การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ในตัวแบบการถดถอยลอจิสติกภายใต้ข้อมูลที่มีมิติสูงแบบบางเบาและตัวแปรทำนายมีความสัมพันธ์กันสูง

Performance Comparison of Penalized Regression Method in Logistic Regression for High-dimensional Sparse Data with Multicollinearity

สุปราณี ลิสวัสดิ์, วรางคณา วัชรเสถียร และเบญจมาศ ตุลยนิติกุล*

สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Supranee Lisawadi, Warangkhana Watcharasatian and Benjamas Tulyanitikul*

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Rangsit Centre, Khlong Nueng, Khlong Luang, Pathum Thani 12120

Received: August 12, 2020; Accepted: November 10, 2020

บทคัดย่อ

ปัจจุบันเทคโนโลยีได้รับการพัฒนาเป็นอย่างมาก ความเจริญก้าวหน้าทางด้านเทคโนโลยีส่งผลต่อกระบวนการทางข้อมูล ทำให้การเก็บรวบรวมข้อมูลมีประสิทธิภาพมากขึ้น ดังนั้นในการวิเคราะห์ข้อมูล นักวิเคราะห์ข้อมูลจึงได้มีการค้นหาวิธีการที่เหมาะสมเพื่อวิเคราะห์ข้อมูลขนาดใหญ่ นักวิเคราะห์นิยมใช้วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ในการวิเคราะห์ข้อมูลที่มีขนาดใหญ่และมีจำนวนตัวแปรเป็นจำนวนมาก ซึ่งวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์เป็นวิธีการหนึ่งที่ใช้ประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย การคัดเลือกตัวแปรเข้าสู่ตัวแบบ และการแก้ไขปัญหาตัวแปรมีความสัมพันธ์เชิงเส้นกัน ในการศึกษาครั้งนี้ได้พิจารณาการประมาณค่าสำหรับตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีมิติสูง ($n < p$) แบบบางเบาและตัวแปรทำนายมีความสัมพันธ์กันสูง โดยพิจารณาตัวประมาณจากวิธีการถดถอยแบบพีนอลไลซ์ ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง ซึ่งตัวประมาณทั้งสามสามารถใช้ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยในข้อมูลที่มีมิติสูงและสามารถแก้ปัญหาตัวแปรทำนายมีความสัมพันธ์กันสูง โดยพิจารณาเปรียบเทียบด้วยค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) จากการจำลองข้อมูลด้วยวิธีมอนติคาร์โล ซึ่งการศึกษาข้อมูลดังกล่าวสามารถสรุปว่าตัวประมาณแลซโซแบบปรับปรุงมีประสิทธิภาพที่ดีที่สุดเมื่อเปรียบเทียบจากทั้งสามตัวประมาณ

คำสำคัญ : วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์; ข้อมูลมิติสูงแบบบางเบา; ตัวแปรทำนายมีความสัมพันธ์กันสูง; ตัวประมาณแบบริดจ์; ตัวประมาณแลซโซ; ตัวประมาณแลซโซแบบปรับปรุง

**Corresponding author: benjamas@mathstat.sci.tu.ac.th*

Abstract

Nowadays, technology is widely developed. The growth in high technology affects data science processes. One effect is that more data can be collected in a shorter time than before. This can be used in analyses. Analysts need to find an appropriate method to analyze the extensive data. The analyst should use the proper methodology for data of considerable size and high dimensions. One approach is penalized regression. That is a method for estimated coefficient parameters, variable selection, and the multicollinearity problem when the predictor variables are correlated. This study considers estimations for a logistic regression model with high-dimensional sparse data ($n < p$) and high correlation. We apply estimators from the penalized regression method: ridge regression, LASSO, and adaptive LASSO. These can be used to estimate coefficient parameters in high-dimensional data and could solve the multicollinearity problem. We compared the performance of these estimators. The performance in terms of the mean of prediction mean square error (mPMSE) using Monte Carlo simulation. The result showed that the adaptive LASSO estimator has the lowest mPMSE. Overall, adaptive LASSO performed better than ridge regression or LASSO.

Keywords: penalized regression; high-dimensional data; high-correlation; ridge regression; LASSO; adaptive LASSO

