf{X}$ และค่าพารามิเตอร์ $\boldsymbol{\beta}$ ในแต่ละกรณี ดังนี้
   3.1 ตัวแปรอิสระที่สัมพันธ์กันแบ่งออกเป็น 1 กลุ่ม

3.1.1 กรณีตัวแปรอิสระมีความสัมพันธ์แบบคงที่ (Constant Model)

3.1.1.1 ตัวแปรอิสระทั้งหมด p ตัวที่สัมพันธ์กัน $\left(x_{i(1)},...,x_{i(p)}\right)$ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)},...,x_{i(p)}\right)$ คือ r เมื่อ $r=0.5, \quad i,j=1,...,p$

$\beta_1=1,\ \beta_2,\beta_3=-0.5,\ \beta_4,...,\beta_6=0.1,\ \beta_7,...,\beta_{10}=0.05,\ \beta_{11},...,\beta_{15}=0.01$ และ $\beta_{16},...,\beta_{p-15}=0$

3.1.1.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.2.1.1 ยกเว้น $r = 0.6, 0.7, 0.8$ และ 0.9 ตามลำดับ

3.1.2 กรณีตัวแปรอิสระมีความสัมพันธ์แบบโทพลิท (Toeplitz Model)

3.1.2.1 ตัวแปรอิสระทั้งหมด p ตัวที่สัมพันธ์กัน $\left(x_{i(1)},...,x_{i(p)}\right)$ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)},...,x_{i(p)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r=0.5, \quad i,j=1,...,p$ เมื่อ $\beta_1=1,\ \beta_2,\beta_3=-0.5,\ \beta_4,...,\beta_6=0.1,\ \beta_7,...,\beta_{10}=0.05,\ \beta_{11},...,\beta_{15}=0.01$ และ $\beta_{16},...,\beta_{p-15}=0$

3.1.2.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1.2.1 ยกเว้น $r = 0.6, 0.7, 0.8$ และ 0.9 ตามลำดับ

3.1.3 กรณีตัวแปรอิสระมีความสัมพันธ์แบบฮับโทพลิท (Hub Toeplitz Model)

3.1.3.1 ตัวแปรอิสระทั้งหมด p ตัวที่สัมพันธ์กัน $\left(x_{i(1)},...,x_{i(p)}\right)$ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)},...,x_{i(p)}\right)$ คือ $\alpha_{k,1}=1$ และ $\alpha_{k,i}=\rho_k-\tau_k\left(i-2\right)$

เมื่อ $\tau_k=\dfrac{\left(r_{\max}-r_{\min}\right)}{g_k-2}$ และ $r_{\max}=0.9,\ r_{\min}=0.5, \quad i,j=1,...,15$

โดยที่ g_k คือ ขนาดของจำนวนตัวแปรอิสระ $\left(g_k>0\right)$

และค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(16)},...,x_{i(p-15)}\right)$ คือ $\alpha_{k,1}=1$ และ $\alpha_{k,i}=\rho_k-\tau_k\left(i-2\right)$

เมื่อ $\tau_k=\dfrac{\left(r_{\max}-r_{\min}\right)}{g_k-2}$ และ $r_{\max}=0.9,\ r_{\min}=0.5, \quad i,j=1,...,15$

โดยที่ g_k คือ ขนาดของจำนวนตัวแปรอิสระ $\left(g_k>0\right)$ เมื่อ

$\beta_1 = 1,\ \beta_2, \beta_3 = -0.5,\ \beta_4, ..., \beta_6 = 0.1,\ \beta_7, ..., \beta_{10} = 0.05,\ \beta_{11}, ..., \beta_{15} = 0.01$ และ $\beta_{16}, ..., \beta_{p-15} = 0$

3.1.3.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1.3.1 ยกเว้น $r_{\min} = 0.6, 0.7$ และ 0.8 ตามลำดับ

3.2 ตัวแปรอิสระที่สัมพันธ์กันแบ่งออกเป็น 2 กลุ่ม

3.2.1 กรณีตัวแปรอิสระมีความสัมพันธ์แบบคงที่ (Constant Model)

3.2.1.1 โดยกลุ่มแรก คือ ตัวแปรอิสระ 15 ตัวที่สัมพันธ์กัน $\left(x_{i(1)}, ..., x_{i(15)}\right)$ และกลุ่มที่สอง คือ ตัวแปรอิสระตัวที่เหลือที่สัมพันธ์กัน $\left(x_{i(16)}, ..., x_{i(p-15)}\right)$ โดยที่ตัวแปรอิสระในกลุ่มที่ 1 และกลุ่มที่ 2 เป็นอิสระต่อกัน ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)}, ..., x_{i(15)}\right)$ คือ r เมื่อ และค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(16)}, ..., x_{i(p-15)}\right)$ คือ r เมื่อ $r = 0.5, \quad i, j = 1, ..., 15$ เมื่อ $\beta_1 = 1,\ \beta_2, \beta_3 = -0.5,\ \beta_4, ..., \beta_6 = 0.1,\ \beta_7, ..., \beta_{10} = 0.05,\ \beta_{11}, ..., \beta_{15} = 0.01$ และ $\beta_{16}, ..., \beta_{p-15} = 0$

3.2.1.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1.1.1 ยกเว้น $r = 0.6, 0.7, 0.8$ และ 0.9 ตามลำดับ

3.2.2 กรณีตัวแปรอิสระมีความสัมพันธ์แบบโทพลิท (Toeplitz Model)

3.2.2.1 โดยกลุ่มแรก คือ ตัวแปรอิสระ 15 ตัวที่สัมพันธ์กัน $\left(x_{i(1)}, ..., x_{i(15)}\right)$ และกลุ่มที่สอง คือ ตัวแปรอิสระตัวที่เหลือที่สัมพันธ์กัน $\left(x_{i(16)}, ..., x_{i(p-15)}\right)$ โดยที่ตัวแปรอิสระในกลุ่มที่ 1 และกลุ่มที่ 2 เป็นอิสระต่อกัน ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)}, ..., x_{i(15)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r = 0.5, \quad i, j = 1, ..., 15$ และค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ$\left(x_{i(16)}, ..., x_{i(p-15)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r = 0.5, \quad i, j = 1, ..., 15$ เมื่อ $\beta_1 = 1,\ \beta_2, \beta_3 = -0.5,\ \beta_4, ..., \beta_6 = 0.1,\ \beta_7, ..., \beta_{10} = 0.05,\ \beta_{11}, ..., \beta_{15} = 0.01$ และ $\beta_{16}, ..., \beta_{p-15} = 0$

3.2.2.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1.2.1 ยกเว้น

$r = 0.6, 0.7, 0.8$ และ 0.9 ตามลำดับ

3.2.3 กรณีตัวแปรอิสระมีความสัมพันธ์แบบฮับโทพลิท (Hub Toeplitz Model)

3.2.3.1 โดยกลุ่มแรก คือ ตัวแปรอิสระ 15 ตัวที่สัมพันธ์กัน $\left(x_{i(1)},...,x_{i(15)}\right)$ และกลุ่มที่สอง คือ ตัวแปรอิสระตัวที่เหลือที่สัมพันธ์กัน $\left(x_{i(16)},...,x_{i(p-15)}\right)$ โดยที่ตัวแปรอิสระในกลุ่มที่ 1 และกลุ่มที่ 2 เป็นอิสระต่อกัน ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)},...,x_{i(15)}\right)$ คือ $\alpha_{k,1}=1$ และ $\alpha_{k,i}=\rho_k-\tau_k\left(i-2\right)$

เมื่อ $\tau_k=\dfrac{\left(r_{\max}-r_{\min}\right)}{g_k-2}$ และ $r_{\max}=0.9,\ r_{\min}=0.5,\quad i,j=1,...,15$

โดยที่ g_k คือ ขนาดของจำนวนตัวแปรอิสระ $\left(g_k>0\right)$

และค่าสัมประสิทธิ์สหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(16)},...,x_{i(p-15)}\right)$ คือ $\alpha_{k,1}=1$ และ $\alpha_{k,i}=\rho_k-\tau_k\left(i-2\right)$

เมื่อ $\tau_k=\dfrac{\left(r_{\max}-r_{\min}\right)}{g_k-2}$ และ $r_{\max}=0.9,\ r_{\min}=0.5,\quad i,j=1,...,15$

โดยที่ g_k คือ ขนาดของจำนวนตัวแปรอิสระ $\left(g_k>0\right)$ เมื่อ

$\beta_1=1,\ \beta_2,\beta_3=-0.5,\ \beta_4,...,\beta_6=0.1,\ \beta_7,...,\beta_{10}=0.05,\ \beta_{11},...,\beta_{15}=0.01$

และ $\beta_{16},...,\beta_{p-15}=0$

3.2.3.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1.3.1 ยกเว้น $r_{\min} = 0.6, 0.7$ และ 0.8 ตามลำดับ

1.4 เกณฑ์ที่ใช้ในการพิจารณา

1. ความแม่นยำในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ เมื่อกำหนด $\boldsymbol{\beta}$ ให้มีลักษณะเป็นข้อมูลแบบบางเบา (Sparse Data) นั่นคือ สัมประสิทธิ์การถดถอยส่วนใหญ่มีค่าเป็นศูนย์และบางส่วนไม่ได้มีค่าเป็นศูนย์ โดยนับการจำนวนตัวแปรที่ถูกคัดเลือกผิดพลาดไปจากตัวแบบโดยความผิดพลาดในการคัดเลือกตัวแปรมี 2 แบบ คือ กรณีที่ค่าพารามิเตอร์สัมประสิทธิ์การถดถอยไม่เท่ากับ 0 แต่ค่าประมาณสัมประสิทธิ์การถดถอยเท่ากับ 0 (Identify Criterion 1 : IC1) และกรณีที่ค่าพารามิเตอร์สัมประสิทธิ์การถดถอยเท่ากับ 0 แต่ค่าประมาณสัมประสิทธิ์การถดถอยไม่เท่ากับ 0 (Identify Criterion 2 : IC2) ดังนี้

$$\text{IC1} = \#\left\{j:\ \beta_j\neq 0,\ \hat{\beta}_j=0\right\} \quad \text{และ} \quad \text{IC2} = \#\left\{j:\ \beta_j=0,\ \hat{\beta}_j\neq 0\right\}$$

แล้วพิจารณาความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปร ดังนี้

1.1 อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) เป็นการวัดความน่าจะเป็นที่จะเกิดความผิดพลาดจาก Identify Criterion 1: $\mathrm{IC1}$ สามารถคำนวณได้ดังนี้

$$P(\mathrm{IC1}) = \frac{\mathrm{IC1}}{15 \times m}$$

1.2 อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FNR) เป็นการวัดความน่าจะเป็นที่จะเกิดความผิดพลาดจาก Identify Criterion 2: $\mathrm{IC2}$ สามารถคำนวณได้ดังนี้

$$P(\mathrm{IC2}) = \frac{\mathrm{IC2}}{(p-15) \times m}$$

เมื่อ m คือ จำนวนครั้งของการจำลองข้อมูล

2. ประสิทธิภาพของการพยากรณ์โดยวัดจากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าพยากรณ์ (median of Prediction Mean Square Error : mPMSE) มีค่าน้อยที่สุด

โดยที่ $\mathrm{PMSE}_r = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$ เมื่อ n ขนาดตัวอย่าง สำหรับ $r = 1, 2, ..., m$

1.5 ประโยชน์ที่คาดว่าจะได้รับ

เนื่องจากในการวิเคราะห์การถดถอยพหุคูณ อาจจะไม่ได้มีเพียงแค่การวิเคราะห์ในรูปแบบเชิงเส้นเพียงอย่างเดียว แต่ในบางครั้งตัวแปรที่เราต้องการศึกษา อาจมีลักษณะข้อมูลที่เป็นจำนวนนับ และอาจจะมีตัวแปรอิสระเป็นจำนวนมากในการพยากรณ์ ซึ่งตัวแบบที่นิยมใช้ คือ การวิเคราะห์การถดถอยแบบปัวซง แต่ในทางตรงข้าม ขนาดตัวอย่างที่รวบรวมมาได้ อาจน้อยกว่าจำนวนตัวแปรอิสระ ซึ่งทำให้เกิดลักษณะของข้อมูลที่มีมิติสูง เช่น ข้อมูลทางเศรษฐกิจ ข้อมูลทางชีววิทยา และข้อมูลทางด้านการแพทย์ เป็นต้น และเมื่อตัวแปรอิสระเป็นจำนวนมาก อาจจะเกิดความสัมพันธ์กันเองขึ้น ดังนั้น ผู้วิจัยจึงหวังว่า วิธีการวิเคราะห์การถดถอยปัวซง ในข้อมูลที่มีมิติสูง ด้วยวิธีการวิเคราะห์แบบ-ริดจ์ วิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง จะสามารถแก้ไขปัญหาในการพยากรณ์ให้มีประสิทธิภาพเพิ่มมากขึ้น เมื่อข้อมูลมีลักษณะแตกต่างกันออกไป โดยเปรียบเทียบประสิทธิภาพจากสถานการณ์ที่จำลองขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ตัวแบบการถดถอยเชิงเส้น (Linear Regression Model)

เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนอง (Response Variable) กับตัวแปรอิสระ (Independent Variable) โดยที่ตัวแปรตอบสนองมีการแจกแจงแบบปกติ หรือตัวแปรตอบสนองเป็นตัวแปรสุ่มชนิดต่อเนื่อง โดยมีจุดมุ่งหมายเพื่อการประมาณค่าพารามิเตอร์สัมประสิทธิ์ถดถอยให้เหมาะสมกับข้อมูล เมื่อได้รูปแบบความสัมพันธ์ของตัวแปรที่เราต้องการศึกษากับตัวแปรอิสระ จะสามารถสร้างตัวแบบที่เหมาะสมกับข้อมูลเพื่ออธิบายความสัมพันธ์ของข้อมูลได้

พิจารณาตัวแบบพื้นฐานของการถดถอยเชิงเส้นพหุคูณ สำหรับตัวแปรอิสระ p ตัว และขนาดของกลุ่มตัวอย่างเท่ากับ n

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{เมื่อ} \qquad n > p \tag{2.1}$$

โดยที่

$\mathbf{Y}$ คือ เวกเตอร์ขนาด $(n \times 1)$ ของตัวแปรตามที่วัดได้จากกลุ่มตัวอย่างขนาด n

$\mathbf{X}$ คือ เมทริกซ์ขนาด $\left(n \times (p+1)\right)$ แสดงค่าตัวทำนายทั้ง p ตัวที่วัดได้จากกลุ่มตัวอย่างขนาด n

$\boldsymbol{\beta}$ คือ เวกเตอร์ขนาด $(n \times 1)$ ของสัมประสิทธิ์การถดถอยของประชากร

$\boldsymbol{\varepsilon}$ คือ เวกเตอร์ขนาด $(n \times 1)$ ของความคลาดเคลื่อนของตัวแปรตาม

โดยมีคุณสมบัติ

$$E(\boldsymbol{\varepsilon}) = 0 \quad \text{และ} \quad Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n > 0$$

เมื่อ เวกเตอร์ของตัวแปรตาม $\mathbf{Y} = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$

เวกเตอร์ของค่าสังเกต $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)^T \in \mathbb{R}^{n \times p}$

และเวกเตอร์ของสัมประสิทธิ์การถดถอย $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T \in \mathbb{R}^p$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

แต่ในบางครั้งตัวแปรตอบสนองอาจไม่เป็นตัวแปรสุ่มชนิดต่อเนื่อง แต่มีลักษณะข้อมูลเป็นจำนวนนับ(Count Data) เช่น ข้อมูลทางการแพทย์ ผู้วิจัยอาจจะต้องการศึกษา ปัจจัยที่มีผลต่อจำนวนครั้งของการกำเริบ (Exacerbation) ในผู้ป่วยโรคหลอดลมอุดตันเรื้อรัง (Chronic Obstructive Pulmonary Disease : COPD) หรือศึกษาจำนวนคนที่เสียชีวิตด้วยโรคเอดส์ในช่วงเวลา 3 เดือน ตั้งแต่เดือนมกราคม 2526 จนถึง เดือนมิถุนายน 2529 ดังนั้น ในการศึกษาความสัมพันธ์ของตัวแปรตอบสนองที่มีลักษณะดังกล่าวกับตัวแปรอิสระต่างๆ ตัวแบบที่นิยมศึกษา คือ ตัวแบบการถดถอยปัวซง (Poisson Regression Model)

2.1.2 ตัวแบบการถดถอยปัวซง (Poisson Regression Model)

กำหนดให้ Y เป็นตัวแปรสุ่มที่มีการแจกแจงปัวซง $\left(Y \sim \text{Poisson}\left(\mu\right)\right)$ และมีฟังก์ชันมวลความน่าจะเป็น (Probability Mass Function : p.m.f) ดังนี้

$$P\left(Y=y\right) = f\left(y\right) = \frac{\mu^{y}\exp\left(-\mu\right)}{y!} \qquad \text{เมื่อ} \quad y = 0,1,2,... \tag{2.2}$$

ซึ่ง $f\left(y\right)$ คือ ค่าฟังก์ชันของตัวแปรสุ่ม y ที่มีค่าความน่าจะเป็นที่ Y มีค่าเท่ากับ y และ $y! = y\left(y-1\right)...3\cdot 2\cdot 1$

ค่าเฉลี่ยและความแปรปรวนของการแจกแจงปัวซง คือ

$$E\left[Y\right] = \mu \qquad \text{และ} \qquad Var\left[Y\right] = \mu$$

ตัวแบบการถดถอยปัวซง (Poisson Regression Model) เป็นตัวแบบที่ใช้ในการวิเคราะห์หาความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรอิสระ ซึ่งจะถูกนำมาใช้บ่อยครั้งสำหรับการวิเคราะห์ข้อมูลแบบนับ โดยตัวแปรตอบสนองเป็นตัวแปรแบบไม่ต่อเนื่อง คือ ข้อมูลเป็นจำนวนนับ เช่น จำนวนของการเกิดเหตุการณ์ในช่วงระยะเวลาที่ต่อเนื่องกันในเวลาใดเวลาหนึ่งหรือเกิดขึ้นในพื้นที่หนึ่งที่ต่อเนื่องกัน ส่วนตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพ

กำหนดให้ $\mathbf{y} = \left(y_1,...,y_n\right)^T$ คือ เวกเตอร์ของตัวแปรตอบสนอง มีขนาดของตัวอย่างเท่ากับ n และเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu} = \left(\mu_1,...,\mu_n\right)^T$

ดังนั้น ฟังก์ชันมวลความน่าจะเป็น (Probability Mass Function : p.m.f) ของตัวแบบการถดถอยปัวซง เขียนได้ดังนี้

$$f\left(y_i;\mu_i,\mathbf{X}_i\right)=\frac{\mu_i^{\,y_i}\exp\left(-\mu_i\right)}{y_i!} \qquad \text{เมื่อ} \quad i=1,2,3,\ldots \text{ และ} \quad \mu_i>0$$

โดยที่ $\mu_i = \mu\left(\mathbf{X}_i,\beta\right) = \exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)$

เมื่อ $\mathbf{X}_i$ คือ เมทริกซ์ของตัวแปรอิสระที่ i และ $\boldsymbol{\beta}$ คือ เวกเตอร์ของสัมประสิทธิ์การถดถอย ซึ่งเป็นพารามิเตอร์ที่ไม่ทราบค่าและสามารถประมาณค่าได้ด้วยวิธีภาวะน่าจะเป็นสูงสุด

ดังนั้น ตัวแบบการถดถอยปัวซง

$$Y_i=\mu_i+\varepsilon_i \qquad \text{เมื่อ} \quad i=1,2,3,\ldots,n$$

โดยที่ $\varepsilon_i \sim N\left(0,\sigma^2\right)$

$$Y_i=\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)+\varepsilon_i$$

$$E\left(Y_i\right)=\exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)$$

$$\log\left(E\left(Y_i\right)\right)=\mathbf{X}_i^T\boldsymbol{\beta} \qquad (2.3)$$

ข้อตกลงของตัวแบบปัวซง มีดังนี้

ความน่าจะเป็นของการเกิดเหตุการณ์เป็นค่าคงที่ทุกจุดในช่วงระยะเวลาที่ต่อเนื่องกันในเวลาใดเวลาหนึ่งและความแปรปรวนของข้อมูลตัวแปรตามมีค่าเท่ากับค่าเฉลี่ย แต่การใช้ตัวแบบการถดถอยปัวซงนั้นอาจพบปัญหาค่าความแปรปรวนของข้อมูลตัวแปรตอบสนองมีค่ามากกว่าค่าเฉลี่ย หรือเรียกว่า เกิดปัญหา overdispersion และกรณีที่ความแปรปรวนน้อยกว่าค่าเฉลี่ย จะเรียกว่า underdispersion ซึ่งไม่เป็นไปตามข้อตกลงเบื้องต้นของตัวแบบการถดถอยปัวซง จากการศึกษาของ Ismail และ Jermain ในปี 2007 พบตัวแบบทวินามลบ และตัวแบบการถดถอยปัวซงวางนัยทั่วไป สามารถจัดการกับปัญหา overdispersion และ underdispersion ได้ดีกว่าตัวแบบการถดถอยปัวซง

สำหรับตัวแบบการถดถอย ทั้งในตัวแบบการถดถอยเชิงเส้น และตัวแบบถดถอยปัวซง เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนอง (Response Variable) กับตัวแปรอิสระ (Independent Variable) โดยมีวัตถุประสงค์เพื่อการประมาณค่าพารามิเตอร์สัมประสิทธิ์ถดถอย $\left(\boldsymbol{\beta}\right)$ ให้เหมาะสมกับข้อมูล เพื่อใช้อธิบายความสัมพันธ์ดังกล่าว โดยสามารถใช้วิธีดังต่อไปนี้

2.1.3 วิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์ถดถอย

2.1.3.1 วิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method)

สำหรับตัวแบบการถดถอยเชิงเส้น (Linear Regression) วิธีกำลังสองน้อยที่สุดเป็นวิธีมาตรฐานในการประมาณค่าสัมประสิทธิ์ถดถอย และวิธีดังกล่าวมีคุณสมบัติไม่เอนเอียง ดังนั้นวิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method : OLS) จึงมีคุณสมบัติเรียกสั้นๆ ว่า **BLUE** (Best Linear Unbiased Estimator)

คุณสมบัติของวิธีกำลังสองน้อยที่สุด (Properties of Best Linear Unbiased Estimator) เมื่อเวกเตอร์สุ่ม $\boldsymbol{\varepsilon}$ มีการแจกแจงแบบปกติ จะได้ว่า $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \sigma^2 \mathbf{I}_N\right)$ นั่นคือ ในกรณีที่ความคลาดเคลื่อนเป็นอิสระต่อกัน การใช้การใช้ประมาณเชิงเส้นที่ไม่เอนเอียง ซึ่งคำนวณด้วยวิธีกำลังสองน้อยที่สุด ตัวประมาณของสัมประสิทธิ์การถดถอย ด้วยวิธีกำลังสองน้อยที่สุด คือ

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.4}$$

หลักการของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด $\left(\hat{\boldsymbol{\beta}}_{ols}\right)$ คือ ทำให้ผลบวกกำลังสองของค่าคลาดเคลื่อนน้อยที่สุด

$$\text{minimize}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)$$

เมื่อ $\boldsymbol{\beta}$ แทน เวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอย
จะเห็นว่า $\hat{\boldsymbol{\beta}}_{ols}$ ที่ประมาณได้จากวิธีกำลังสองน้อยที่สุด เป็นตัวประมาณที่ไม่เอนเอียงสำหรับ $\boldsymbol{\beta}$

พิจารณาจาก

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}_{ols}\right] &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E\left[\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

เมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}_{ols}$ คือ

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

พิจารณา กำหนดให้ $a=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ จะได้

$$Var\left(\hat{\boldsymbol{\beta}}\right)=Var\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right)=aVar\left[\mathbf{Y}\right]a^T$$
$$=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T$$
$$=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T X\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \;=\; \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ols}$ คือ

$$MSE\left(\hat{\boldsymbol{\beta}}_{ols}\right)=\sigma^2 Trace\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

พิจารณา ระยะห่างระหว่างเวกเตอร์ของตัวประมาณ $\left(\hat{\boldsymbol{\beta}}_{ols}\right)$ กับเวกเตอร์ของค่าประชากรจริง $\left(\boldsymbol{\beta}\right)$

$$L_2=\left(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}\right)^T\left(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}\right)$$
$$E\left[L_2\right]=E\left[\left(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}\right)^T\left(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}\right)\right]$$
$$=E\left[\hat{\boldsymbol{\beta}}_{ols}^T\hat{\boldsymbol{\beta}}_{ols}-2\hat{\boldsymbol{\beta}}_{ols}^T\boldsymbol{\beta}+\boldsymbol{\beta}^T\boldsymbol{\beta}\right]$$
$$=E\left[\hat{\boldsymbol{\beta}}_{ols}^T\hat{\boldsymbol{\beta}}_{ols}\right]-2\boldsymbol{\beta}E\left[\hat{\boldsymbol{\beta}}_{ols}^T\right]+\boldsymbol{\beta}^T\boldsymbol{\beta}$$

เมื่อ $E\left[\hat{\boldsymbol{\beta}}_{ols}^T\hat{\boldsymbol{\beta}}_{ols}\right]=\boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}+\sigma^2 Trace\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

$$\therefore\quad MSE\left(\hat{\boldsymbol{\beta}}_{ols}\right) \;=\; \boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}+\sigma^2 Trace\left(\mathbf{X}^T\mathbf{X}\right)^{-1}-2\boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}^T\boldsymbol{\beta} \;=\; \sigma^2 Trace\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

นอกจากนี้การประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) มีคุณสมบัติใกล้เคียงกับวิธีกำลังสองน้อยที่สุดเช่นกัน สำหรับใน**ตัวแบบถดถอยปัวซง (Poisson Regression)** วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) เป็นวิธีมาตรฐานในการประมาณค่าสัมประสิทธิ์ถดถอย

2.1.3.2 วิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

การประมาณพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีนี้ จะอาศัยหลักการของความน่าจะเป็น โดยหาฟังก์ชันน่าจะเป็น (Likelihood Function) ของตัวแปรสุ่ม แล้วจึงหาค่าสูงสุดของฟังก์ชันน่าจะเป็นนี้ โดยเทียบกับตัวประมาณที่ยังไม่ทราบค่า ดังนี้

จากฟังก์ชันมวลความน่าจะเป็น (Probability Mass Function : p.m.f) ของตัวแบบการถดถอยปัวซง

$$f\left(y_i;\mu_i,\mathbf{X}_i\right)=\frac{\mu_i^{\,y_i}\exp\left(-\mu_i\right)}{y_i!} \qquad \text{เมื่อ} \quad i=1,2,3,\ldots \text{ และ} \quad \mu_i>0$$

โดยที่ $\mu_i = \mu\left(\mathbf{X}_i,\beta\right) = \exp\left(\mathbf{X}_i^T\boldsymbol{\beta}\right)$

และตัวแบบการถดถอยปัวซง จะได้ ฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) คือ

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\frac{\left[\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)\right]^{y_i}\exp\left[-\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)\right]}{y_i!} \\ &= \frac{\left[\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)\right]^{\sum_{i=1}^{n}y_i}\exp\left[-\sum_{i=1}^{n}\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)\right]}{\prod_{i=1}^{n}y_i!} \end{aligned}$$

ใส่ ln ทั้งสองข้างของสมการ จะได้

$$l(\boldsymbol{\mu};\mathbf{y}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}y_i\ln\left[\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)\right]-\sum_{i=1}^{n}\mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\ln y_i!$$

แทนค่า $\mu_i = \mu\left(\mathbf{X}_i,\boldsymbol{\beta}\right) = \exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)$ จะได้

$$\begin{aligned} &= \sum_{i=1}^{n}y_i\ln\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\ln y_i! \\ &= \sum_{i=1}^{n}y_i\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)-\sum_{i=1}^{n}\ln y_i! \\ &= \sum_{i=1}^{n}\left(y_i\mathbf{X}_i^{\,T}\boldsymbol{\beta}-\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)-\ln y_i!\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}}\ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}y_i\mathbf{X}_i^{\,T}-\sum_{i=1}^{n}\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)\mathbf{X}_i \\ &= \sum_{i=1}^{n}\left(y_i-\exp\left(\mathbf{X}_i^{\,T}\boldsymbol{\beta}\right)\right)\mathbf{X}_i \qquad = \quad 0 \end{aligned}$$

เมื่อพิจารณาสมการข้างต้น สำหรับตัวแบบการถดถอยปัวซง จะเห็นว่ารูปแบบของตัวแบบการถดถอยปัวซงเป็นรูปแบบไม่เชิงเส้น (Non-Linear) ดังนั้น ในการประมาณค่า $\boldsymbol{\beta}$ จะมีคำตอบเช่นเดียวกับวิธี Iterative Weighted Least Square (IWLS) ในกรณีที่ μ_i เป็นฟังก์ชันของความแปรปรวน เนื่องจากวิธี IWLS ถูกพัฒนามาจากวิธีภาวะน่าจะเป็นสูง

และ ในปี 2011 บทความของ Kristofer Mansson และ Ghazi Shukur ได้นำเสนอตัวประมาณสัมประสิทธิ์ถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด ดังนี้

$$\hat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\hat{\mathbf{z}} \tag{2.5}$$

โดยที่ $\hat{\mathbf{W}} = diag\left[\hat{\mu}_i\right]$ คือ เมทริกซ์ทแยงมุมของความแปรปรวนในแต่ละค่าสังเกตที่ i ซึ่งในตัวแบบนี้ ความแปรปรวนมีค่าเท่ากับค่าเฉลี่ย (μ_i) จึงทำให้เส้นทแยงมุมของเมทริกซ์ มีค่าเท่ากับ μ_i และ $\hat{\mathbf{z}}$ คือ เวกเตอร์ของ $\hat{z}_i = \log\left(\hat{\mu}_i\right) + \dfrac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$

โดยตัวประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator: MLE) จะมีการแจกแจง Asymptotically Normal

ดังนั้นเราสามารถหาความแปรปรวน และค่าคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด ดังนี้
เมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}_{ML}$

$$Cov\left(\hat{\boldsymbol{\beta}}_{ML}\right) = \left[-E\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k^T}\right)\right]^{-1} = \left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}$$

ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ML}$

$$E\left(L^2{}_{ML}\right) = E\left(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}\right)^T\left(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}\right) = tr\left[\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\right] = \sum_{j=1}^{p}\frac{1}{\lambda_j^*}$$

โดยที่ λ_j^* คือ ค่าไอเกนของเมทริกซ์ $\mathbf{X}^T\hat{\mathbf{W}}\hat{\mathbf{z}}$

แต่ในบางครั้งในการวิเคราะห์การถดถอยพหุคูณ เมื่อตัวแปรอิสระมีจำนวนมาก อาจทำให้ตัวแปรอิสระมีความสัมพันธ์กันสูงหรือเกิดปัญหาภาวะร่วมเชิงเส้น ดังนั้น การประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบเชิงเส้น (Linear Regression) จะทำให้ $\det\left(\mathbf{X}^T\mathbf{X}\right)$ มีค่าเข้าใกล้ 0 และเมทริกซ์ผกผันของ $\mathbf{X}^T\mathbf{X}$ มีค่ามากๆ เข้าใกล้ ∞ ส่งผลทำให้ $Var\left(\hat{\boldsymbol{\beta}}_{ols}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ เข้าใกล้ อนันต์(∞) ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด จึงมีความไม่เสถียร (Uncertainty) และไม่มีประสิทธิภาพ (Inefficiency)

และการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบถดถอยปัวซง (Poisson Regression) เมทริกซ์น้ำหนัก $\left(\hat{\mathbf{W}}\right)$ ของเมทริกซ์ $\mathbf{X}^T\hat{\mathbf{W}}\hat{\mathbf{z}}$ จะไม่เป็นไปตามเงื่อนไข ทำให้เกิดความไม่เสถียร (Instability) และความแปรปรวนของตัวประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุดสูง ซึ่งในสถานการณ์เช่นนี้ ทำให้การอธิบายผลจากการประมาณค่าพารามิเตอร์เป็นไปได้ยาก เนื่องจากเวกเตอร์ของการประมาณค่าสัมประสิทธิ์การถดถอย (Coefficient) มีค่ามาก ดังนั้น จึงมีผู้วิจัยหลายท่านได้เสนอวิธีการประมาณค่าสัมประสิทธิ์การถดถอยวิธีอื่น ซึ่งจะกล่าวในหัวข้อถัดไป

2.1.3.3 วิธีการหดค่าลง (Shrinkage Method)

วิธีการหดค่าลง เป็นวิธีการเปลี่ยนแปลงค่าประมาณสัมประสิทธิ์การถดถอยให้มีขนาดเล็กลง หรือควบคุมไม่ให้มีค่าใหญ่เกินไป และค่าประมาณสัมประสิทธิ์การถดถอยอาจจะมีค่าเท่ากับศูนย์ได้ ซึ่งจะแตกต่างจากวิธีการคัดเลือกตัวแปร (Selection Variable) ยกตัวอย่างเช่น วิธี Stepwise จะเป็นวิธีการคัดเลือกตัวแปรอิสระ เฉพาะตัวที่มีอิทธิผลกับตัวแปรตามมากที่สุด โดยพิจารณาจากสถิติสอบ F ว่ามีนัยสำคัญ เข้าสู่ตัวแบบ และยังสามารถคัดเลือกตัวแปรออกในภายหลังได้ ถ้าพบว่าตัวแปรอิสระนั้น มีความสัมพันธ์กับตัวแปรอื่นๆ ซึ่งวิธี stepwise จะช่วยสามารถแก้ไขปัญหาตัวแปรอิสระมีความสัมพันธ์กันสูงได้ (Multicollinearity) แต่วิธีการหดค่าลง จะนำตัวแปรจะไม่ใช่การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ เพราะตัวแปรอิสระทุกตัว จะต้องนำมาคำนวณในตัวแบบทั้งหมด หลังจากนั้นวิธีการหดค่าลง จะทำการลงขนาดค่าประมาณสัมประสิทธิ์การถดถอย ให้มีขนาดเล็กจนมีค่าเข้าใกล้ศูนย์ หรือเท่ากับศูนย์ ซึ่งเป็นการแสดงความสัมพันธ์ว่าตัวแปรอิสระนั้น เป็นตัวแปรที่มีผล (Active Variables) หรือเป็นตัวแปรไม่มีผล (Inactive Variable) กับตัวแปรตาม จึงทำให้ค่าสัมประสิทธิ์การถดถอยด้วยวิธีหดค่าลง จะมีความต่อเนื่องมากกว่า วิธีการคัดเลือกตัวแปร แต่จะไม่สามารถลดความแปรปรวน และความเคลื่อนของการพยากรณ์ได้ นอกวิธีการหดค่าลงจะสามารถแก้ไขปัญหาตัวแปรอิสระมีความสัมพันธ์กันสูงได้ ยังสามารถแก้ไขปัญหาข้อมูลที่มิติสูงได้อีกด้วย โดยในงานวิจัยนี้ สนใจศึกษาการหดค่าลง (Shrinkage) ของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge Regression) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และ วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO)

