Research Article

Received: April 1, 2019; Accepted: August 6, 2019

การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษภายใต้ข้อมูลที่มีมิติสูง

Comparing Methods of Parameter Estimation with Penalized Regression Analysis under High-Dimensional Data

เบญจมาส รุ่งศรานนท์* และอัชฌา อระวีพร

ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพมหานคร 10520

Benjamas Rungsaranon* and Autcha Araveeporn

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Chalongkrung Road, Ladkrabang, Bangkok 10520

บทคัดย่อ

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยที่ปรับด้วยฟังก์ชันลงโทษ 5 วิธี ได้แก่ การถดถอยริดจ์ การถดถอยลาสโซ่ การถดถอยอิลาสติคเน็ต การถดถอยลาสโซ่แบบปรับปรุง และการถดถอยอิลาสติคเน็ตแบบปรับปรุงของตัวแบบการถดถอยเชิงเส้นพหุคูณ ซึ่งตัวแบบนี้ประกอบไปด้วยตัวแปรตามและตัวแปรอิสระ กรณีที่จำนวนตัวแปรอิสระมีจำนวนมากกว่าขนาดตัวอย่างหรือที่เรียกว่าข้อมูลที่มีมิติสูง การเปรียบเทียบประสิทธิภาพทั้ง 5 วิธี ใช้เกณฑ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย ข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นการจำลองข้อมูล โดยกำหนดขนาดตัวอย่างเล็ก (n = 5, 10 และ 15) จำนวนตัวแปรอิสระ 16 ตัวแปร ขนาดตัวอย่างปานกลาง (n = 20, 30 และ 40) จำนวนตัวแปรอิสระ 50 ตัวแปร และขนาดตัวอย่างใหญ่ (n = 60, 70 และ 80) จำนวนตัวแปรอิสระ 100 ตัวแปร โดยตัวแปรอิสระสร้างมาจากการแจกแจกปรกติ และค่าความคลาดเคลื่อนของตัวแบบการถดถอยเชิงเส้นพหุคูณสร้างมาจากการแจกแจงปรกติ การแจกแจงปรกติปลอมปน และการแจกแจงไวบูล โดยข้อมูลจากการจำลองใช้เทคนิคมอนติคาร์โล ซึ่งแต่ละกรณีจะทำซ้ำ 1,000 ครั้ง ผลการวิจัยพบว่าวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุงให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำที่สุดในทุกกรณี นอกจากนี้ผู้วิจัยยังนำทั้ง 5 วิธี มาประยุกต์ใช้กับข้อมูลจริงที่ขนาดตัวอย่างเล็ก จำนวนตัวแปรอิสระ 16 ตัวแปร ซึ่งวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุงเป็นวิธีการที่ดีกว่าวิธีอื่น ๆ เช่นเดียวกับข้อมูลจำลอง

***ผู้รับผิดชอบบทความ :** benjamas.rung99@gmail.com **doi:** 10.14456/tstj.2020.108

คำสำคัญ : วิธีการถดถอยริดจ์; วิธีการถดถอยลาสโซ่; วิธีการถดถอยอิลาสติคเน็ต; วิธีการถดถอยลาสโซ่แบบปรับปรุง; วิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง

Abstract

The objective of this research is to compare the efficiency of coefficient parameter estimation by using penalized regression analysis on five methods namely the ridge regression, the lasso regression, the elastic net regression, the adaptive lasso regression, and the adaptive elastic net regression methods. This research uses the multiple linear regression model, which is consisted of a dependent variable and independent variables. In case the number of independent variables is larger than number of sample sizes called high-dimensional data. For comparison the efficiency of five methods, the criterion is based on the average mean square errors. The data of this research is simulated by the small sample sizes (n = 5, 10, and 15) when the number of independent variables is specified by 16. For medium sample sizes (n = 20, 30, and 40), the number of independent variables is specified by 50. For large sample sizes (n = 60, 70, and 80), the number of independent variables is defined 100. The independent variable distribution is generated from the normal distribution, and the residuals are generated from the normal distribution, contaminated normal distribution, and Weibull distribution The data are obtained through simulation using a Monte Carlo technique with 1,000 replications for each case. The results are found that the adaptive elastic net regression is the minimum average mean square error in all cases. Furthermore, we apply five methods for real data based on the small sample sizes when the number of independent variables is considered on 16. The results of real data show that the adaptive elastic net regression outperforms the other methods as the simulation data.