1. คำนำ

การวิเคราะห์การถดถอย (regression analysis) เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตอบสนองกับตัวแปรอิสระหรือตัวแปรทำนาย ในอดีตการวิเคราะห์การถดถอยมีเงื่อนไขว่าขนาดตัวอย่างจะต้องมากกว่าจำนวนตัวแปรอิสระ แต่เนื่องด้วยปัจจุบันมีความก้าวหน้าทางเทคโนโลยี มีเทคโนโลยีที่ทันสมัยทำให้การจัดเก็บข้อมูลมีการพัฒนาไปอย่างรวดเร็วและมีประสิทธิภาพมากขึ้น ในการวิเคราะห์ข้อมูลผู้วิเคราะห์จึงอาจต้องเผชิญกับข้อมูลที่มีขนาดใหญ่และมีความซับซ้อนมากขึ้น บ่อยครั้งที่ข้อมูลเหล่านั้นมีจำนวนตัวแปรมากกว่าขนาดตัวอย่าง หรือเรียกว่าข้อมูลมีมิติสูง (high-dimensional data) ตัวอย่าง เช่น ข้อมูลทางดาราศาสตร์ที่ได้จากกล้องโทรทรรศน์ที่ใช้เทคโนโลยีขั้นสูง ทำให้ได้ข้อมูลจากภาพของวัตถุที่ได้รับการวัดค่าตัวแปรในจำนวนหลักสิบหรือหลักร้อยตัวแปร ซึ่งถือเป็นตัวแปรที่มีจำนวนมากหากนำไปใช้ในการจำแนกประเภทของวัตถุ หรือข้อมูลสำหรับการจำแนกประเภทเอกสารเป็นข้อมูลที่มีจำนวนมิติมากโดยตัวแปรที่เป็นไปได้ในการใช้จำแนกเอกสาร คือ คำ หรือวลี ซึ่งมีจำนวนมากในแต่ละเอกสาร นอกจากนี้ข้อมูลปัจจุบันจำนวนมากมีรูปแบบของข้อมูลเป็นแบบทวิภาค (binary data) จึงจำเป็นต้องใช้ตัวแบบการถดถอยลอจิสติกในการวิเคราะห์ข้อมูล และกรณีที่เป็นข้อมูลแบบทวิภาคที่มีขนาดใหญ่ จึงจำเป็นที่จะต้องใช้วิธีการที่เหมาะสมที่สามารถวิเคราะห์ข้อมูลดังกล่าวนี้ได้ งานวิจัยนี้จึงสนใจศึกษาเพื่อเปรียบเทียบวิธีการที่เหมาะสมที่ใช้สำหรับวิเคราะห์ข้อมูลที่มีขนาดใหญ่หรือมีมิติสูงโดยใช้ตัวแบบการถดถอยลอจิสติก ตัวอย่างข้อมูลขนาดใหญ่ที่ใช้ตัวแบบการถดถอยลอจิสติก เช่น ข้อมูลไมโครอาร์เรย์ ซึ่งเป็นข้อมูลที่ได้จากการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน โดยยีนที่ศึกษามีจำนวนเป็นหลักพันหรือหลักหมื่น สามารถนำมาใช้

ในการจำแนกเนื้อเยื่อมะเร็งและเนื้อเยื่อปกติ หรือข้อมูลทางวิศวกรรมศาสตร์ เช่น การเก็บตัวอย่างดิน เป็นการศึกษาคุณภาพของดินเพื่อนำไปใช้ประโยชน์ในการก่อสร้าง ทำถนน หรือการสร้างเขื่อน ข้อมูลที่ได้จากตัวอย่างดิน ได้แก่ ค่าแร่ธาตุในดิน ค่าความชื้นในมวลดิน ค่าการรับน้ำหนักของดิน ค่าการไหลซึมน้ำผ่านมวลดิน ค่าความหนาแน่นของดิน ค่าการบดอัด และค่าอื่น ๆ ซึ่งนำไปใช้ในการพิจารณาการนำไปใช้ได้หรือไม่ ซึ่งวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์สามารถใช้ในการวิเคราะห์กับข้อมูลที่มีขนาดใหญ่และตัวแปรมีความสัมพันธ์กันสูง งานวิจัยนี้เป็นการศึกษาและเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอย สำหรับกรณีที่ข้อมูลมีมิติสูงแบบบางเบา ในตัวแบบการถดถอยลอจิสติก

2. วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ ในตัวแบบการถดถอยลอจิสติก กรณีที่ข้อมูลมีมิติสูงแบบบางเบา ซึ่งวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ที่จะเปรียบเทียบมีด้วยกัน 3 วิธี ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (ridge regression) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง (adaptive LASSO) และพิจารณาร่วมกับกรณีที่ตัวแปรทำนายมีความสัมพันธ์กันสูง ซึ่งมี 3 รูปแบบ ได้แก่ ความสัมพันธ์แบบคงที่ (constant correlation) ความสัมพันธ์แบบโทพลิท (Toeplitz correlation) และความสัมพันธ์แบบฮับโทพลิท (Hub Toeplitz correlation)

3. ขอบเขตของการศึกษา

งานวิจัยครั้งนี้ศึกษาในตัวแบบการถดถอยลอจิสติก กรณีที่ข้อมูลมีมิติสูงและตัวแปรมีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปร โดยกำหนดขนาดตัวอย่าง 50 และ 100 ที่มีจำนวนตัวแปร 100, 200, 500 และตัวแปรทำนายมีความสัมพันธ์กันที่ระดับ 0.5, 0.6 ,0.7, 0.8, 0.9 ด้วยความสัมพันธ์ 3 รูปแบบ ได้แก่ ความสัมพันธ์แบบคงที่ ความสัมพันธ์แบบโทพลิท และความสัมพันธ์แบบฮับโทพลิท ในตัวแปรทำนาย 1 กลุ่ม และ 2 กลุ่ม และศึกษากรณีที่เป็นแบบบางเบา จึงกำหนดให้ค่าสัมประสิทธิ์การถดถอย q ตัวแรกมีค่าไม่เท่ากับศูนย์ และ p-q ตัวมีค่าเท่ากับศูนย์