2.1.3.4. วิธีการวิเคราะห์การถดถอบแบบริดจ์ (Ridge Regression)

สำหรับตัวแบบการถดถอยเชิงเส้น (Linear Regression) ในปี 1970 Hoerl และ Kennard ได้แนะนำวิธีการประมาณค่าสัมประสิทธิ์การถดถอยในกรณีที่เกิดปัญหาภาวะร่วมเชิง

เส้น (Multicollinearity) นั่นคือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ เพื่อแก้ไขปัญหาความไม่เพียงพอของวิธีกำลังสองน้อยที่สุด สำหรับปัญหาในลักษณะที่เมทริกซ์ $\mathbf{X}^T\mathbf{X}$ อยู่ในรูปเมทริกซ์สหสัมพันธ์ (Correlation Matrix) หรือเป็นเมทริกซ์ไม่เป็นเชิงตั้งฉาก (Non – Orthogonal) ที่ไม่ใกล้เคียงกับ เมทริกซ์เอกลักษณ์ (Unit Matrix) ดังนั้นวิธีการถดถอยด้วยวิธีกำลังสองน้อยที่สุด จึงแสดงความไว (Sensitive) ต่อความผิดพลาดของการประมาณพารามิเตอร์สัมประสิทธิ์ถดถอย ทำให้ผลลัพธ์ที่ได้ไม่แม่นยำ ซึ่งเราสามารถหาตัวประมาณสัมประสิทธิ์การถดถอย $\left(\hat{\boldsymbol{\beta}}\right)$ ด้วยวิธีการวิเคราะห์แบบริดจ์ ได้จาก

$$\hat{\boldsymbol{\beta}}^{*}_{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \mathbf{Y} - \sum_{j=1}^{p} X_j \beta_j \right\|^2 \quad \text{ภายใต้เงื่อนไข} \quad \boldsymbol{\beta}^T\boldsymbol{\beta} = r^2 < \lambda \qquad (2.6)$$

พิจารณา

ถ้าให้ $\boldsymbol{\beta}$ แทน เวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอยใดๆ

หลักการของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธี การวิเคราะห์แบบริดจ์ $\left(\hat{\boldsymbol{\beta}}^{*}_{ridge}\right)$ คือ ทำให้ผลบวกกำลังสองของค่าคลาดเคลื่อนน้อยที่สุด (ϕ)

$$\begin{aligned}
\phi &= \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) \\
&= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}\right)^T \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}\right) + \left(\mathbf{B} - \hat{\boldsymbol{\beta}}_{ols}\right)^T \mathbf{X}^T\mathbf{X}\left(\mathbf{B} - \hat{\boldsymbol{\beta}}_{ols}\right) \\
&= \phi_{\min} + \phi_0
\end{aligned}$$

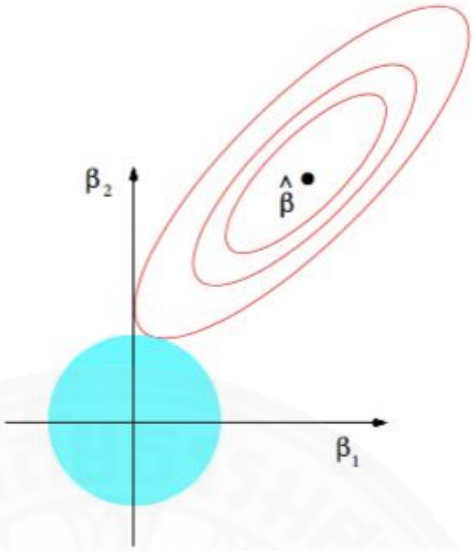
เมื่อคอนทัวร์ของค่าคงที่ ϕ คือ พื้นผิวของ hyperellipsoids ที่มีจุดศูนย์กลางที่ $\hat{\boldsymbol{\beta}}_{ols}$

โดยที่ $\phi_{\min}$ คือ ค่าน้อยที่สุดของ

ϕ_0 คือ กำลังสอง (Quadratic Form) ของ $\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ols}\right)$

ดังนั้น จึงทำให้เกิดความสัมพันธ์ $\phi = \phi_{\min} + \phi_0$ ซึ่ง $\phi_0 > 0$ เป็นส่วนที่ทำให้มีค่าเพิ่มขึ้น

วิธีการวิเคราะห์การถดถอยแบบริดจ์ ในตัวแบบเชิงเส้น สามารถแสดลักษณะพื้นที่ผิวตอบสนองของ $\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)^T \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)$ ด้วยภาพที่ 2.1 ดังนี้



ภาพที่ 2.1 แสดงของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีกำลังสองน้อยที่สุด (OLS) และของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge Regression)

ในเชิงคณิตศาสตร์ จุดต่ำสุดด้วยวิธีของริดจ์ (Minimum Ridge) สามารถหาได้ ดังนี้

$$\text{minimize } (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad \text{ภายใต้เงื่อนไข } \boldsymbol{\beta}^T\boldsymbol{\beta} = r^2 < \lambda$$

ถ้าให้ $\boldsymbol{\beta}$ แทน เวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอยใดๆ โดยต้องการให้ผลบวกกำลังสองของพื้นผิว $(\boldsymbol{\beta}^T\boldsymbol{\beta})$ มีขนาดเล็กที่สุด โดยการหาอนุพันธ์ เทียบกับ $\boldsymbol{\beta}$ ภายใต้ข้อจำกัด (Constraint) คือ $\boldsymbol{\beta}^T\boldsymbol{\beta} = r^2 < \lambda$

เนื่องจาก $\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ols}\right)^T\mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ols}\right) = \phi_0$

และแก้สมการหาคำตอบโดยใช้วิธีของลากรานจ์ (Lagrangrian) จะได้ว่า

กำหนดให้ $\text{minimize } F = \boldsymbol{\beta}^T\boldsymbol{\beta} + \left(\frac{1}{\lambda}\right)\left[\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ols}\right)^T\mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ols}\right) - \phi_0\right]$

เมื่อ $\frac{1}{\lambda}$ คือ ตัวคูณ ดังนั้น

$$\frac{\partial F}{\partial \mathbf{B}} = 2\boldsymbol{\beta} + \left(\frac{1}{\lambda}\right)\left[2\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} - 2\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ols}\right] = 0$$

จะได้ว่า $$2\boldsymbol{\beta} = -\left(\frac{1}{\lambda}\right)\left[2\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} - 2\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ols}\right]$$

$$\boldsymbol{\beta} = -\frac{1}{\lambda}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} + \frac{1}{\lambda}\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ols}$$

$$\left(\mathbf{I} + \frac{1}{\lambda}\left(\mathbf{X}^T\mathbf{X}\right)\right)\boldsymbol{\beta} = \frac{1}{\lambda}\left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ols}$$

$$\left(\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ols}$$

เนื่องจาก $\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$

$$\left(\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

จะได้ $$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^*_{ridge} = \left[\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}\mathbf{X}^T\mathbf{y}$$

ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ คือ

$$\hat{\boldsymbol{\beta}}^*_{ridge} = \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}\mathbf{X}^T\mathbf{Y} \quad \text{เมื่อ} \quad \lambda \geq 0$$
$$= \mathbf{W}\mathbf{X}^T\mathbf{Y} \tag{2.7}$$

และ $\mathbf{W} = \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}$

$\mathbf{I}_p$ คือ เป็นเมทริกซ์เอกลักษณ์ขนาด $(p \times p)$

จะเห็นได้ว่า ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ ในตัวแบบเชิงเส้น จะอยู่ภายใต้เงื่อนไข $\mathbf{B}^T\mathbf{B} = r^2 < \lambda$ ซึ่งจะเป็นการควบคุมขนาดของค่าประมาณสัมประสิทธิ์การถดถอย $\left(\hat{\boldsymbol{\beta}}\right)$ไม่ให้มีขนาดใหญ่เกินไป $\left(\hat{\boldsymbol{\beta}} \to 0\right)$ โดยจะเพิ่มส่วนของ $\lambda\mathbf{I}_p$ เข้ามา เพื่อทำให้เมทริกซ์ $\mathbf{X}^T\mathbf{X}$ มีความเสถียรมากขึ้น เมื่อเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์กันและข้อมูลที่มีมิติสูง และเมื่อ $\lambda = 0$ ค่าประมาณสัมประสิทธิ์การถดถอยที่ได้ จะมีค่าเท่ากับวิธีกำลังสองน้อยที่สุด นอกจากนี้ ในขณะที่ $\lambda \to \infty$, $\hat{\boldsymbol{\beta}} \to 0$ นั่นคือ เมื่อ λ เพิ่มมากขึ้น จะส่งผลทำให้ $\hat{\boldsymbol{\beta}}$ มีค่าน้อยลง

ความสัมพันธ์ระหว่างตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุด และวิธีการวิเคราะห์แบบริดจ์

เนื่องจาก $\hat{\boldsymbol{\beta}}^*_{ridge} = \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}\mathbf{X}^T\mathbf{Y}$ และ $\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$

พิจารณา $$\hat{\boldsymbol{\beta}}^*_{ridge} = \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}\mathbf{X}^T\mathbf{Y}$$
$$= \left[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right]^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

$$= \left[\mathbf{I}_p + \lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]^{-1}\hat{\boldsymbol{\beta}}_{ols}$$

$$= \mathbf{Z}\hat{\boldsymbol{\beta}}_{ols}$$

เมื่อ $\mathbf{Z}=\left[\mathbf{I}_p + \lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]^{-1}$

จะเห็นได้ว่า ถึงแม้จะไม่เกิดภาวะร่วมเชิงเส้น (Multicollinearity) ค่าประมาณโดยวิธีการวิเคราะห์แบบริดจ์ มีค่าไม่เท่ากับ ค่าประมาณโดยวิธีกำลังสองน้อยที่สุด

จากความสัมพันธ์ต่างๆ สามารถขยายออกได้ในเวลาต่อมา บางคุณสมบัติของตัวประมาณด้วยวิธีการวิเคราะห์แบบริดจ์ คือ $\hat{\boldsymbol{\beta}}^*_{ridge}, \mathbf{W}$ และวิธีกำลังสองน้อยที่สุด คือ $\mathbf{Z}$ ซึ่งสามารถแสดงได้ดังนี้

ให้ค่าไอเกน $e_i(\mathbf{W})$ และ $e_i(\mathbf{Z})$ ดังนี้

$$e_i(\mathbf{W}) = \frac{1}{\lambda_i^* + \lambda}$$

$$e_i(\mathbf{Z}) = \frac{\lambda_i^*}{\lambda_i^* + \lambda}$$

ซึ่ง λ_i^* คือค่าไอเกนของเมทริกซ์ $\mathbf{X}^T\mathbf{X}$ โดยที่ $\lambda^*_{\max} = \lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_p^* = \lambda^*_{\min} > 0$

ดังนั้น เราสามารถหาค่าเฉลี่ย ความแปรปรวน และค่าคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ ดังนี้

ค่าคาดหวังของ $\hat{\boldsymbol{\beta}}^*_{ridge}$

พิจารณา $E\left[\hat{\boldsymbol{\beta}}^*_{ridge}\right] = E\left[\mathbf{Z}\hat{\boldsymbol{\beta}}_{ols}\right] = \mathbf{Z}E\left[\hat{\boldsymbol{\beta}}_{ols}\right]$

เนื่องจาก $E\left[\hat{\boldsymbol{\beta}}^*_{ridge}\right] = \boldsymbol{\beta}$

$\therefore \quad E\left[\hat{\boldsymbol{\beta}}^*_{ridge}\right] = \mathbf{Z}\boldsymbol{\beta}$

จะเห็นว่า $\hat{\boldsymbol{\beta}}^*_{ridge}$ เป็นตัวประมาณที่เอนเอียง (Biased Estimator) สำหรับ $\boldsymbol{\beta}$

เมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}^*_{ridge}$

พิจารณา $Cov\left(\hat{\boldsymbol{\beta}}^*_{ridge}\right) = \mathrm{Cov}\left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ols}\right)$

เนื่องจาก $Cov\left(\hat{\boldsymbol{\beta}}_{ols}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

$\therefore \quad Cov\left(\hat{\boldsymbol{\beta}}^*_{ridge}\right) = \sigma^2\mathbf{Z}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{Z}^T$

$$= \sigma^2 \left(\frac{\lambda_i^*}{\lambda_i^* + \lambda} \right)^2 \left(\mathbf{X}^T \mathbf{X} \right)^{-1}$$

ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ridge}^*$

พิจารณา $MSE\left(\hat{\boldsymbol{\beta}}_{ridge}^*\right) = E\left[L^2(\lambda) \right]$

$$= E\left[\left(\hat{\boldsymbol{\beta}}_{ridge}^* - \boldsymbol{\beta} \right)^T \left(\hat{\boldsymbol{\beta}}_{ridge}^* - \boldsymbol{\beta} \right) \right]$$

$$= E\left[\left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \right)^T \left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \right) \right]$$

$$= E\left[\left(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \right)^T \mathbf{Z}'\mathbf{Z} \left(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \right) \right] + \left(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta} \right)^T \left(\mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\beta} \right)$$

$$= \sigma^2 Trace\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{Z}^T \mathbf{Z} + \boldsymbol{\beta}^T \left(\mathbf{Z} - \mathbf{I} \right)^T \left(Z - \mathbf{I} \right) \boldsymbol{\beta}$$

เนื่องจาก $\mathbf{Z} = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{X}$ หรือ $\mathbf{Z} = \mathbf{I} - \lambda \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1}$

$$= \sigma^2 Trace\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{X} \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{X} \quad + \quad \boldsymbol{\beta}^T \left(-\lambda \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \right)^2 \boldsymbol{\beta}$$

$$= \sigma^2 \left[Trace\left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} - \lambda Trace\left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-2} \right] \quad + \quad \lambda^2 \boldsymbol{\beta}^T \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-2} \boldsymbol{\beta}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i^*}{\left(\lambda_i^* + \lambda \right)^2} \quad + \quad \lambda^2 \boldsymbol{\beta}^T \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p \right)^{-2} \boldsymbol{\beta}$$

$$= \gamma_1(\lambda) \quad + \quad \gamma_2(\lambda)$$

ดังนั้น ความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ridge}^*$ คือ ผลบวกของเทอมฟังก์ชัน $\gamma_1(\lambda)$ และ $\gamma_2(\lambda)$

เทอม $\gamma_1(\lambda)$ คือ ผลรวมของค่าความแปรปรวนของตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\boldsymbol{\beta}}_{ridge}^*$ แต่ละตัว ซึ่งเป็นผลรวมของสมาชิกแนวทแยงของความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}_{ridge}^*$ หรือ $Cov\left(\hat{\boldsymbol{\beta}}_{ridge}^*\right)$

ส่วนเทอม $\gamma_2(\lambda)$ คือ ค่ากำลังสองของความเอนเอียง (ระยะห่างระหว่าง $\mathbf{Z}\boldsymbol{\beta}$ กับ $\boldsymbol{\beta}$) ซึ่งจะมีค่าเป็น 0 ถ้า $\lambda = 0$ เพราะว่า $\mathbf{Z} = \mathbf{I_p}$ ดังนั้น $\gamma_2(\lambda)$ จึงสามารถพิจารณาเป็นค่ากำลังสองของความเอนเอียงของ $\hat{\boldsymbol{\beta}}_{ridge}^*$ และฟังก์ชัน $\gamma_1(\lambda)$ เป็นฟังก์ชันลดทางเดียว (Monotonic Decreasing

Function) ของ λ ฟังก์ชัน $\gamma_2(\lambda)$ เป็นฟังก์ชันเพิ่มทางเดียว (Monotonic Increasing Function) ของ λ

สำหรับตัวแบบการถดถอยปัวซง (Poisson Regression Model) ในปี 2011 Kristofer และ Ghazi ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ โดยตัวประมาณแบบริดจ์สำหรับตัวแบบการถดถอยปัวซง เป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง (Multicollinearity) สามารถคำนวณหาค่าสัมประสิทธิ์การถดถอยจากการหาค่าต่ำสุด ของฟังก์ชันลบของลอการิทึมของฟังก์ชันน่าจะเป็น (Log Likelihood) จากวิธีภาวะน่าจะเป็นสูงสุด ภายใต้เงื่อนไข L_2 penalty บน $\boldsymbol{\beta}$

โดยที่ $L_2 = \|\boldsymbol{\beta}\|_2 = \sum_{j=1}^{p} \beta_j^{\,2}$

และ $P_\lambda(\boldsymbol{\beta})$ ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ คือ $\lambda \sum_{j=1}^{p} \beta_j^2$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบริดจ์ ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ridge} &= \arg\min_{\beta} \left(-l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1^2 \right) \\ &= \arg\min_{\beta} \left(-\sum_{i=1}^{n} \left(y_i \mathbf{X}_i^{\,T} \boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{\,T} \boldsymbol{\beta} \right) - \ln y_i ! \right) + \lambda \|\boldsymbol{\beta}\|_1^2 \right) \end{aligned} \tag{2.8}$$

เมื่อ λ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ซึ่งควบคุมขนาดการหดตัว (Shrinkage) ของตัวประมาณ $\hat{\boldsymbol{\beta}}_{ridge}$ โดยที่ $\lambda \to \infty$, $\hat{\boldsymbol{\beta}} \to 0$

พิจารณา ถ้าให้ $\boldsymbol{\beta}$ แทน เวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอยใดๆ หลักการของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธี การวิเคราะห์แบบริดจ์ $\left(\hat{\boldsymbol{\beta}}_{ridge}\right)$ คือ ทำให้ผลบวกกำลังสองของค่าคลาดเคลื่อนน้อยที่สุด (ϕ)

$$\begin{aligned} \phi &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}\right)^T \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}\right)^T \mathbf{X}^T \hat{\mathbf{W}} \mathbf{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}\right) \\ &= \phi_{\min} + \phi_0 \end{aligned}$$

เมื่อคอนทัวร์ของค่าคงที่ ϕ คือ พื้นผิวของ hyperellipsoids ที่มีจุดศูนย์กลางที่ $\hat{\boldsymbol{\beta}}_{ML}$

โดยที่ $\phi_{\min}$ คือ ค่าน้อยที่สุดของ ϕ

ϕ_0 คือ กำลังสอง (Quadratic Form) ของ $\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ML}\right)$

ดังนั้น จึงทำให้เกิดความสัมพันธ์ $\phi=\phi_{\min}+\phi_0$ ซึ่ง $\phi_0>0$ เป็นส่วนที่ทำให้มีค่าเพิ่มขึ้น เมื่อ $\hat{\boldsymbol{\beta}}_{ML}$ ถูกแทนที่ด้วย $\boldsymbol{\beta}$

ในการหาตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบริดจ์ โดยต้องการให้ผลบวกกำลังสองของค่าคลาดเคลื่อนต่ำที่สุด ดังนั้น โดยการหาอนุพันธ์ เทียบกับ $\boldsymbol{\beta}$ ภายใต้ข้อจำกัด (Constraint) คือ $\boldsymbol{\beta}^T\boldsymbol{\beta}=r^2<\lambda$

เนื่องจาก $\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T\mathbf{X}^T\mathbf{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right) = \phi_0$

และแก้สมการหาคำตอบโดยใช้วิธีของลากรานจ์ (Lagrangrian) จะได้ว่า

กำหนดให้ minimize $F = \boldsymbol{\beta}^T\boldsymbol{\beta} + \left(\frac{1}{\lambda}\right)\left[\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ML}\right)^T\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{ML} - \phi_0\right)\right]$

เมื่อ $\frac{1}{\lambda}$ คือ ตัวคูณ ดังนั้น

$$\frac{\partial F}{\partial \mathbf{B}} = 2\mathbf{B} + \left(\frac{1}{\lambda}\right)\left[2\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{B}-2\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}\right] = 0$$

จะได้ว่า

$$2\mathbf{B} = -\left(\frac{1}{\lambda}\right)\left[2\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{B}-2\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}\right]$$

$$\mathbf{B} = -\frac{1}{\lambda}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{B}+\frac{1}{\lambda}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}$$

$$\left(\mathbf{I}+\frac{1}{\lambda}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\right)\mathbf{B} = \frac{1}{\lambda}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}$$

$$\left(\lambda\mathbf{I}+\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{B} = \left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}$$

$$\mathbf{B} = \left(\lambda\mathbf{I}+\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML}$$

จะได้ $$\mathbf{B} = \hat{\boldsymbol{\beta}}_{ridge} = \left(\lambda\mathbf{I}+\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}\right)\hat{\boldsymbol{\beta}}_{ML} \quad (2.9)$$

เราสามารถหาคุณสมบัติของตัวประมาณจากต่างความหวัง ความแปรปรวน และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีริดจ์ ดังนี้

เมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}_{ridge}$

$$Cov\left(\hat{\boldsymbol{\beta}}_{ridge}\right) = Cov\left(\mathbf{Z}\hat{\boldsymbol{\beta}}_{ML}\right) = \mathbf{Z}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{Z}^T$$

ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ridge}$

$$\therefore E\left(L^2_{Ridge}\right) = E\left(\hat{\boldsymbol{\beta}}_{ridge}-\boldsymbol{\beta}\right)^T\left(\hat{\boldsymbol{\beta}}_{ridge}-\boldsymbol{\beta}\right)$$

$$= E\left[\left(\hat{\boldsymbol{\beta}}_{ridge}-\boldsymbol{\beta}\right)^T\mathbf{Z}^T\mathbf{Z}\left(\hat{\boldsymbol{\beta}}_{ridge}-\boldsymbol{\beta}\right)\right]+\left(\mathbf{Z}\boldsymbol{\beta}-\boldsymbol{\beta}\right)^T\left(\mathbf{Z}\boldsymbol{\beta}-\boldsymbol{\beta}\right)$$

$$= \sum_{j=1}^{p}\frac{\lambda_j^*}{\left(\lambda_j^*+\lambda\right)^2}+\boldsymbol{\beta}^T\left(\left(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X}+\lambda I\right)^{-1}\left(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X}\right)-\mathbf{I}\right)^T\left(\left(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X}+\lambda I\right)^{-1}\left(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X}\right)-\mathbf{I}\right)\boldsymbol{\beta}$$

$$= \sum_{j=1}^{p}\frac{\lambda_j^*}{\left(\lambda_j^*+\lambda\right)^2}+\boldsymbol{\beta}^T\lambda^2\left(\mathbf{X^T}\hat{\mathbf{W}}\mathbf{X}+\lambda I\right)^{-2}\boldsymbol{\beta}$$

$$= \gamma_1(\lambda)+\gamma_2(\lambda)$$

ดังนั้น ความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}_{ridge}$ คือ ผลบวกของเทอมฟังก์ชัน $\gamma_1(\lambda)$ และ $\gamma_2(\lambda)$

ในปี 1970 Hoerl และ Kennard ได้ให้ความหมายของเทอม $\gamma_1(\lambda)$ คือผลรวมของค่าความแปรปรวนของตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\boldsymbol{\beta}}_{ridge}$ แต่ละตัว ซึ่งเป็นผลรวมของสมาชิกแนวทแยงของความแปรปรวนและความแปรปรวนร่วมของ $\hat{\boldsymbol{\beta}}_{ridge}$ หรือ $Cov\left(\hat{\boldsymbol{\beta}}_{ridge}\right)$

ส่วนเทอม $\gamma_2(\lambda)$ คือ ค่ากำลังสองของความเอนเอียง (ระยะห่างระหว่าง $\mathbf{Z}\boldsymbol{\beta}$ กับ $\boldsymbol{\beta}$) ซึ่งจะมีค่าเป็น 0 ถ้า λ=0 เพราะว่า $\mathbf{Z}$=$\mathbf{I_p}$ ดังนั้น $\gamma_2(\lambda)$ จึงสามารถพิจารณาเป็นค่ากำลังสองของความเอนเอียงของ $\hat{\boldsymbol{\beta}}_{ridge}$ และฟังก์ชัน $\gamma_1(\lambda)$ เป็นฟังก์ชันลดทางเดียว (Monotonic-Decreasing Function) ของ λ ฟังก์ชัน $\gamma_2(\lambda)$ เป็นฟังก์ชันเพิ่มทางเดียว (Monotonic-Increasing Function) ของ λ ความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำ เมื่อ λ มีค่าเข้าใกล้ในบริเวณจุดกำเนิด

เนื่องจากในการวิเคราะห์การถดถอย จะแบ่งลักษณะของข้อมูลในการศึกษาออกเป็น 2 ลักษณะ คือ ข้อมูลที่มีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ $(n > p)$ ซึ่งเรียกข้อมูลที่มีลักษณะนี้ว่า ข้อมูลที่มีมิติต่ำ (Low – Dimensional) ซึ่งเป็นลักษณะข้อมูลที่พบได้ทั่วไป แต่ในบางครั้งข้อมูลอาจจะไม่เป็นไปตามเงื่อนไขดังกล่าว นั่นคือ ขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ $(n < p)$ และเรียกข้อมูลที่มีลักษณะนี้ว่า ข้อมูลที่มีมิติสูง (High – Dimensional) โดยข้อมูลที่มีลักษณะนี้ได้แก่ ข้อมูลทางด้านการแพทย์ เศรษฐศาสตร์ วิศวกรรมศาสตร์ เป็นต้น ซึ่งสามารถศึกษาได้ในหัวข้อถัดไป

2.1.4 ข้อมูลที่มีมิติสูง (High – Dimensional)

พิจารณา สำหรับตัวแบบการถดถอยเชิงเส้น พิจารณาตัวแบบพื้นฐานของการถดถอยเชิงเส้นพหุคูณ สำหรับตัวแปรอิสระ p ตัว และขนาดของกลุ่มตัวอย่างเท่ากับ n

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad \text{เมื่อ} \quad n<p \tag{2.10}$$

โดยที่

$\mathbf{Y}$ คือ เวกเตอร์ขนาด $(n\times 1)$ ของตัวแปรตามที่วัดได้จากกลุ่มตัวอย่างขนาด n

$\mathbf{X}$ คือ เมทริกซ์ขนาด $\left(n\times\left(p+1\right)\right)$ แสดงค่าตัวทำนายทั้ง n ตัวที่วัดได้จากกลุ่มตัวอย่างขนาด n

$\boldsymbol{\beta}$ คือ เวกเตอร์ขนาด $(n\times 1)$ ของสัมประสิทธิ์การถดถอยของประชากร

$\boldsymbol{\varepsilon}$ คือ เวกเตอร์ขนาด $(n\times 1)$ ของความคลาดเคลื่อนของตัวแปรตาม

โดยมีคุณสมบัติ

$$E\left(\boldsymbol{\varepsilon}\right)=\mathbf{0} \quad \text{และ} \quad Var\left(\boldsymbol{\varepsilon}\right)=\sigma^2\mathbf{I}_n$$

เมื่อ เวกเตอร์ของตัวแปรตาม $\mathbf{Y}=\left(y_1,y_2,...,y_n\right)\in\mathbb{R}^n$

เวกเตอร์ของค่าสังเกต $\mathbf{X}=\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_p\right)^T\in\mathbb{R}^{n\times p}$

และเวกเตอร์ของสัมประสิทธิ์การถดถอย $\boldsymbol{\beta}=\left(\beta_1,\beta_2,...,\beta_p\right)^T\in\mathbb{R}^p$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

เครื่องมือในการวิเคราะห์การถดถอยทางสถิติแบบเดิมหลายๆ เครื่องมือไม่สามารถนำมาใช้วิเคราะห์ได้ ในข้อมูลที่มีลักษณะ ข้อมูลที่มีมิติสูง (High – Dimensional) เนื่องจากผู้วิเคราะห์จะต้องเผชิญกับปัญหาในการวิเคราะห์การถดถอย ดังนี้

1) ในปี 2009 Johnstone และ Titterington กล่าวว่า เมื่อกรณีที่ข้อมูลมิติสูง จะไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอย $\left(\hat{\boldsymbol{\beta}}\right)$ ด้วยวิธีกำลังสองน้อยที่สุดได้ เนื่องจาก เมื่อข้อมูลที่มีมีมิติสูง $\left(n<p\right)$ เมทริกซ์ $\mathbf{X}^T\mathbf{X}$ จะเป็นเมทริกซ์เอกฐาน (Singular Matrix) ซึ่งไม่สามารถหาเมทริกซ์ผกผัน (Inverse Matrix) ได้ จึงไม่สามารถหาค่าประมาณสัมประสิทธิ์การถดถอย $\left(\hat{\boldsymbol{\beta}}\right)$ โดยวิธีกำลังสองน้อยที่สุด

2) ในปี 2012 วิฐรา พึ่งพาพงศ์ ได้อธิบายว่า เมื่อตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูง เนื่องจากตัวแปรอิสระมีเป็นจำนวนมาก อาจทำให้เกิดปัญหาภาวะร่วมเชิงเส้น (Multicollinearity) นี้

ขึ้น และจะส่งผลให้ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด มีความแปรปรวนและส่วนเบี่ยงเบนมาตรฐานมีค่าสูง เนื่องจาก เมทริกซ์ $\mathbf{X}^T\mathbf{X}$ จะมีความไม่เสถียร

3) ในปี 2012 วิฐรา พึ่งพาพงศ์ ได้อธิบายว่า เมื่อข้อมูลมีมิติสูง การแปลผลลัพธ์ของตัวแบบมีความยากและสลับซับซ้อน เนื่องจากกรณีที่ตัวแปรอิสระมีจำนวนมาก ตัวแบบที่ดีควรมีเฉพาะตัวแปรอิสระที่มีความสำคัญกับตัวแปรตามในตัวแบบเท่านั้น ทั้งนี้เพื่อให้ได้ตัวแบบที่สามารถแปลผลได้โดยง่ายและไม่สลับซับซ้อน แนวคิดในการคัดเลือกตัวแปรเข้าตัวแบบจึงมีความสำคัญเป็นอย่างมาก ในการวิเคราะห์การถดถอยเชิงเส้นแบบดั้งเดิม การคัดเลือกตัวแปรสามารถทำได้โดยการทดสอบสมมติฐานว่าสัมประสิทธิ์การถดถอยเท่ากับศูนย์หรือไม่ โดยใช้ตัวสถิติทดสอบทีและตัวสถิติทดสอบเอฟ อย่างไรก็ตาม การทดสอบสมมติฐานเหล่านี้ไม่สามารถใช้งานได้กับข้อมูลที่มีมิติสูง ทั้งนี้เนื่องมาจากตัวสถิติทดสอบดังกล่าวถูกพัฒนาขึ้นจากแนวคิดการทดสอบ ด้วยอัตราส่วนภาวะน่าจะเป็น (Likelihood Ratio Test) ซึ่งอ้างอิงตัวประมาณจากวิธีกำลังสองน้อยที่สุด

ดังนั้น เมื่อเกิดปัญหาข้อมูลที่มีลักษณะเป็นข้อมูลที่มีมิติสูง จึงได้มีการนำเสนอวิธีต่างๆ เพื่อแก้ปัญหาที่ไม่สามารถใช้วิธีการถดถอยแบบเดิมได้ วิธีที่เป็นที่นิยมวิธีหนึ่ง คือ วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ (Penalized Regression)

2.1.5 ข้อมูลที่มีมิติสูงแบบบางเบา (High – Dimensional Sparse data)

พิจารณา สำหรับตัวแบบการถดถอยเชิงเส้น พิจารณาตัวแบบพื้นฐานของการถดถอยเชิงเส้นพหุคูณ สำหรับตัวแปรอิสระ p ตัว และขนาดของกลุ่มตัวอย่างเท่ากับ n

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{เมื่อ} \quad n < p \tag{2.11}$$

สมมติให้ $\mathbf{x}_i$ สามารถแยกออกเป็น $\mathbf{x}_i = \left(\mathbf{x}_{iA}, \mathbf{x}_{iB}\right)$

โดยที่ $\mathbf{x}_{iA} = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(p-q)}\right)^T \in \mathbb{R}^{p-q}$

$\mathbf{x}_{iB} = \left(\mathbf{x}_{i(p-q+1)}, \ldots, \mathbf{x}_{ip}\right)^T \in \mathbb{R}^{q}$

โดยที่

p คือ จำนวนพารามิเตอร์ทั้งหมด

q คือ จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอย มีค่าเท่ากับ 0 (Inactive Parameter)

$p-q$ คือ จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอย มีค่าไม่เท่ากับ 0 (Active Parameter) และส่วนมากมีค่าเป็นศูนย์

ถ้าจำนวนพารามิเตอร์สัมประสิทธิ์การถดถอยที่มีค่าเท่ากับศูนย์ มีจำนวนมากกว่า จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอยที่มีค่าไม่เท่ากับศูนย์ หรือ $q > p-q$ จะเรียกข้อมูลที่มีลักษณะนี้ว่า “ข้อมูลบางเบา”

ให้ $\mathbf{X}_A = \left(\mathbf{x}_{1A},...,\mathbf{x}_{nA}\right)^T \in \mathbb{R}^{n\times(p-q)}$ และ $\mathbf{X}_B = \left(\mathbf{x}_{1B},...,\mathbf{x}_{nB}\right)^T \in \mathbb{R}^{n\times q}$ คือ เมทริกซ์ของ $\mathbf{x}_{iA}$ และ $\mathbf{x}_{iB}$ ตามลำดับ

จะได้ว่า $\mathbf{X} = \left(\mathbf{X}_A, \mathbf{X}_B\right)^T \in \mathbb{R}^{n\times p}$ คือ เมทริกซ์ของตัวแปรอิสระทั้งหมด และสามารถเขียนให้อยู่ในตัวแบบ ได้ดังนี้

$$\mathbf{Y} = \mathbf{X}_A\boldsymbol{\beta}_A + \mathbf{X}_B\boldsymbol{\beta}_B + \boldsymbol{\varepsilon} \tag{2.12}$$

โดยที่ $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_A, \boldsymbol{\beta}_B\right)^T \in \mathbb{R}^p$, $\boldsymbol{\beta}_A^T \in \mathbb{R}^{p-q}$ และ $\boldsymbol{\beta}_B^T \in \mathbb{R}^q$ คือ เวกเตอร์ที่ไม่ทราบค่าของสัมประสิทธิ์การถดถอย

2.1.6 การวิเคราะห์การถดถอยแบบพีนอลไลซ์สำหรับข้อมูลที่มีมิติสูง

สำหรับการวิเคราะห์การถดถอย เราจะพิจารณาในกรณีที่ขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ $(n > p)$ ในทางตรงกันข้าม เมื่อ ขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ $(n < p)$ เราจะใช้วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ เพื่อแก้ไขปัญหาที่เกิดจากเมทริกซ์ $\mathbf{X}^T\mathbf{X}$ ไม่เป็นเมทริกซ์เอกฐาน หรือเกิดความไม่เสถียร

วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ เป็นวิธีที่มีแนวคิดแบบ frequentist หรือ non-Bayesian ซึ่งจุดมุ่งหมายหลัก คือ การประมาณค่าสัมประสิทธิ์การถดถอยซึ่งสามารถใช้ได้กับข้อมูลที่มีมิติสูง และสามารถลดปัญหาความไม่เพียงพอของวิธีการวิเคราะห์การถดถอยแบบดั้งเดิมในตัวแบบการถดถอยปัวซงให้มีประสิทธิภาพ ดังนั้น วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ เป็นวิธีที่นิยมใช้กันอย่างแพร่หลายในการประมาณค่า $\boldsymbol{\beta}$ เมื่อข้อมูลมีมิติสูง