Keywords: ridge regression method; lasso regression method; elastic net regression method; adaptive lasso regression method; adaptive elastic net regression method

1. บทนำ

การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยเพื่อสร้างตัวแบบความสัมพันธ์ระหว่างตัวแปรตาม (dependent variable) และตัวแปรอิสระ (independent variable) ซึ่งโดยทั่วไปข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ เรียกว่า ข้อมูลที่มีมิติต่ำ (low-dimensional data) โดยมีข้อกำหนดเบื้องต้นที่สำคัญในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ (multiple linear regression) คือ ค่าของตัวแปรอิสระแต่ละตัวเป็นค่าคงที่ และตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กับตัวแปรอิสระอื่น ๆ และตัวแปรตามเป็นข้อมูลเชิงปริมาณ ซึ่งวิธีที่นิยมในการประมาณค่า คือ วิธีกำลังสองน้อยที่สุด (ordinary least square method, OLS) แต่ในบางกรณีข้อมูลมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง เรียกว่า

ข้อมูลที่มีมิติสูง (high-dimensional data) ทำให้ข้อมูลที่ได้ไม่สามารถใช้วิธีกำลังสองน้อยที่สุดในการประมาณค่าสัมประสิทธิ์การถดถอย

ดังนั้นในงานวิจัยนี้ผู้วิจัยได้ศึกษาวิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (penalized regression) เพื่อใช้สำหรับวิเคราะห์ข้อมูลที่มีมิติสูงในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ซึ่งมีนักสถิติหลาย ๆ ท่าน ได้นำเสนอวิธีการถดถอยที่ปรับด้วยฟังก์ชันลงโทษ ได้แก่ Hoerl และ Kennard [1] ได้เสนอวิธีการถดถอยริดจ์ (ridge regression) ใช้สำหรับการประมาณค่าสัมประสิทธิ์การถดถอย เพื่อแก้ปัญหาพหุสัมพันธ์ (multicollinearity) โดยไม่ต้องตัดตัวแปรอิสระออกจากตัวแบบ ต่อมา Tibshirani [2] เสนอวิธีการถดถอยลาสโซ่ (least absolute shrinkage and selection operator regression, lasso) วิธีนี้ไม่เพียงแต่ประมาณค่าสัมประสิทธิ์การถดถอย แต่ยังสามารถใช้คัดเลือกตัวแปรอิสระที่เกี่ยวข้องเข้าสู่ตัวแบบ แต่ยังมีข้อจำกัดในการประมาณค่าเมื่อเกิดปัญหาพหุสัมพันธ์ ถัดมา Zou และ Hastie [3] ได้เสนอวิธีการถดถอยอิลาสติคเน็ต (elastic net regression) เป็นวิธีที่พัฒนาขึ้นมาเพื่อแก้ไขข้อจำกัดของวิธีการถดถอยลาสโซ่ ซึ่งวิธีนี้รวมระหว่างวิธีการถดถอยริดจ์และวิธีการถดถอยลาสโซ่เข้าด้วยกัน และใช้เมื่อเกิดปัญหาพหุสัมพันธ์ จากนั้น Zou [4] ได้ศึกษาและพัฒนาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณที่สามารถแก้ปัญหาการคัดเลือกตัวประมาณค่าที่เอนเอียงในตัวแบบวิธีการถดถอยลาสโซ่ ซึ่งเรียกวิธีนี้ว่าวิธีการถดถอยลาสโซ่แบบปรับปรุง (adaptive least absolute shrinkage and selection operator regression) และต่อมา Zou และ Zhang [5] เสนอวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง (adaptive elastic net regression) โดยรวมระหว่างวิธีการถดถอยอิลาสติคเน็ตและวิธีการถดถอยลาสโซ่แบบปรับปรุงเข้าด้วยกัน เป็นวิธีที่พัฒนาขึ้นมาเพื่อใช้ในการคัดเลือกหรือละเว้นตัวแปรอิสระเข้าสู่ตัวแบบ เมื่อเกิดปัญหาพหุสัมพันธ์กันในกรณีที่ค่าสัมประสิทธิ์ใกล้กัน