4. เกณฑ์ที่ใช้พิจารณา

พิจารณาประสิทธิภาพของการพยากรณ์ด้วยค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าพยากรณ์ (mean of prediction mean square error, mPMSE)

$$\text{mPMSE} = \frac{1}{1,000}\sum_{i=1}^{1000}\sum_{j=1}^{n}\frac{\left(y_j^{(i)} - y_j^{(i)}\right)^2}{n}$$

และพิจารณาประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ จะพิจารณาที่อัตราความผิดพลาดในการตรวจจับเชิงบวก (false positive rate, FPR) และอัตราความผิดพลาดในการตรวจจับเชิงลบ (false negative rate, FNR) ซึ่ง FPR คือ ความผิดพลาดกรณีที่ค่าพารามิเตอร์ไม่เท่ากับศูนย์ แต่ตัวประมาณพารามิเตอร์มีค่าเท่ากับศูนย์ (ตัวแปรทำนายไม่ถูกคัดเลือกเข้าในตัวแบบ) และ FNR คือ ความผิดพลาดกรณีที่ค่าพารามิเตอร์เท่ากับศูนย์ แต่ตัวประมาณพารามิเตอร์มีค่าไม่เท่ากับศูนย์

ความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปร เป็นดังนี้

$$\text{FNR} = \frac{\text{IC1}}{\text{q}\times\text{m}} \quad \text{และ} \quad \text{FPR} = \frac{\text{IC2}}{(\text{p}-\text{q})\times\text{m}}$$

เมื่อ q คือ จำนวนพารามิเตอร์สัมประสิทธิ์การถดถอย ที่มีค่าไม่เท่ากับศูนย์; m คือ จำนวนครั้งของการจำลองข้อมูล

5. การวิเคราะห์การถดถอยแบบพีนอลไลซ์

วิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์เป็นวิธีการทางสถิติที่ใช้หาค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอยอีกวิธีหนึ่ง โดยมีจุดมุ่งหมายหลัก คือ เพื่อประมาณค่าสัมประสิทธิ์การถดถอยที่ใช้ได้กับข้อมูลที่มีมิติสูง และลดปัญหาการเกิดภาวะร่วมเชิงเส้น (multicollinearity) ได้ โดยค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอย จะหาได้จากการหาค่าประมาณ β ที่ทำให้ฟังก์ชันเป้าหมาย (objective function) $\hat{\boldsymbol{\beta}} = \arg\min_{\beta}\left(-l\left(\boldsymbol{\beta}\right)\right) + P_{\lambda}\left(\boldsymbol{\beta}\right)$ มีค่าต่ำที่สุด จากฟังก์ชันเป้าหมายจะเห็นว่าสมการมีความคล้ายคลึงกับฟังก์ชันภาวะน่าจะเป็นสูงสุดที่ใช้กันโดยทั่วไปในการหาค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอย β แต่จะมีส่วนที่แตกต่างกันคือมี $P_{\lambda}(\boldsymbol{\beta})$ เพิ่มขึ้นมา ซึ่งเรียกว่าฟังก์ชันพีนอลที (penalty function) โดยมีพารามิเตอร์ λ ซึ่งมีค่ามากกว่าหรือเท่ากับศูนย์ โดยทั่วไปจะใช้วิธี cross-validation ในการหาค่า λ ที่เหมาะสมสำหรับข้อมูลที่ต้องการวิเคราะห์ ซึ่งฟังก์ชันพีนอลทีนั้นจะมีหลายรูปแบบที่ต่างกันไป ดังนี้

วิธีการถดถอยแบบริดจ์(Hoerl and Kennard, 1970) เป็นวิธีการหนึ่งของการวิเคราะห์การถดถอยแบบพีนอลไลซ์ โดยมีฟังก์ชันพีนอลที ดังนี้

$$P_{\lambda}\left(\boldsymbol{\beta}\right) = \lambda\sum_{i=1}^{p}\beta_i^2$$

ซึ่งฟังก์ชันพีนอลทีดังกล่าวจะใช้วิธีการดึงค่าประมาณเข้าสู่ศูนย์กลาง ทำให้ได้ค่า $\hat{\boldsymbol{\beta}}$ ทุกตัวมีขนาดเล็ก ดังนั้นวิธีการถดถอยแบบริดจ์จะเป็นวิธีที่เหมาะสมสำหรับการวิเคราะห์ข้อมูลในลักษณะที่ตัวแปรทำนายทุกตัวมีความสัมพันธ์กับตัวแปรตอบสนอง ซึ่งตัวประมาณพารามิเตอร์สัมประสิทธิ์การถดถอยที่ได้จะมีความเสถียร จึงเป็นอีกวิธีหนึ่งที่นิยมใช้เพื่อแก้ไขปัญหาตัวแปรทำนายมีความสัมพันธ์กันสูง หรือเกิดปัญหาภาวะร่วมเชิงเส้น แต่วิธีนี้อาจขาดสมบัติในการคัดเลือกตัวแปร ทำให้การแปลผลในตัวแบบทำได้ยาก