โดยตัวประมาณดังกล่าว จะหาได้จากการหาค่า $\boldsymbol{\beta}$ ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) ให้มีค่าน้อยที่สุด และความแตกต่างของวิธีพีนอลไลซ์แต่ละวิธี คือ ส่วนของฟังก์ชันพีนอลตี้ (Penalty Function) แทนด้วย $P_\lambda\left(\boldsymbol{\beta}\right)$ โดยฟังก์ชันนี้จะอยู่ในรูปของ $\boldsymbol{\beta}$ และมีพารามิเตอร์ λ ซึ่งมีค่ามากกว่าหรือเท่ากับศูนย์ เพื่อใช้ในการให้น้ำหนักของฟังก์ชันพีนอลตี้ดังกล่าว

สำหรับฟังก์ชันพีนอลตี้นั้น มีอยู่ด้วยกันหลายรูปแบบตามแต่วิธีที่ใช้ในการวิเคราะห์ ซึ่งจะให้ค่าประมาณที่แตกต่างกัน ส่วนพารามิเตอร์ λ โดยทั่วไปแล้วจะใช้วิธี cross-validation ในการหาค่า λ ที่เหมาะสมสำหรับข้อมูลที่ต้องการวิเคราะห์

ในงานวิจัยนี้ สนใจวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge Regression) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO) โดยวิธีการวิเคราะห์การถดถอยแบบริดจ์ ได้กล่าวไปก่อนหน้านี้แล้ว เนื่องจาก วิธีการวิเคราะห์การถดถอยแบบริดจ์ เป็นที่นิยมมากกว่าในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง จึงได้แยกส่วนในการอธิบายออกไป แต่อีกสองวิธีจะมีความแตกต่างจากวิธีริดจ์ จะอธิบายในหัวข้อถัดไป

2.1.6.1 วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO)

สำหรับตัวแบบการถดถอยเชิงเส้น ในปี 1996 Tibshirani ได้นำเสนอวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบแลซโซ โดยวิธีการนี้มีคุณสมบัติเช่นเดียวกับตัวประมาณที่ได้จากวิธีการวิเคราะห์แบบริดจ์ ซึ่งมีคุณสมบัติเป็นตัวประมาณที่เอนเอียง แต่สามารถลดความแปรปรวนลงได้ เมื่อไม่สามารถใช้วิธีการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้ โดยวิธีการวิเคราะห์การถดถอยแบบแลซโซ มีข้อดีที่แตกต่างจากวิธีการวิเคราะห์การถดถอยแบบริดจ์ นั่นคือ นอกจากวิธีการนี้จะสามารถประมาณค่าพารามิเตอร์การถดถอยได้ เมื่อตัวแปรอิสระมีเป็นจำนวนมาก ยังสามารถคัดเลือกตัวแปร (Variable Selection) ที่เหมาะสมหรือมีอิทธิพลสูงเข้าสู่ตัวแบบได้ ซึ่งสามารถทำให้การอธิบายผลลัพธ์ของตัวแบบได้ง่ายขึ้น
ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีแบบแลซโซ $\left(\hat{\boldsymbol{\beta}}^{*}_{lasso}\right)$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}^{*}_{lasso} = \arg\min_{\beta} \left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|^{2} + P_{\lambda}\left(\boldsymbol{\beta}\right) \tag{2.13}$$

จะหาได้จากการหาค่า $\hat{\boldsymbol{\beta}}^{*}_{lasso}$ ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) ดังสมการข้างต้นให้มีค่าน้อยที่สุด

โดยที่ $P_{\lambda}\left(\boldsymbol{\beta}\right)$ ของวิธีการวิเคราะห์การถดถอยแบบแลซโซ คือ $\lambda\sum_{j=1}^{p}\left|\beta_{j}\right|$

จะได้ว่า $\hat{\boldsymbol{\beta}}^{*}_{lasso} = \arg\min_{\beta} \left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|^{2} + \lambda\sum_{j=1}^{p}\left|\beta_{j}\right|$ เมื่อ $\lambda > 0$

พิจารณา ในกรณีของเมทริกซ์การตั้งฉากปกติ (Orthogonal Design) ของเมทริกซ์ $\mathbf{X}^{T}\mathbf{X}$ นั่นคือ สมมติให้ $\mathbf{X}^{T}\mathbf{X} = \mathbf{I}_{p}$

เนื่องจาก $\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{Y} = \mathbf{X}^{T}\mathbf{Y}$

เราต้องการหา $\hat{\boldsymbol{\beta}}^{*}_{lasso} = \arg\min_{\beta} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p}|\beta_j|$

พิจารณา $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$ จะได้ว่า $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}$

เนื่องจาก $\mathbf{y}^T\mathbf{y}$ ไม่มีตัวแปรที่เราสนใจศึกษา $(\boldsymbol{\beta})$ จึงตัดออก เหลือแค่

$$\hat{\boldsymbol{\beta}}^{*}_{lasso} = \arg\min_{\beta}\left(-2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}\right) + \lambda\sum_{j=1}^{p}|\beta_j|$$

เนื่องจาก $\hat{\boldsymbol{\beta}}_{ols} = \mathbf{X}^T\mathbf{Y}$ จะได้ว่า

$$\hat{\boldsymbol{\beta}}^{*}_{lasso} = \arg\min_{\beta}\left(-2\hat{\boldsymbol{\beta}}^T_{ols}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}\right) + \lambda\sum_{j=1}^{p}\beta_j$$

$$= \arg\min_{\beta}\sum_{j=1}^{p}\left(-2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda|\beta_j|\right) \quad = \quad L_j$$

ถ้า $\hat{\beta}^{ols}_j \geq 0$ จะได้ว่า $\beta_j \geq 0$

$\hat{\beta}^{ols}_j \leq 0$ จะได้ว่า $\beta_j \leq 0$

กรณีที่ 1 $\hat{\beta}^{ols}_j \geq 0$ ดังนั้น $\beta_j \geq 0$

$$L_j = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda\beta_j$$

$$\frac{\partial L_j}{\partial \beta_j} = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda\beta_j = 0$$

เมื่อ $\beta_j = \hat{\beta}^{lasso^*}_j$

$$2\hat{\beta}^{lasso^*}_j = 2\hat{\beta}^{ols}_j - \lambda$$

$$\hat{\beta}^{lasso^*}_j = \hat{\beta}^{ols}_j - \frac{\lambda}{2}$$

กรณีที่ 2 $\hat{\beta}^{ols}_j \leq 0$ ดังนั้น $\beta_j \leq 0$

$$L_j = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 - \lambda\beta_j$$

$$\frac{\partial L_j}{\partial \beta_j} = -2\hat{\beta}^{ols}_j + \beta_j - \lambda = 0$$

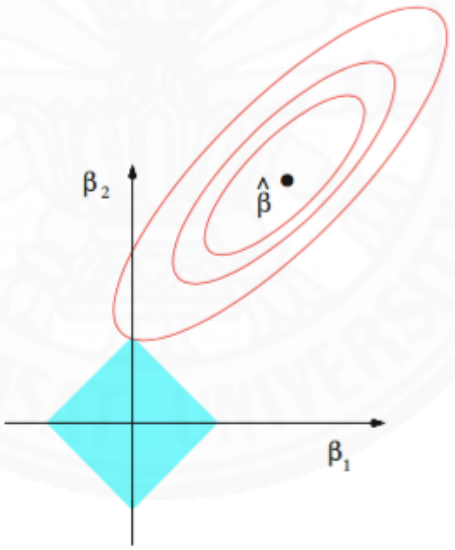
เมื่อ $\beta_j = \hat{\beta}^{lasso^*}_j$

$$2\hat{\beta}_j^{lasso^*} = 2\hat{\beta}_j^{ols} + \lambda$$
$$\hat{\beta}_j^{lasso^*} = \hat{\beta}_j^{ols} + \frac{\lambda}{2}$$

ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีแบบแลซโซ $\left(\hat{\boldsymbol{\beta}}_{lasso}\right)$

$$\hat{\boldsymbol{\beta}}_{lasso}^{*} = \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & , \quad \hat{\beta}_j^{ols} > \frac{\lambda}{2} \\ 0 & , \quad \frac{-\lambda}{2} \le \hat{\beta}_j^{ols} \le \frac{\lambda}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2} & , \quad \hat{\beta}_j^{ols} < \frac{-\lambda}{2} \end{cases} \tag{2.14}$$

วิธีการวิเคราะห์การถดถอยแบบแลซโซ ในตัวแบบเชิงเส้น สามารถแสดลักษณะพื้นที่ผิวตอบสนองของ $\left(\mathbf{Y}-\mathbf{XB}\right)^T\left(\mathbf{Y}-\mathbf{XB}\right)$ ด้วยภาพที่ 2.2 ดังนี้



ภาพที่ 2.2 แสดงของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีกำลังสองน้อยที่สุด (OLS) และของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO)

สำหรับตัวแบบการถดถอยปัวซง ในปี 2007 Park และ Hastie ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซโซ สำหรับตัวแบบการถดถอยปัวซง โดยวิธีการวิเคราะห์นี้เป็นวิธีที่นิยมสำหรับการประมาณค่าและการคัดเลือกตัวแปรในคราวเดียวกัน

สามารถคำนวณหาค่าสัมประสิทธ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 - norm penalty บน $\boldsymbol{\beta}$

โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ ดังนั้น $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบ LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \left(-l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right) \tag{2.15}$$

สามารถหา $l(\boldsymbol{\beta})$ ได้ดังนี้

ตัวแบบการถดถอยปัวซง จะได้ฟังก์ชันน่าจะเป็น (Likelihood Function) คือ

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]^{y_i} \exp\left[-\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]}{y_i!}$$

$$= \frac{\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]^{\sum_{i=1}^{n} y_i} \exp\left[-\sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta})\right]}{\prod_{i=1}^{n} y_i!}$$

ใส่ ln ทั้งสองข้างของสมการ จะได้

$$l(\boldsymbol{\mu}; \mathbf{y}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right] - \sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln y_i!$$

แทนค่า $\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i^T \boldsymbol{\beta})$ จะได้

$$= \sum_{i=1}^{n} y_i \ln \exp(\mathbf{X}_i^T \boldsymbol{\beta}) - \sum_{i=1}^{n} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln y_i!$$

$$= \sum_{i=1}^{n} y_i (\mathbf{X}_i^T \boldsymbol{\beta}) - \sum_{i=1}^{n} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln y_i!$$

$$= \sum_{i=1}^{n} \left(y_i \mathbf{X}_i^T \boldsymbol{\beta} - \exp(\mathbf{X}_i^T \boldsymbol{\beta}) - \ln y_i! \right)$$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบ LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \left(-\sum_{i=1}^{n} \left(y_i \mathbf{X}_i^T \boldsymbol{\beta} - \exp(\mathbf{X}_i^T \boldsymbol{\beta}) - \ln y_i! \right) + \lambda \sum_{j=1}^{p} \beta_j \right) \tag{2.16}$$

ซึ่ง $\lambda > 0$ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดของการหดตัว(Shrinkage) ของตัวประมาณ $\boldsymbol{\beta}$ ถ้า $\lambda_0 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} \beta_j$ เมื่อกำหนด $\lambda_0 > \lambda$ จำทำให้ตัวประมาณสัมประสิทธิ์ถูกดึงหดลงเข้าหาค่าศูนย์ และตัวประมาณบางตัวอาจเท่ากับศูนย์

แต่เมื่อกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง วิธีการวิเคราะห์การถดถอยด้วยวิธีแลซโซที่มีคุณสมบัติในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ โดยไม่ได้สนใจว่าตัวแปรใดในกลุ่มนั้นมีความสำคัญมากที่สุด จึงทำให้ความสามารถในการคัดเลือกตัวแปร ยังขาดคุณสมบัติความคงเส้นคงวา ต่อมาจึงได้มีผู้วิจัยคิดค้นการเคราะห์การถดถอยที่มีประสิทธิภาพในกรคัดเลือกตัวแปรเพิ่มมากขึ้น

2.1.6.2. วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO)

สำหรับตัวแบบการถดถอยเชิงเส้น ในปี 2006 Zou ได้นำเสนอวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แลซโซแบบปรับปรุง เพื่อที่จะแก้ไขข้อจำกัดของวิธีการวิเคราะห์การถดถอยแบบแลซโซ ให้มีคุณสมบัติในการคัดเลือกตัวแปรที่แม่นยำและมีประสิทธิภาพมากขึ้น โดยมีการถ่วงน้ำหนักให้กับตัวแปรอิสระที่คัดเลือกเข้าสู่ตัวแบบ ซึ่งทำให้เกิดความคงเส้นคงวาในการคัดเลือกตัวแปร โดยวิธีการนี้มีคุณสมบัติเช่นเดียวกับตัวประมาณที่ได้จากวิธีการวิเคราะห์แบบริดจ์ ซึ่งมีคุณสมบัติเป็นตัวประมาณที่เอนเอียง แต่สามารถลดความแปรปรวนลงได้ เมื่อไม่สามารถใช้วิธีการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้
ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีแลซโซแบบปรับปรุง $\left(\hat{\boldsymbol{\beta}}^{*}_{adaplasso}\right)$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}^{*}_{adaplasso} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + P_{\lambda}(\boldsymbol{\beta}) \qquad (2.17)$$

จะหาได้จากการหาค่า $\hat{\boldsymbol{\beta}}^{*}_{adaplasso}$ ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) ดังสมการข้างต้นให้มีค่าน้อยที่สุด ภายใต้ L_1 - norm penalty บน $\boldsymbol{\beta}$

โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

และ $P_{\lambda}(\boldsymbol{\beta})$ ของวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง คือ $\lambda \sum_{j=1}^{p} w_j |\beta_j|$

จะได้ว่า $\hat{\boldsymbol{\beta}}^{*}_{adaplasso} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|$ เมื่อ $\lambda > 0$

พิจารณา ในกรณีของเมทริกซ์การตั้งฉากปกติ (Orthogonal Design) ของเมทริกซ์ $\mathbf{X}^T\mathbf{X}$ นั่นคือ

สมมติให้ $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$

เนื่องจาก $\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{Y}$

เราต้องการหา $\hat{\boldsymbol{\beta}}^*_{adaplasso} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p} w_j |\beta_j|$

พิจารณา $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ จะได้ว่า $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}$

เนื่องจาก $\mathbf{y}^T\mathbf{y}$ ไม่มีตัวแปรที่เราสนใจศึกษา $(\boldsymbol{\beta})$ จึงตัดออก เหลือแค่

$$\hat{\boldsymbol{\beta}}^*_{adaplasso} = \arg\min_{\beta}\left(-2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}\right) + \lambda\sum_{j=1}^{p} w_j |\beta_j|$$

เนื่องจาก $\hat{\boldsymbol{\beta}}_{ols} = \mathbf{X}^T\mathbf{Y}$ จะได้ว่า

$$\hat{\boldsymbol{\beta}}^*_{adaplasso} = \arg\min_{\beta}\left(-2\hat{\boldsymbol{\beta}}^T_{ols}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta}\right) + \lambda\sum_{j=1}^{p} w_j\beta_j$$

$$= \arg\min_{\beta}\sum_{j=1}^{p}\left(-2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda w_j|\beta_j|\right) = L_j$$

ถ้า $\hat{\beta}^{ols}_j \geq 0$ จะได้ว่า $\beta_j \geq 0$

$\hat{\beta}^{ols}_j \leq 0$ จะได้ว่า $\beta_j \leq 0$

กรณี 1 $\hat{\beta}^{ols}_j \geq 0$ ดังนั้น $\beta_j \geq 0$

$$L_j = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda w_j\beta_j$$

$$\frac{\partial L_j}{\partial \beta_j} = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 + \lambda w_j\beta_j = 0$$

เมื่อ $\beta_j = \hat{\beta}^{adaplasso^*}_j$

$$2\hat{\beta}^{adaplasso^*}_j = 2\hat{\beta}^{ols}_j - \lambda w_j$$

$$\hat{\beta}^{adaplasso^*}_j = \hat{\beta}^{ols}_j - \frac{\lambda w_j}{2}$$

กรณี 2 $\hat{\beta}^{ols}_j \leq 0$ ดังนั้น $\beta_j \leq 0$

$$L_j = -2\hat{\beta}^{ols}_j\beta_j + \beta_j^2 - \lambda w_j\beta_j$$

$$\frac{\partial L_j}{\partial \beta_j} = -2\hat{\beta}_j^{ols} + \beta_j - \lambda w_j = 0$$

เมื่อ $\beta_j = \hat{\beta}_j^{adaplasso^*}$

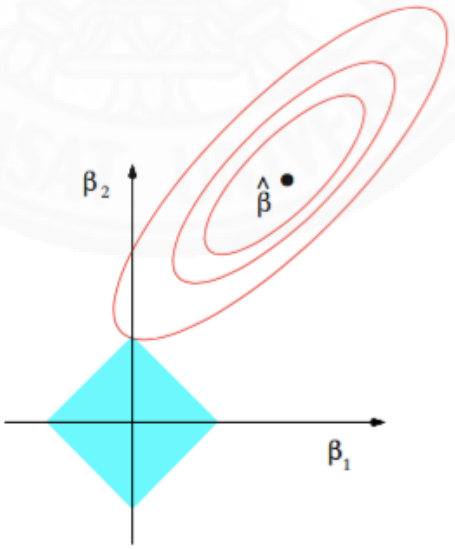
$$2\hat{\beta}_j^{adaplasso^*} = 2\hat{\beta}_j^{ols} + \lambda w_j$$

$$\hat{\beta}_j^{adaplasso^*} = \hat{\beta}_j^{ols} + \frac{\lambda w_j}{2}$$

ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีแลซโซแบบปรับปรุง $\left(\hat{\boldsymbol{\beta}}^*_{adaplasso}\right)$

$$\hat{\boldsymbol{\beta}}^*_{adaplasso} = \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda w_j}{2} & , \quad \hat{\beta}_j^{ols} > \frac{\lambda w_j}{2} \\ 0 & , \quad \frac{-\lambda w_j}{2} \le \hat{\beta}_j^{ols} \le \frac{\lambda w_j}{2} \\ \hat{\beta}_j^{ols} + \frac{\lambda w_j}{2} & , \quad \hat{\beta}_j^{ols} < \frac{-\lambda w_j}{2} \end{cases} \tag{2.18}$$

วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง ในตัวแบบเชิงเส้น สามารถแสดงลักษณะพื้นที่ผิวตอบสนองของ $\left(\mathbf{Y}-\mathbf{XB}\right)^T\left(\mathbf{Y}-\mathbf{XB}\right)$ ด้วยภาพที่ 2.3 ดังนี้



ภาพที่ 2.3 แสดงของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีกำลังสองน้อยที่สุด (OLS) และของเขตของตัวประมาณ $\hat{\boldsymbol{\beta}}$ ด้วยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO)

สำหรับตัวแบบการถดถอยแบบปัวซง ในปี 2007 Park และ Hastie ได้ศึกษาต่อจากงานของ Fan และ Li ในปี 2001 โดยใช้วิธีการวิเคราะห์ด้วยวิธีแลซโซแบบปรับปรุง ในตัวแบบถดถอยปัวซง เพื่อเพิ่มประสิทธิภาพในการการประมาณค่าพารามิเตอร์และการคัดเลือกตัวแปรมากขึ้น โดยมีการให้ความสำคัญกับตัวแปร โดยการถ่วงน้ำหนักให้กับตัวแปร สามารถคำนวณหาค่าสัมประสิทธิ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 penalty บน $\boldsymbol{\beta}$

โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

และ $P_\lambda(\boldsymbol{\beta})$ ของวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง คือ $\lambda \sum_{j=1}^{p} w_j |\beta_j|$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{adaplasso} = \arg\min_{\beta} \left(-l(\beta) + \lambda w_j \|\beta\|_1 \right) \tag{2.19}$$

สามารถหา $l(\boldsymbol{\beta})$ ได้ดังนี้

ตัวแบบการถดถอยปัวซง จะได้ฟังก์ชันน่าจะเป็น (Likelihood Function) คือ

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]^{y_i} \exp\left[-\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]}{y_i!}$$

$$= \frac{\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right]^{\sum_{i=1}^{n} y_i} \exp\left[-\sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta})\right]}{\prod_{i=1}^{n} y_i!}$$

ใส่ ln ทั้งสองข้างของสมการ จะได้

$$l(\boldsymbol{\mu}; \mathbf{y}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln\left[\mu(\mathbf{X}_i, \boldsymbol{\beta})\right] - \sum_{i=1}^{n} \mu(\mathbf{X}_i, \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln y_i!$$

แทนค่า $\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right)$ จะได้

$$= \sum_{i=1}^{n} y_i \ln \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right) - \sum_{i=1}^{n} \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right) - \sum_{i=1}^{n} \ln y_i!$$

$$= \sum_{i=1}^{n} y_i \left(\mathbf{X}_i^T \boldsymbol{\beta}\right) - \sum_{i=1}^{n} \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right) - \sum_{i=1}^{n} \ln y_i!$$

$$= \sum_{i=1}^{n} \left(y_i \mathbf{X}_i^T \boldsymbol{\beta} - \exp\left(\mathbf{X}_i^T \boldsymbol{\beta}\right) - \ln y_i! \right)$$

ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{adaplasso} = \arg\min_{\beta}\left(-\sum_{i=1}^{n}\left(\mathbf{y}_i\mathbf{X}_i^{T}\boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{T}\boldsymbol{\beta}\right) - \ln y_i!\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| w_j \right) \quad (2.20)$$

ซึ่ง w_j คือ การถ่วงน้ำหนักให้กับตัวแปร (Adaptive Weight) นิยามโดย $w_j = \left|\hat{\beta}_j\right|^{-\tau}$ สำหรับ $\tau > 0$ และ $\hat{\beta}_j$ คือ ค่าสูงสุดของ log likelihood $l(\boldsymbol{\beta})$ โดยจะให้ weight มาก กับตัวแปร inactive และเพื่อให้เกิดการหดตัวมากของค่าค่าสัมประสิทธ์ถดถอยมากขึ้น ในทางตรงข้ามจะให้ weight น้อย กับตัวแปร active เพื่อให้เกิดการหดตัวน้อยลงของค่าค่าสัมประสิทธ์ถดถอย ในทางทฤษฎีตัวประมาณด้วนวิธี Adaptive LASSO มีคุณสมบัติในการพยากรณ์ แต่ตัวประมาณ LASSO ไม่มี เมื่อกำหนด k และ $n \to \infty$ และเลือกค่า λ ตัวประมาณด้วยวิธี Adaptive LASSO เลือกตัวแบบที่ถูกต้องและแม่นยำ

2.2 งานวิจัยที่เกี่ยวข้อง

ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ การประมาณค่าพารามิเตอร์สัมประสิทธิ์ถดถอย วิธีที่นิยมใช้มากที่สุด คือ วิธีกำลังสองน้อยที่สุด เนื่องจากตัวประมาณด้วยวิธีนี้มีคุณสมบัติเป็นตัวประมาณไม่เอนเอียง และมีความแปรปรวนต่ำสุด แต่ในบางครั้งวิธีการนี้ไม่สามารถหาตัวประมาณสัมประสิทธิ์ถดถอยได้ เมื่อเกิดภาวะร่วมเชิงเส้น หรือตัวแปรอิสระมีความสัมพันธ์กันสูง ถ้าใช้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด จะทำให้ความแปรปรวนของตัวประมาณที่ได้เพิ่มสูงขึ้น จึงทำให้วิธีการวิเคราะห์ในวิธีการนี้ไม่เหมาะสม

ในปี 1970 Hoerl และ Kennard ได้แนะนำวิธีการประมาณค่าสัมประสิทธิ์การถดถอยในกรณีที่เกิดปัญหาภาวะร่วมเชิงเส้น สำหรับตัวแบบการถดถอยเชิงเส้น นั่นคือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ เพื่อแก้ไขปัญหาความไม่เพียงพอของวิธีกำลังสองน้อยที่สุด โดยตัวประมาณที่ได้จากวิธีการวิเคราะห์การถดถอยแบบริดจ์ จะเป็นตัวประมาณที่มีความเอนเอียง ซึ่งมีคุณสมบัติแตกต่างจากวิธีกำลังสองน้อยที่สุด แต่จะสามารถช่วยลดความแปรปรวนจากการประมาณค่าด้วยวิธีนี้ได้ โดยเชื่อว่า การประมาณบนพื้นฐานของ $\left[\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right]$ เมื่อ $k \geq 0$ ดีกว่า $\mathbf{X}^T\mathbf{X}$ โดยกระบวนการประมาณค่าด้วยวิธีการวิเคราะห์แบบริดจ์นี้ สามารถช่วยหลีกเลี่ยงความยุ่งยาก ด้วยการประมาณโดยใช้วิธีการวิเคราะห์ด้วยวิธีกำลังสองน้อยที่สุด กระบวนการวิเคราะห์นี้ช่วยให้เห็นการความไว ของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย แต่เนื่องจากวิเคราะห์มีคุณสมบัติในการประมาณ

ค่าพารามิเตอร์สัมประสิทธิ์การถดถอยเพียงอย่างเดียว ยังขาดคุณสมบัติในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ

นอกจากนี้ ในการศึกษาการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ได้แบ่งลักษณะของข้อมูลที่ต้องการศึกษาออกเป็น 2 แบบ คือ ข้อมูลที่มีมิติต่ำ $(n > p)$ และข้อมูลที่มีมิติสูง $(n < p)$ โดยที่ในการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด เหมาะสำหรับข้อมูลที่มีมิติต่ำ ดังนั้น ในกรณีที่เกิดข้อมูลในลักษณะตรงข้าม ทำให้การวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด อาจจะเกิดปัญหาในการวิเคราะห์ได้ นั่นคือ ปัญหาในการประมาณค่าสัมประสิทธิ์การถดถอย เนื่องจากข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ แล้วเมทริกซ์ $\mathbf{X}^T\mathbf{X}$ จะเป็นเมทริกซ์เอกฐาน ซึ่งไม่สามารถหาเมทริกซ์ผกผันได้ ดังนั้น วิธีกำลังสองน้อยที่สุดจึงไม่เหมาะสม ซึ่งสำหรับการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง โดยวิธีที่นิยมใช้คือ วิธีพีนอลไลซ์รีเกรสชั่น ซึ่งมีจุดมุ่งหมายเพื่อประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย ที่สามารถใช้กับข้อมูลที่มีมิติสูง เพื่อช่วยแก้ปัญหาดังกล่าวๆได้

ต่อมา ในปี 1996 Tibshirani ได้นำเสนอวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบแลซโซ โดยวิธีการนี้มีคุณสมบัติเช่นเดียวกับตัวประมาณที่ได้จากวิธีการวิเคราะห์แบบริดจ์ ซึ่งมีคุณสมบัติเป็นตัวประมาณที่เอนเอียง แต่สามารถลดความแปรปรวนลงได้ เมื่อไม่สามารถใช้วิธีการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้ โดยวิธีการวิเคราะห์การถดถอยแบบแลซโซ มีข้อดีที่แตกต่างจากวิธีการวิเคราะห์การถดถอยแบบริดจ์ นั่นคือ นอกจากวิธีการนี้จะสามารถประมาณค่าพารามิเตอร์การถดถอยได้ เมื่อตัวแปรอิสระมีเป็นจำนวนมาก ยังสามารถคัดเลือกตัวแปรที่เหมาะสมหรือมีอิทธิพลสูงเข้าสู่ตัวแบบได้ ซึ่งสามารถทำให้การอธิบายผลลัพธ์ของตัวแบบได้ง่ายขึ้นในกรณีที่ตัวแปรอิสระมีเป็นจำนวนมาก แต่วิธีการนี้ก็ยังมีข้อเสียที่ว่า การคัดเลือกตัวแปรเข้าสู่ตัวแบบ ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง ทำให้การคัดเลือกตัวแปร จะเลือกตัวแปรจากกลุ่มใดกลุ่มหนึ่งมาเพียงหนึ่งตัว โดยไม่สนใจว่าตัวแปรอิสระตัวนั้นมีความสำคัญกับตัวแบบมากที่สุดหรือไม่

ในปี 2006 Zou ได้นำเสนอวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แลซโซแบบปรับปรุง เพื่อที่จะแก้ไขข้อจำกัดของวิธีการวิเคราะห์การถดถอยแบบแลซโซ ให้มีคุณสมบัติในการคัดเลือกตัวแปรที่แม่นยำและมีประสิทธิภาพมากขึ้น โดยมีการถ่วงน้ำหนัก (w_j) ให้กับตัวแปรอิสระที่คัดเลือกเข้าสู่ตัวแบบ ซึ่งทำให้เกิดความคงเส้นคงวาในการคัดเลือกตัวแปร โดยวิธีการนี้มีคุณสมบัติเช่นเดียวกับตัวประมาณที่ได้จากวิธีการวิเคราะห์แบบริดจ์ ซึ่งมีคุณสมบัติเป็น

ตัวประมาณที่เอนเอียง แต่สามารถลดความแปรปรวนลงได้ เมื่อไม่สามารถใช้วิธีการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดได้

ในปี 2007 Park และ Hastie ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซโซ สำหรับตัวแบบการถดถอยปัวซง โดยวิธีการวิเคราะห์นี้เป็นวิธีที่นิยมสำหรับการประมาณค่าและการคัดเลือกตัวแปรในคราวเดียวกัน ในปี 2007 Park และ Hastie ได้ศึกษาต่อจากงานของ Fan และ Li ในปี 2001 โดยใช้วิธีการวิเคราะห์ด้วยวิธีแลซโซแบบปรับปรุง ในตัวแบบถดถอยปัวซง เพื่อเพิ่มประสิทธิภาพในการการประมาณค่าพารามิเตอร์และการคัดเลือกตัวแปรมากขึ้น โดยมีการให้ความสำคัญกับตัวแปร โดยการถ่วงน้ำหนักให้กับตัวแปร

ในปี 2010 Yüzbaşı, Arashi และ Ahmed จึงได้มีการเปรียบเทียบประสิทธิภาพการประมาณค่าสัมประสิทธิ์การถดถอย ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กัน กำหนดให้ความสัมพันธ์มีขนาด $(r = 0.5, 0.9)$ ในตัวแบบเชิงเส้น ในข้อมูลที่มีมิติต่ำและมิติสูง ซึ่งมีวิธีการวิเคราะห์การถดถอยแบบริดจ์ การวิเคราะห์การถดถอยแบบแลซโซ และการวิเคราะห์การถดถอยแลซโซปรับปรุงรวมอยู่ด้วย ซึ่งให้ผลการวิจัยว่า สำหรับกรณีที่ข้อมูลมีมิติต่ำ $(n > p)$ ประสิทธิภาพในพยากรณ์และประสิทธิภาพในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย โดยเปรียบเทียบประสิทธิภาพเพียง 3 วิธีดังกล่าว จะเห็นว่า ตัวประมาณด้วยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง จะให้ประสิทธิภาพดีที่สุด

แต่ในบางครั้งในการวิเคราะห์การถดถอยพหุคูณ ตัวแปรตอบสนองไม่เป็นตัวแปรสุ่มชนิดต่อเนื่อง แต่มีลักษณะเป็นข้อมูลจำนวนนับ (Count Data) เช่น ข้อมูลทางการแพทย์ ผู้วิจัยอาจจะต้องการศึกษา ปัจจัยที่มีผลต่อจำนวนครั้งของการกำเริบ (Exacerbation) ในผู้ป่วยในผู้ป่วยโรคหลอดลมอุดตันเรื้อรัง (Chronic Obstructive Pulmonary Disease : COPD) หรือศึกษาจำนวนคนที่เสียชีวิตด้วยโรคเอดส์ในช่วงเวลา 3 เดือน ตั้งแต่เดือนมกราคม 2526 จนถึง เดือนมิถุนายน 2529 ดังนั้น ในการศึกษาความสัมพันธ์ของตัวแปรตอบสนองที่มีลักษณะดังกล่าวกับตัวแปรอิสระต่างๆ ตัวแบบที่นิยมศึกษา คือ ตัวแบบการถดถอยปัวซง (Poisson Regression) ดังนั้น จึงมีผู้วิจัยสนใจที่จะศึกษาการวิเคราะห์การถดถอยปัวซง ซึ่งวิธีการวิเคราะห์การถดถอยปัวซงแบบเดิมนั้น วิธีที่นิยมนำมาใช้ คือ วิธีภาวะน่าจะเป็นสูงสุด โดยลักษณะของข้อมูลที่ศึกษาแบ่งออกได้เป็น 2 แบบ คือ ข้อมูลที่มีมิติต่ำ $(n > p)$ และข้อมูลที่มีมิติสูง $(n < p)$ เช่นเดียวกับการวิเคราะห์การถดถอยเชิงเส้น ถ้าในกรณีที่ข้อมูลมีมิติสูง และตัวแปรอิสระมีความสัมพันธ์กันหรือเกิดปัญหาภาวะร่วมเชิงเส้น ทำให้วิธีการวิเคราะห์สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดไม่มีประสิทธิภาพมากพอ จึงได้มี

ผู้วิจัยหาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบอื่นๆ เพื่อเพิ่มประสิทธิภาพในการประมาณเมื่อเกิดกรณีดังกล่าว

ในปี 2012 Hossain และ Ahmed ได้ศึกษาการเปรียบเทียบประสิทธิภาพการการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีพีนอลไลซ์และวิธีการหดลงของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย (Shrinkage) ในตัวแบบการถดถอยปัวซง โดยศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยด้วยกัน 6 วิธี เทียบกับวิธีการวิเคราะห์สัมประสิทธิ์การถดถอยแบบวิธีภาวะน่าจะเป็นสูงสุด นั่นคือ วิธีภาวะความน่าจะเป็นสูงสุดแบบเต็ม (Unrestricted Maximum Likelihood Estimator : UE) และวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ศึกษา คือ 1. วิธีภาวะความน่าจะเป็นสูงสุดแบบจำกัด (Restricted Maximum Likelihood Estimator : RE) 2. วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบหดค่าลง (Shrinkage Estimator : SE) ซึ่งเป็นวิธีที่มีความคลาดเคลื่อนหรือความแปรปรวนต่ำกว่า ตัวประมาณสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด ในตัวแบบถดถอย 3. วิธีการประมาณแบบหดค่าลงทางบวก (Positive Part Shrinkage Estimator : PSE) วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบหดค่าลงไม่เป็นผลรวมเชิง convex ของตัวประมาณด้วยวิธีภาวะความน่าจะเป็นสูงสุดแบบเต็ม และวิธีภาวะความน่าจะเป็นสูงสุดแบบจำกัด ดังนั้น ตัวประมาณด้วยวิธีการประมาณแบบหดค่าลงทางบวก จะหลีกเลี่ยงความผิดปกติของตัวประมาณด้วยวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบหดค่าลง (Shrinkage Estimator) ได้ 4. วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบแลซโซ เป็นวิธีที่นิยมสำหรับการประมาณค่าและการคัดเลือกตัวแปรในคราวเดียวกัน 5. วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแลซโซแบบปรับปรุง และ 6. วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบ SCAD โดยพบว่า ในการจำลองข้อมูลที่มีมิติต่ำ $(n < p)$ จะเปรียบเปรียบประสิทธิภาพของตัวประมาณด้วยวิธีตัวประมาณด้วยวิธีพีนอลไลซ์ นั่นคือ วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบแลซโซ และวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแลซโซแบบปรับปรุง พบว่า ตัวประมาณนี้จะมีประสิทธิภาพ เมื่อตัวแปรอิสระที่ไม่มีผล $(\beta = 0)$ มีเป็นจำนวนมากอยู่ในตัวแบบ ในทางตรงกันข้าม ตัวประมาณด้วยวิธีการประมาณแบบหดค่าลงทางบวก (PSE) จะมีประสิทธิภาพ เมื่อตัวแปรอิสระที่ไม่มีผล มีเป็นจำนวนปานกลางหรือมากอยู่ในตัวแบบ ยิ่งไปกว่านั้น ตัวประมาณด้วยวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบหดค่าลง (SE) จะมีประสิทธิภาพกว่า ตัวประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุด สำหรับในกรณีที่ตัวแปรอิสระที่ไม่มีผลมากกว่า 2 ตัวขึ้นไป