การศึกษางานวิจัยที่เกี่ยวข้อง Phakdee [6] ได้เปรียบเทียบวิธีการคัดเลือกตัวแปรและการประมาณค่าสัมประสิทธิ์การถดถอย โดยเปรียบเทียบวิธีการถดถอยริดจ์ วิธีการถดถอยริดจ์แบบทางเลือก (alternative ridge regression) และวิธีการถดถอยริดจ์แบบมีข้อจำกัด (restricted ridge regression) พบว่าวิธีการถดถอยริดจ์ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (average mean squared error, AMSE) ต่ำที่สุด ถัดมา Choosawat และ Lisawadi [7] ได้เปรียบเทียบประสิทธิภาพการวิเคราะห์สัมประสิทธิ์การถดถอยด้วยวิธีการถดถอยริดจ์ วิธีการถดถอยลาสโซ่ และวิธีการถดถอยลาสโซ่แบบปรับปรุง กรณีข้อมูลมีมิติสูงและตัวแปรอิสระมีความสัมพันธ์กันค่อนข้างสูง พบว่าวิธีการถดถอยลาสโซ่แบบปรับปรุงจะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของค่าพยากรณ์ (prediction mean square error, PMSE) ต่ำที่สุด เมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มสูงจากนั้น Algamal และ Lee [8] ประยุกต์ใช้ข้อมูล DNA microarray จากปัญหาเกี่ยวกับการจำแนกมะเร็งที่มีข้อมูลมิติสูง ซึ่งได้ศึกษาการถดถอยโลจิสติกส์วิธีอิลาสติคเน็ต (elastic net logistic regression) และวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง นำมาประยุกต์ใช้ในการจำแนกประเภทมะเร็งในมิติสูงเพื่อใช้ในการประมาณค่าสัมประสิทธิ์ของยีน และการคัดเลือกยีนไปพร้อมกัน โดยในงานวิจัยศึกษาเปรียบเทียบวิธีการถดถอยโลจิสติกส์อิลาสติคเน็ต และวิธีการถดถอยโลจิสติกส์อิลาสติคเน็ตแบบปรับปรุงเปรียบเทียบกับวิธีใหม่ที่นำเสนอ คือ วิธีการปรับของการถดถอยโลจิสติกส์อิลาสติคเน็ตแบบปรับปรุง (adjusted adaptive regularized logistic regres-

sion, AAElastic) พบว่าวิธีการที่นำเสนอให้ผลลัพธ์ดีกว่าวิธีการถดถอยโลจิสติกส์อิลาสติคเน็ตและวิธีการถดถอยโลจิสติกส์อิลาสติคเน็ตแบบปรับปรุง

การศึกษางานวิจัยต่าง ๆ ผู้วิจัยจึงสนใจศึกษาการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ในการประมาณค่าพารามิเตอร์เพื่อใช้พยากรณ์ตัวแปรตามโดยเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำที่สุด กรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง และตัวแปรตามเป็นข้อมูลเชิงปริมาณในลักษณะข้อมูลที่มีมิติสูง งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการวิเคราะห์การถดถอยเชิงเส้นพหุคูณแบบปรับด้วยฟังก์ชันการลงโทษ 5 วิธี คือ วิธีการถดถอยริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่แบบปรับปรุง และวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง เพื่อหาวิธีดีที่สุดในการประมาณค่าพารามิเตอร์ที่อยู่ภายใต้สถานการณ์ที่แตกต่างกันในหลายเงื่อนไข โดยเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย และใช้โปรแกรม R ในการจำลองและการวิเคราะห์ข้อมูล

2. ตัวสถิติที่ใช้ในงานวิจัย

งานวิจัยนี้ศึกษาตัวแบบการถดถอยเชิงเส้นพหุคูณเป็นตัวแบบที่มีตัวแปรตาม (y) เป็นข้อมูลเชิงปริมาณ ซึ่งมีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ (X) เป็นข้อมูลเชิงปริมาณจำนวน p ตัวแปร เขียนอยู่ในรูปของเมทริกซ์ได้ [9] ดังนี้