วิธีการถดถอยแบบแลซโซ (least absolute shrinkage and selection operator, LASSO) (Tibshirani, 1996) คิดค้นโดย Robert Tibshirani โดยมีวัตถุประสงค์เพื่อให้คัดเลือกตัวแปรเข้าสู่ตัวแบบได้ รวมถึงสามารถประมาณค่าในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงในคราวเดียวกัน โดยวิธีการถดถอยแบบแลซโซมีฟังก์ชันพีนอลทีดังนี้

$$P_{\lambda}\left(\boldsymbol{\beta}\right) = \lambda\sum_{i=1}^{p}\left|\beta_i\right|$$

การประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยตัวประมาณที่ได้จากวิธีแลซโซจะมีค่า $\hat{\boldsymbol{\beta}}$ ส่วนใหญ่เป็นศูนย์และค่า $\hat{\boldsymbol{\beta}}$ บางส่วนไม่เท่ากับศูนย์ (sparse estimator) ดังนั้นวิธีแลซโซนี้จึงเป็นวิธีการที่สามารถคัดเลือกตัวแปรได้โดยอัตโนมัติ และมีวิธีการคำนวณที่ไม่ยุ่งยาก จึงทำให้วิธีการนี้ได้รับความนิยม และปัจจุบันมีโปรแกรมสถิติหลายโปรแกรมที่มีฟังก์ชันในการคำนวณตัวประมาณ LASSO อย่างไรก็ตาม วิธีการถดถอยแบบแลซโซยังมีข้อจำกัดในการวิเคราะห์ข้อมูลที่มีมิติสูง เนื่องจากเลือกตัวแปรเข้าได้มากสุดจำนวน n ตัว ดังนั้นหากข้อมูลมีจำนวนตัวแปรทำนายมากกว่าขนาดตัวอย่างเป็นจำนวนมาก หรือตัวแปรทำนายมีความสัมพันธ์กันสูง วิธีการถดถอยแบบแลซโซมีแนวโน้มที่จะเลือกตัวแปรทำนายเข้าสู่ตัวแบบเพียงตัวเดียวจากกลุ่มตัวแปรทำนายที่มีความสัมพันธ์กันสูง โดยไม่สนว่าจะเป็นตัวแปรใดในกลุ่ม

วิธีการถดถอยแลซโซแบบปรับปรุง (adaptive LASSO) (Zou, 2006) พัฒนามาจากการถดถอยแบบแลซโซ โดย Hui Zou นำเสนอวิธีนี้ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย เพื่อแก้ไขข้อจำกัดของวิธีการถดถอยแบบแลซโซ ให้

มีสมบัติในการคัดเลือกตัวแปรที่แม่นยำและมีประสิทธิภาพมากขึ้น โดยมีการถ่วงน้ำหนักให้กับตัวแปรทำนายที่คัดเลือกเข้าสู่ตัวแบบ ซึ่งทำให้เกิดความคงเส้นคงวา อีกทั้งวิธีการถดถอยแบบแลซโซแบบปรับปรุงยังมีสมบัติเช่นเดียวกับวิธีการถดถอยแบบริดจ์ ซึ่งทำให้ตัวประมาณที่ได้จากวิธีการถดถอยแบบแลซโซแบบปรับปรุงเป็นตัวประมาณที่ไม่เอนเอียง และสามารถลดความแปรปรวนลงได้ โดยวิธีการถดถอยแบบแลซโซแบบปรับปรุง มีฟังก์ชันพีนอลที คือ $P_\lambda(\boldsymbol{\beta}) = \lambda\sum_{i=1}^{p}|\beta_i|w_i$ โดยที่ w_i คือ การถ่วงน้ำหนักให้กับตัวแปร (adaptive weight) นิยามโดย $w_i = \left|\beta_i\right|^{-\gamma}$ เมื่อ $\gamma > 0$ และ β_i คือ ค่าพารามิเตอร์สัมประสิทธิ์ที่ได้จากวิธีกำลังสองน้อยที่สุด ซึ่งหากไม่สามารถหาค่าจากวิธีดังกล่าวสามารถใช้ค่าประมาณสัมประสิทธิ์ที่ได้จากวิธีการถดถอยแบบริดจ์แทน

ตารางที่ 1 เปรียบเทียบข้อดีข้อเสียของการประมาณพารามิเตอร์ของการวิเคราะห์การถดถอยแบบพีนอลไลซ์แต่ละวิธี