ในปี 2012 นิศาชล งามประเสริฐสิทธิ์ ได้ศึกษาการเปรียบเทียบประสิทธิภาพของการคัดเลือกตัวแปรอิสระในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ที่ตัวแบบมีตัวแปรอิสระที่เกี่ยวข้อง

และไม่เกี่ยวข้องกับตัวแปรตาม โดยตัวแปรอิสระที่เกี่ยวข้องมีความสัมพันธ์กันสูง 0.95, 0.99, 0.999 และ 0.9999 การคัดเลือกตัวแปรอิสระใช้วิธีการถดถอยด้วยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ โดยใช้วิธีการประมาณค่าพารามิเตอร์ริดจ์ 4 วิธี คือ วิธีโฮเอิร์ล, เคนนาร์ด และ บาลด์วิน (Hoerl, Kennard and Baldwin) วิธีลอว์เลสและแวง (Lawless and Wang) วิธีนอมูระ (Nomura) และวิธีคาลาฟและชูเกอร์ (Khalaf and Shukur) กับการค้นหาแบบต้องห้ามที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาด เคลื่อนกำลังสองเฉลี่ย(MSE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันพีนอลตี้ (Penalty Function) เกณฑ์ที่ใช้ในการเปรียบเทียบการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ คือ ร้อยละของจำนวนครั้งที่แต่ละวิธีสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ตามตัวแบบจริง, ตัวแบบ Overspecification, ตัวแบบ Underspecification และตัวแบบ Misspecification การศึกษาใช้วิธีการจำลองข้อมูล กำหนดขนาดตัวอย่าง (n) เท่ากับ 20, 60 และ 100

จากการศึกษาพบว่า วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันพีนอลตี้ มีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องตามตัวแบบจริงมากกว่าวิธีการถดถอยด้วยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ในทุกขนาดตัวอย่าง เมื่อสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.95, 0.99 และ 0.999 แต่เมื่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระสูงขึ้นเป็น 0.9999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและวิธีการถดถอยแบบขั้นตอนที่มีการประมาณค่าพารามิเตอร์กำลังสองน้อยที่สุดและแบบริดจ์ มีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องมีค่าต่ำ เมื่อขนาดตัวอย่างเท่ากับ 20 $(n=20)$ แต่จะค่อยๆเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น และมีร้อยละของตัวแบบ Underspecification ลดลงอย่างเห็นได้ชัด ในขณะที่วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันพีนอลตี้ มีร้อยละการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องสูงและค่อนข้างคงที่ โดยไม่ขึ้นกับขนาดตัวอย่างและค่าสัมประสิทธิ์สหสัมพันธ์ นอกจากนั้นผลการศึกษาไม่พบตัวแบบ Underspecification และ Misspecification เลยมีเพียงตัวแบบ Overspecification ซึ่งเป็นปัญหาที่รุนแรงน้อยกว่าการที่ตัวแบบมีตัวแปรอิสระที่เกี่ยวข้องขาดหายไป

ในปี 2014 ฑิฆัมพร สาระกอ และ นัท กุลวานิช ได้ศึกษาการเปรียบเทียบประสิทธิภาพวิธีการคัดเลือกและการพยากรณ์ที่เหมาะสม สำหรับตัวแบบการถดถอยเชิงเส้น โดยทำการเปรียบเทียบจากวิธีที่แตกต่าง ซึ่งในงานวิจัยนี้จะเปรียบเทียบผลที่ได้จาก วิธี Stepwise, LASSO, Adaptive LASSO และ Elastic net โดยมีการจำลองข้อมูลให้มีขอบเขตที่แตกต่างกัน ดังนี้ ขนาดตัวอย่างเท่ากับ 40 120 240 และมีอัตราส่วนของจำนวนตัวแปรต่อขนาดตัวอย่างเป็น 0.3 และ 0.7

จำนวนของค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่มีค่าเป็นศูนย์คิดเป็นร้อยละ 10 50 และ 90 ของจำนวนตัวแปร โดยเกณฑ์ที่ใช้ในการวัดประสิทธิภาพของผลที่ได้จากการประมาณค่าของแต่ละวิธี ได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) อัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) และค่าคลาดเคลื่อนในการทำนาย (PE) ผลการศึกษาพบว่า อัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) และค่าคลาดเคลื่อนในการทำนาย (PE) นั้น ให้ผลไปในทิศทางเดียวกัน คือ วิธี Adaptive LASSO นั้นมีประสิทธิภาพที่เหมาะสมกับข้อมูลที่มีขนาดเล็ก และมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์มากที่สุด

ในปี 2015 Algamal และ Lee ได้ศึกษาในข้อมูลที่มีมิติต่ำ ในตัวแบบการถดถอยปัวซง นอกจากนี้ตัวแปรอิสระมีความสัมพันธ์กันสูงอีกด้วย ดังนั้นผู้วิจัยจึงใช้วิธีการวิเคราะห์แบบพีนอลไลซ์เพื่อแก้ปัญหาที่เกิดขึ้น โดยมีการแบ่งข้อมูลออกเป็น 2 ชุด นั่นคือ ชุดทดสอบ และชุดฝึกฝน เพื่อเปรียบเทียบค่ามัธยฐานของความคลาดเคลื่อนเฉลี่ยในชุดของมูลฝึกฝน และชุดทดสอบ ตามลำดับ นอกจากนี้ได้ศึกษาการคัดเลือกตัวแปรของวิธีการวิเคราะห์การถดถอยของทั้ง 3 วิธี ได้แก่ วิธีการวิเคราะห์แบบแลซโซ, วิธีการวิเคราะห์แลซโซแบบปรับปรุง, วิธีการวิเคราะห์แบบลาแลซโซ (RALASSO) ซึ่งเป็นการรวมวิธีการวิเคราะห์แบบแลซโซและแลซโซแบบปรับปรุงเข้าด้วยกัน เพื่อเป็นการแก้ปัญหาการเกิดข้อมูลที่มีมิติสูง โดยในงานวิจัยนี้ ได้ใช้ในการจำลองข้อมูล เพื่อเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี และให้ผลสรุปว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุงมีประสิทธิภาพสูงสุด ในตัวแบบการถดถอยปัวซง ในข้อมูลที่มีมิติต่ำ แต่วิธีการวิเคราะห์แบบลาแลซโซ จะแสดงให้เห็นถึงประสิทธิภาพของการคัดเลือกตัวแปรได้อย่างมีประสิทธิภาพ โดยวิธีการนี้จะมีความแม่นยำมากขึ้น

ในปี 2015 Oyeyemi, Ogunjobi และ Folorunsho ได้ศึกษาการวิเคราะห์การถดถอยในกรณีที่เกิดปัญหาที่ตัวแปรอิสระมีความสัมพันธ์กันสูง ในตัวแบบเชิงเส้น และกรณีที่ขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ $(n > p)$ โดยคำนึงถึงปัญหาของการค่าสัมประสิทธิ์การถดถอย (β) ด้วยวิธีกำลังน้อยที่สุด เมื่อตัวแปรอิสระอยู่ในลักษณะดังกล่าว ดังนั้นจึงได้ศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีแลซโซ วิธีแลซโซแบบปรับปรุง และวิธี Elastic net โดยเปรียบเทียบประสิทธิภาพของวิธีดังกล่าวด้วยการจำลองข้อมูล ซึ่งเกณฑ์ที่ใช้ในการเปรียบเทียบ คือ Akaike Information Criterion (AIC) และ Bayesian Information Criterion (BIC) เป็นวิธีการทดสอบการคัดเลือกตัวแปรเข้าสู่ตัวแบบ และจากการจำลองข้อมูลจะได้ว่า ประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบแลซโซมีประสิทธิภาพสูงที่สุด แต่วิธีการวิเคราะห์การถดถอยแบบ Elastic net มีแนวโน้มที่จะมีความแม่นยำในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวมากขึ้น เมื่อขนาดตัวอย่างเพิ่มมากขึ้น

ในปี 2016 Ivanoff, Picard และ Rivoirard ได้ศึกษาข้อมูลที่มิติสูง ในตัวแบบการถดถอยปัวซง โดยในงานวิจัยนี้จะประมาณค่าสัมประสิทธิ์การถดถอย ด้วยวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แบบกรุ๊ปแลซโซ (Group LASSO) ซึ่งจะเป็นการรวมวิธีการวิเคราะห์ทั้ง 2 วิธีดังกล่าว เพื่อใช้ในการวิเคราะห์การถดถอยหรือประมาณค่าสัมประสิทธิ์การถดถอย ในส่วนของการคัดเลือกตัวแปรอิสระ จะขึ้นอยู่กับการถ่วงน้ำหนักในฟังก์ชันพีนอลตี้ ที่ต้องใช้วิธีการนี้ เนื่องจากวิธีพื้นฐานสามารถใช้ในตัวแบบเกาส์เซี่ยน แต่ไม่สามารถนำมาใช้ได้โดยตรงกับตัวแบบปัวซง เพราะชนิดของข้อมูลที่ใช้แตกต่างกัน ดังนั้นในงานวิจัยนี้จะแสดงเห็นถึงประสิทธิภาพในการวิเคราะห์การถดถอยด้วยวิธีการวิเคราะห์แบบ LASSO และ วิธี Group LASSO เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ในตัวแบบปัวซง

ในปี 2016 Gao, Ahmed และ Feng ได้ศึกษาการวิเคราะห์การถดถอยเชิงเส้น ในกรณีที่ข้อมูลที่มีมิติสูง หรือจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมากๆ $(n < p)$ เมื่อข้อมูลมีลักษณะดังกล่าว วิธีที่นิยมใช้ คือ วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ ซึ่งมีอยู่ด้วยกันหลายวิธี โดยศึกษาทั้งการประมาณค่าสัมประสิทธิ์การถดถอยและการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ แต่ก็ยังไม่ได้ให้ประสิทธิภาพมากเท่าที่ควร ดังนั้น งานวิจัยนี้จึงได้นำเสนอวิธี Post Selection Shrinkage Estimation for High-Dimensional Data Analysis (PSE) เพื่อที่จะพัฒนาประสิทธิภาพของการพยากรณ์ จากการจำลองข้อมูล จะได้ว่าประสิทธิภาพของการวิธีการวิเคราะห์แบบ PSE มีประสิทธิภาพดีกว่าวิธีการวิเคราะห์แบบริดจ์ที่มีการถ่วงน้ำหนัก (Weighted Ridge) ไม่ว่าจะเป็นประสิทธิภาพในการพยากรณ์และประสิทธิภาพในการคัดเลือกตัวแปรอิสระ จะให้ผลลัพธ์ที่ดีกว่า

บทที่ 3
วิธีการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการวิเคราะห์การถดถอยปัวซง (Poisson Regression) สำหรับข้อมูลที่มีมิติสูงแบบบางเบา ในกรณีที่ตัวแปรอิสระเกิดภาวะร่วมเชิงเส้นทั้ง 3 รูปแบบ นั่นคือ Constant model , Toeplitz model, และ Hub Toeplitz โดยเปรียบเทียบประสิทธิภาพของเครื่องมือในการวิเคราะห์การถดถอย แบบ penalized regression 3 วิธี ได้แก่ Ridge regression, LASSO และ Adaptive LASSO

3.1 ข้อมูลที่มีมิติสูงแบบบางเบา สำหรับตัวแบบการถดถอยปัวซง

พิจารณา สำหรับตัวแบบการถดถอยปัวซงพิจารณาตัวแบบพื้นฐานของการถดถอยเชิงเส้นพหุคูณ สำหรับตัวแปรอิสระ p ตัว และขนาดของกลุ่มตัวอย่างเท่ากับ n

$$\mathbf{Y} = \exp\left(\mathbf{X\beta}\right) + \boldsymbol{\varepsilon} \qquad \text{เมื่อ} \quad n < p$$

สมมติให้ $\mathbf{x}_i$ สามารถแยกออกเป็น $\mathbf{x}_i = \left(\mathbf{x}_{iA}, \mathbf{x}_{iB}\right)$

โดยที่ $\mathbf{x}_{iA} = \left(\mathbf{x}_{i1}, ..., \mathbf{x}_{i(p-q)}\right)^T \in \mathbb{R}^{p-q}$

$\mathbf{x}_{iB} = \left(\mathbf{x}_{i(p-q+1)}, ..., \mathbf{x}_{ip}\right)^T \in \mathbb{R}^{q}$

โดยที่

p คือ จำนวนพารามิเตอร์ทั้งหมด

q คือ จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอย มีค่าเท่ากับ 0 (Inactive Parameter)

$p-q$ คือ จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอย มีค่าไม่เท่ากับ 0 (Active Parameter) และส่วนมากมีค่าเป็นศูนย์

ถ้าจำนวนพารามิเตอร์สัมประสิทธิ์การถดถอยที่มีค่าเท่ากับศูนย์ มีจำนวนมากกว่าจำนวนพารามิเตอร์สัมประสิทธิ์การถดถอยที่มีค่าไม่เท่ากับศูนย์ หรือ $q > p-q$ จะเรียกข้อมูลที่มีลักษณะนี้ว่า “ข้อมูลบางเบา”

ให้ $\mathbf{X}_A = \left(\mathbf{x}_{1A}, ..., \mathbf{x}_{nA}\right)^T \in \mathbb{R}^{n\times(p-q)}$ และ $\mathbf{X}_B = \left(\mathbf{x}_{1B}, ..., \mathbf{x}_{nB}\right)^T \in \mathbb{R}^{n\times q}$ คือ เมทริกซ์ของ $\mathbf{x}_{iA}$ และ $\mathbf{x}_{iB}$ ตามลำดับ

จะได้ว่า $\mathbf{X}=\left(\mathbf{X}_A,\mathbf{X}_B\right)^T\in\mathbb{R}^{n\times p}$ คือ เมทริกซ์ของตัวแปรอิสระทั้งหมด และสามารถเขียนให้อยู่ในตัวแบบ ได้ดังนี้

$$\mathbf{Y}=\exp\left(\mathbf{X}_A\boldsymbol{\beta}_A+\mathbf{X}_B\boldsymbol{\beta}_B\right)+\boldsymbol{\varepsilon}$$

โดยที่ $\boldsymbol{\beta}=\left(\boldsymbol{\beta}_A,\boldsymbol{\beta}_B\right)^T\in\mathbb{R}^p$, $\boldsymbol{\beta}_A^T\in\mathbb{R}^{p-q}$ และ $\boldsymbol{\beta}_B^T\in\mathbb{R}^q$ คือ เวกเตอร์ที่ไม่ทราบค่าของสัมประสิทธิ์การถดถอย

3.2 การหาตัวประมาณสัมประสิทธิ์การถดถอยปัวซง

3.2.1 วิธีการวิเคราะห์แบบริดจ์

ในปี 2011 Kristofer และ Ghazi ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ โดยตัวประมาณแบบริดจ์สำหรับตัวแบบการถดถอยปัวซง เป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง (Multicollinearity) สามารถคำนวณหาค่าสัมประสิทธิ์การถดถอย จากการหาค่าต่ำสุด ของฟังก์ชันลบของลอการิทึมของฟังก์ชันน่าจะเป็น (Log Likelihood) จากวิธีภาวะน่าจะเป็นสูงสุด ภายใต้เงื่อนไข L_2 penalty บน $\boldsymbol{\beta}$ โดยที่ $L_2=\left\|\boldsymbol{\beta}\right\|_2=\sum_{j=1}^{p}\beta_j{}^2$

และ $P_\lambda\left(\boldsymbol{\beta}\right)$ ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ คือ $\lambda\sum_{j=1}^{p}\beta_j^2$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบริดจ์ ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ridge} &= \arg\min_{\beta}\left(-l\left(\boldsymbol{\beta}\right)+\lambda\left\|\boldsymbol{\beta}\right\|_2\right)\\ &= \arg\min_{\beta}\left(-\sum_{i=1}^{n}\left(y_i\mathbf{X}_i{}^T\boldsymbol{\beta}-\exp\left(\mathbf{X}_i{}^T\boldsymbol{\beta}\right)-\ln y_i!\right)+\lambda\sum_{j=1}^{p}\beta_j^2\right)\end{aligned}$$

เมื่อ λ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดการหดตัว (Shrinkage) ของตัวประมาณ $\hat{\boldsymbol{\beta}}_{ridge}$ โดยที่ $\lambda\to\infty$, $\hat{\boldsymbol{\beta}}\to 0$

3.2.2 วิธีการวิเคราะห์แบบแลซโซ

ในปี 2007 Park และ Hastie ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซโซ สำหรับตัวแบบการถดถอยปัวซง โดยวิธีการวิเคราะห์นี้เป็นวิธีที่นิยม

สำหรับการประมาณค่าและการคัดเลือกตัวแปรในคราวเดียวกัน สามารถคำนวณหาค่าสัมประสิทธ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 penalty บน $\boldsymbol{\beta}$ โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

และ $P_\lambda(\boldsymbol{\beta})$ ของวิธีการวิเคราะห์การถดถอยแบบแลซโซ คือ $\lambda \sum_{j=1}^{p} |\beta_j|$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบ LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \left(-l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right)$$

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \left(-\sum_{i=1}^{n} \left(y_i \mathbf{X_i}^T \boldsymbol{\beta} - \exp\left(\mathbf{X_i}^T \boldsymbol{\beta} \right) - \ln y_i ! \right) + \lambda \sum_{j=1}^{p} \beta_j \right)$$

ซึ่ง $\lambda > 0$ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดของการหดตัว (Shrinkage) ของตัวประมาณ $\boldsymbol{\beta}$ ถ้า $\lambda_0 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} \beta_j$ เมื่อกำหนด $\lambda_0 > \lambda$ จะทำให้ตัวประมาณสัมประสิทธิ์ถูกดึงหดลงเข้าหาค่าศูนย์ และตัวประมาณบางตัวอาจเท่ากับศูนย์

3.2.3 วิธีการวิเคราะห์แลซโซแบบปรับปรุง

ในปี 2007 Park และ Hastie ได้ศึกษาต่อจากงานของ Fan และ Li ในปี 2001 โดยใช้วิธีการวิเคราะห์ด้วยวิธีแลซโซแบบปรับปรุง ในตัวแบบถดถอยปัวซง เพื่อเพิ่มประสิทธิภาพในการการประมาณค่าพารามิเตอร์และการคัดเลือกตัวแปรมากขึ้น โดยมีการให้ความสำคัญกับตัวแปร โดยการถ่วงน้ำหนักให้กับตัวแปร สามารถคำนวณหาค่าสัมประสิทธ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 penalty บน $\boldsymbol{\beta}$

โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

และ $P_\lambda(\boldsymbol{\beta})$ ของวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง คือ $\lambda \sum_{j=1}^{p} w_j |\beta_j|$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบ Adaptive LASSO ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

$$\hat{\boldsymbol{\beta}}_{adaplasso} = \arg\min_{\beta} \left(-l(\boldsymbol{\beta}) + \lambda w_j \|\boldsymbol{\beta}\|_1 \right)$$

$$\hat{\boldsymbol{\beta}}_{adaplasso} = \underset{\beta}{\operatorname{argmin}}\left(-\sum_{i=1}^{n}\left(\mathbf{y}_i \mathbf{X}_i^{T}\boldsymbol{\beta} - \exp\left(\mathbf{X}_i^{T}\boldsymbol{\beta}\right) - \ln y_i !\right) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| w_j \right)$$

ซึ่ง $\lambda > 0$ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดของการหดตัว (Shrinkage) ของตัวประมาณ $\boldsymbol{\beta}$

3.3 การสร้างรูปแบบความสัมพันธ์ของตัวแปรอิสระ

3.3.1. รูปแบบความสัมพันธ์แบบ Constant correlation model

เป็นรูปแบบความสัมพันธ์แรก ที่ค่าความสัมพันธ์เป็นค่าคงที่ ทั้งความสัมพันธ์ในแต่ละกลุ่มและระหว่างกลุ่ม และสามารถอธิบายโครงสร้างของรูปแบบความสัมพันธ์แบบคงที่ ได้ดังนี้

$$\Sigma_k = \begin{bmatrix} 1 & \rho_k & \rho_k & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \rho_k & \rho_k & \cdots & \rho_k \\ \rho_k & \rho_k & 1 & \rho_k & \cdots & \rho_k \\ \rho_k & \rho_k & \rho_k & 1 & \cdots & \rho_k \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k & \rho_k & \rho_k & \cdots & 1 \end{bmatrix}$$

โดยที่

Σ_k คือ เมทริกซ์สหสัมพันธ์ขนาด $(k \times k)$

k คือ จำนวนกลุ่มการจำแนกหรือจำนวนตัวแปรอิสระ $(k > 0)$

g_k คือ ขนาดของจำนวนกลุ่มการจำแนกหรือขนาดตัวอย่าง $(g_k > 0)$

ρ_k คือ ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม โดยมีค่าอยู่ระหว่าง $0 \le \rho_k \le 1$

ตัวอย่างเช่น มีจำนวนตัวแปรอิสระทั้งหมด 5 ตัวแปร และกำหนดให้ $\rho_k = 0.5$

$$\Sigma = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$$

3.3.2. รูปแบบความสัมพันธ์แบบ Toeplitz correlation model

เป็นรูปแบบความสัมพันธ์ที่มีลักษณะ คือ ตัวแปรอิสระที่อยู่ใกล้กัน จะมีความสัมพันธ์กันสูง และความสัมพันธ์จะลดน้อยลงเมื่อตัวแปรอิสระอยู่ห่างกันมากขึ้น ในการสร้างตัวแบบความสัมพันธ์นี้ Guoet al. (2007) ได้อธิบายโครงสร้างของรูปแบบความสัมพันธ์แบบ Toeplitz ได้ดังนี้

$$\Sigma_k = \begin{bmatrix} 1 & \rho_k & \rho_k^{\,2} & \rho_k^{\,3} & \cdots & \rho_k^{\,g_k-1} \\ \rho_k & 1 & \rho_k & \rho_k^{\,2} & \cdots & \rho_k^{\,g_k-2} \\ \rho_k^{\,2} & \rho_k & 1 & \rho_k & \cdots & \rho_k^{\,g_k-3} \\ \rho_k^{\,3} & \rho_k^{\,2} & \rho_k & 1 & \cdots & \rho_k^{\,g_k-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_k^{\,g_k-1} & \rho_k^{\,g_k-2} & \rho_k^{\,g_k-3} & \rho_k^{\,g_k-4} & \cdots & 1 \end{bmatrix}$$

โครางสร้างที่มีลักษณะแบบ Toeplitz ได้ถูกนำมาใช้อย่างแพร่หลายในการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) ความสัมพันธ์ในแต่ละกลุ่ม (Group Correlation) (Paneg et al.,2009; Dabney และ Storey, 2007; Zuber และ Strimmer, 2009; Tibshirani, 2009; Guo etal., 2007) นอกจากนี้ โครงสร้างที่มีลักษณะนี้ยังได้ถูกนำมาใช้ในเรื่องของอนุกรมเวลา (Time Series) (Ng และ Joe,2010; Joe, 2006) โดยโครงสร้างนี้ ได้นิยามความสัมพันธ์ระหว่างค่าสังเกตตัวที่ i และ j คือ $\rho^{|i-j|}$ ซึ่งความสัมพันธ์นี้จะมีลักษณะเป็นเอกซ์โปแนนเชียล (Exponential)

โดยที่

Σ_k คือ เมทริกซ์สหสัมพันธ์ขนาด $(k \times k)$

k คือ จำนวนกลุ่มการจำแนกหรือจำนวนตัวแปรอิสระ $(k > 0)$

g_k คือ ขนาดของจำนวนกลุ่มการจำแนกหรือขนาดตัวอย่าง $(g_k > 0)$

ρ_k คือ ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม หรือความสัมพันธ์ของตัวแปรอิสระ โดยมีค่าอยู่ระหว่าง $0 \le \rho_k \le 1$

ตัวอย่างเช่น มีจำนวนตัวแปรอิสระทั้งหมด 5 ตัวแปร และกำหนดให้ $\rho_k = 0.5$

$$\Sigma = \begin{bmatrix} 1 & 0.5^{|1-2|} & 0.5^{|1-3|} & 0.5^{|1-4|} & 0.5^{|1-5|} \\ 0.5^{|2-1|} & 1 & 0.5^{|2-3|} & 0.5^{|2-4|} & 0.5^{|2-5|} \\ 0.5^{|3-1|} & 0.5^{|3-2|} & 1 & 0.5^{|3-4|} & 0.5^{|3-5|} \\ 0.5^{|4-1|} & 0.5^{|4-2|} & 0.5^{|4-3|} & 1 & 0.5^{|4-5|} \\ 0.5^{|5-1|} & 0.5^{|5-2|} & 0.5^{|5-3|} & 0.5^{|5-4|} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.5 & 0.025 & 0.125 & 0.0625 \\ 0.5 & 1 & 0.5 & 0.025 & 0.125 \\ 0.025 & 0.5 & 1 & 0.5 & 0.025 \\ 0.125 & 0.025 & 0.5 & 1 & 0.5 \\ 0.0625 & 0.125 & 0.025 & 0.5 & 1 \end{bmatrix}$$

3.3.3. รูปแบบความสัมพันธ์แบบ Hub Toeplitz correlation model

โครงสร้างแบบ Hub จะมีลักษณะที่เรารู้ความสัมพันธ์ระหว่างจุดศูนย์รวมค่าสังเกต (Hub Observation) หรือค่าสังเกตแรก กับแต่ละค่าสังเกตอื่นๆ อย่างไรก็ตาม หนึ่งในรูปแบบความสัมพันธ์แบบ Hub จะสมมติว่า ความสัมพันธ์ระหว่างค่าสังเกตตัวแรก กับค่าสังเกตตัวที่ i จะค่อยๆลดลงตามลำดับที่ i

สามารถอธิบายรูปแบบความสัมพันธ์ ได้ดังนี้ โดยสมมติว่า แถวแรกและหลักแรกของเมทริกซ์สหสัมพันธ์ (A) ที่มีขนาด $(g \times g)$ สามารถคำนวณได้จาก

$$A_{11} = 1, \qquad A_{1i} = \rho_{\max} - (\rho_{\max} - \rho_{\min})\left(\frac{i-2}{g-2}\right)^{\gamma}$$

ซึ่งค่าความสัมพันธ์ของค่าสังเกตจะลดลง จาก $A_{12} = \rho_{\max}$ ไปจนถึง $A_{1g} = \rho_{\min}$ เมื่อ $2 \le i \le g$ สำหรับความสัมพันธ์ของค่าสังเกตที่มีลักษณะเชิงเส้น $(\gamma = 1)$

ตัวอย่างเช่น ตัวแบบในงานวิจัยของ (Zhang และ Horvath, 2005; Langfelder et.al., 2008; Langfelder และ Horvath, 2008) โดยมีวัตถุประสงค์ที่ศึกษาในกรณีอย่างง่าย โดยตัวแปรอิสระอยู่ในรูปแบบเชิงเส้น $(\gamma = 1)$ และนำมาใช้อธิบายได้อย่างสะดวก ยิ่งไปกว่านั้น $\rho_{\max}$ และ $\rho_{\min}$ จะพิจารณาเพียง $\rho_{\max}$ และแทนที่ $\tau = \dfrac{(\rho_{\max} - \rho_{\min})}{g-2}$

หลังจากที่เราสามารถหาค่าความสัมพันธ์ของค่าสังเกตในแถวแรกแล้ว เราหาค่าความสัมพันธ์ของค่าสังเกตที่เหลือ ของเมทริกซ์สหสัมพันธ์ โดยใช้โครงสร้างของ Hub ในการสร้าง สามารถหาได้จากอัลกอริทึม ดังต่อไปนี้

$$\Sigma_k = \begin{bmatrix} 1 & \alpha_{k,2} & \alpha_{k,3} & \alpha_{k,4} & \cdots & \alpha_{k,g_k} \\ \alpha_{k,2} & 1 & \alpha_{k,2} & \alpha_{k,3} & \cdots & \alpha_{k,g_k-1} \\ \alpha_{k,3} & \alpha_{k,2} & 1 & \alpha_{k,2} & \cdots & \alpha_{k,g_k-2} \\ \alpha_{k,4} & \alpha_{k,3} & \alpha_{k,2} & 1 & \cdots & \alpha_{k,g_k-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{k,g_k-1} & \alpha_{k,g_k-1} & \alpha_{k,g_k-2} & \alpha_{k,g_k-3} & \cdots & 1 \end{bmatrix}$$

โดยที่

Σ_k คือ เมทริกซ์สหสัมพันธ์ขนาด $(k \times k)$

k คือ จำนวนกลุ่มการจำแนกหรือจำนวนตัวแปรอิสระ $(k > 0)$

g_k คือ ขนาดของจำนวนกลุ่มการจำแนกหรือขนาดตัวอย่าง $(g_k > 0)$

ρ_k คือ ความสัมพันธ์ขององค์ประกอบภายใน k กลุ่ม หรือความสัมพันธ์ของตัวแปรอิสระ โดยมีค่าอยู่ระหว่าง $0 \le \rho_k \le 1$

$\alpha_{k,1} = 1$ และ $\alpha_{k,i} = \rho_k - \tau_k(i-2)$ เมื่อ $\tau_k = \dfrac{(\rho_{\max} - \rho_{\min})}{g_k - 2}$

ตัวอย่างเช่น มีจำนวนตัวแปรอิสระทั้งหมด 5 ตัวแปร และกำหนดให้ $\rho_k = 0.5$

$$\Sigma = \begin{bmatrix} 1 & 0.95 & 0.8 & 0.65 & 0.5 \\ 0.95 & 1 & 0.95 & 0.8 & 0.65 \\ 0.8 & 0.95 & 1 & 0.95 & 0.8 \\ 0.65 & 0.8 & 0.95 & 1 & 0.95 \\ 0.5 & 0.65 & 0.8 & 0.95 & 1 \end{bmatrix}$$

3.4 การดำเนินการวิจัย

ทำการจำลองข้อมูลโดยใช้โปรแกรม R เวอร์ชั่น 3.3.1

1. ทำการสร้างข้อมูลค่าสังเกต $(y_i, \mathbf{X}_i),\ \ i = 1,...,n$ ของชุดข้อมูลอย่างละ n โดยมีขั้นตอนดังนี้
   1.1 สร้างเมทริกซ์ของตัวแปรอิสระ $(\mathbf{X})$ โดย $\mathbf{X}$ มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนหรือความแปรปรวนร่วม $\mathbf{\Sigma}$ นั่นคือ $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma})$ ซึ่งเมทริกซ์ความแปรปรวนหรือความแปรปรวนร่วม $\mathbf{\Sigma}$ จะถูกกำหนดแตกต่างกันตามรูปแบบของความสัมพันธ์ และตามสถานการณ์ต่างๆ ตามขอบเขตของงานวิจัย โดยรูปแบบความสัมพันธ์ของ $\mathbf{\Sigma}$ แบ่งออกเป็น 3 แบบ ดังนี้
      1.1.1 รูปแบบความสัมพันธ์แบบ Constant โดยสามารถหาได้จาก $\rho_k = r$
      1.1.2 รูปแบบความสัมพันธ์แบบ Toeplitz โดยสามารถหาได้จาก $\rho_k = r^{|i-j|}$
      1.1.3 รูปแบบความสัมพันธ์แบบ Hub Toeplitz โดยสามารถหาได้ ρ_k

      $$\alpha_{k,i} = r_k - \tau_k (i-2) \quad \text{เมื่อ} \quad \tau_k = \frac{(r_{\max} - r_{\min})}{g_k - 2}$$

   1.2 สร้างเวกเตอร์ความคลาดเคลื่อน $\boldsymbol{\varepsilon} = [\varepsilon_1, ..., \varepsilon_n]^T$ โดยที่ ε_i มีการแจกแจงปรกติมาตรฐาน โดยมีค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ 1 สำหรับ $i = 1, 2, ..., n$
   1.3 สร้างเมทริกซ์ค่าเฉลี่ย $(\boldsymbol{\mu})$ จากความสัมพันธ์ $\mu_i = \mu(\mathbf{X}_i, \boldsymbol{\beta}) = \exp(\mathbf{X}_i^T \boldsymbol{\beta})$
   1.4 สร้างเวกเตอร์ตัวแปรตาม $(\mathbf{y})$ ที่มีการแจกแจงปัวซง โดยมีเวกเตอร์ค่าเฉลี่ยเท่ากับ $\boldsymbol{\mu}$ ซึ่งสามารถหาได้จากข้อ 1.2 และ 1.3 ตามลำดับ จากตัวแบบ $y_i = \exp(\mathbf{X}_i^T \boldsymbol{\beta}) + \varepsilon_i$
2. นำ $\mathbf{X}$ และ $\mathbf{y}$ จากชุดข้อมูล สามารถหา tuning parameter (λ) โดยวิธี V-fold cross–validation โดยใช้ package glmnet ในโปรแกรม R ซึ่งมีขั้นตอน ดังนี้
   2.1 ทำการแปลงตัวอย่างสุ่ม $\mathbf{X}_i$ จากชุดข้อมูลให้เป็นค่ามาตรฐาน (Standardize)
   2.2 แบ่งข้อมูลออกเป็น 5 ส่วน $(V=5)$ ถ้าให้ ข้อมูลทั้งหมดแทนด้วย T จะได้ T_1, T_2, T_3, T_4, T_5 ให้หนึ่งส่วนเป็น test set (T_v) และอีก 4 ส่วนที่เป็น training set $(T - T_v)$ เมื่อ $v = 1, 2, ..., 5$
   2.3 กำหนดค่าพารามิเตอร์ปรับแต่ง (Tuning Parameter) ของวิธีการประมาณค่าแต่ละวิธี นั่นคือ $\lambda \geq 0$ เมื่อ λ เป็นค่าคงใดๆ โดยที่ หา $\hat{\boldsymbol{\beta}}$ ที่สอดคล้องกับเงื่อนไขของแต่ละวิธีการหาตัวประมาณ และทำให้ $\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left(-l(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1^2 \right)$ ทำเช่นนี้ในทุกๆ ส่วนย่อยของ training set $(T - T_v)$ เมื่อ $v = 1, 2, ..., 5$