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

โดยที่ $\underset{\sim}{y}$ แทน เวกเตอร์ของตัวแปรตามขนาด n x 1; X แทน เมทริกซ์ของค่าสังเกตของตัวแปรอิสระขนาด n x $(p+1)$; $\underset{\sim}{\beta}$ แทน เวกเตอร์ของค่าสัมประสิทธิ์การถดถอยพหุคูณขนาด $(p+1)$ x 1; $\underset{\sim}{\varepsilon}$ แทน เวกเตอร์ของค่าความคลาดเคลื่อนที่เกิดขึ้นขนาด n x 1; p แทน จำนวนตัวแปรอิสระ; n แทน ขนาดตัวอย่าง

โดยมีข้อตกลงเบื้องต้นของตัวแบบการถดถอยพหุคูณ ดังนี้

(1) $\underset{\sim}{\varepsilon} \sim N(0, \sigma^2 I)$ เมื่อ I คือ เมทริกซ์เอกลักษณ์ขนาด n x n

(2) ค่าของตัวแปรอิสระแต่ละตัวเป็นค่าคงที่

(3) ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กับตัวแปรอิสระอื่น ๆ

โดยทั่วไปวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ใช้วิธีกำลังสองน้อยที่สุด โดยตัวประมาณค่าสัมประสิทธิ์การถดถอย ดังนี้

$$\underset{\sim}{\hat{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2 = (X^T X)^{-1} X^T \underset{\sim}{y};$$

$i = 1, 2, ..., n$, $j = 1, 2, ..., p$

แต่สำหรับข้อมูลที่มีมิติสูง เมื่อจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง เมทริกซ์ $X^T X$ จะเป็นเมทริกซ์เอกฐาน (singular matrix) ซึ่งไม่สามารถหาเมทริกซ์ผกผัน (inverse of a matrix) ทำให้ไม่สามารถหาค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด จึงใช้วิธีการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ โดยในการวิจัยครั้งนี้สนใจเปรียบเทียบ 5 วิธี ประกอบด้วย

2.1 วิธีการถดถอยริดจ์

Hoerl และ Kennard [1] เสนอวิธีการถดถอยริดจ์ (ridge regression method, ridge) เพื่อแก้ปัญหาการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ ซึ่งจะลดความคลาดเคลื่อนกำลังสองของค่าเฉลี่ยให้มีค่าต่ำลง หลักการของวิธีกำลังสองน้อยที่สุดพบว่าถ้าต้องการลดความคลาดเคลื่อนกำลังสองเฉลี่ยต้องทำให้ $\left|X^T X\right|$ มีค่าเพิ่มขึ้น โดยบวกค่าคงที่ λ เข้ากับสมาชิกทุกตัวบนเส้นทแยงมุมของ เมทริกซ์ $\left|X^T X\right|$ ดังนั้นตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีการถดถอยริดจ์อยู่ในรูปดังต่อไปนี้

$$\hat{\underset{\sim}{\beta}}_R = (X^T X + \lambda I_{p+1})^{-1} X^T \underset{\sim}{y};\ 0 < \lambda < 1$$

เมื่อ λ เป็นค่าคงที่ ซึ่งอยู่ระหว่าง 0 ถึง 1 และ I_{p+1} เป็นเมทริกซ์เอกลักษณ์ (identity matrix) ขนาด $(p+1)\times(p+1)$

โดยตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณวิธีการถดถอยริดจ์ เป็นดังนี้

$$\hat{\underset{\sim}{\beta}}_R = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right];\ \lambda > 0$$

เมื่อ $\lambda\sum_{j=1}^{p}\beta_j^2$ คือ ฟังก์ชันการลงโทษ และ λ คือ พารามิเตอร์ปรับแต่ง (tuning parameter) ซึ่งควบคุมขนาดการหดตัว (shrinkage) ของ $\hat{\underset{\sim}{\beta}}_R$ โดยใช้วิธีการตรวจสอบไขว้ (cross-validation) ในที่นี้จะใช้วิธีการตรวจสอบไขว้ (generalized cross-validation) คิดค้นโดย Boonstra และคณะ [10] สำหรับการหาค่าพารามิเตอร์ปรับแต่งที่น้อยที่สุด