	ข้อดี	ข้อเสีย
Ridge	ตัวประมาณที่ได้มีความเสถียรและสามารถแก้ไขปัญหาตัวแปรมีความสัมพันธ์เชิงเส้นกัน	ขาดสมบัติในการเลือกตัวแปรเข้าสู่ตัวแบบ ทำให้แปรผลได้ยาก
ASSO	สามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่าได้ในคราวเดียวกัน อีกทั้งสามารถลดความแปรปรวน	ขาดสมบัติคงเส้นคงวา กรณีที่ตัวทำนายมีความสัมพันธ์เชิงเส้นสูง วิธี LASSO จะเลือกตัวแปรเข้าเพียงตัวแปรเดียวจากกลุ่มตัวแปรทำนายที่มีความสัมพันธ์เชิงเส้นสูงโดยไม่สนใจว่าจะเป็นตัวแปรใดในกลุ่ม
Adaptive LASSO	ลดความเอนเอียงในการคัดเลือกตัวแปร เมื่อเปรียบเทียบกับวิธี LASSO แบบเดิม	

6. รูปแบบความสัมพันธ์ของตัวแปรทำนาย

6.1 ความสัมพันธ์แบบคงที่

เป็นรูปแบบความสัมพันธ์ของตัวแปรทำนายคู่ใด ๆ ที่มีความสัมพันธ์เท่ากันหมด ซึ่งสามารถเขียนในรูปเมทริกซ์ ได้ดังนี้

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \rho & \dots & \rho \\ \rho & \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \rho & \dots & 1 \end{bmatrix}$$

โดยที่ Σ คือ เมทริกซ์สหสัมพันธ์ขนาด $p \times p$; ρ มีค่าอยู่ระหว่าง 0 ถึง 1

6.2 ความสัมพันธ์แบบโทพลิท

เป็นความสัมพันธ์ของตัวแปรทำนายที่อยู่ใกล้เคียงกันจะมีความสัมพันธ์กันสูง แต่ตัวแปรทำนายที่อยู่ไกลกันจะมีความสัมพันธ์กันน้อยลง

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{p-3} \\ \rho^3 & \rho^2 & \rho & 1 & \dots & \rho^{p-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \rho^{p-4} & \dots & 1 \end{bmatrix}$$

โดยที่ Σ คือ เมทริกซ์สหสัมพันธ์ขนาด $k \times k$; ρ คือ ความสัมพันธ์ของตัวแปรทำนาย โดยมีค่าอยู่ระหว่าง $0 \le \rho \le 1$

6.3 ความสัมพันธ์แบบฮับโทพลิท

เป็นความสัมพันธ์ของตัวแปรทำนาย ในแถวและหลักที่ 1 ของเมทริกซ์สหสัมพันธ์ที่ได้จะเรียงจากมากไปน้อย ซึ่งอธิบายความสัมพันธ์ของตัวแปรทุกตัวได้หมดภายในแถวหรือหลักที่ 1

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

7. วิธีการดำเนินงานวิจัย

การศึกษาครั้งนี้ใช้วิธีการจำลองข้อมูลด้วยวิธีมอนติคาร์โล เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณการถดถอยแบบพีนอลไลซ์ทั้งสามตัวประมาณ ได้แก่ ตัวประมาณแบบริดจ์ ตัวประมาณแลซโซ และตัวประมาณแลซโซแบบปรับปรุง โดยพิจารณาจากค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ (mPMSE) จากการจำลองซ้ำทั้งหมด 1,000 ครั้ง ซึ่งค่า mPMSE สามารถคำนวณได้จาก $\text{PMSE}_v = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n_\lambda}$ โดยการจำลองจะใช้ตัวแบบการถดถอยลอจิสติกที่มีความสัมพันธ์สามรูปแบบ และมีขนาดตัวอย่าง คือ 50 และ 100 และมีจำนวนตัวแปร คือ 100, 200 และ 500 และมีค่าสหสัมพันธ์ คือ 0.5, 0.6, 0.7, 0.8, 0.9 ในกรณีที่เป็นหนึ่งกลุ่มและสองกลุ่ม ตามลำดับ

8. ผลการวิจัย

8.1 กรณีที่ตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบคงที่ ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 2 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายมี 1 กลุ่ม และมีความสัมพันธ์แบบคงที่

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.22540	0.18034	0.17093*	0.22603	0.18749	0.18544*
	200	0.20964	0.17701	0.16761*	0.22356	0.18835	0.17294*
	500	0.20731	0.18447	0.16153*	0.20693	0.18638	0.16171*
0.6	100	0.22431	0.18499	0.17804*	0.23008	0.19071*	0.19282
	200	0.21857	0.18487	0.17295*	0.22459	0.18866	0.17917*
	500	0.20736	0.18815	0.16662*	0.21285	0.18950	0.16976*
0.7	100	0.23027	0.19523	0.18907*	0.23333	0.19431	0.19423*
	200	0.21983	0.18793	0.18066*	0.22586	0.19292	0.18676*
	500	0.21318	0.19218	0.17409*	0.21810	0.19463	0.17335*
0.8	100	0.22987	0.19638	0.19340*	0.23524	0.19545*	0.20164
	200	0.22555	0.19179	0.18861*	0.22997	0.19932	0.19816*
	500	0.21642	0.19162	0.17842*	0.22044	0.19650	0.18190*
0.9	100	0.22772	0.20103	0.19075*	0.23528	0.20286*	0.21188
	200	0.22941	0.20005	0.19412*	0.23110	0.20232	0.20123*
	500	0.22371	0.19814	0.19037*	0.22814	0.20527	0.19534*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