2.4 เมื่อได้ค่าประมาณ $\hat{\boldsymbol{\beta}}$ ที่คำนวณได้จากแต่ละ training set ย่อย ไปหาค่าพยากรณ์ของ $\mathbf{y}$ ใน test set $\left(T_v\right)$ เมื่อ $v = 1,2,...,5$ จากสมการ $\hat{\mathbf{y}} = \exp\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)$ แล้วคำนวณค่า $\text{PMSE}_v = \sum_{i=1}^{n} \frac{\left(y_i - y_i\right)^2}{n_\lambda}$ โดยที่ n_λ เป็นจำนวนตัวอย่างของ $\left(T_v\right)$ จาก $y_i \in T_v$

2.5 หาค่าเฉลี่ย $\text{PMSE}_v = \frac{1}{5}\sum_{v=1}^{5} \text{PMSE}_v$ ที่ λ จุดนั้น

2.6 เลือกค่า λ ที่ให้ค่าเฉลี่ยของ PMSE ต่ำสุด ค่า λ ค่านั้น ก็จะเป็น tuning parameter

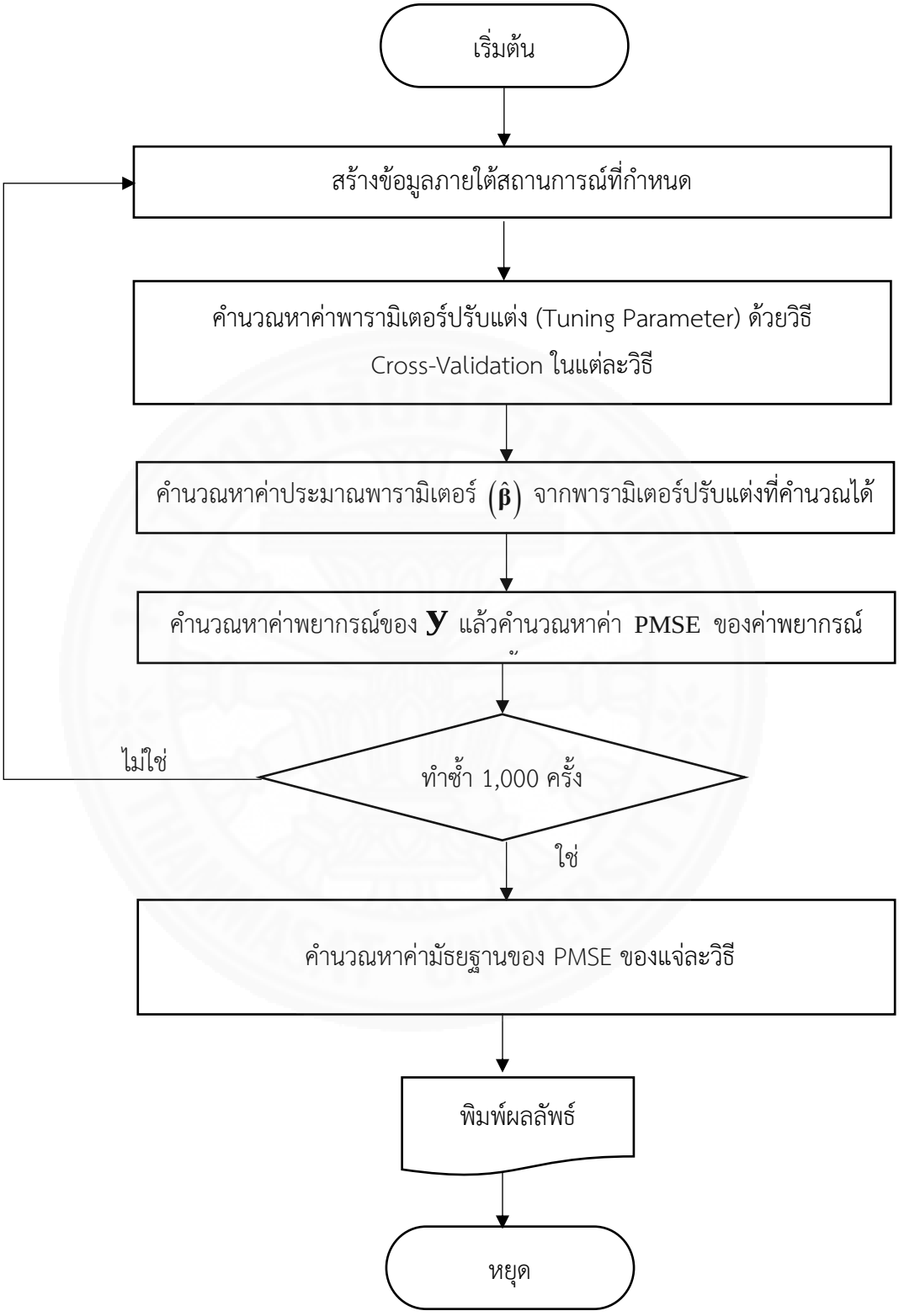
3. หาตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\boldsymbol{\beta}}$ จากชุดข้อมูล ของวิธีการประมาณแต่ละวิธี ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO) โดยใช้ค่า λ ที่ได้จากข้อ 2.6 มาใช้
4. หลังจากได้ตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\boldsymbol{\beta}}$ นำค่า $\hat{\boldsymbol{\beta}}$ ที่ได้ไปหาค่าพยากรณ์ของ $\mathbf{y}$ ที่อยู่ในชุดข้อมูล จากสมการถดถอย $\hat{\mathbf{y}} = \exp\left(\mathbf{X}\hat{\boldsymbol{\beta}}\right)$ แล้วคำนวณค่า PMSE จากชุดข้อมูล
5. คำนวณหาความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ โดยนับจำนวนตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ในเวกเตอร์ $\hat{\boldsymbol{\beta}}$ จากข้อ 3 ที่มีค่าเท่ากับ 0 แต่ค่าพารามิเตอร์ β_j ไม่เท่ากับ 0 (Identify Criterion 1: IC1) และนับจำนวนตัวประมาณสัมประสิทธิ์การถดถอย $\hat{\beta}_j$ ที่มีค่าไม่เท่ากับ 0 แต่ค่าพารามิเตอร์ β_j เท่ากับ 0 (Identify Criterion 2: IC2)
6. ทำซ้ำขั้นตอน 1-5 จำนวน 1,000 ครั้ง
7. หาค่ามัธยฐานของค่า PMSE และความน่าจะเป็นของ IC1 และ IC2 จากผลการทดลองทั้งหมด 1,000 ครั้ง
8. เปรียบเทียบค่าที่คำนวณได้จากขั้นตอนที่ 8 ของวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยทั้ง 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO)
9. สรุปผลการวิจัย

3.5 แผนการจำลองข้อมูล

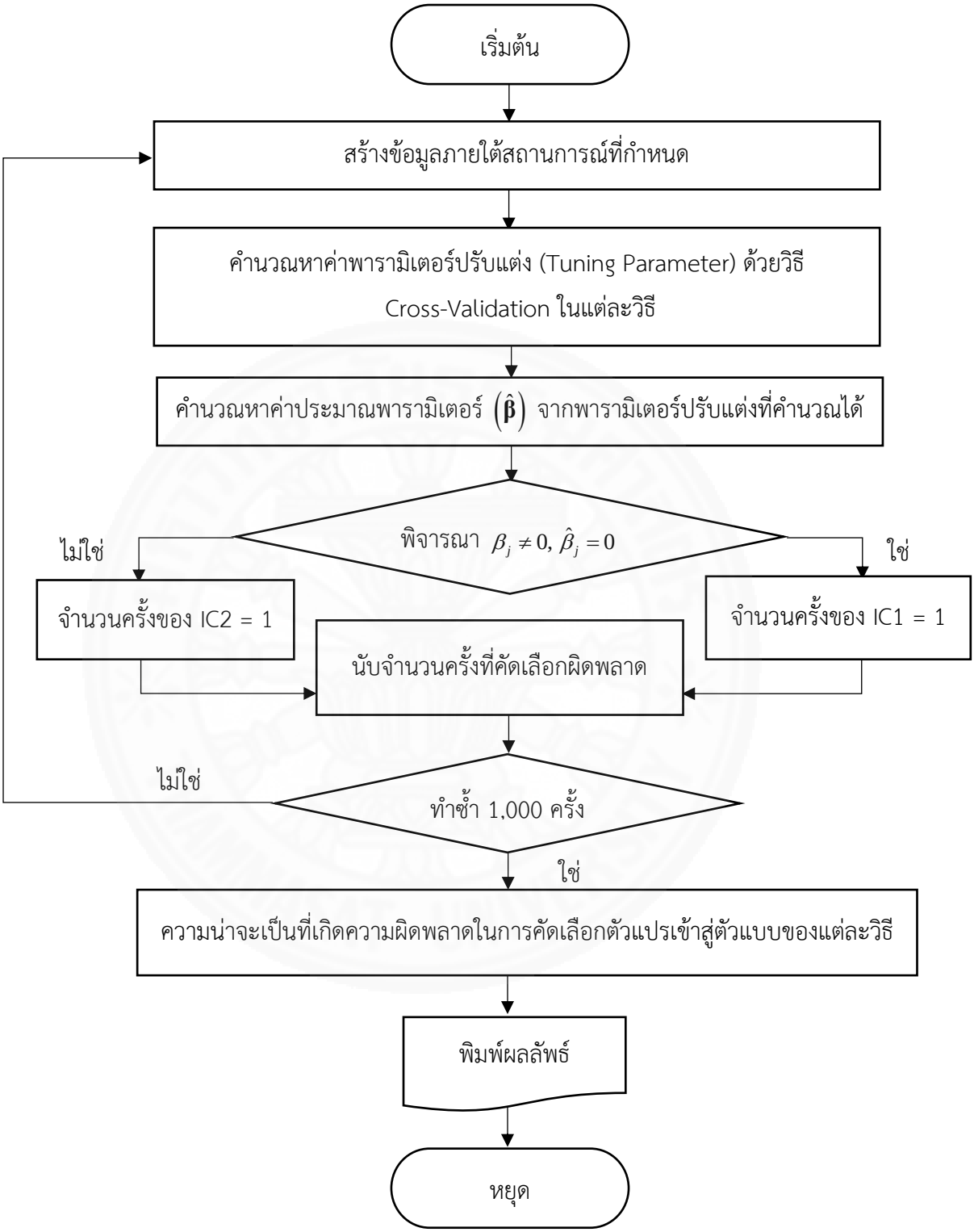
การวิจัยนี้เป็นการวิจัยเชิงทดลอง โดยวิธีการจำลอง (Simulation Study) ซึ่งการวิจัยครั้งนี้ได้จำลอง (Simulation) ข้อมูลด้วยโปรแกรม R เวอร์ชั่น 3.1.2 เพื่อคำนวณค่าความคลาดเคลื่อนจากวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์ และความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ ของการถดถอยทั้ง 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์(Ridge) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO) โดยแบ่งการทำงานออกเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 เปรียบเทียบประสิทธิภาพการพยากรณ์จากวิธีการวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์ของทั้ง 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง

ส่วนที่ 2 เปรียบเทียบความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรเข้าสู่ตัวแบบของวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์ของทั้ง 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง



ภาพที่ 3.1 เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี



ภาพที่ 3.2 เปรียบเทียบความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO

บทที่ 4
ผลการวิจัยและอภิปรายผล

ในส่วนนี้จะนำเสนอผลการดำเนินการวิจัย โดยแบ่งการศึกษาออกเป็น 2 ส่วน เพื่อเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์ในแต่ละวิธี ได้แก่

1. ประสิทธิภาพของการพยากรณ์ โดยเปรียบเทียบวิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง โดยแบ่งตัวแปรอิสระออกเป็น 1 กลุ่ม และ 2 กลุ่ม และตัวแปรอิสระมีความสัมพันธ์ในรูปแบบ Constant model, Toeplitz model และ Hub Toeplitz model ในแต่ละระดับความสัมพันธ์ต่างๆ $r = 0.5, ..., 0.9$

2. ประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ โดยเปรียบเทียบวิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง เท่านั้น และแบ่งลักษณะของตัวแปรอิสระเช่นเดียวกับการเปรียบเทียบประสิทธิภาพของการพยากรณ์

และศึกษาทั้งในการจำลองสถานการณ์ (Simulation) และข้อมูลจริง (Real Data) มาประยุกต์ใช้ โดยจะจำลองสถานการณ์ต่างๆ ดังต่อไปนี้

1) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่แต่ละระดับความสัมพันธ์ต่างๆ

2) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่แต่ละระดับความสัมพันธ์ต่างๆ

3) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่แต่ละระดับความสัมพันธ์ต่างๆ

4) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่แต่ละระดับความสัมพันธ์ต่างๆ

5) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่แต่ละระดับความสัมพันธ์ต่างๆ

6) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่แต่ละระดับความสัมพันธ์ต่างๆ

4.1 ผลลัพธ์จากการจำลองสถานการณ์

4.1.1 ประสิทธิภาพของการพยากรณ์ในแต่ละวิธี

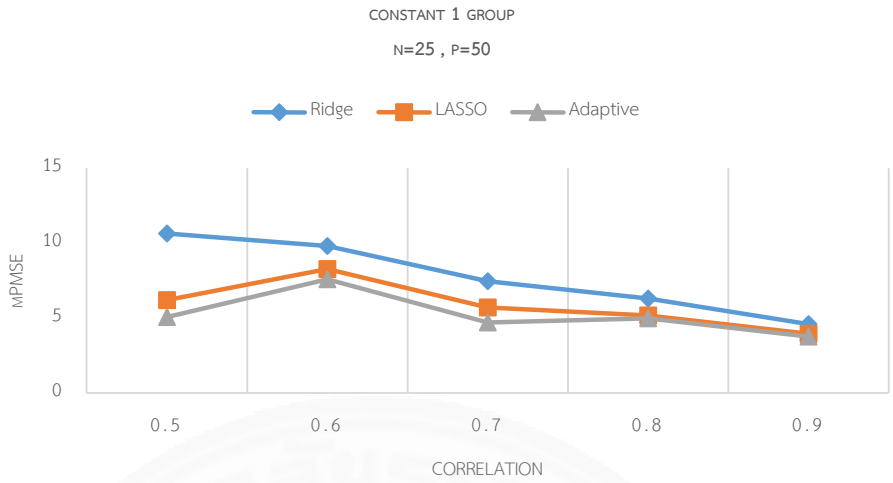
เนื่องจากรูปแบบความสัมพันธ์แบบ Constant ที่ค่าความสัมพันธ์ของตัวแปรอิสระเป็นค่าคงที่ และรูปแบบความสัมพันธ์แบบ Toeplitz และ Hub Toeplitz ได้ถูกนำมาใช้อย่างแพร่หลายในการวิเคราะห์การจำแนกประเภท (Discriminant Analysis) หรือใช้ในการวิเคราะห์เกี่ยวกับข้อมูลอนุกรมเวลา โดยแบ่งตัวแปรอิสระในการศึกษาออกเป็น 2 ลักษณะ นั่นคือ 1 กลุ่ม และ 2 กลุ่ม ตามลำดับ โดยพิจารณาจากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) เพื่อดูประสิทธิภาพในการพยากรณ์ของวิธีการวิเคราะห์การถดถอยทั้ง 3 วิธี ในลักษณะข้อมูลที่มีมิติสูงแบบบางเบา ว่ามีความแตกต่างกันหรือไม่ โดยศึกษาจากจำนวนข้อมูลและจำนวนตัวแปรอิสระที่แตกต่างกัน รวมไปถึงศึกษาในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูงในหลายระดับดังต่อไปนี้ $r = 0.5, 0.6, 0.7, 0.8$ และ $r = 0.9$ ซึ่งวิธีที่มีประสิทธิภาพสูงที่สุด จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์น้อยที่สุด

1) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่ระดับความสัมพันธ์ต่างๆ

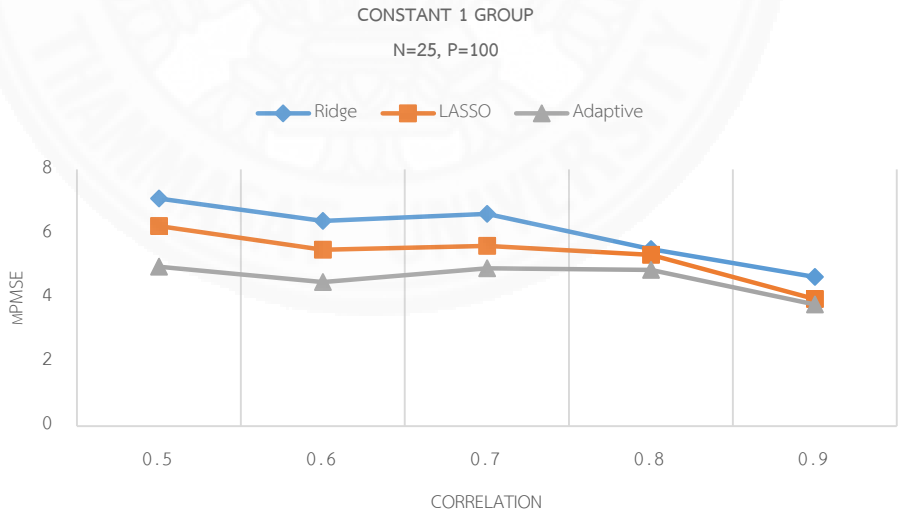
ตารางที่ 4.1 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Constant

		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
r	p		$n=25$			$n=50$	
0.5	50	10.6524	6.189153	5.076029*	23.89061	14.69172	13.18459*
	100	7.086174	6.230699	4.965397*	10.47410	8.08080	6.840008*
	200	11.21742	10.12269	7.63458*	13.36298	10.86931	8.42348*
0.6	50	9.80650	8.26983	7.57658*	15.07056	11.30258	10.61303*
	100	6.39466	5.49901	4.48624*	8.37779	6.92433	5.51997*
	200	11.15964	9.37587	7.03877*	12.90637	11.90031	8.94083*
0.7	50	7.45606	5.69710	4.69269*	12.66095	9.72264	9.82975*
	100	6.61012	5.60962	4.91888*	8.33500	6.60421	5.65492*
	200	6.44326	5.57225	4.78896*	8.90323	7.79999	6.37740*
0.8	50	6.31462	5.16679	4.98767*	11.63907	8.37228	8.12134*
	100	5.516635	5.339169	4.864279*	13.00762	9.97526	9.38871*
	200	3.45510	2.99071	2.71107*	9.47735	7.84162	6.93040*
0.9	50	4.59058	3.95301	3.76701*	7.18089	5.97884	6.07353*
	100	4.64637	3.95966	3.78637*	9.28073	7.48053	7.17576*
	200	4.33225	3.65982	3.53580*	6.66773	5.65949	5.26692*

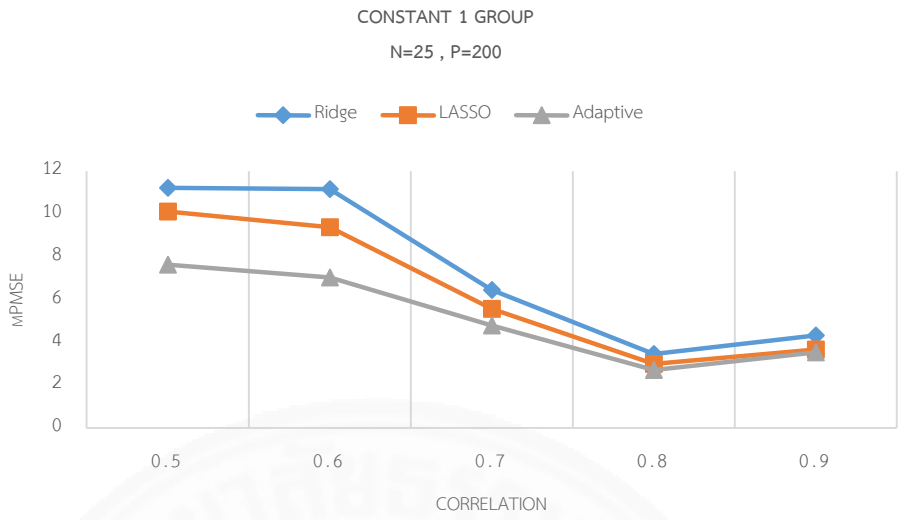
หมายเหตุ * แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



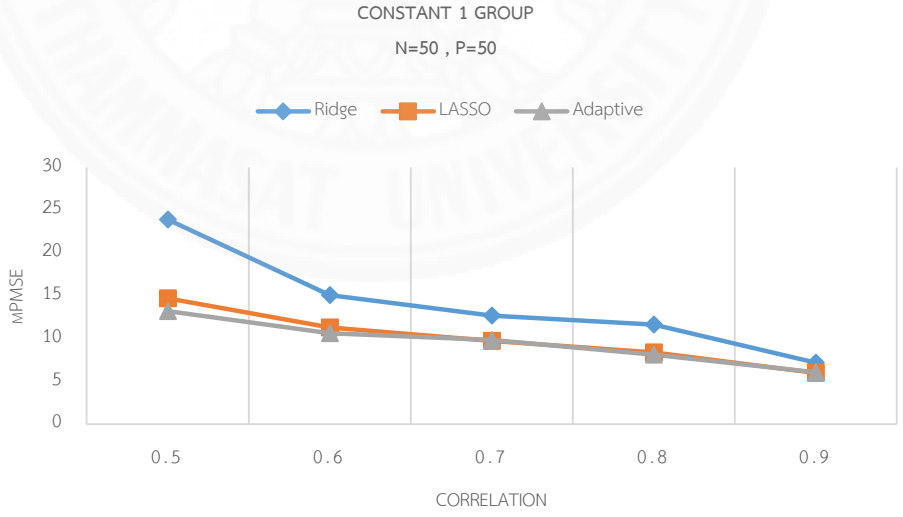
ภาพที่ 4.1 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 50$



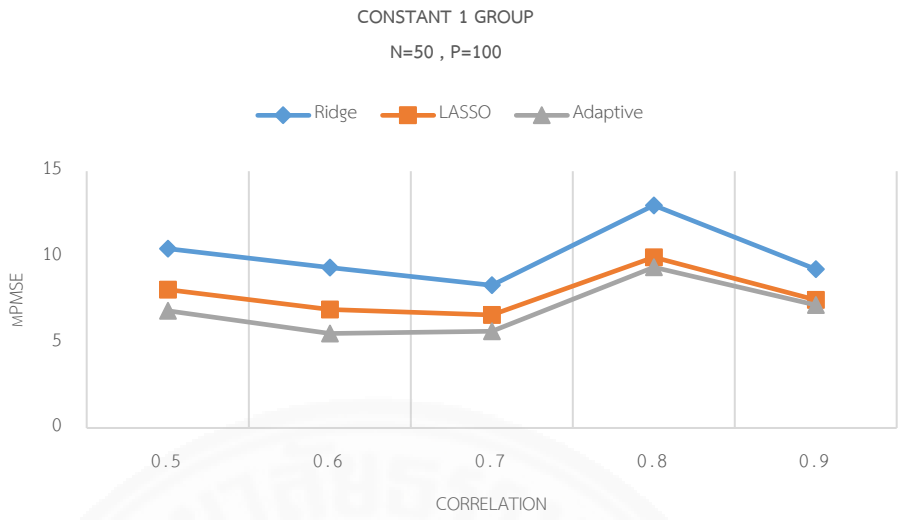
ภาพที่ 4.2 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 100$



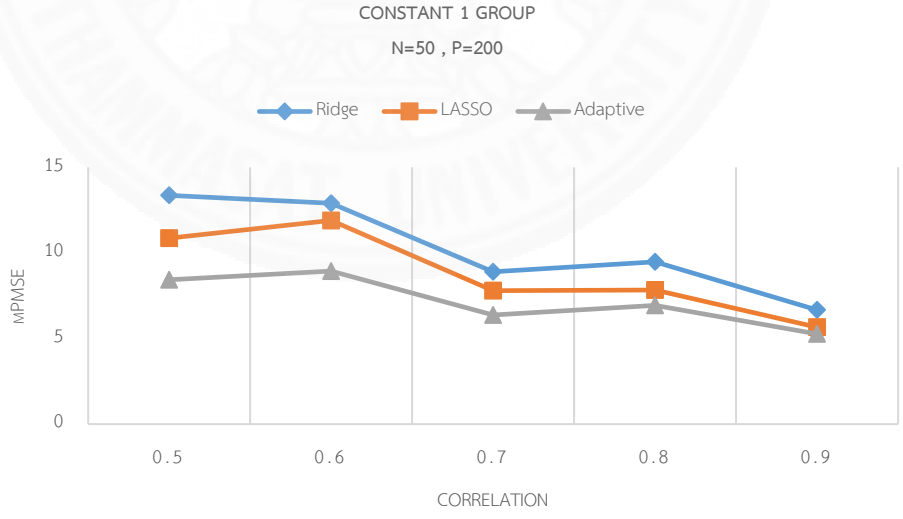
ภาพที่ 4.3 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 200$



ภาพที่ 4.4 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 50$



ภาพที่ 4.5 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 100$



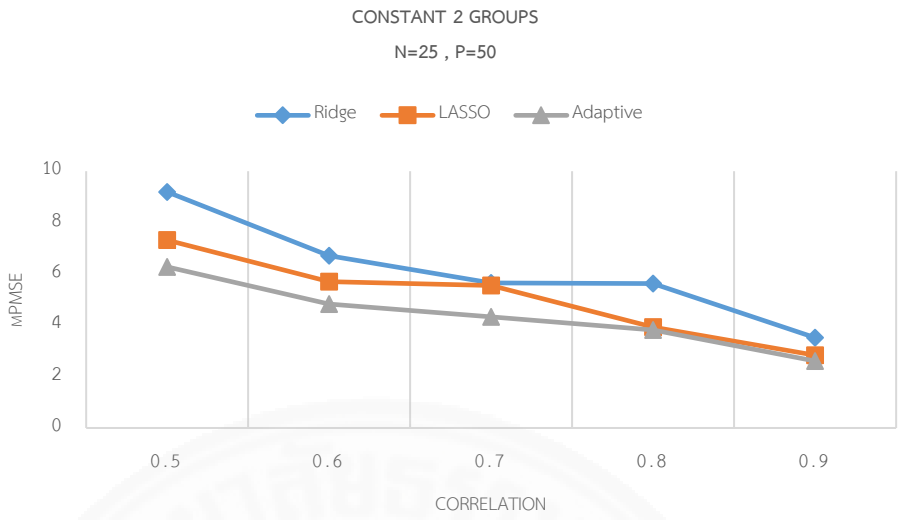
ภาพที่ 4.6 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 200$

2) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่ระดับความสัมพันธ์ต่างๆ

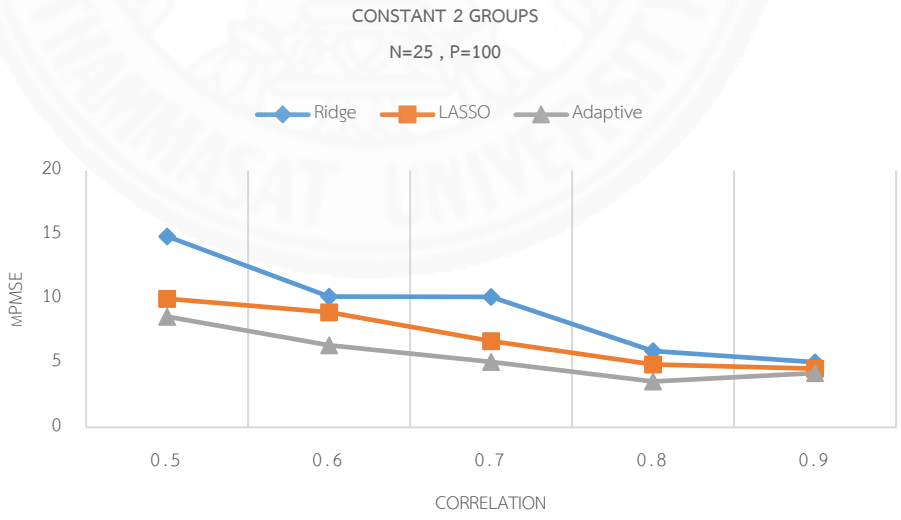
ตารางที่ 4.2 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Constant

		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
r	p		$n=25$			$n=50$	
0.5	50	9.193113	7.315495	6.273284*	17.0052	10.86592	8.985308*
	100	14.87306	10.01593	8.624981*	10.63253	8.413216	6.737952*
	200	10.44676	9.60953	6.869651*	11.62968	9.861743	7.209156*
0.6	50	6.713646	5.69325	4.827821*	11.36982	8.003056	6.918903*
	100	10.17809	8.974986	6.389409*	12.81477	10.50288	8.412309*
	200	7.207811	6.612675	4.877901*	11.34478	9.782164	7.938626*
0.7	50	5.649507	5.548511	4.333519*	8.210524	6.548858	6.604283*
	100	10.15141	6.719304	5.090199*	8.666649	7.805727	6.776083*
	200	6.105768	5.026212	4.441009*	9.097253	9.02425	6.828234*
0.8	50	5.61927	3.922529	3.813159*	8.01197	6.352199	6.341005*
	100	5.942702	4.889642	3.57325*	8.128593	6.376282	6.201446*
	200	6.080402	5.015287	3.7425*	7.841283	6.704962	4.693072*
0.9	50	3.516828	2.832158	2.606889*	7.983044	6.146937	5.742676*
	100	5.072896	4.570755	4.219117*	7.369058	6.107473	5.450372*
	200	4.42025	3.665096	3.500912*	4.897087	3.461501	3.307722*

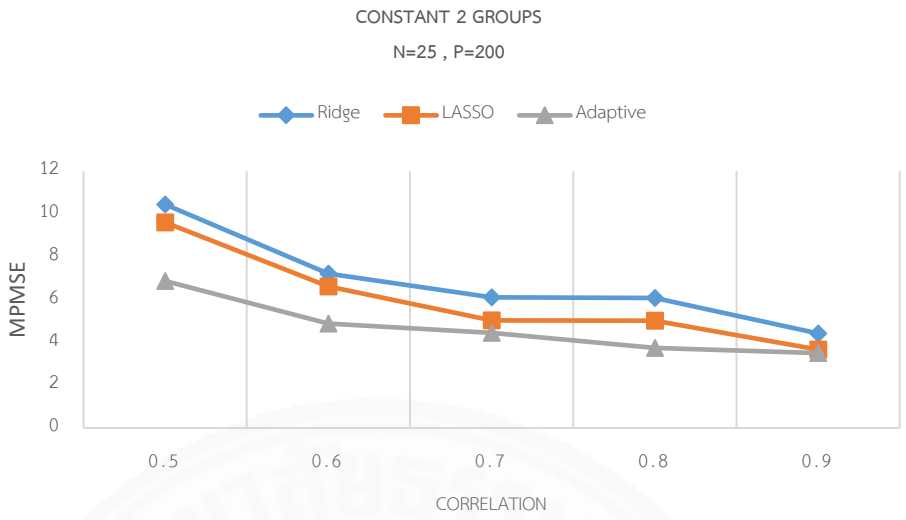
หมายเหตุ ***** แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



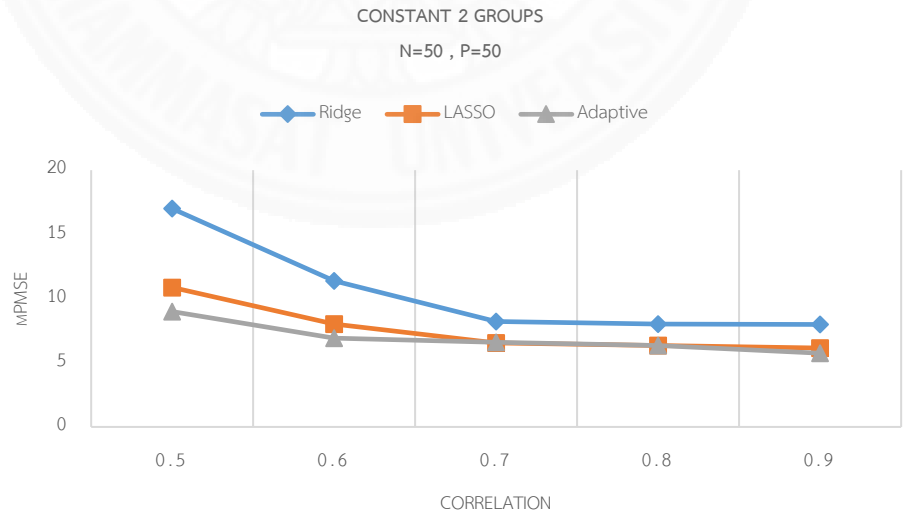
ภาพที่ 4.7 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 50$



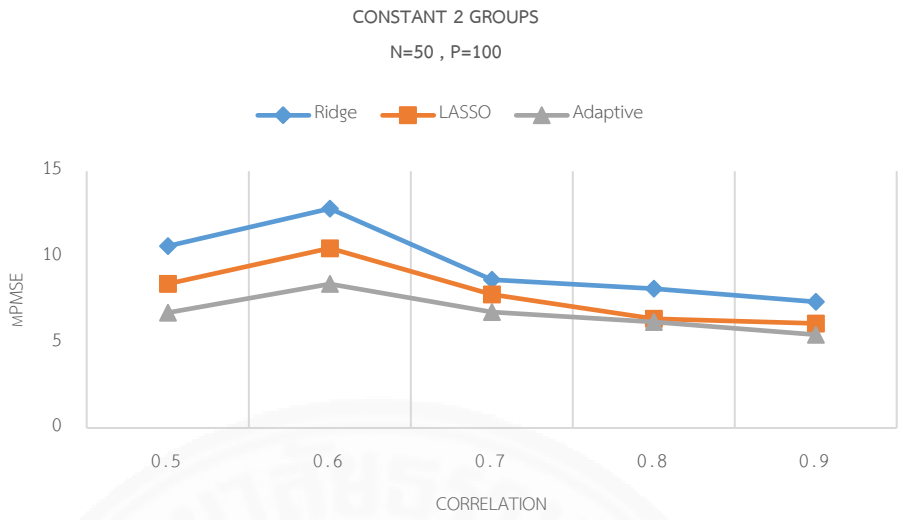
ภาพที่ 4.8 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 100$