2.2 วิธีการถดถอยลาสโซ่

Tibshirani [2] เสนอวิธีการวิเคราะห์การถดถอยลาสโซ่ (least absolute shrinkage and selection operator regression method, lasso) ใช้สำหรับการประมาณค่าและการคัดเลือกตัวแปรเข้าสู่ตัวแบบในคราวเดียวกัน กรณีที่จำนวนตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามจำนวนไม่มากนักในตัวแบบ จะทำให้วิธีลาสโซ่มีประสิทธิภาพสูงสุดในการพยากรณ์ โดยค่าสัมประสิทธิ์การถดถอยพหุคูณของวิธีลาสโซ่ จะอยู่ในรูปผลบวกระหว่างผลรวมความคลาดเคลื่อนกำลังสองและผลรวมสัมบูรณ์ของค่าสัมประสิทธิ์ถ่วงน้ำหนักให้มีค่าต่ำสุด หากค่าถ่วงน้ำหนักมาก ๆ จะทำให้สัมประสิทธิ์นั้นหดตัวเหลือเท่ากับศูนย์หรือหายไป กล่าวคือ วิธีนี้จะทำให้ค่าสัมประสิทธิ์ส่วนใหญ่เป็นศูนย์และค่าสัมประสิทธิ์บางส่วนไม่เท่ากับศูนย์ (sparse estimator) โดยตัวประมาณสัมประสิทธิ์การถดถอยวิธีลาสโซ่ ดังนี้

$$\hat{\underset{\sim}{\beta}}_L = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right];\ \lambda > 0$$

เมื่อ $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ คือ ฟังก์ชันการลงโทษ และ λ คือ พารามิเตอร์ปรับแต่ง ซึ่งควบคุมขนาดการหดตัวของ $\hat{\underset{\sim}{\beta}}_L$ โดยใช้วิธีการตรวจสอบไขว้ ในที่นี้จะใช้วิธีอัลกอริทึม LARS (LARS algorithm) คิดค้นโดย Efron และคณะ [11] สำหรับการหาค่าพารามิเตอร์ปรับแต่งที่น้อยที่สุด

2.3 วิธีการถดถอยอิลาสติคเน็ต

Zou และ Hastie [3] เสนอวิธีการถดถอยอิลาสติคเน็ต (elastic net regression method, elastic net) โดยเป็นการรวมกันระหว่างวิธีการถดถอยริดจ์และวิธีลาสโซ่ จึงสามารถคัดเลือกตัวแปรอิสระและประมาณค่าไปพร้อมกัน วิธีนี้เหมาะสำหรับการวิเคราะห์ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมาก ๆ และตัวแปรอิสระที่มีความสัมพันธ์กันสูง โดยค่าสัมประสิทธิ์การถดถอยพหุคูณของวิธีอิลาสติคเน็ตจะอยู่ในรูปผลบวกของผลรวมความคลาดเคลื่อนกำลังสอง ผลรวมสัมบูรณ์ของค่าสัมประสิทธิ์ถ่วงน้ำหนัก และผลรวมกำลังสองของค่าสัมประสิทธิ์ถ่วงน้ำหนักให้มีค่าต่ำสุด ซึ่งสามารถประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีอิลาสติคเน็ตดังนี้

$$\hat{\underset{\sim}{\beta}}_E = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right];$$

$$0 < \lambda_1 + \lambda_2 < 0$$

ฟังก์ชันการลงโทษของวิธีการถดถอยอิลาสติคเน็ตเป็นการรวมกันของฟังก์ชันการลงโทษของวิธีการถดถอยริดจ์และวิธีลาสโซ่ และเมื่อ $\lambda_1 = 0$ วิธีการถดถอยอิลาสติคเน็ตจะเป็นวิธีการถดถอยริดจ์อย่างง่าย (simple ridge regression) ซึ่งฟังก์ชันการลงโทษของวิธีการถดถอยอิลาสติคเน็ตจะมี λ_1 และ

λ_2 คือพารามิเตอร์ปรับแต่ง ซึ่งควบคุมขนาดการหดตัวของ $\hat{\beta}_{E}$ โดยจะใช้วิธีการตรวจสอบไขว้ คิดค้นโดย Hastie et al. [12] ในการหาพารามิเตอร์ปรับแต่ง