ตารางที่ 2 เป็นการพิจารณาประสิทธิภาพของการพยากรณ์ ในกรณีที่ตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบคงที่ ที่ระดับความสัมพันธ์ 0.50, 0.6, 0.7, 0.8 และ 0.9 พบว่าวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุดในเกือบทุกกรณี ยกเว้นกรณีที่ n = 100 และ p = 100 ที่ระดับความสัมพันธ์ 0.6, 0.8 และ 0.9 ซึ่งกรณีดังกล่าววิธีการวิเคราะห์การถดถอยแบบแลซโซจะให้ค่า mPMSE ที่ต่ำกว่าวิธีอื่น ๆ

8.2 กรณีที่ตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบโทพลิท ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 3 พบว่าค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ เมื่อตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบโทพลิท โดยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าต่ำที่สุดในแต่ละระดับความสัมพันธ์ ยกเว้นกรณีที่ n = 100 ที่ระดับความสัมพันธ์ 0.9 เมื่อ p = 100 วิธีการวิเคราะห์การถดถอยแบบแลซโซจะให้ค่า mPMSE ที่ต่ำที่สุด

ตารางที่ 3 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบโทพลิท

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.21888	0.18349	0.17013*	0.22299	0.017839	0.16986*
	200	0.20671	0.18228	0.1603*	0.20996	0.17472	0.16250*
	500	0.19467	0.1748	0.15262*	0.2039	0.18588	0.15684*
0.6	100	0.22149	0.1869	0.17466*	0.22586	0.18347	0.17952*
	200	0.21052	0.18314	0.16516*	0.21565	0.18417	0.16984*
	500	0.2007	0.1849	0.15884*	0.19998	0.17861	0.15543*
0.7	100	0.22385	0.18887	0.1779*	0.22745	0.18546	0.18417*
	200	0.21104	0.18893	0.17363*	0.21922	0.18377	0.17631*
	500	0.20376	0.19016	0.16172*	0.20613	0.18418	0.16453*
0.8	100	0.22370	0.18925	0.18317*	0.23134	0.19556	0.19565*
	200	0.21235	0.19007	0.17466*	0.22229	0.18951	0.18332*
	500	0.20628	0.19050	0.16834*	0.20745	0.18619	0.16590*
0.9	100	0.2264	0.19337	0.19165*	0.23176	0.19681*	0.20479
	200	0.21690	0.19145	0.18051*	0.22725	0.19735	0.195936*
	500	0.20742	0.18933	0.17375*	0.21210	0.19771	0.18097*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

8.3 กรณีที่ตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบฮับโทพลิท ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 4 พบว่าค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ เมื่อตัวแปรทำนายมี 1 กลุ่ม และมีรูปแบบความสัมพันธ์แบบฮับโทพลิท โดยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าต่ำที่สุด ในแต่ละระดับความสัมพันธ์ ยกเว้นกรณีที่ n = 100 ที่ระดับความสัมพันธ์ 0.9 เมื่อ p = 100 วิธีการวิเคราะห์การถดถอยแบบแลซโซจะให้ค่า mPMSE ที่ต่ำที่สุด

8.4 กรณีที่ตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบคงที่ ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 5 พบว่าค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ เมื่อตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบคงที่ โดยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าต่ำที่สุดในแต่ละระดับความสัมพันธ์ ยกเว้นกรณีที่มีระดับความสัมพันธ์ 0.9 เมื่อ n = 50, p = 100 และ n = 100, p = 100 วิธีการวิเคราะห์การถดถอยแบบแลซโซจะให้ค่า mPMSE ที่ต่ำที่สุด

ตารางที่ 4 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายแบ่งออกเป็น 1 กลุ่ม และมีความสัมพันธ์แบบฮับโทพลิท

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.21964	0.18514	0.17423*	0.22299	0.17839	0.16986*
	200	0.20470	0.18011	0.16382*	0.20996	0.17472	0.16250*
	500	0.20123	0.18124	0.16029*	0.20394	0.18588	0.15684*
0.6	100	0.22124	0.18886	0.17732*	0.22586	0.18347	0.17952*
	200	0.21008	0.18245	0.16669*	0.21565	0.18417	0.16984*
	500	0.20068	0.18674	0.15978*	0.19998	0.17861	0.15543*
0.7	100	0.22281	0.19001	0.17889*	0.22745	0.18546	0.18417*
	200	0.21104	0.18893	0.17363*	0.21922	0.18377	0.17631*
	500	0.20376	0.19016	0.16172*	0.20613	0.18418	0.16453*
0.8	100	0.22371	0.18925	0.18317*	0.23134	0.19556	0.19565*
	200	0.21235	0.19007	0.17466*	0.22222	0.18951	0.18332*
	500	0.20628	0.19050	0.16834*	0.20745	0.18619	0.16590*
0.9	100	0.22648	0.19337	0.19165*	0.23176	0.19681*	0.20479
	200	0.21690	0.19145	0.18051*	0.22725	0.19735	0.19593*
	500	0.20742	0.18932	0.17375*	0.21210	0.19771	0.18097*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