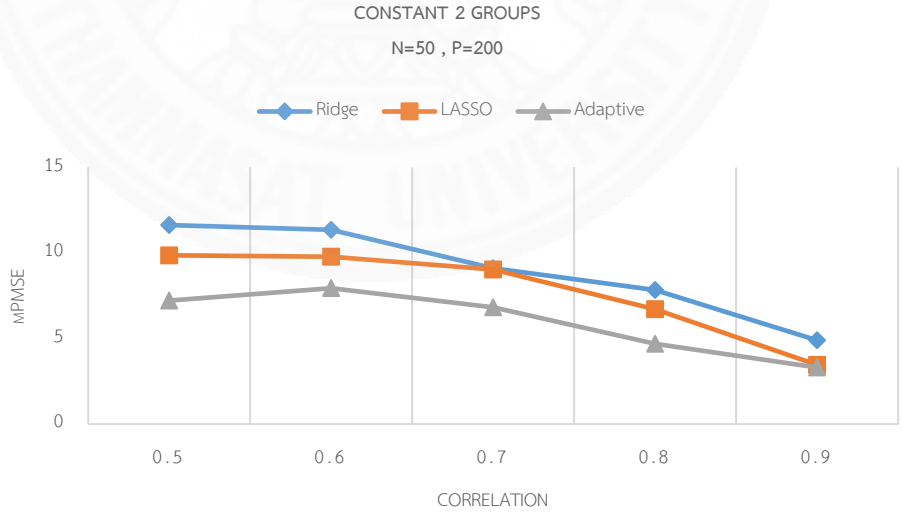
ภาพที่ 4.9 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 200$



ภาพที่ 4.10 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธแบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 50$



ภาพที่ 4.11 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 100$



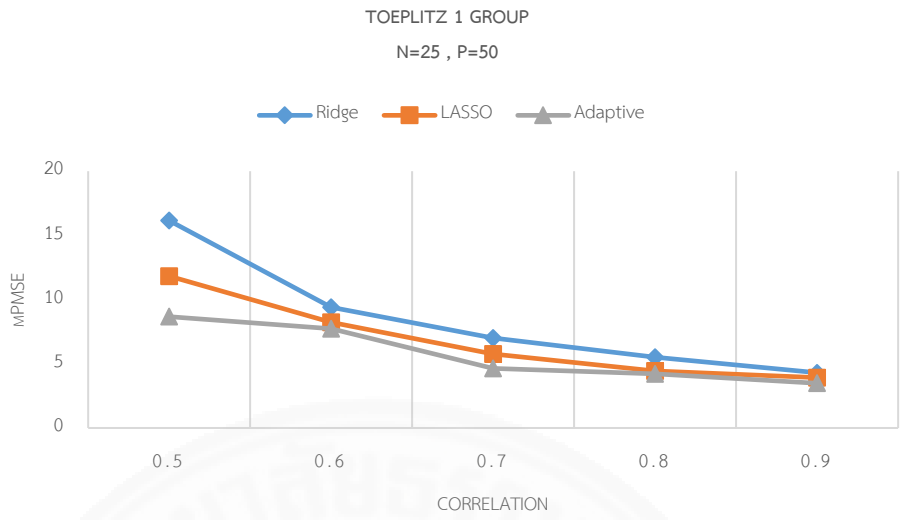
ภาพที่ 4.12 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Constant และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 200$

3) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

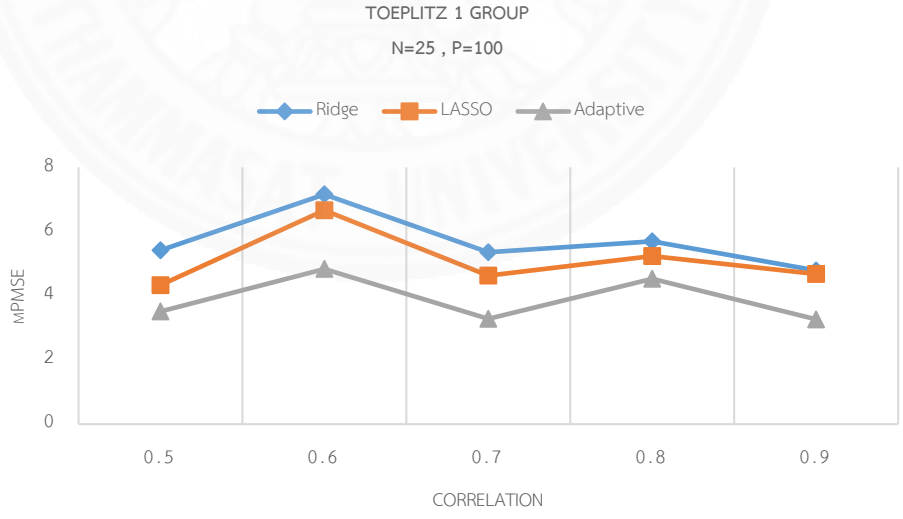
ตารางที่ 4.3 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Toeplitz

		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
r	p		$n=25$			$n=50$	
0.5	50	16.15398	11.82943	8.670539*	29.17474	20.41302	14.77027*
	100	5.411574	4.330981	3.507906*	21.0602	15.97994	12.22886*
	200	8.194067	8.51634	5.741609*	20.8608	17.23301	12.59359*
0.6	50	9.404797	8.2385	7.731062*	11.12831	8.843067	8.115946*
	100	7.156612	6.662911	4.832732*	13.03712	9.175676	7.816185*
	200	7.575133	7.6305	5.468474*	13.0495	13.77537	9.601196*
0.7	50	7.019531	5.752669	4.630747*	7.534662	6.704349	4.632464*
	100	5.352946	4.624005	3.279594*	7.877103	7.321964	5.105933*
	200	4.112259	3.702483	3.025093*	9.166551	8.427082	6.591876*
0.8	50	5.524409	4.449157	4.204694*	6.486302	5.884532	5.433853*
	100	5.69925	5.235287	4.528435*	7.395702	6.749447	5.683808*
	200	7.441	6.440762	5.798644*	10.36842	9.454582	7.000082*
0.9	50	4.303445	3.923764	3.483195*	5.583285	5.517457	4.928481*
	100	4.786392	4.68125	3.264642*	6.60735	6.140578	5.245554*
	200	5.65511	5.745	4.40743*	7.230346	6.871357	5.275375*

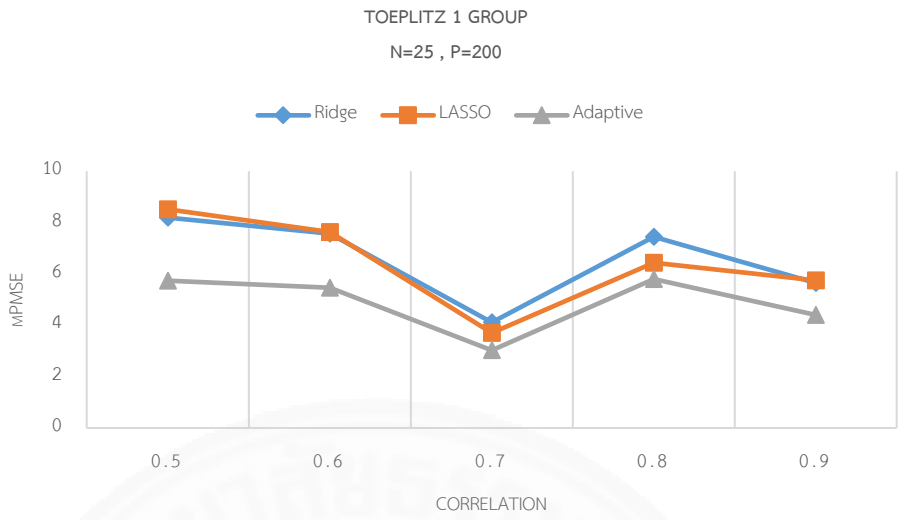
หมายเหตุ ***** แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



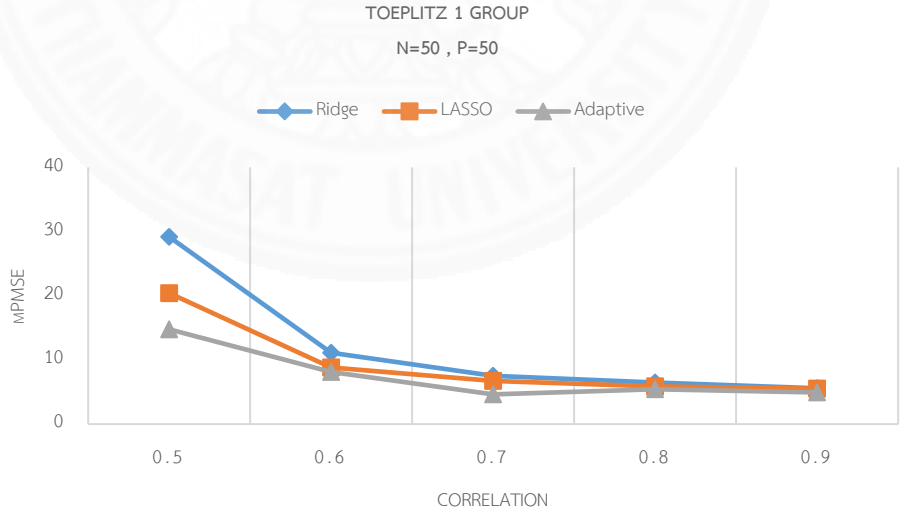
ภาพที่ 4.13 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 50$



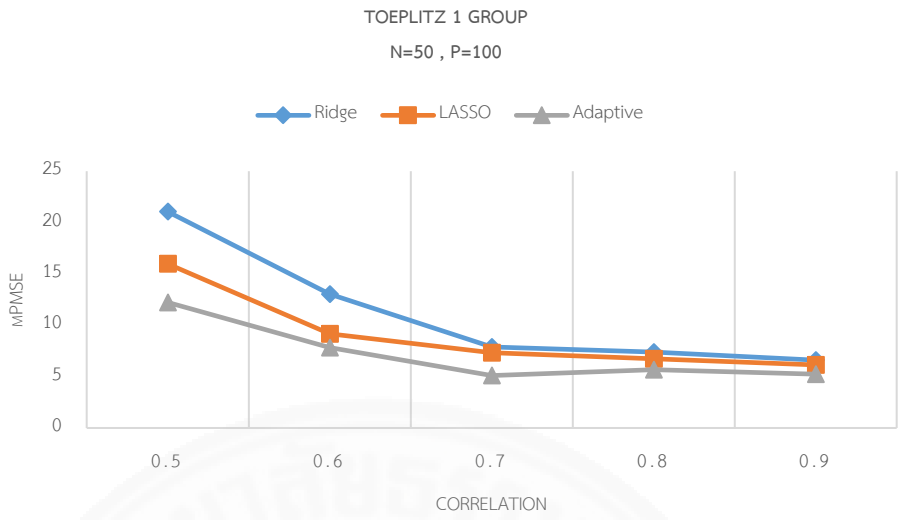
ภาพที่ 4.14 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 100$



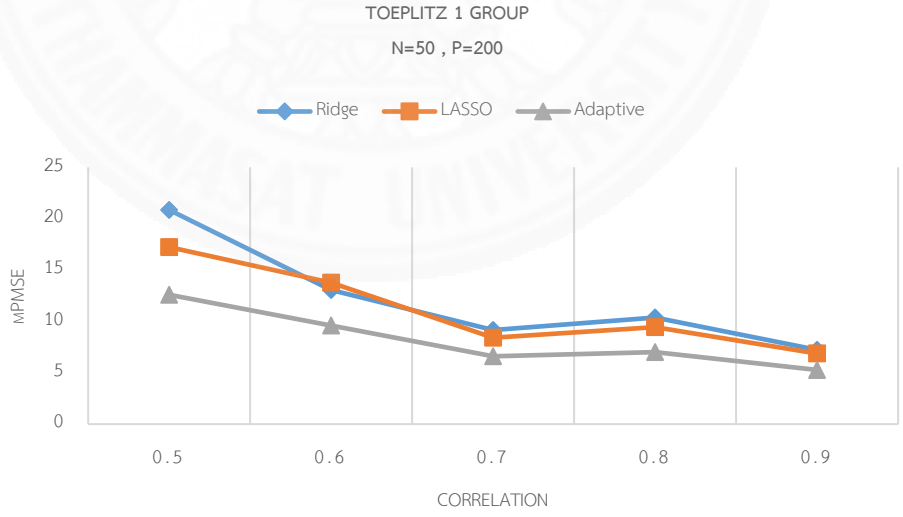
ภาพที่ 4.15 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 200$



ภาพที่ 4.16 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 50$



ภาพที่ 4.17 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 100$



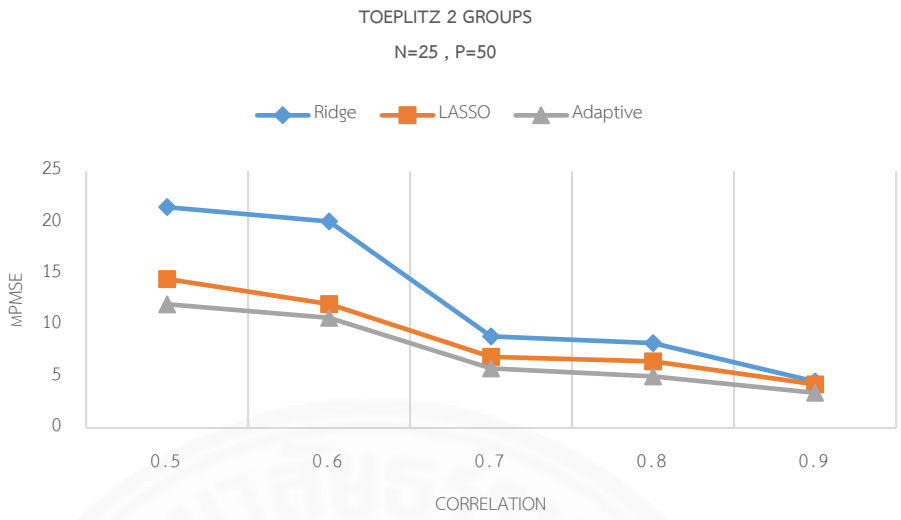
ภาพที่ 4.18 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 200$

4) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

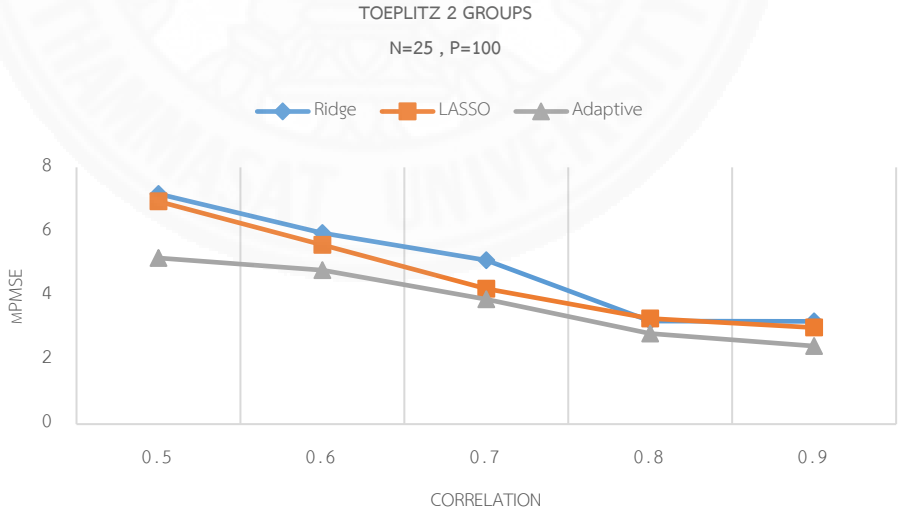
ตารางที่ 4.4 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Toeplitz

		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
r	p		$n=25$			$n=50$	
0.5	50	21.49514	14.50181	12.04918*	15.70018	10.45752	9.403252*
	100	7.173	6.938153	5.175408*	16.843	10.79561	8.288666*
	200	8.346166	9.250106	5.660197*	19.1793	16.3648	10.47843*
0.6	50	20.09852	12.05908	10.72716*	12.3677	10.87173	9.680359*
	100	5.950882	5.582327	4.792026*	17.27558	13.61563	9.493351*
	200	7.4207	7.527739	3.179241*	17.96085	15.93033	11.75516*
0.7	50	8.92925	6.911095	5.795728*	10.60313	9.869713	7.721668*
	100	5.101996	4.219367	3.886607*	9.123451	8.015267	6.180892*
	200	6.529189	7.118164	4.412802*	11.5711	12.21688	7.479142*
0.8	50	8.26525	6.484199	5.013742*	6.569777	6.098478	5.47677*
	100	3.214252	3.288736	2.826307*	8.393946	7.906238	6.23852*
	200	6.309313	6.8925	4.813641*	6.604733	7.432	5.406157*
0.9	50	4.524129	4.248	3.42357*	6.22906	5.248282	4.718587*
	100	3.201618	3.007115	2.431623*	6.146114	5.65	4.329029*
	200	5.307936	4.83777	4.139178*	6.549703	6.637454	5.035816*

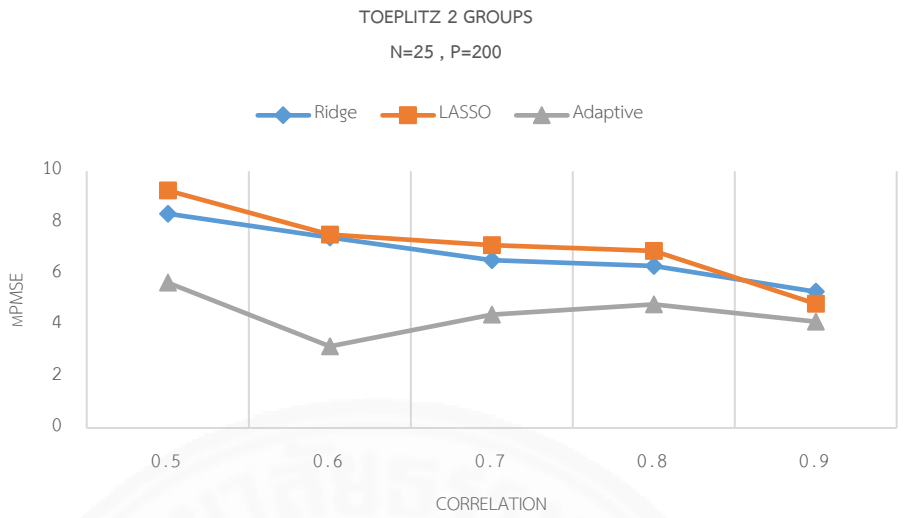
หมายเหตุ ***** แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



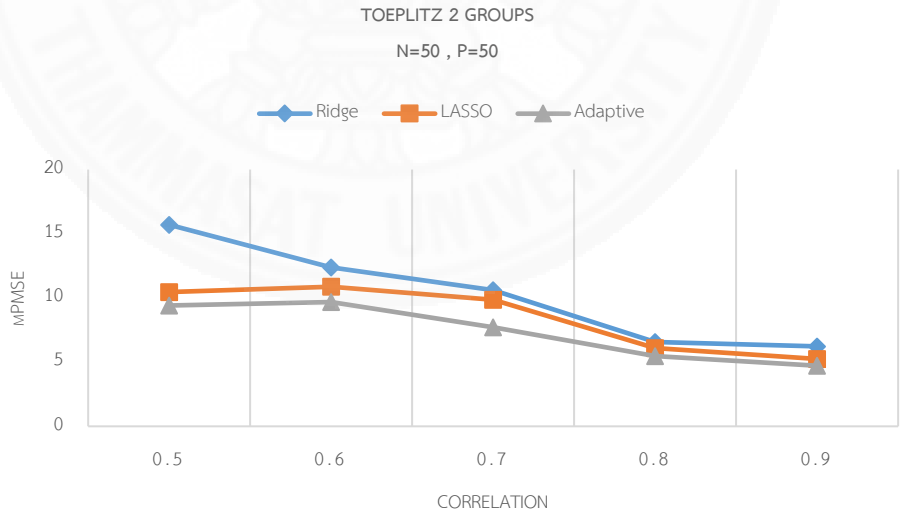
ภาพที่ 4.19 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 50$



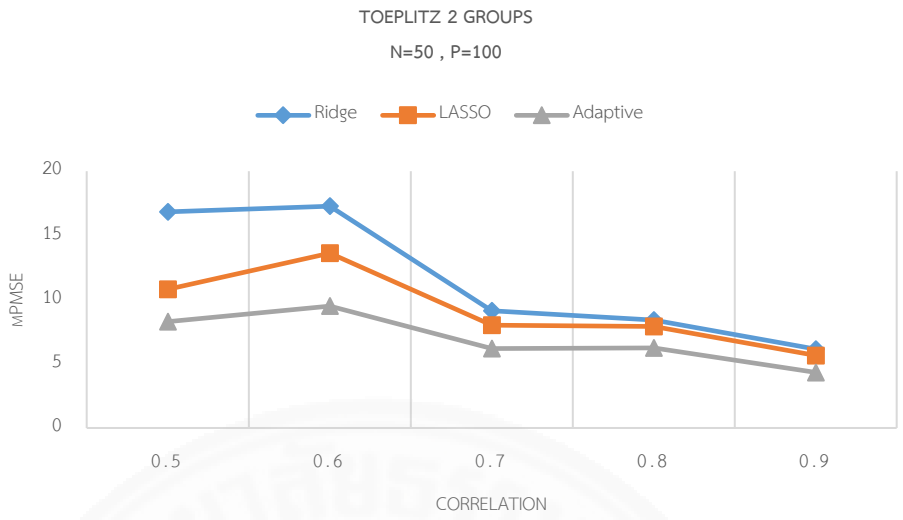
ภาพที่ 4.20 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 100$



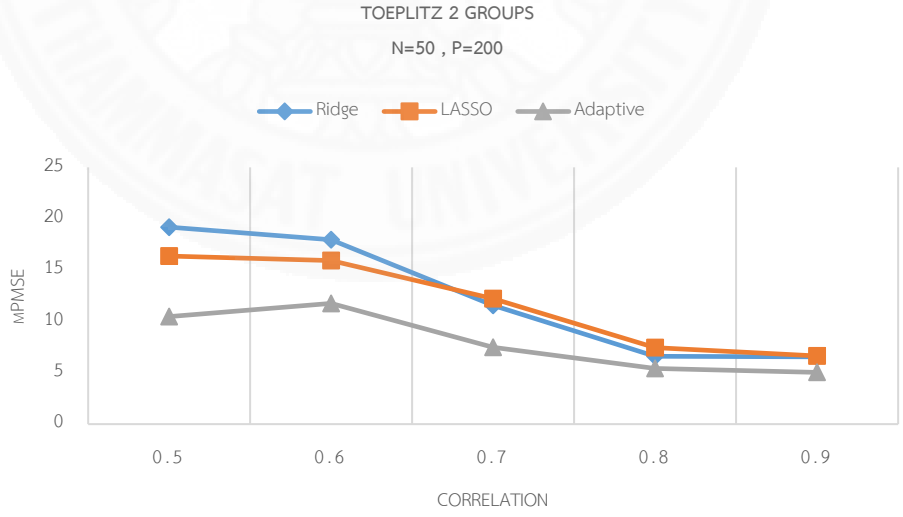
ภาพที่ 4.21 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25, p = 200$



ภาพที่ 4.22 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 50$



ภาพที่ 4.23 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 100$



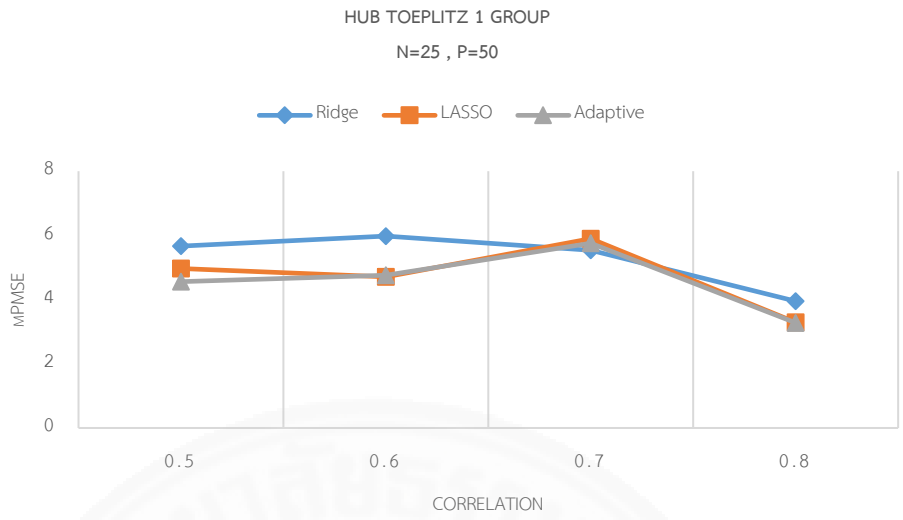
ภาพที่ 4.24 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50, p = 200$

5) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

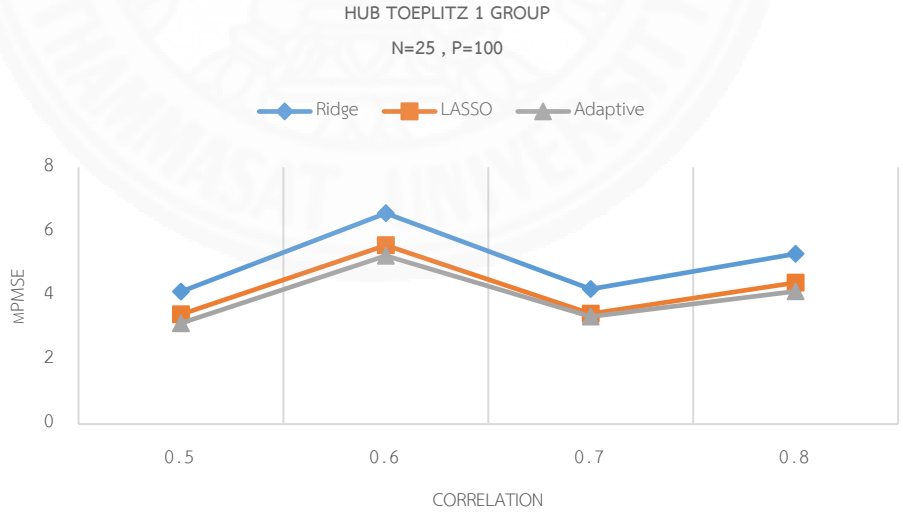
ตารางที่ 4.5 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Hub Toeplitz ที่มีความสัมพันธ์สูงสุด คือ $r_{\max} = 0.9$

r	p	Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
			$n = 25$			$n = 50$	
0.5	50	5.663284	4.969779	4.558758*	7.452506	6.105046	6.224579*
	100	4.126709	3.41684	3.144514*	7.397727	7.853712	7.709643*
	200	5.339324	5.004806	4.378678*	7.572047	6.458922	5.999585*
0.6	50	5.97937	4.705151	4.762273*	6.465611	5.439784	5.305445*
	100	6.569274	5.571595	5.246697*	7.162625	6.239826	5.838298*
	200	4.57	3.952026	3.590337*	7.316807	6.498869	6.003882*
0.7	50	5.539403	5.898214	5.755828*	8.221824	6.855845	6.994952*
	100	4.205279	3.436321	3.351131*	7.709394	6.455817	6.303988*
	200	5.276949	5.012136	4.332429*	6.274708	5.312827	5.147458*
0.8	50	3.950863	3.293	3.27679*	7.220565	6.26025	6.280313*
	100	5.309332	4.41025	4.140341*	7.167214	6.072605	5.916763*
	200	5.010201	4.32625	3.963443*	4.577	4.076177	3.898372*

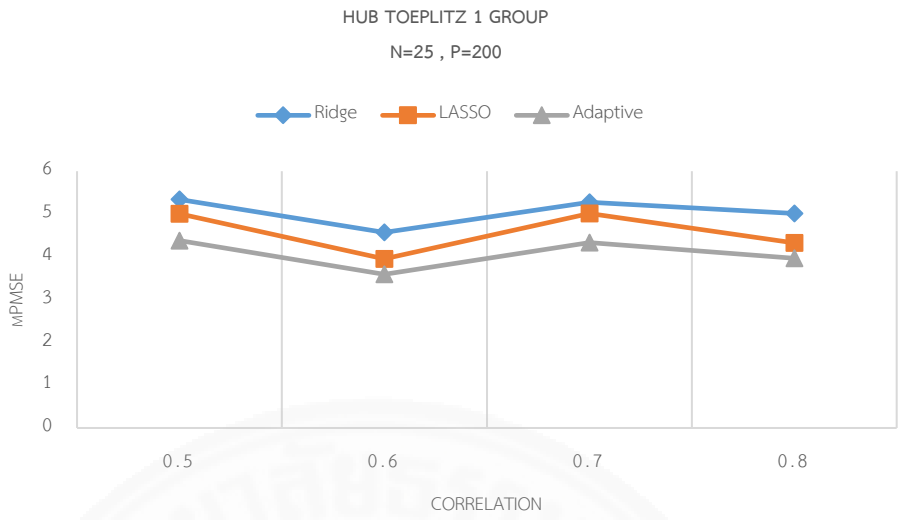
หมายเหตุ ***** แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



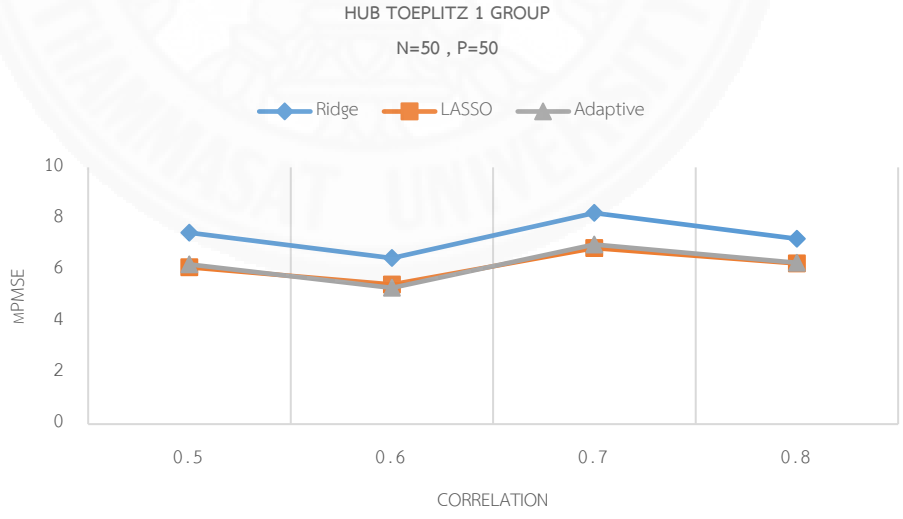
ภาพที่ 4.25 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 50$



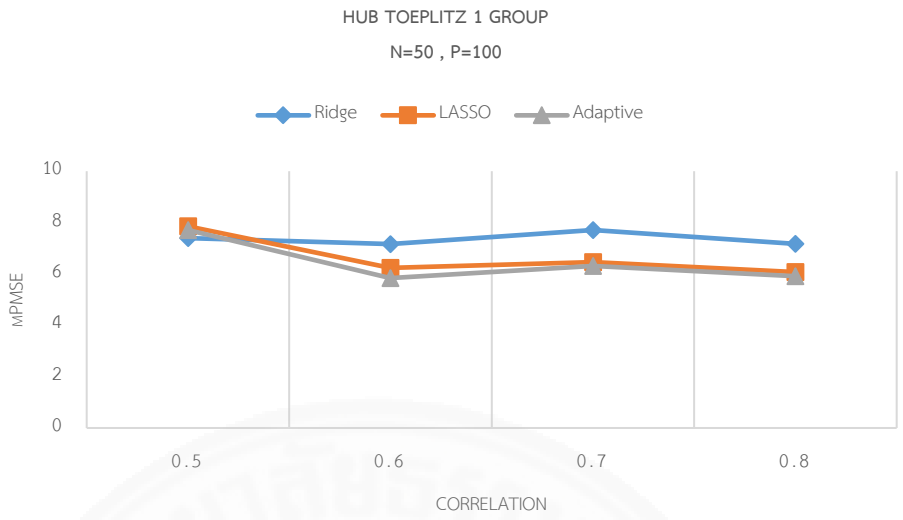
ภาพที่ 4.26 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 100$



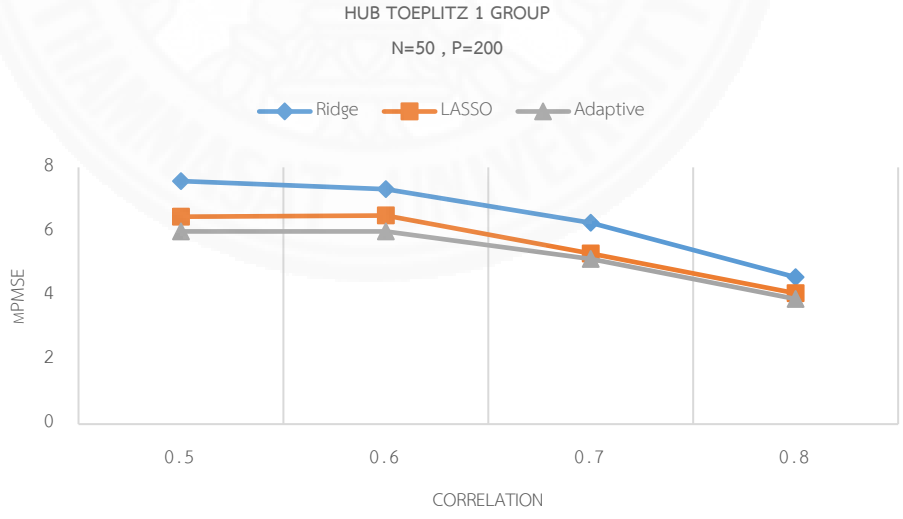
ภาพที่ 4.27 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 200$



ภาพที่ 4.28 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 50$



ภาพที่ 4.29 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 100$



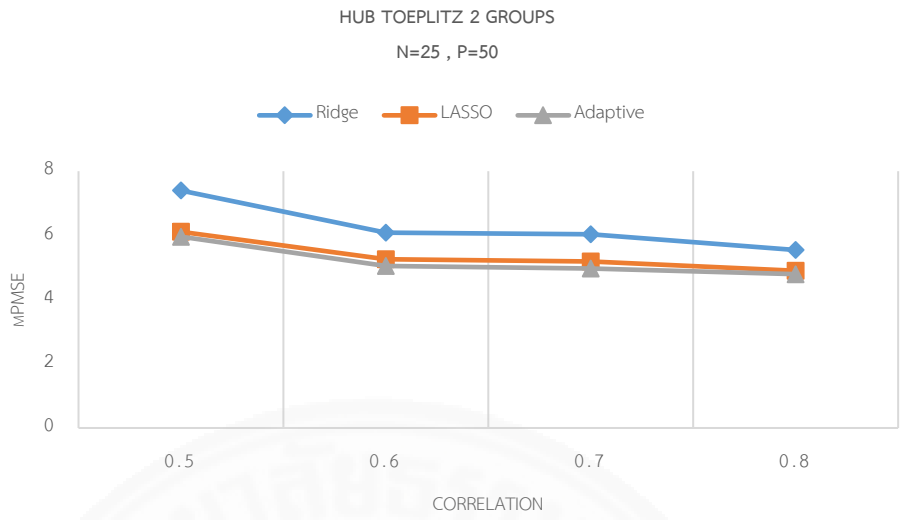
ภาพที่ 4.30 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 200$

6) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

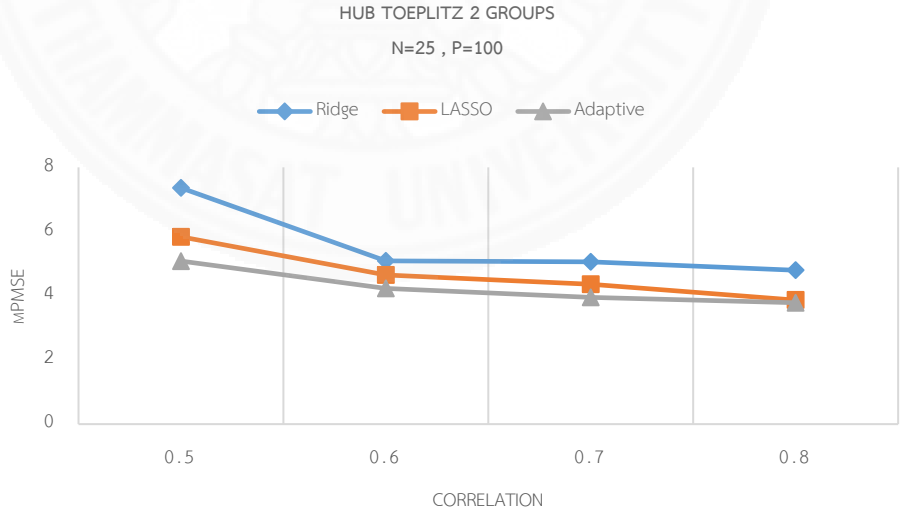
ตารางที่ 4.6 ค่ามัธยฐานของ PMSE ของแต่ละวิธี เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Hub Toeplitz ที่มีความสัมพันธ์สูงสุด คือ $r_{\max} = 0.9$

r	p	Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
			$n = 25$			$n = 50$	
0.5	50	7.39815	6.111601	5.959504*	8.662857	7.574599	7.28274*
	100	7.36089	5.836827	5.079702*	6.9151	6.382655	5.860514*
	200	6.683906	5.97525	5.55325*	7.392311	6.585958	5.960586*
0.6	50	6.089352	5.262101	5.051646*	6.358564	5.617563	5.550303*
	100	5.0846	4.644216	4.23125*	6.299727	5.504977	5.238714*
	200	5.361805	4.746037	4.430046*	6.968477	6.246467	5.653002*
0.7	50	6.039138	5.1865	4.970384*	5.505935	4.601287	4.649058*
	100	5.056463	4.359547	3.950028*	5.524701	4.779396	4.516602*
	200	3.755971	3.39325	3.030539*	6.855877	6.089826	5.608774*
0.8	50	5.545045	4.898264	4.792709*	6.033837	5.919316	6.033541*
	100	4.794218	3.864274	3.781592*	5.182348	4.570781	4.278688*
	200	3.30336	2.941	2.723301*	6.183978	5.669062	5.151912*

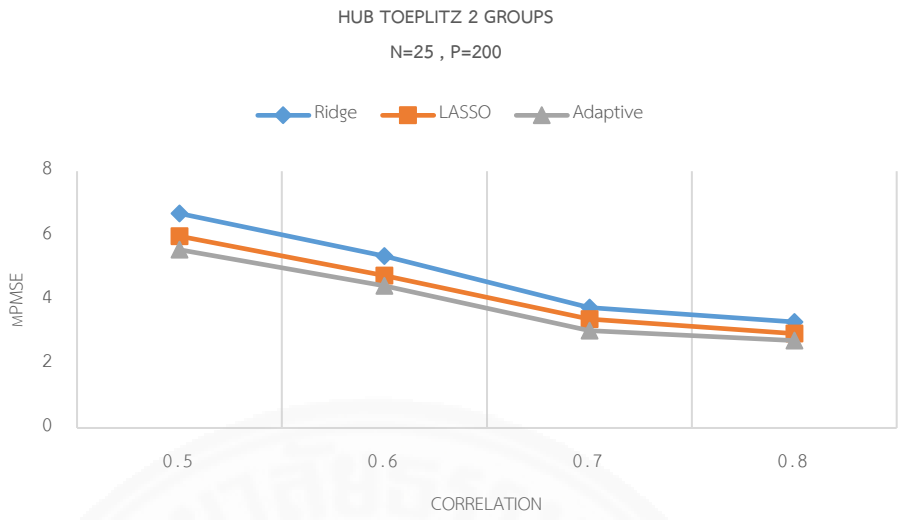
หมายเหตุ * แทน วิธีที่ให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด



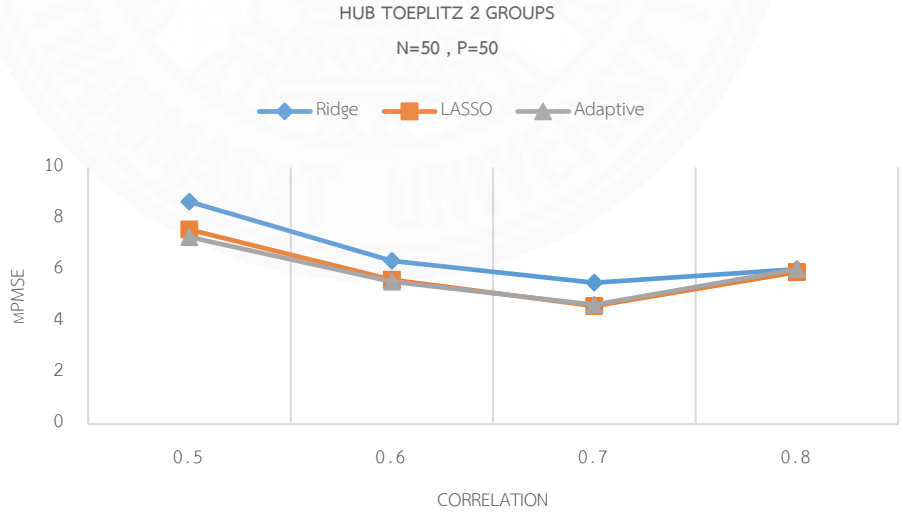
ภาพที่ 4.31 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 50$



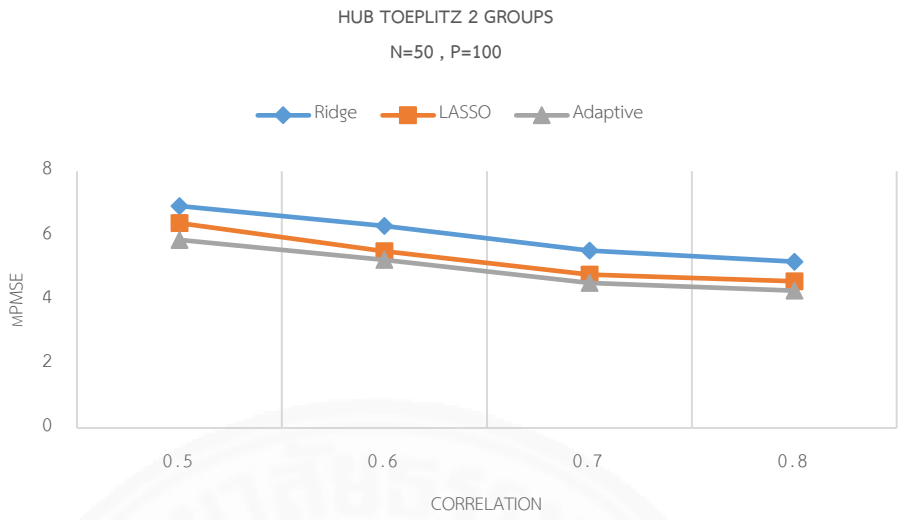
ภาพที่ 4.32 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 100$



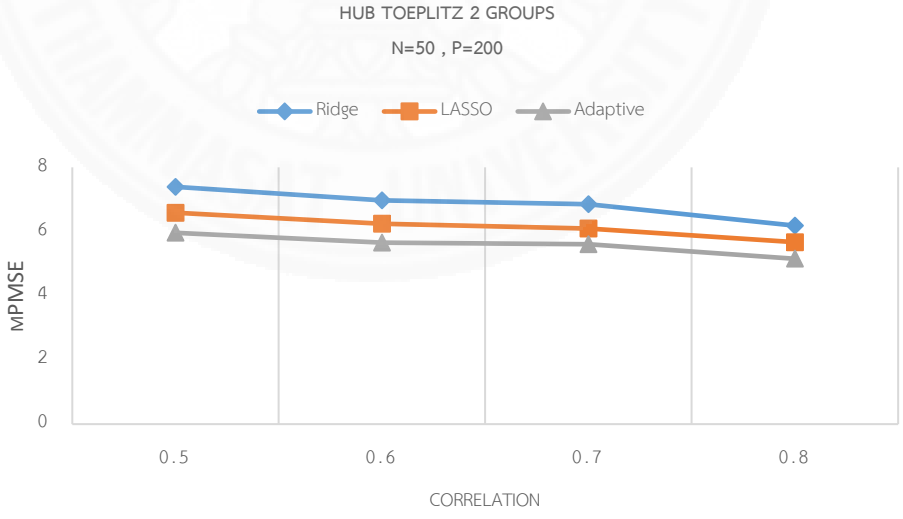
ภาพที่ 4.33 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 25,$ $p = 200$



ภาพที่ 4.34 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 50$



ภาพที่ 4.35 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 100$



ภาพที่ 4.36 ค่า mPMSE ของการพยากรณ์ในแต่ละวิธี ที่ตัวแปรอิสระมีความสัมพันธ์แบบ Hub Toeplitz และ ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ที่ระดับความสัมพันธ์ต่างๆ เมื่อ $n = 50,$ $p = 200$

จากตารางที่ 4.1-4.6 พบว่า กรณีที่ $r = 0.5, 0.6, 0.7, 0.8$ วิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) ต่ำที่สุดในทุกกรณี รองลงมา คือ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแบบริจ์ ตามลำดับ ทั้งในรูปแบบความสัมพันธ์แบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz และรูปแบบความสัมพันธ์แบบ Hub Toeplitz ทั้งกรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และ 2 กลุ่ม โดยผลลัพธ์ที่ได้ไปในทิศทางเดียวกัน

และเมื่อ n เพิ่มขึ้น จาก 25 เป็น 50 ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ แลซโซ และแลซโซแบบปรับปรุง จะมีค่าเพิ่มมากขึ้นด้วย

และจากภาพที่ 4.1-4.36 จะเห็นว่า เมื่อ p เพิ่มสูงขึ้น ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ แลซโซ และแลซโซแบบปรับปรุง จะมีแนวโน้มที่ค่าลดน้อยลง

นอกจากนี้ เมื่อความสัมพันธ์ของตัวแปรอิสระมีค่าเพิ่มขึ้น ยกตัวอย่างเช่น จาก $r = 0.5$ เป็น $r = 0.9$ ค่ามัธยฐานของ PMSE ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ แลซโซ และแลซโซแบบปรับปรุง จะมีค่าลดน้อยลง ทั้งในรูปแบบความสัมพันธ์แบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz และรูปแบบความสัมพันธ์แบบ Hub Toeplitz ทั้งกรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และ 2 กลุ่ม โดยผลลัพธ์ที่ได้ไปในทิศทางเดียวกัน

4.1.2 ประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ

1) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.7 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Constant

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n=25$				$n=50$			
0.5	50	0.1717	0.3913	0.1892	0.3886	0.3049	0.3441	0.2561	0.3413
	100	0.1759	0.1699	0.2521	0.1679	0.3095	0.1600	0.3780	0.1546
	200	0.2327	0.0792	0.3030	0.0789	0.3321	0.0771	0.4984	0.0747
0.6	50	0.1449	0.3922	0.1789	0.3907	0.2321	0.3580	0.2163	0.3579
	100	0.1761	0.1705	0.2457	0.1691	0.2705	0.1609	0.3181	0.1568
	200	0.2115	0.0779	0.2853	0.0773	0.3533	0.0746	0.4695	0.0720
0.7	50	0.1129	0.3995	0.1239	0.3936	0.2009	0.3737	0.2417	0.3659
	100	0.1855	0.1694	0.2007	0.1685	0.2956	0.1598	0.3077	0.1579
	200	0.2085	0.0792	0.2585	0.0789	0.3159	0.0786	0.4307	0.0776
0.8	50	0.1135	0.3996	0.0997	0.3983	0.1919	0.3882	0.1443	0.3795
	100	0.1795	0.1685	0.1931	0.1681	0.2979	0.1631	0.2829	0.1630
	200	0.1977	0.0795	0.2103	0.0791	0.3391	0.0770	0.3659	0.0767
0.9	50	0.1071	0.4090	0.1002	0.4087	0.1679	0.3837	0.1235	0.3930
	100	0.1451	0.1719	0.1465	0.1719	0.2599	0.1653	0.2437	0.1658
	200	0.1631	0.0793	0.1773	0.0792	0.2953	0.0786	0.3083	0.0785

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

2) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Constant ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.8 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Constant

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n = 25$				$n = 50$			
0.5	50	0.1093	0.4000	0.1467	0.3920	0.2453	0.3549	0.2927	0.3566
	100	0.1640	0.1652	0.1693	0.1611	0.3133	0.1587	0.3060	0.1514
	200	0.1887	0.0779	0.3013	0.0764	0.3207	0.0760	0.4673	0.0744
0.6	50	0.1113	0.3943	0.1220	0.3929	0.2093	0.3609	0.1747	0.3551
	100	0.1673	0.1604	0.1793	0.1602	0.2327	0.1555	0.3573	0.1496
	200	0.2007	0.0763	0.2347	0.0736	0.2933	0.0729	0.4173	0.0697
0.7	50	0.1200	0.4051	0.1247	0.3960	0.1507	0.3749	0.1900	0.3720
	100	0.1380	0.1633	0.2087	0.1621	0.2013	0.1641	0.3200	0.1578
	200	0.1727	0.0781	0.2147	0.0750	0.3347	0.0736	0.4600	0.0718
0.8	50	0.0880	0.3866	0.1093	0.3837	0.1813	0.3760	0.1933	0.3709
	100	0.1420	0.1649	0.1647	0.1612	0.1907	0.1596	0.1960	0.1575
	200	0.1413	0.0750	0.1800	0.0715	0.2620	0.0739	0.4480	0.0711
0.9	50	0.0573	0.3911	0.0973	0.3909	0.0707	0.3734	0.1267	0.3694
	100	0.0720	0.1652	0.1167	0.1640	0.1440	0.1578	0.1507	0.1519
	200	0.1413	0.0759	0.1720	0.0750	0.1933	0.0732	0.2507	0.0700

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

3) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.9 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Toeplitz

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n=25$				$n=50$			
0.5	50	0.1753	0.3846	0.2147	0.3769	0.3320	0.3343	0.3307	0.3311
	100	0.1787	0.1666	0.2713	0.1632	0.3900	0.1561	0.5080	0.1499
	200	0.2267	0.0792	0.3367	0.0770	0.4087	0.0736	0.5567	0.0706
0.6	50	0.1107	0.4049	0.1607	0.3977	0.1747	0.3771	0.2327	0.3577
	100	0.2087	0.1696	0.2913	0.1659	0.3260	0.1629	0.4740	0.1584
	200	0.2367	0.0786	0.3647	0.0779	0.3633	0.0776	0.6120	0.0769
0.7	50	0.1547	0.3863	0.1680	0.3760	0.2360	0.3649	0.2540	0.3617
	100	0.1500	0.1707	0.2147	0.1666	0.3033	0.1629	0.4327	0.1564
	200	0.1760	0.0804	0.3220	0.0791	0.2807	0.0772	0.5440	0.0726
0.8	50	0.1033	0.3983	0.1340	0.3989	0.1473	0.3889	0.1380	0.3809
	100	0.1113	0.1726	0.1960	0.1695	0.2233	0.1666	0.3660	0.1593
	200	0.1807	0.0789	0.2940	0.0775	0.2513	0.0781	0.4967	0.0760
0.9	50	0.0973	0.4060	0.1293	0.4029	0.1460	0.3943	0.1493	0.3897
	100	0.1513	0.1714	0.2400	0.1699	0.1787	0.1664	0.2347	0.1618
	200	0.1547	0.0790	0.2467	0.0777	0.2093	0.0781	0.4220	0.0756

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

4) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.10 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Toeplitz

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n=25$				$n=50$			
0.5	50	0.1753	0.3794	0.2213	0.3800	0.2380	0.3466	0.2420	0.3389
	100	0.1933	0.1667	0.2473	0.1612	0.4007	0.1519	0.4707	0.1456
	200	0.2107	0.0794	0.3553	0.0782	0.4247	0.0738	0.6580	0.0724
0.6	50	0.1240	0.3951	0.1647	0.3829	0.1887	0.3791	0.2527	0.3569
	100	0.1947	0.1668	0.2573	0.1654	0.2980	0.1596	0.4987	0.1520
	200	0.2167	0.0798	0.3347	0.0785	0.4027	0.0755	0.6427	0.0732
0.7	50	0.1160	0.3960	0.1900	0.3874	0.1893	0.3874	0.2293	0.3797
	100	0.1460	0.1722	0.2187	0.1651	0.2387	0.1609	0.3613	0.1559
	200	0.2327	0.0792	0.3487	0.0779	0.3400	0.0771	0.5693	0.0746
0.8	50	0.1247	0.3946	0.1473	0.3817	0.1520	0.3974	0.1733	0.3971
	100	0.1680	0.1702	0.2493	0.1700	0.1547	0.1695	0.3140	0.1602
	200	0.1840	0.0786	0.3127	0.0766	0.2380	0.0768	0.5127	0.0729
0.9	50	0.0967	0.4100	0.1133	0.4009	0.1387	0.3840	0.1347	0.3794
	100	0.1073	0.1720	0.1627	0.1672	0.2100	0.1658	0.2733	0.1629
	200	0.1947	0.0786	0.2673	0.0786	0.2567	0.0771	0.3627	0.0757

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

5) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.11 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบ Hub Toeplitz ที่มีความสัมพันธ์สูงสุด คือ $r_{\max} = 0.9$

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n = 25$				$n = 50$			
0.5	50	0.0902	0.4053	0.0935	0.4006	0.1733	0.3874	0.1347	0.3847
	100	0.1218	0.1690	0.1223	0.1665	0.1948	0.1647	0.1668	0.1641
	200	0.1426	0.0798	0.2052	0.0794	0.2603	0.0778	0.2793	0.0773
0.6	50	0.1043	0.3965	0.1240	0.4023	0.1184	0.3864	0.1477	0.3866
	100	0.1311	0.1689	0.1407	0.1678	0.1971	0.1657	0.2007	0.1650
	200	0.1577	0.0796	0.1804	0.0794	0.2719	0.0784	0.3133	0.0784
0.7	50	0.1013	0.3971	0.0891	0.3967	0.1286	0.3912	0.1645	0.3845
	100	0.1437	0.1697	0.1222	0.1697	0.2332	0.1686	0.2577	0.1664
	200	0.1557	0.0785	0.1945	0.0784	0.2410	0.0785	0.2911	0.0783
0.8	50	0.1201	0.4047	0.0962	0.4045	0.1132	0.3913	0.1499	0.3913
	100	0.1364	0.1707	0.1501	0.1697	0.2305	0.1686	0.2553	0.1676
	200	0.1590	0.0797	0.1852	0.0796	0.2585	0.0772	0.2726	0.0767

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

6) กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบ Hub Toeplitz ที่ระดับความสัมพันธ์ต่างๆ

ตารางที่ 4.12 ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบ Hub Toeplitz ที่มีความสัมพันธ์สูงสุด คือ $r_{\max} = 0.9$

		LASSO		Adaptive LASSO		LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
r	p	$n = 25$				$n = 50$			
0.5	50	0.0557	0.3928	0.0721	0.3871	0.1350	0.3800	0.1489	0.3776
	100	0.1283	0.1639	0.1375	0.1615	0.1531	0.1623	0.1555	0.1583
	200	0.1295	0.0778	0.1583	0.0768	0.2016	0.0743	0.1991	0.0723
0.6	50	0.0553	0.3925	0.0693	0.3867	0.1417	0.3825	0.1036	0.3809
	100	0.1035	0.1658	0.1069	0.1637	0.1778	0.1618	0.1905	0.1594
	200	0.1193	0.0766	0.1423	0.0749	0.1931	0.0752	0.2471	0.0735
0.7	50	0.0640	0.3894	0.0840	0.3866	0.1362	0.3719	0.0854	0.3680
	100	0.0842	0.1680	0.1039	0.1646	0.1741	0.1660	0.2033	0.1629
	200	0.1151	0.0771	0.1297	0.0756	0.1959	0.0722	0.1771	0.0694
0.8	50	0.0777	0.3969	0.0618	0.3921	0.1544	0.3761	0.1010	0.3742
	100	0.1083	0.1651	0.1049	0.1631	0.1810	0.1602	0.1619	0.1559
	200	0.1215	0.0774	0.1280	0.0758	0.1936	0.0739	0.1921	0.0709

หมายเหตุ : FNR คือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR)
FPR คือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR)

จากตารางที่ 4.7-4.12 พบว่า วิธีการวิเคราะห์การถดถอยแบบแลซโซ มีความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรซึ่งเมื่อกำหนดให้พารามิเตอร์มีค่าไม่เท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าเป็น 0 (IC1) หรือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) น้อยกว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง

แต่เมื่อพิจารณา IC2 ซึ่งเป็นความผิดพลาดที่เมื่อกำหนดให้พารามิเตอร์มีค่าเท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าไม่เป็น 0 หรือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง มีโอกาสเกิดความผิดพลาดในการคัดเลือกตัวแปรน้อยกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซ ในทุกกรณี และความน่าจะเป็นที่จะเกิด IC2 ลดลงเมื่อจานวนตัวแปรอิสระ p เพิ่มขึ้น แต่เมื่อจำนวนตัวแปรอิสระ p เท่ากับขนาดตัวอย่าง n จะเห็นว่า โอกาสที่จะเกิด IC2 จะมีค่าสูงที่สุด

4.2 ตัวอย่างการประยุกต์ใช้กับข้อมูลจริง

4.2.1 Software Engineering

ตัวอย่างข้อมูลจาก Fulda University (2016) ได้ศึกษาข้อมูลเกี่ยวกับการเรียนรู้การทำงานเป็นทีมของวิศวกร Software Engineering (SE) และสามารถศึกษาข้อมูลได้จากเว็ปไซต์ (http://archive.ics.uci.edu/ml/datasets) โดยศึกษาตัวอย่างจากวิศวกรในแต่ละทีม จากทั้งหมด 64 ทีม โดยในแต่ละทีมจะมีจำนวนวิศวกรแตกต่างกันออกไป และศึกษาร่วมกับตัวแปรอิสระทั้งหมด 79 ตัวแปร ยกตัวอย่างเช่น ร้อยละของจำนวนชั่วโมงในการประชุมของวิศวกรในแต่ละทีม ค่าเฉลี่ยของชั่วโมงการทำงานในแต่ละสัปดาห์ เป็นต้น จะเห็นได้ว่าตัวแปรที่ต้องการศึกษาเป็นจำนวนนับ นั่นคือ จำนวนสมาชิกของวิศวกรในแต่ละทีม กับตัวแปรอิสระต่างๆ ทั้งที่เป็นจำนวนจริงและจำนวนนับ และข้อมูลมีลักษณะเป็นข้อมูลที่มีมิติสูง จึงมีความเหมาะสมกับงานวิจัยนี้

ผลลัพธ์ที่ได้

ตารางที่ 4.13 ค่ามัธยฐานของ PMSE ในแต่ละวิธี สำหรับข้อมูล Software Engineering

Median of Predictive Mean Square Error (mPMSE)

n	p	Ridge	LASSO	Adaptive LASSO
64	79	0.3449383	0.2414236	0.2170461

จากตารางที่ 4.13 พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของ PMSE ต่ำที่สุด รองลงมา คือ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และสุดท้าย คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ ตามลำดับ

ตารางที่ 4.14 ความน่าจะเป็นในการคัดเลือกผิดพลาดของตัวแปรอิสระ โดยเปรียบเทียบวิธีการวิเคราะห์แบบแลซโซ และแลซโซแบบปรับปรุง สำหรับข้อมูล Software Engineering

Probability of Incorrect Selection

n	p	LASSO		Adaptive LASSO	
		FNR	FPR	FNR	FPR
64	79	0.3333333	0.1875	0.7333333	0.1875

จากตารางที่ 4.14 พบว่า วิธีการวิเคราะห์การถดถอยแบบแลซโซ มีความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรซึ่งเมื่อกำหนดให้พารามิเตอร์มีค่าไม่เท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าเป็น 0 (IC1) หรือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) น้อยกว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง แต่เมื่อพิจารณา IC2 ซึ่งเป็นความผิดพลาดที่เมื่อกาหนดให้พารามิเตอร์มีค่าเท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าไม่เป็น 0 หรือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง มีโอกาสเกิดความผิดพลาดในการคัดเลือกตัวแปร น้อยกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซ

4.2.2 Lung Cancer

ตัวอย่างข้อมูลจาก Fulda University (1991) ได้ศึกษาข้อมูลเกี่ยวกับการจำแนกผู้ป่วยมะเร็งปอดใน 3 ลักษณะ จากผู้ป่วยทั้งหมด 25 คน และสามารถศึกษาข้อมูลได้จากเว็บไซต์ (http://archive.ics.uci.edu/ml/datasets) โดยศึกษาร่วมกับตัวแปรอิสระต่างๆ อีก 52 ตัวแปร โดยที่ตัวแปรตอบสนองและตัวแปรอิสระเป็นข้อมูลจำนวนนับ และข้อมูลมีลักษณะเป็นข้อมูลที่มีมิติสูง จึงมีความเหมาะสมกับงานวิจัยนี้ เช่นกัน

ผลลัพธ์ที่ได้

ตารางที่ 4.15 ค่ามัธยฐานของ PMSE ของการพยากรณ์ในแต่ละวิธี สำหรับข้อมูล Lung Cancer

Median of Predictive Mean Square Error (mPMSE)

n	p	Ridge	LASSO	Adaptive LASSO
25	52	0.2106633	0.207631	0.1391923

จากตารางที่ 4.15 พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของ PMSE ต่ำที่สุด รองลงมา คือ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และสุดท้าย คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ ตามลำดับ

ตารางที่ 4.16 ความน่าจะเป็นในการคัดเลือกผิดพลาดของตัวแปรอิสระ โดยเปรียบเทียบวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง สำหรับข้อมูล Lung Cancer

Probability of Incorrect Selection

n	p	LASSO		Adaptive LASSO	
		FNR	**FPR**	**FNR**	**FPR**
25	52	0.4	0.270270	0.666666	0.270270

จากตารางที่ 4.16 พบว่า วิธีการวิเคราะห์การถดถอยแบบแลซโซ มีความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรซึ่งเมื่อกำหนดให้พารามิเตอร์มีค่าไม่เท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าเป็น 0 (IC1) หรือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) น้อยกว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง แต่เมื่อพิจารณา IC2 ซึ่งเป็นความผิดพลาดที่เมื่อกาหนดให้พารามิเตอร์มีค่าเท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าไม่เป็น 0 พบว่า หรือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง มีโอกาสเกิดความผิดพลาดในการคัดเลือกตัวแปร น้อยกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซ ซึ่งในผลลัพธ์ไปในทิศทางเดียวกับการจำลองข้อมูล

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้เป็นการศึกษาเปรียบเทียบวิธีการวิเคราะห์การถดถอยปัวซง (Poisson Regression) ด้วยวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ (Penalized Regression) 3 วิธี ได้แก่ วิธีการวิเคราะห์แบบริดจ์ (Ridge) วิธีการวิเคราะห์แบบแลซโซ (LASSO) และวิธีการวิเคราะห์แลซโซแบบปรับปรุง (Adaptive LASSO) ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง ทั้ง 3 รูปแบบ นั่นคือ รูปแบบความสัมพันธ์แบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz และ รูปแบบความสัมพันธ์แบบ Hub Toeplitz โดยกำหนดให้ตัวแปรอิสระมีความสัมพันธ์ในระดับ 0.5, 0.6, 0.7, 0.8 และ 0.9 โดยจะกล่าวถึงสรุปผลการวิจัยและข้อเสนอแนะตามลำดับนี้

5.1 สรุปผลการวิจัย

จากการวัดประสิทธิภาพการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีการวิเคราะห์แบบ penalized regression 3 วิธี โดยพิจารณาจากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยในการพยากรณ์ และความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของวิธีการวิเคราะห์แบบแลซโซ และแลซโซแบบปรับปรุง

5.1.1 ผลสรุปจากการจำลองสถานการณ์

5.1.1.1 ประสิทธิภาพในการพยากรณ์

1. ถ้าพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา หรือกรณีที่ขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ $(n < p)$ จะพบว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุงดีที่สุดในทุกกรณี ทั้งในรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Constant, Toeplitz และ Hub Toeplitz เนื่องจากวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยในการพยากรณ์ต่ำที่สุด โดยพิจารณาจากความถูกต้องในการพยากรณ์

2. เมื่อขนาดตัวอย่างเพิ่มขึ้นจาก 25 เป็น 50 และที่จำนวนตัวแปรอิสระเท่ากัน จะพบว่า วิธีการวิเคราะห์แบบ penalized regression ทั้ง 3 วิธี ได้แก่ วิธีการวิเคราะห์แบบริดจ์ วิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง จะมีประสิทธิภาพลดลง สามารถแสดง

ได้จากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์จะมีเพิ่มมากขึ้น เมื่อขนาดตัวอย่างเพิ่มมากขึ้น

3. ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันเพิ่มมากขึ้น ตัวอย่างเช่น เมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นจาก 0.5 เป็น 0.9 โดยที่วิธีการวิเคราะห์แบบริดจ์ ก็จะมีประสิทธิภาพเพิ่มมากขึ้นด้วยเช่นกัน ทั้งรูปแบบความสัมพันธ์ของตัวแปรอิสระ Constant, Toeplitz และ Hub Toeplitz ดังนั้น วิธีการวิเคราะห์แบบ penalized regression ทั้ง 3 วิธี สามารถแก้ไขปัญหาภาวะร่วมเชิงเส้น (Multicollinearity) ซึ่งสอดคล้องกับผลงานวิจัยของ Hoerl และ Kennard ในปี 1970 ที่ว่า วิธีการวิเคราะห์แบบริดจ์ สามารถแก้ไขปัญหาภาวะร่วมเชิงเส้นได้

และวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง ก็มีประสิทธิภาพเพิ่มมากขึ้นเช่นกัน เมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มมากขึ้น โดยสามารถดูได้จากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ลดลง

4. ถ้าพิจารณาเปรียบเทียบรูปแบบความสัมพันธ์ของตัวแปรอิสระทั้ง 3 รูปแบบข้างต้น โดยพิจารณาจากความถูกต้องในการพยากรณ์ ด้วยธีการวิเคราะห์แบบริดจ์ จะพบว่า รูปแบบความสัมพันธ์แบบ Hub Toeplitz จะให้ประสิทธิภาพดีที่สุด รองลงมา คือ รูปแบบความสัมพันธ์แบบ Constant และสุดท้าย คือ รูปแบบความสัมพันธ์แบบ Toeplitz ให้ประสิทธิภาพน้อยที่สุด เมื่อเปรียบเทียบในกรณีที่ตัวแปรอิสระมีความสัมพันธ์เท่ากัน ที่เป็นเช่นนี้ เนื่องจากรูปแบบความสัมพันธ์แบบ Toeplitz และ Hub Toeplitz จะสร้างความสัมพันธ์ของตัวแปรอิสระ โดยที่ ตัวแปรอิสระที่อยู่ใกล้กันจะมีความสัมพันธ์กันสูง แต่ตัวที่อยู่ห่างกัน ก็จะมีความสัมพันธ์กันลดน้อยลง ดังนั้น จะทำให้ความสัมพันธ์ลดลงเรื่อยๆ สำหรับความสัมพันธ์ของตัวแปรอิสระที่อยู่ติดกัน แต่รูปแบบความสัมพันธ์แบบ Hub Toeplitz จะสร้างความสัมพันธ์ของตัวแปรอิสระ 2 ตัวใดๆ ให้มีค่ามากกว่ารูปแบบความสัมพันธ์แบบ Toepltz ที่ระดับความสัมพันธ์เดียวกัน แต่ในทางตรงกันข้าม รูปแบบความสัมพันธ์แบบ Constant จะให้ความสัมพันธ์ของตัวแปรอิสระทุกตัวมีค่าเท่ากันหมด ดังนั้น เมื่อพิจารณาที่ระดับความสัมพันธ์ของตัวแปรอิสระที่เท่ากันแล้ว รูปแบบความสัมพันธ์แบบ Hub Toeplitz จะให้ความสัมพันธ์ของตัวแปรอิสระ 2 ตัวใดๆสูงที่สุด รองลงมา คือ รูปแบบความสัมพันธ์แบบ Constant และสุดท้าย คือ รูปแบบความสัมพันธ์แบบ Toeplitz ทำให้ได้ว่า วิธีการวิเคราะห์การถดถอยแบบริดจ์ สามารถแก้ไขปัญหาการเกิดภาวะร่วมเชิงเส้นของตัวแปรอิสระได้ ซึ่งสอดคล้องกับผลงานวิจัยของ Hoerl และ Kennard ในปี 1970 เช่นเดียวกัน

และวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง ในรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Hub Toeplitz จะมีประสิทธิภาพดีที่สุด รองลงมา คือ รูปแบบความสัมพันธ์แบบ Constant และสุดท้าย คือ รูปแบบความสัมพันธ์แบบ Toepltz เช่นเดียวกับวิธีการวิเคราะห์แบบริดจ์ ดังนั้น จึงสรุปได้ว่า วิธีการวิเคราะห์การถดถอยแบบแลซโซและแลซโซแบบปรับปรุง สามารถแก้ไขปัญหาภาวะร่วมเชิงเส้นของตัวแปรอิสระได้เช่นกัน

ผลการวิจัยในตัวแปรอิสระ 1 กลุ่ม และ 2 กลุ่ม ให้ผลไปในทิศทางเดียวกัน แต่ในกรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม จะทำให้ความสัมพันธ์ของตัวแปรอิสระมีความสัมพันธ์กันสูงกว่า ในกรณีที่ตัวแปรอิสระมี 1 กลุ่ม ที่ระดับความสัมพันธ์เดียวกัน เนื่องจากกรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ความสัมพันธ์ของตัวแปรอิสระในกลุ่มแรกของตัวแปรอิสระที่อยู่ใกล้เคียงกัน จะมีความสัมพันธ์กันสูง และตัวแปรอิสระกที่อยู่ไกลกันจะมีความสัมพันธ์ลดลง ซึ่งเป็นเช่นเดียวกันกับตัวแปรอิสระในกลุ่มที่สอง ซึ่งจะเห็นได้ว่า ถ้าในกรณีที่ตัวแปรอิสระมีเพียงกลุ่มเดียว ก็จะทำให้ความสัมพันธ์ของตัวแปรอิสระที่อยู่ไกลกันลดลงเรื่อย ดังนั้น กรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม จึงมีความสัมพันธ์ของตัวแปรอิสระ 2 ตัวใดๆ สูงกว่า ที่ระดับความสัมพันธ์เดียวกัน กรณีที่ตัวแปรอิสระมีเพียงกลุ่มเดียว จึงสรุปได้ว่า วีธีการวิเคราะห์แบบ penalized regression ทั้ง 3 วิธี สามารถแก้ไขปัญหาภาวะร่วมเชิงเส้นได้ แสดงให้เห็นได้จากประสิทธิภาพในการพยากรณ์ เพราะเมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม จะทำให้ค่ามัธยฐานของคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าลดลง จากกรณีที่ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม ทั้งในรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Constant, รูปแบบความสัมพันธ์แบบ Toeplitz และรูปแบบความสัมพันธ์แบบ Hub Toeplitz