2.4 วิธีการถดถอยลาสโซ่แบบปรับปรุง

วิธีการถดถอยลาสโซ่ยังมีความเอนเอียงในการคัดเลือกตัวประมาณสัมประสิทธิ์การถดถอย Zou [4] จึงเสนอวิธีการถดถอยลาสโซ่แบบปรับปรุง (adaptive least absolute shrinkage and selection operator regression method, Alasso) โดยเพิ่มค่าถ่วงน้ำหนัก (weight) อีกหนึ่งพารามิเตอร์เข้ามาในวิธีลาสโซ่ ซึ่งการให้ค่าถ่วงน้ำหนักกับพารามิเตอร์แต่ละตัวแตกต่างกันในฟังก์ชันการลงโทษ $(P_{\lambda}(\beta))$ โดยการกำหนดค่าถ่วงน้ำหนักให้มีค่าสูงสำหรับค่าสัมประสิทธิ์ที่มีค่าน้อย และกำหนดค่าถ่วงน้ำหนักให้มีค่าน้อยสำหรับค่าสัมประสิทธิ์ที่มีค่าสูง เพื่อลดความไม่คงเส้นคงวาที่ทำให้เกิดความเอนเอียงในการประมาณค่าสัมประสิทธิ์การถดถอยในวิธีการถดถอยลาสโซ่ การใช้ฟังก์ชันการลงโทษนี้ยังมีสมบัติอีก 1 ข้อ คือ เมื่อขนาดตัวอย่างมีจำนวนมากพอ วิธีการถดถอยลาสโซ่แบบปรับปรุงจะมีความสามารถในการเลือกตัวแปรเสมือนกับว่าทราบตัวแบบที่แท้จริง (true model) ซึ่งสมบัตินี้ไม่มีในวิธีลาสโซ่ โดยสามารถประมาณค่าสัมประสิทธิ์การถดถอยวิธีลาสโซ่แบบปรับปรุงดังนี้

$$\hat{\beta}_{AL} = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ij})^2+\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right]$$

โดยที่ $\hat{w}_j = \dfrac{1}{|\hat{\beta}_{j(\mathrm{L})}|^{\gamma}}; j=1,2,...,p, \gamma>0$

โดยทั่วไปกำหนดให้ $\gamma=1$ และ $\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|$ คือ ฟังก์ชันการลงโทษของวิธีการถดถอยลาสโซ่แบบปรับปรุง โดยเป็นฟังก์ชันการลงโทษของวิธีการถดถอยลาสโซ่ที่เพิ่มค่าถ่วงน้ำหนักเข้ามา โดยใช้วิธีการตรวจสอบไขว้แบบวิธีเกณฑ์ข้อสนเทศของเบส์ (BIC cross-validation) คิดค้นโดย Zou และคณะ [13] สำหรับการหาค่าพารามิเตอร์ปรับแต่งที่น้อยที่สุด

2.5 วิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง

เนื่องจากวิธีการถดถอยอิลาสติคเน็ตไม่มีสมบัติที่ว่าเมื่อขนาดตัวอย่างมีจำนวนมากพอจะมีความสามารถในการเลือกตัวแปรเสมือนกับว่าทราบตัวแบบที่แท้จริง จึงเสนอวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง (adaptive elastic net regression method, Aelastic net) โดย Zou และ Zhang [5] ได้พัฒนาวิธีการประมาณค่าสัมประสิทธิ์การถดถอย เพื่อเพิ่มประสิทธิภาพในการประมาณค่าสัมประสิทธิ์และการคัดเลือกตัวแปรมากขึ้น โดยรวมระหว่างวิธีถดถอยอิลาสติคเน็ตและวิธีลาสโซ่แบบปรับปรุงเข้าด้วยกัน สามารถประมาณค่าสัมประสิทธิ์การถดถอยวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุงดังนี้

$$\hat{\beta}_{AE} = \arg\min_{\beta}\left[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ij})^2+\lambda_1\sum_{j=1}^{p}\hat{w}_j|\beta_j|+\lambda_2\sum_{j=1}^{p}\beta_j^2\right]$$