ตารางที่ 5 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบคงที่

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.22185	0.18075	0.17220*	0.22960	0.18991	0.18173*
	200	0.21497	0.18287	0.16714*	0.22136	0.18867	0.17194*
	500	0.20672	0.18548	0.15933*	0.20910	0.19008	0.16278*
0.6	100	0.22236	0.18066	0.17543*	0.23059	0.19219	0.18663*
	200	0.21589	0.18339	0.16920*	0.22174	0.19179	0.17661*
	500	0.20489	0.18291	0.16130*	0.21014	0.19159	0.16478*
0.7	100	0.22710	0.18615	0.18028*	0.23326	0.19618	0.19304*
	200	0.21903	0.18216	0.17169*	0.22780	0.19465	0.18494*
	500	0.20873	0.18427	0.16451*	0.21601	0.19427	0.16999*
0.8	100	0.23297	0.19498	0.19278*	0.23448	0.19812	0.19747*
	200	0.22765	0.19354	0.18442*	0.22919	0.19814	0.18685*
	500	0.21802	0.19398	0.17731*	0.21829	0.19696	0.17505*
0.9	100	0.22964	0.19363*	0.19443	0.23607	0.20365*	0.20525
	200	0.22704	0.19024	0.18690*	0.23248	0.20365	0.19537*
	500	0.21944	0.19239	0.18188*	0.22615	0.20310	0.18530*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

8.5 กรณีที่ตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบโทพลิท ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 6 พบว่าค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ เมื่อตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบโทพลิท โดยส่วนมากวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าที่สุด

8.6 กรณีที่ตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบฮับโทพลิท ที่ระดับความสัมพันธ์ต่าง ๆ

ตารางที่ 7 พบว่าค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ เมื่อตัวแปรทำนายมี 2 กลุ่ม และมีรูปแบบความสัมพันธ์แบบฮับโทพลิท โดยวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุงจะให้ค่าต่ำที่สุด ในแต่ละระดับความสัมพันธ์ ยกเว้นกรณี n = 50, p = 100 และที่ระดับความสัมพันธ์ 0.5, 0.6 และ 0.9 และกรณี n = 100, p = 100 วิธีการวิเคราะห์การถดถอยแบบแลซโซจะให้ค่า mPMSE ที่ต่ำที่สุด

ตารางที่ 6 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบโทพลิท

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.20743	0.18809	0.16046*	0.21764	0.14789*	0.14989
	200	0.20297	0.15478	0.13855*	0.20489	0.16935	0.13422*
	500	0.20339	0.18705	0.14188*	0.19699	0.19303	0.15162*
0.6	100	0.22427	0.18701	0.16656*	0.22076	0.17621*	0.18102
	200	0.20308	0.18623	0.18047*	0.22652	0.18609	0.17673*
	500	0.18615	0.19459	0.17143*	0.20443	0.17896	0.14923*
0.7	100	0.22291	0.19176	0.17086*	0.23262	0.18331*	0.18460
	200	0.20794	0.17422	0.17166*	0.21462	0.18278	0.16597*
	500	0.18414	0.18821	0.16576*	0.20420	0.19446	0.17695*
0.8	100	0.22745	0.19391	0.18026*	0.22766	0.19281*	0.19403
	200	0.19925	0.15462	0.14728*	0.21965	0.18903	0.17289*
	500	0.20268	0.19117	0.165272*	0.20914	0.18323	0.16601*
0.9	100	0.22012	0.17693*	0.18293	0.23199	0.19714*	0.19796
	200	0.21390	0.17742	0.16758*	0.22387	0.18411	0.18239*
	500	0.20100	0.18384	0.18260*	0.20087	0.18692	0.16730*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

9. สรุป

การประมาณค่าสำหรับตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีมิติสูง แบบบางเบา และตัวแปรทำนายมีความสัมพันธ์กันสูง โดยพิจารณาตัวประมาณจากวิธีการถดถอยแบบพีนอลไลซ์ ได้แก่ วิธีการวิเคราะห์การถดถอยแบบริดจ์ วิธีการวิเคราะห์การถดถอยแบบแลซโซ และวิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง โดยพิจารณาเปรียบเทียบด้วยค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ จากการจำลองข้อมูลด้วยวิธีมอนติคาร์โล สามารถสรุปว่าตัวประมาณแลซโซแบบปรับปรุงมีประสิทธิภาพที่ดีที่สุดเมื่อเปรียบเทียบจากทั้งสามตัวประมาณ ทั้งนี้มีข้อสังเกตว่ากรณีที่ n = p = 100 พบว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซจะมีประสิทธิภาพดีกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง ทั้งนี้เนื่องมาจากขนาดของ n ไม่ได้น้อยกว่าขนาดของ p เหมือนกรณี อื่น ๆ (ตารางที่ 8)