5.1.1.2 ประสิทธิภาพในการคัดเลือกตัวแปร

ถ้าพิจารณาเปรียบเทียบความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จะพบว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุง มีโอกาสในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆที่ตัวแปรอิสระนั้นไม่ควรอยู่ในตัวแบบ หรือ อัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) ซึ่งเป็นการวัดความน่าจะเป็นที่จะเกิดความผิดพลาดจาก Identify Criterion 1 $(\mathrm{IC1})$ มากกว่า วิธีการวิเคราะห์แบบแลซโซ

แต่มีโอกาสที่จะไม่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆ ที่ตัวแปรอิสระนั้นควรอยู่ในตัวแบบ หรือ อัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) เป็นการวัด

ความน่าจะเป็นที่จะเกิดความผิดพลาดจาก Identify Criterion 2 $(\mathrm{IC2})$ น้อยกว่าวิธีการวิเคราะห์แบบแลซโซ และเมื่อตัวแปรอิสระ (p) เพิ่มมากขึ้น โอกาสที่จะเกิดความผิดพลาดจะลดลงน้อยลง

และเมื่อตัวแปรอิสระมีความสัมพันธ์กันเพิ่มสูงขึ้น อัตราความผิดพลาดในการตรวจจับเชิงลบและเชิงบวกของวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง โอกาสที่จะเกิดความผิดพลาดจะลดลงน้อยลงเช่นเดียวกัน

เมื่อเปรียบเทียบประสิทธิภาพของการคัดเลือกตัวแปรอิสระของทั้ง 2 วิธี นั่นคือ วิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แลซโซแบบปรับปรุง จะพบว่า เมื่อขนาดตัวอย่าง (n) เพิ่มมากขึ้น อัตราความผิดพลาดทั้ง 2 ชนิด จะเพิ่มมากขึ้นด้วย แต่ในขณะที่ตัวแปรอิสระมีความสัมพันธ์กันสูงขึ้น อัตราความผิดพลาดทั้ง 2 ชนิดของวิธีการวิเคราะห์แลซโซแบบปรับปรุง จะมีความเสถียรมากกว่า วิธีการวิเคราะห์แบบแลซโซ

และผลการวิจัยในตัวแปรอิสระ 1 กลุ่ม และ 2 กลุ่ม ให้ผลไปในทิศทางเดียวกัน

ดังนั้น จากการเปรียบเทียบประสิทธิภาพในการพยากรณ์ และการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระ ทั้ง 2 เกณฑ์ จะพบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะมีประสิทธิภาพดีที่สุด ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง ในตัวแบบการถดถอยปัวซง

5.1.2 ผลสรุปจากการประยุกต์ใช้กับข้อมูลจริง

จากการประยุกต์ใช้กับข้อมูลจริง ได้ผลสรุปว่า ในการเปรียบเทียบประสิทธิภาพของการพยากรณ์ วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงมีประสิทธิภาพดีที่สุด รองลงมา คือ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และสุดท้าย คือ วิธีการวิเคราะห์การถดอถอยแบบริดจ์

นอกจากนี้ ประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ วิธีการวิเคราะห์แลซโซแบบปรับปรุง มีโอกาสในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆที่ตัวแปรอิสระนั้นไม่ควรอยู่ในตัวแบบมากกว่าวิธีการวิเคราะห์แบบแลซโซแต่มีโอกาสที่จะไม่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆ ที่ตัวแปรอิสระนั้นควรอยู่ในตัวแบบน้อยกว่าวิธีการวิเคราะห์แบบแลซโซ ถึงแม้ว่าลักษณะของข้อมูลที่ใช้ในการวิเคราะห์จะมีลักษณะต่างกัน ก็จะให้ผลเช่นเดียวกับการจำลองสถานการณ์ต่างๆ

5.2 ข้อเสนอแนะ

ผลจากการจำลองสถานการณ์ต่างๆ และการประยุกต์ใช้กับข้อมูลจริง ผู้วิจัยได้ข้อเสนอแนะ ดังนี้

5.2.1 ข้อเสนอแนะเกี่ยวกับการวิจัย

เนื่องจากผู้วิจัยได้ศึกษาประสิทธิภาพของการวิเคราะห์การถดถอยแบบ penalized regression 3 วิธี นั่นคือ วิธีการวิเคราห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง โดยการนำวิธีเหล่านี้มาวิเคราะห์ จำเป็นต้องมีการกำหนดให้ตัวแปรอิสระส่วนน้อยให้มีผลต่อตัวแบบ หรือเรียกว่า ตัวแบบบางเบา (Sparse Model) ถึงเหมาะกับการเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยทั้ง 3 วิธี และในการกำหนดขนาดตัวอย่าง ไม่ควรที่จะกำหนดขนาดตัวอย่างน้อยเกิดไป เนื่องจาก วิธีการวิเคราะห์การถดถอยแบบ penalized จะต้องใช้วิธี 5-fold cross validation ในการคำนวณหาพารามิเตอร์ปรับแต่ง จึงต้องมีการแบ่งข้อมูลออกเป็นส่วนๆ ดังนั้น ถ้าเรามีการกำหนดขนาดตัวอย่างน้อยเกินไปอาจจะไม่สามารถวิเคราะห์ข้อมูลออกมาได้ นอกจากนี้ เนื่องจากตัวแปรอิสระมีเป็นจำนวนมาก โดยบางครั้ง ผู้วิจัยไม่สามารถทราบได้ว่า หน่วยของตัวแปรอิสระที่นำมาศึกษานั้น มีหน่วยเดียวกันหรือไม่ ดังนั้น ผู้วิจัยจึงต้องมีการ standardized ข้อมูล ก่อนนำมาวิเคราะห์ เพื่อทำให้ผลลัพธ์ที่ได้มีประสิทธิภาพดีที่สุด

5.2.2 ข้อเสนอแนะเกี่ยวกับการนำไปประยุกต์ใช้

จากการประยุกต์ใช้กับข้อมูลจริง ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบ penalized ทั้ง 3 วิธี ได้แก่ วิธีการวิเคราห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง ในกรณีที่ข้อมูลที่มีมิติสูงและตัวแปรอิสระมีความสัมพันธ์กัน ดังนั้น ผู้วิจัยจึงจำเป็นต้องใช้ข้อมูลในกรณีที่ข้อมูลมีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระ ซึ่งให้ผลไปในทิศทางเดียวกับการจำลองข้อมูล แต่ข้อมูลจริงอาจจะมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างไม่มากเท่ากับการจำลองข้อมูล ดังนั้น ผู้วิจัยจึงอาจขยายการศึกษาเพิ่มเติม ในกรณีที่ตัวแปรอิสระมีมากกว่าขนาดตัวอย่าง มากกว่าข้อมูลจริงที่นำมาประยุกต์ใช้

รายการอ้างอิง

หนังสือและบทความในหนังสือ

สายชล สินสมบูรณ์ทอง. (2011). สถิติคณิตศาสตร์ 1 (พิมพ์ครั้งที่ 5). กรุงเทพฯ : โรงพิมพ์ จามจุรีโปรดักส์.

Travor H., & Robert T., and Jerome F. (2008). The Elements of Statistical Learning, Second Edition: Springer Series in Statistics.

Raymond H., Douglas C., Geoffrey G., and Timothy J. (2010). Generalized Linear Models, Second Edition: John Wiley & Sons, Inc.

Julian J. (2015). Faraway Linear Model with R second edition. Shrinkage method, Chapter 11, 161-177.

บทความวารสาร

วิฐรา พึ่งพาพงศ์. (2012). บทความวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. *วารสารวิทยาศาสตร์และเทคโนโลยี ปีที่ 23 ฉบับที่ 2 (เมษายน-มิถุนายน)*, ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

ทิฆัมพร สาระกอ และ นัท กุลวานิช. (2014). การเปรียบเทียบประสิทธิภาพการพยากรณ์และการคัดเลือกตัวแปรอิสระของวิธีเพิ่มลดตัวแปรในขั้นตอน วิธีแลซโซ วิธีอีลาสติคเน็ต และวิธีแลซโซปรับปรุง สำหรับผลกระทบขนาดล็ก และมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์. *การประชุมสัมมนาวิชาการ มทร.ตะวันออก มรภ.กลุ่มศรีอยุธยา และราชนครินทร์วิชาการและวิจัย* (14-16พฤษภาคม).

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

Park, M. Y., and Hastie, T. (2007). L1‐regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 659-677.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

Johnstone, I. M., and Titterington, D. M. (2009). Statistical challenges of high-dimensional data: 4237-4253.

Månsson, K., and Shukur, G. (2011). A Poisson ridge regression estimator. *Economic Modelling*, *28*(4), 1475-1481.

Hossain, S., and Ahmed, S. E. (2012). Shrinkage and penalty estimators of a Poisson regression model. *Australian & New Zealand Journal of Statistics*, *54*(3), 359-373.

Hardin, J., Garcia, S. R., and Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 1733-1762.

Pungpapong, V. (2014). A Brief review on high-dimensional Linear regression. *Scholarly Article of Statistical*, ChulalongkornUniversity, Bangkok.

Oyeyemi, G. M., Ogunjobi, E. O., and Folorunsho, A. I. (2015). On performance of shrinkage methods–a Monte Carlo Study. *International Journal of Statistics and Applications*, *5*(2), 72-76.

Algamal, Z. Y., and Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional poisson regression model. *Modern Applied Science*, *9*(4), 170.

Ahmed, S. E., and Yüzbaşı, B. (2016). Big data analytics: integrating penalty strategies. *International Journal of Management Science and Engineering Management*, *11*(2), 105-115.

Ivanoff, S., Picard, F., and Rivoirard, V. (2016). Adaptive Lasso and group-Lasso for functional Poisson regression. *Journal of Machine Learning Research*, *17*(55), 1-46.

Yüzbaşı, B., Arashi, M., and Ahmed, S. E. (2017). Shrinkage Estimation Strategies in Generalized Ridge Regression Models Under Low/High-Dimension Regime. *arXiv preprint arXiv:1707.02331*.

Gao, X., Ahmed, S. E., and Feng, Y. (2017). Post selection shrinkage estimation for high-dimensional data analysis. *Applied Stochastic Models in Business and Industry*, *33*(2), 97-120.

Yüzbası, B., Ahmed, S. E., and Güngör, M. (2017). Improved Penalty Strategies in Linear Regression Models. *REVSTAT–Statistical Journal*, *15*(2), 251-276.

วิทยานิพนธ์

นิศาชล งามประเสริฐสิทธิ์ (2012). การเปรียบเทียบการคัดเลือกตัวแปรอิสระที่มีปัญหาสหสัมพันธ์เชิงเส้นพหุด้วยวิธีการถดถอยแบบริดจ์และการค้นหาแบบต้องห้าม.

ภาคผนวก ก

ภาคผนวก ก
โปรแกรมที่ใช้ในการสร้างและวิเคราะห์ข้อมูล

1. โปรแกรมที่ใช้ในการจำลองข้อมูล ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ และความผิดพลาดในการคัดเลือกแปร เมื่อตัวแปรอิสระมีรูแปบบความสัมพันธ์แบบ Constant เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม

```
library(MASS)  # Package needed to generate correlated predictors
library(glmnet)  # Package to fit ridge/lasso/elastic net models
```

สร้างฟังก์ชันหลัก GetConstant2Groups ในการจำลองข้อมูล

```
GetConstant1Group <- function(n,p,r)    {
 N <- 1000
 beta.1 <- array(data=NA,dim = c(p,1,N))
 x <- array(data=NA,dim = c(n,p,N))
 Y <- array(data=NA,dim = c(n,1,N))
 y_k <- array(data=NA,dim = c(n,1,N))
 y <- array(data=NA,dim = c(n,1,N))
 e <- array(data=NA,dim = c(n,1,N))
 e.t <- array(data=NA,dim = c(n,1,N))
 x.t <- array(data=NA,dim = c(n,p,N))
 Y.t <- array(data=NA,dim = c(n,1,N))
 y_k.t <- array(data=NA,dim = c(n,1,N))
 y.t <- array(data=NA,dim = c(n,1,N))
 train_rows <- array(data=NA,dim = c(0.8*n,1,N))
 x.train <- array(data=NA,dim = c(0.8*n,p,N))
 x.test <- array(data=NA,dim = c(0.2*n,p,N))
 y.train <- array(data=NA,dim = c(0.8*n,1,N))
 y.test <- array(data=NA,dim = c(0.2*n,1,N))
 muS <- array(data=NA,dim = c(n,1,N))
 muS.t <- array(data=NA,dim = c(n,1,N))
 lamda <- array(data=NA,dim = c(n,1,N))
 muX <- array(data=NA,dim = c(p,1,N))
```

```
sdX <- array(data=NA,dim = c(p,1,N))
x.star <- array(data=NA,dim = c(n,p,N))
lam.ridge <- array(data=NA,dim = c(1,1,N))
lam.lasso <- array(data=NA,dim = c(1,1,N))
lam.adaplasso <- array(data=NA,dim = c(1,1,N))
ridge <- array(data=NA,dim = c(p+1,1,N))
lasso <- array(data=NA,dim = c(p+1,1,N))
adaplasso <- array(data=NA,dim = c(p+1,1,N))
IC1.lasso <- array(data=NA,dim = c(1,1,N))
IC2.lasso <- array(data=NA,dim = c(1,1,N))
IC1.adaplasso <- array(data=NA,dim = c(1,1,N))
IC2.adaplasso <- array(data=NA,dim = c(1,1,N))
sum.IC1.lasso <- array(data=NA,dim = c(1,1,N))
sum.IC2.lasso <- array(data=NA,dim = c(1,1,N))
sum.IC1.adaplasso <- array(data=NA,dim = c(1,1,N))
sum.IC2.adaplasso <- array(data=NA,dim = c(1,1,N))
pmse.ridge <- array(data=NA,dim = c(1,1,N))
pmse.lasso <- array(data=NA,dim = c(1,1,N))
pmse.adaplasso <- array(data=NA,dim = c(1,1,N))
```

*****สร้างรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Constant เมื่อตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม**:

```
R <- matrix(rep(r,p*p),ncol=p,nrow = p)
diag(R) <- 1
CorMat1 <- matrix(c(R),ncol = p)
```

<u>หมายเหตุ</u> : *** ในแต่ละรูแปบบความสัมพันธ์ของตัวแปรอิสระ จะต้องสร้างเมทริกซ์ความแปรปรวนที่แตกต่างกันออกไป และเนื่องจากต้องการให้ตัวแปรอิสระแบ่งออกเป็น 1 กลุ่ม จึงสร้างเมทริกซ์ความแปรปรวน 1 ชุด นั่นคือ CorMat1

เช่นเดียวกับ กรณีรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Toeplitz เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม
จะใช้

```
CorMat1 <- outer(1:p, 1:p, function(x,y) {r^abs(x-y)})
```

และกรณีรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Hub Toeplitz เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม

```
R <- outer(1:p, 1:p, function(x,y) {rmax-(((rmax-rmin)/(p-2))*(abs(x-y)-1))})
diag(R) <- 1
CorMat <- matrix(c(R),ncol = p)
```

สร้างเมทริกซ์ของตัวแปรอิสระ จากการแจกแจงปกติ โดยกำหนดให้เมทริกซ์ความแปรปรวน แตกต่างกันในแต่ละรูปแบบความสัมพันธ์

```
X <- mvrnorm(n, rep(0,p),CorMat1)

beta <- as.vector(c(1,-0.5,-0.5,0.1,0.1,
                    0.1,0.05,0.05,0.05,0.05,
                    0.01,0.01,0.01,0.01,0.01,rep(0,p-15)))
 for(i in 1:N)   {
```

<u>ขั้นตอนที่ 1</u> : สร้างข้อมูลค่าสังเกต $(y_i, \mathbf{X}_i)$

```
x[,,i] <- X[sample(nrow(X),size=n,replace = TRUE),]
e[,,i]  <- rpois(n,1)
muS[,,i] <- exp(x[,,i]%*%beta+ e[,,i])
y[,,i] <- rpois(n=n, lambda = muS[,,i])
data <- data.frame(y=y[,,i], x=x[,,i])
```

ขั้นตอนที่ 2 : แบ่งชุดข้อมูลค่าสังเกตออกเป็น 2 ชุด นั่นคือ test set และ training set

```
train_rows[,,i] <- matrix(sample(1:n,0.8*n),ncol = 1)

x.train[,,i] <- x[c(train_rows[,1,i]),,i]
x.test[,,i]  <- x[-c(train_rows[,1,i]),,i]

y.train[,,i] <- y[train_rows[,1,i],,i]
y.test[,,i]  <- y[-c(train_rows[,1,i]),,i]
```

ขั้นตอนที่ 3 : **หาพารามิเตอร์ปรับแต่ง (tuning parameter) ด้วยวิธี 5-fold cross validation**

```
### 1. Ridge Regression
lam.ridge[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,alpha=0,standardize=TRUE,
                            family="poisson")$lambda.min

### 2. LASSO
lam.lasso[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,nfolds=5,alpha=1,
                            family="poisson")$lambda.min

### 3. Adaptive LASSO
#### Ridge Regression to screate the Adaptive Weights Vector
w3 <- 1/abs(matrix(coef(cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,standardize=TRUE,
                                  alpha=0,family="poisson"), s=lam.ridge[,,i])
                                  [, 1][2:(ncol(x)+1)] ))^1
w3[w3[,1] == Inf] <- 999999999
## Replacing values estimated as Infinite for 999999999
lam.adaplasso[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,standardize=TRUE,
                                alpha=1, parallel=TRUE,family="poisson",
                                penalty.factor=w3)$lambda.min
```

<u>ขั้นตอนที่ 4</u> : หาค่าสัมประสิทธิ์การถดถอยในแต่ละวิธี

```
### 1. Ridge Regression
ridge[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],alpha=0,standardize=TRUE,
                                          family="poisson",intercept =TRUE,nfolds=5),
                                          x.test,s=lam.ridge[,,i],type="coefficients"))
### 2. LASSO
lasso[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,alpha=1,
                                          family="poisson",intercept = TRUE,nfolds=5),
                                          x.test,s=lam.lasso[,,i],type="coefficients"))

### 3. Adaptive LASSO
adaplasso[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,
                                          alpha=1,nfolds = 5,penalty.factor=w3,
                                          family="poisson",intercept = TRUE),
                                          x.test,s=lam.adaplasso[,,i],
                                          type="coefficients"))
```

<u>ขั้นตอนที่ 5</u> : หาค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (PMSE) ในแต่ละวิธี

```
### 1. Ridge Regression
pmse.ridge[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%ridge[,,i])) %*% (y[,,i]-
                    exp(cbind(rep(1,n),x[,,i])%*%ridge[,,i])) )/n

### 2. LASSO
pmse.lasso[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%lasso[,,i])) %*% (y[,,i]-
                    exp(cbind(rep(1,n),x[,,i])%*%lasso[,,i])) )/n
```

```
### 3. Adaptive LASSO
pmse.adaplasso[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%adaplasso[,,i])) %*% (y[,,i]-
                        exp(cbind(rep(1,n),x[,,i])%*%adaplasso[,,i])) )/n
```

<u>ขั้นตอนที่ 6</u> : พิจารณาค่าประมาณสัมประสิทธิ์ถดถอยที่ได้ กับค่าพารามิเตอร์สัมสิทธิ์การถดถอยที่กำหนด (IC1,IC2)

```
beta.1[,,i] <- as.vector(c(1,-0.5,-0.5,0.1,0.1,
                           0.1,0.05,0.05,0.05,0.01,
                           0.01,0.01,0.01,0.01,0.01,rep(0,p-15)))
### 1. LASSO
IC1.lasso[,,i] <- length(which(beta.1[,,i] ==0 & lasso[,,i] !=0))
IC2.lasso[,,i] <- length(which(beta.1[,,i] !=0 & lasso[,,i] ==0))

### 2. Adaptive LASSO
IC1.adaplasso[,,i] <- length(which(beta.1[,,i]==0 & adaplasso[,,i] !=0))
IC2.adaplasso[,,i] <- length(which(beta.1[,,i] !=0 & adaplasso[,,i] ==0))

                    } #end for
```

<u>ขั้นตอนที่ 7</u> : หาค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) ในแต่ละวิธี

```
### 1. Ridge Regression
sim.pmse.ridge <- apply(pmse.ridge,2,median)


### 2. LASSO
sim.pmse.lasso <- apply(pmse.lasso,2,mean)
```

```
### 3. Adaptive LASSO
sim.pmse.adaplasso <- apply(pmse.adaplasso,2,mean)
```

<u>ขั้นตอนที่ 8</u> : คำนวณหาอัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) และอัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) ของวิธี LASSO และ Adaptive LASSO

```
### 1. LASSO
sum.IC1.lasso <- apply(IC1.lasso,2,sum)
sum.IC2.lasso <- apply(IC2.lasso,2,sum)

prob.IC1.lasso <-  sum.IC1.lasso /(15*N)
prob.IC2.lasso <-  sum.IC2.lasso  /((p-15)*N)

### 2. Adaptive LASSO

sum.IC1.adaplasso <- apply(IC1.adaplasso,2,sum)
sum.IC2.adaplasso <- apply(IC2.adaplasso,2,sum)

prob.CI1.adaplasso  <-  sum.CI1.adaplasso  /(15*N)
prob.CI2.adaplasso  <-  sum.CI2.adaplasso  /((p-15)*N)

cat(sim.pmse.ridge,'\t',sim.pmse.lasso,'\t',sim.pmse.adaplasso,'\t',
    prob.IC1.lasso,'\t', prob.IC2.lasso,'\t',prob.IC1.adaplasso,'\t',prob.IC2.adaplasso,'\n' )

                        } # end GetConstant1Group
#####################################################################
```

2. โปรแกรมที่ใช้ในการจำลองข้อมูล ค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ และความผิดพลาดในการคัดเลือกแปร เมื่อตัวแปรอิสระมีรูแปบบความสัมพันธ์แบบ Constant เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม

```
library(MASS)  # Package needed to generate correlated predictors
library(glmnet)  # Package to fit ridge/lasso/elastic net models
```

สร้างฟังก์ชันหลัก GetConstant2Groups ในการจำลองข้อมูล

```
GetConstant2Groups <- function(n,p,r)     {
 N <- 1000
 beta.1 <- array(data=NA,dim = c(p,1,N))
 x <- array(data=NA,dim = c(n,p,N))
 Y <- array(data=NA,dim = c(n,1,N))
 y_k <- array(data=NA,dim = c(n,1,N))
 y <- array(data=NA,dim = c(n,1,N))
 e <- array(data=NA,dim = c(n,1,N))
 e.t <- array(data=NA,dim = c(n,1,N))
 x.t <- array(data=NA,dim = c(n,p,N))
 Y.t <- array(data=NA,dim = c(n,1,N))
 y_k.t <- array(data=NA,dim = c(n,1,N))
 y.t <- array(data=NA,dim = c(n,1,N))
 train_rows <- array(data=NA,dim = c(0.8*n,1,N))
 x.train <- array(data=NA,dim = c(0.8*n,p,N))
 x.test <- array(data=NA,dim = c(0.2*n,p,N))
 y.train <- array(data=NA,dim = c(0.8*n,1,N))
 y.test <- array(data=NA,dim = c(0.2*n,1,N))
 muS <- array(data=NA,dim = c(n,1,N))
 muS.t <- array(data=NA,dim = c(n,1,N))
 lamda <- array(data=NA,dim = c(n,1,N))
 muX <- array(data=NA,dim = c(p,1,N))
 sdX <- array(data=NA,dim = c(p,1,N))
 x.star <- array(data=NA,dim = c(n,p,N))
```

```
lam.ridge <- array(data=NA,dim = c(1,1,N))
lam.lasso <- array(data=NA,dim = c(1,1,N))
lam.adaplasso <- array(data=NA,dim = c(1,1,N))
ridge <- array(data=NA,dim = c(p+1,1,N))
lasso <- array(data=NA,dim = c(p+1,1,N))
adaplasso <- array(data=NA,dim = c(p+1,1,N))
IC1.lasso <- array(data=NA,dim = c(1,1,N))
IC2.lasso <- array(data=NA,dim = c(1,1,N))
IC1.adaplasso <- array(data=NA,dim = c(1,1,N))
IC2.adaplasso <- array(data=NA,dim = c(1,1,N))
sum.IC1.lasso <- array(data=NA,dim = c(1,1,N))
sum.IC2.lasso <- array(data=NA,dim = c(1,1,N))
sum.IC1.adaplasso <- array(data=NA,dim = c(1,1,N))
sum.IC2.adaplasso <- array(data=NA,dim = c(1,1,N))
pmse.ridge <- array(data=NA,dim = c(1,1,N))
pmse.lasso <- array(data=NA,dim = c(1,1,N))
pmse.adaplasso <- array(data=NA,dim = c(1,1,N))
```

*****สร้างรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Constant เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม**:

```
R1 <- matrix(rep(r,(15*15)),ncol=15,nrow = 15)
diag(R1) <- 1
CorMat1 <- matrix(c(R1),ncol = 15)
R2 <- matrix(rep(r,(p-15)*(p-15)),ncol= (p-15),nrow = (p-15))
diag(R2) <- 1
CorMat2 <- matrix(c(R2),ncol = p-15)
```

<u>หมายเหตุ</u> : *** ในแต่ละรูแปบบความสัมพันธ์ของตัวแปรอิสระ จะต้องสร้างเมทริกซ์ความแปรปรวนที่แตกต่างกันออกไป และเนื่องจากต้องการให้ตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม จึงสร้างเมทริกซ์ความแปรปรวน 2 ชุด นั่นคือ CorMat1 และ CorMat2

เช่นเดียวกับ กรณีรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Toeplitz เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม
จะใช้

```
CorMat1 <- outer(1:15, 1:15, function(x,y) {r^abs(x-y)})
CorMat2 <- outer(16:p, 16:p, function(x,y) {r^abs(x-y)})
```

และกรณีรูปแบบความสัมพันธ์ของตัวแปรอิสระแบบ Hub Toeplitz เมื่อตัวแปรอิสระแบ่งออกเป็น 2 กลุ่ม

```
R1 <- outer(1:15, 1:15, function(x,y) {rmax-(((rmax-rmin)/(15-2))*(abs(x-y)-
                                                    1))})
diag(R1) <- 1
CorMat1 <- matrix(c(R1),ncol = 15)

R2 <- outer(16:p, 16:p, function(x,y) {rmax-(((rmax-rmin)/((p-15)-2))*(abs(x-
                                                    y)-1))})
diag(R2) <- 1
CorMat2 <- matrix(c(R2),ncol = p-15)
```

สร้างเมทริกซ์ของตัวแปรอิสระ จากการแจกแจงปกติ โดยกำหนดให้เมทริกซ์ความแปรปรวน แตกต่างกันในแต่ละรูปแบบความสัมพันธ์

```
x1 <- mvrnorm(n, rep(0,15),CorMat1)
x2 <- mvrnorm(n, rep(0,p-15),CorMat2)
X <- cbind(x1,x2)
beta <- as.vector(c(1,-0.5,-0.5,0.1,0.1,
                    0.1,0.05,0.05,0.05,0.05,
                    0.01,0.01,0.01,0.01,0.01,rep(0,p-15)))
 for(i in 1:N)   {
```

<u>ขั้นตอนที่ 1</u> : สร้างข้อมูลค่าสังเกต $(y_i, \mathbf{X}_i)$

```
x[,,i] <- X[sample(nrow(X),size=n,replace = TRUE),]
e[,,i]  <- rnorm(0,1)
muS[,,i] <- exp(x[,,i]%*%beta+ e[,,i]  )
y[,,i] <- rpois(n=n, lambda = muS[,,i])
data <- data.frame(y=y[,,i], x=x[,,i])
```

<u>ขั้นตอนที่ 2</u> : แบ่งชุดข้อมูลค่าสังเกตออกเป็น 2 ชุด นั่นคือ test set และ training set

```
train_rows[,,i] <- matrix(sample(1:n,0.8*n),ncol = 1)

x.train[,,i] <- x[c(train_rows[,1,i]),,i]
x.test[,,i]  <- x[-c(train_rows[,1,i]),,i]

y.train[,,i] <- y[train_rows[,1,i],,i]
y.test[,,i]  <- y[-c(train_rows[,1,i]),,i]
```

<u>ขั้นตอนที่ 3</u> : หาพารามิเตอร์ปรับแต่ง (tuning parameter) ด้วยวิธี 5-fold cross validation

```
### 1. Ridge Regression
lam.ridge[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,alpha=0,standardize=TRUE,
                            family="poisson")$lambda.min
### 2. LASSO
lam.lasso[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,nfolds=5,alpha=1,
                            family="poisson")$lambda.min
### 3. Adaptive LASSO
#### Ridge Regression to screate the Adaptive Weights Vector
w3 <- 1/abs(matrix(coef(cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,standardize=TRUE,
                                  alpha=0,family="poisson"), s=lam.ridge[,,i])
```

```
                                        [, 1][2:(ncol(x)+1)] ))^1
w3[w3[,1] == Inf] <- 999999999
## Replacing values estimated as Infinite for 999999999
lam.adaplasso[,,i] <- cv.glmnet(x.train[,,i],y.train[,,i],nfolds=5,standardize=TRUE,
                                alpha=1, parallel=TRUE,family="poisson",
                                penalty.factor=w3)$lambda.min
```

<u>ขั้นตอนที่ 4</u> : หาค่าสัมประสิทธิ์การถดถอยในแต่ละวิธี

```
### 1. Ridge Regression
ridge[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],alpha=0,standardize=TRUE,
                                family="poisson",intercept =TRUE,nfolds=5),
                                x.test,s=lam.ridge[,,i],type="coefficients"))

### 2. LASSO
lasso[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,alpha=1,
                                family="poisson",intercept = TRUE,nfolds=5),
                                x.test,s=lam.lasso[,,i],type="coefficients"))

### 3. Adaptive LASSO
adaplasso[,,i] <- as.vector(predict(cv.glmnet(x.train[,,i],y.train[,,i],standardize=TRUE,
                                alpha=1,nfolds = 5,penalty.factor=w3,
                                family="poisson",intercept = TRUE),
                                x.test,s=lam.adaplasso[,,i],
                                type="coefficients"))
```

<u>ขั้นตอนที่ 5</u> : หาค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (PMSE) ในแต่ละวิธี

```
### 1. Ridge Regression
pmse.ridge[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%ridge[,,i])) %*% (y[,,i]-
```

```
                    exp(cbind(rep(1,n),x[,,i])%*%ridge[,,i])) )/n
### 2. LASSO
pmse.lasso[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%lasso[,,i])) %*% (y[,,i]-
                    exp(cbind(rep(1,n),x[,,i])%*%lasso[,,i])) )/n

### 3. Adaptive LASSO
pmse.adaplasso[,,i] <- ( t(y[,,i]-exp(cbind(rep(1,n),x[,,i])%*%adaplasso[,,i])) %*% (y[,,i]-
                        exp(cbind(rep(1,n),x[,,i])%*%adaplasso[,,i])) )/n
```

<u>ขั้นตอนที่ 6</u> : พิจารณาค่าประมาณสัมประสิทธิ์ถดถอยที่ได้ กับค่าพารามิเตอร์สัมสิทธิ์การถดถอยที่กำหนด (IC1,IC2)

```
beta.1[,,i] <- as.vector(c(1,-0.5,-0.5,0.1,0.1,
                            0.1,0.05,0.05,0.05,0.01,
                            0.01,0.01,0.01,0.01,0.01,rep(0,p-15)))

### 1. LASSO
IC1.lasso[,,i] <- length(which(beta.1[,,i] ==0 & lasso[,,i] !=0))
IC2.lasso[,,i] <- length(which(beta.1[,,i] !=0 & lasso[,,i] ==0))

### 2. Adaptive LASSO
IC1.adaplasso[,,i] <- length(which(beta.1[,,i]==0 & adaplasso[,,i] !=0))
IC2.adaplasso[,,i] <- length(which(beta.1[,,i] !=0 & adaplasso[,,i] ==0))

                    } #end for
```

<u>ขั้นตอนที่ 7</u> : หาค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) ในแต่ละวิธี

```
### 1. Ridge Regression
sim.pmse.ridge <- apply(pmse.ridge,2,median)
```

```
### 2. LASSO
sim.pmse.lasso <- apply(pmse.lasso,2,mean)

### 3. Adaptive LASSO
sim.pmse.adaplasso <- apply(pmse.adaplasso,2,mean)
```

<u>ขั้นตอนที่ 8</u> : คำนวณหาอัตราความผิดพลาดในการตรวจจับเชิงลบ (False Negative Rate: FNR) และอัตราความผิดพลาดในการตรวจจับเชิงบวก (False Positive Rate: FPR) ของวิธี LASSO และ Adaptive LASSO

```
### 1. LASSO
sum.IC1.lasso <- apply(IC1.lasso,2,sum)
sum.IC2.lasso <- apply(IC2.lasso,2,sum)

prob.IC1.lasso <-  sum.IC1.lasso /(15*N)
prob.IC2.lasso <-  sum.IC2.lasso  /((p-15)*N)

### 2. Adaptive LASSO

sum.IC1.adaplasso <- apply(IC1.adaplasso,2,sum)
sum.IC2.adaplasso <- apply(IC2.adaplasso,2,sum)

prob.CI1.adaplasso  <-  sum.CI1.adaplasso  /(15*N)
prob.CI2.adaplasso  <-  sum.CI2.adaplasso  /((p-15)*N)

cat(sim.pmse.ridge,'\t',sim.pmse.lasso,'\t',sim.pmse.adaplasso,'\t',
    prob.IC1.lasso,'\t', prob.IC2.lasso,'\t',prob.IC1.adaplasso,'\t',prob.IC2.adaplasso,'\n'
                        } # end GetConstant2Groups

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

ประวัติผู้เขียน

ชื่อ	นางสาวชุติกาญจน์ ชูสวัสดิ์
วันเดือนปีเกิด	8 ตุลาคม พ.ศ.2536
ประวัติการศึกษา	ปีการศึกษา 2558 วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ทุนการศึกษา	ปี พ.ศ. 2559 : ทุนบัณฑิตเรียนดี ประเภท ข คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ผลงานทางวิชาการ

ชุติกาญจน์ ชูสวัสดิ์ และ สุปราณี ลิสวัสดิ์. (มีนาคม 2561). “การเปรียบเทียบประสิทธิภาพของวิธีวิเคราะห์การถดถอยแบบริดจ์ แลซโซ และแลซโซแบบปรับปรุง ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง.” การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติ ครั้งที่ 19, มหาวิทยาลัยขอนแก่น, ขอนแก่น, 9 มีนาคม 2561