โดยที่ $\hat{w}_j = \left(|\hat{\beta}_{j(\mathrm{E})}|+\dfrac{1}{n}\right)^{-\gamma}$; $\gamma>0$

ฟังก์ชันการลงโทษของวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง เป็นการรวมกันของฟังก์ชันการลงโทษของวิธีการถดถอยอิลาสติคเน็ต และวิธีลาสโซ่แบบปรับปรุง ซึ่งหาค่าพารามิเตอร์ปรับแต่งที่น้อยที่สุดโดยใช้วิธีการตรวจสอบไขว้แบบวิธีเกณฑ์ข้อสนเทศของเบส์

3. วิธีการดำเนินการวิจัย

การวิจัยครั้งนี้มีขั้นตอนดังนี้ คือ

3.1 กำหนดค่าความคลาดเคลื่อน (residual) ของตัวแบบการถดถอยเชิงเส้นพหุคูณ (ε) สุ่มมาจาก 3 การแจกแจง ได้แก่

3.1.1 การแจกแจงปรกติ (normal distribution) ที่มีค่าพารามิเตอร์ คือ ค่าเฉลี่ย (μ) และ

ความแปรปรวน (σ^2) หรือเขียนได้ว่า $N\left(\mu,\sigma^2\right)$ กำหนดให้เป็น $N(0,1)$ และ $N(0,9)$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น (probability density function, pdf) ดังนี้

$$f(\varepsilon)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\varepsilon-\mu}{\sigma}\right)^2} \quad ; \quad -\infty<\varepsilon<\infty,-\infty<\mu<\infty,\sigma>0$$

3.1.2 การแจกแจงปรกติปลอมปน (contaminated normal distribution) เป็นการผสมระหว่างการแจกแจงปรกติและการแจกแจงปรกติที่มีความแปรปรวนสูงมาก [14] ในที่นี้กำหนดให้การแจกแจงปรกติ $N\left(\mu,\sigma_1^2\right)$ เป็น $N(0,1)$ และค่าความน่าจะเป็นที่ข้อมูลมีการปลอมปนการแจกแจงปรกติเป็น 0.1 $(p'=0.1)$ หรือ 10 % ของการแจกแจงปรกติ $N\left(\mu,\sigma_2^2\right)$ เป็น $N(0,25)$ และ $N(0,100)$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น ดังนี้

$$f(\varepsilon)=(1-p')N(\mu,\sigma_1^2)+p'N(\mu,\sigma_2^2)\,;$$
$$-\infty<\varepsilon<\infty\,,-\infty<\mu<\infty\,,\ \sigma_1^2>0,\sigma_2^2>0$$

3.1.3 การแจกแจงไวบูล (Weibull distribution) ประกอบด้วยพารามิเตอร์บอกอัตราส่วน (scale parameter, α) และพารามิเตอร์บอกรูปร่าง (shape parameter, β) หรือเขียนได้ว่า $W(\alpha,\beta)$ กำหนดให้เป็น $W(1,1.5)$ และ $W(1,3)$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น ดังนี้

$$f(\varepsilon)=\left(\frac{\beta}{\alpha}\right)\left(\frac{\varepsilon}{\alpha}\right)^{\beta-1}e^{-\left(\frac{\varepsilon}{\alpha}\right)^{\beta}} \quad ; \quad 0<\varepsilon<\infty,\beta>0,\alpha>0$$

3.2 ตัวแปรอิสระ (X) สุ่มมาจากการแจกแจงปรกติ ด้วยค่าเฉลี่ย (μ) และความแปรปรวน (σ^2) หรือเขียนได้ว่า $N\left(\mu,\sigma^2\right)$ กำหนดให้เป็น $N(0,2)$

3.3 กำหนดให้จำนวนตัวแปรอิสระ (p) ในตัวแบบการถดถอยเชิงเส้นพหุคูณมากกว่าขนาดตัวอย่าง (n) ดังนี้

3.3.1 กำหนดขนาดตัวอย่าง (n) เท่ากับ 5, 10 และ 15 เมื่อจำนวนตัวแปรอิสระ (p) เท่ากับ 16 ตัวแปร

3.3.2 กำหนดขนาดตัวอย่าง (n) เท่ากับ 20 30 และ 40 เมื่อจำนวนตัวแปรอิสระ (p) เท่ากับ 50 ตัวแปร