ตารางที่ 7 ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ของแต่ละวิธี เมื่อตัวแปรทำนายแบ่งออกเป็น 2 กลุ่ม และมีความสัมพันธ์แบบฮับโทพลิท

r	p	n = 50			n = 100		
		Ridge	LASSO	Adaptive LASSO	Ridge	LASSO	Adaptive LASSO
0.5	100	0.23235	0.19582*	0.19663	0.23448	0.20153*	0.204481
	200	0.22776	0.19173*	0.19201	0.23033	0.20168	0.196508*
	500	0.22072	0.19485	0.18683*	0.22397	0.20202	0.189589*
0.6	100	0.23254	0.19659*	0.19670	0.23420	0.20236*	0.205008
	200	0.22765	0.19422	0.19215*	0.23016	0.20100	0.196648*
	500	0.22096	0.19484	0.18422*	0.22391	0.20097	0.187323*
0.7	100	0.23237	0.19343*	0.19703	0.21630	0.19846	0.17788*
	200	0.22774	0.19145	0.19086*	0.22967	0.20200	0.196227*
	500	0.22044	0.19637	0.18482*	0.22391	0.20107	0.186288*
0.8	100	0.23250	0.19437*	0.19783	0.21562	0.19859	0.176905*
	200	0.22768	0.19212	0.19037*	0.22956	0.20257	0.194111*
	500	0.22056	0.19425	0.18285*	0.22362	0.20115	0.186467*
0.9	100	0.23235	0.19582*	0.19663	0.23448	0.20153*	0.204481
	200	0.22776	0.19173*	0.19201	0.23033	0.20168	0.196508*
	500	0.22072	0.19485	0.18683*	0.22397	0.20202	0.189589*

* แทนวิธีที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ต่ำที่สุด

10. รายการอ้างอิง

Hardin, J., Garcia, S.R. and Golan, D., 2013, A method for generating realistic correlation matrices, Ann. Appl. Stat. 7: 1733-1762.

Hoerl, A.E. and Kennard, R.W., 1970, Ridge regression: Biased estimation for non-orthogonal problems, Technometrics 12: 55-67.

Hossain, S. and Ahmed, S.E., 2012, Shrinkage and penalty estimators of a Poisson regression model, Aust. N. Z. J. Stat. 54: 359-373.

Hossain, S. and Ahmed, S., 2014, Shrinkage estimation and selection for a logistic regression model, CRM Proc. Contemp. Math. 622: 159-176.

Honboonherm, O. and Pungpapong, V., 2013, Empirical bayes variable selection and estimation for the COX's proportional hazard model with high dimensional data, The 4th Hatyai National Conference, Hatyai University, Songkhla. (in Thai)

Pungpapong, V., 2015, A brief review on high-dimensional linear regression, Thai Sci. Technol. J. 23(2): 212-223. (in Thai)

Sarakor, T. and Kulvanich, N., 2014, Comparing the

ตารางที่ 8 สรุปผลตัวประมาณที่มีประสิทธิภาพจากวิธีการถดถอยแบบพีนอลไลซ์ 3 วิธี โดยพิจารณาจากค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ที่ดีที่สุดในแต่ละสถานการณ์

กรณีตัวแปรทำนาย 1 กลุ่ม		
รูปแบบความสัมพันธ์	กรณี n = 50	กรณี n = 100
คงที่	adaptive Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.6, 0.8 และ 0.9 → Lasso
โทพลิท	adaptive Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.9 → Lasso
ฮับโทพลิท	adaptive Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.9 → Lasso
กรณีตัวแปรทำนาย 2 กลุ่ม		
รูปแบบความสัมพันธ์	กรณี n = 50	กรณี n = 100
คงที่	adaptive Lasso ยกเว้น กรณี p = 100 r = 0.9 → Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.9 → Lasso
โทพลิท	adaptive Lasso ยกเว้น กรณี p = 100 r = 0.9 → Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.6, 0.7, 0.8, 0.9 → Lasso
ฮับโทพลิท	adaptive Lasso ยกเว้น กรณี p = 100 ทุกระดับความสัมพันธ์ → Lasso	adaptive Lasso ยกเว้นกรณี p = 100 r = 0.5, 0.6, 0.9 → Lasso

prediction accuracy and subset selection performances of stepwise, Lasso, elastic net and adaptive Lasso for small and sparse signals, Rajamangala University of Technology Tawan-ok Research Conference, Rajamangala University, Nakhon Nayok. (in Thai)

Singruang, S. and Pungpapong, V., 2017, A method comparison of gene set enrichment analysis and binary logistic regression for investigating the relationship between gene sets and a binary phenotype, Thai Sci. Technol J. 25(5): 778-790. (in Thai)

Tibshirani, R., 1996, Regression shrinkage and selection *via* the LASSO, J. Royal Stat. Soc. Ser. B 58: 267-288.

Zou, H., 2006, The adaptive LASSO and its oracle properties, J. Am. Stat. Assoc. 101: 1418-1429.