3.3.3 กำหนดขนาดตัวอย่าง (n) เท่ากับ 60 70 และ 80 เมื่อจำนวนตัวแปรอิสระ (p) เท่ากับ 100 ตัวแปร

3.4 ค่าสัมประสิทธิ์การถดถอย (β_j) ทุกตัวมีค่าเป็น 1 ในทุกสถานการณ์

3.5 กำหนดตัวแปรตามใช้ตัวแบบการถดถอยเชิงเส้นพหุคูณ ดังนี้ $\underset{\sim}{y}=X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$

3.6 โปรแกรมที่ใช้ในการวิจัยครั้งนี้ทั้งหมดเขียนด้วยโปรแกรม R เวอร์ชั่น 3.5.2 ซึ่งทดลองซ้ำ 1,000 ครั้ง (m) ในแต่ละสถานการณ์

3.7 เกณฑ์ที่ใช้ในการวิจัยสำหรับข้อมูลที่จำลอง คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (mean squared error, MSE) ของตัวแบบการถดถอยเชิงเส้นพหุคูณของการทดลองในแต่ละรอบ (m) และนำมาหาค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (average mean squared error, AMSE) โดยค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำที่สุดคือวิธีที่ประมาณค่าสัมประสิทธิ์การถดถอยที่ดีที่สุด สามารถคำนวณดังนี้

$$MSE_h(\underset{\sim}{\hat{y}})=E(\underset{\sim}{\hat{y}}-\underset{\sim}{y})^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}\,;$$
$$i=1,2,\ldots,n;\ h=1,2,\ldots,m$$
$$AMSE(\underset{\sim}{\hat{y}})=\frac{\sum_{h=1}^{m}MSE_h(\underset{\sim}{\hat{y}})}{m}$$

3.8 นำชุดข้อมูลจริงซึ่งเป็นชุดข้อมูลการสร้างอาคารที่อยู่อาศัยในเมืองเตหะราน ประเทศอิหร่าน ปี ค.ศ. 2018 [15] มาประยุกต์ใช้เพื่อเปรียบเทียบวิธีการถดถอยแบบปรับด้วยฟังก์ชันการลงโทษ 5 วิธี ซึ่งข้อมูลที่ใช้ประกอบด้วยตัวแปรตาม (y) และตัวแปรอิสระ

(X) โดยชุดข้อมูลตัวแปรอิสระมีทั้งหมด 27 ตัวแปร แต่ผู้วิจัยคัดเลือกตัวแปรอิสระมาทั้งหมด 16 ตัวแปร เนื่องจากคัดเลือกตัวแปรอิสระที่เป็นข้อมูลเชิงปริมาณ โดยขนาดตัวอย่างที่ใช้ (n) เป็น 5, 10 และ 15 ซึ่งได้จากการสุ่มตัวอย่างแบบง่าย (simple random sampling, SRS) จากข้อมูลทั้งหมด 372 ค่า และหน่วยของตัวแปรส่วนใหญ่ขึ้นอยู่กับค่า ***IRR*** (internal rate of return) คือ อัตราผลการตอบแทนภายใน เป็นการประเมินว่าผลตอบแทนให้อัตราผลตอบแทนเท่าใด มีรายละเอียดดังตารางที่ 1 และเกณฑ์ที่ใช้ในการวิจัยสำหรับข้อมูลจริง คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE)

Table 1 The descriptions of independent variables (X) and dependent variable (y) of building residential apartments data set in Tehran, Iran (2018)

Variables	Descriptions	Units
X_1	Lot area	square meter (m^2)
X_2	Total preliminary estimated construction cost based on the prices at the beginning of the project	10,000,000 ***IRR***
X_3	Preliminary estimated construction cost based on the prices at the beginning of the project	10,000 ***IRR***
X_4	Equivalent preliminary estimated construction cost based on the prices at the beginning of the project in a selected base year	10,000 ***IRR***
X_5	Duration of construction	month
X_6	Price of the unit at the beginning of the project per m^2	10,000 ***IRR***
X_7	Total floor areas of building permits issued by the city/municipality	square meter (m^2)
X_8	Cumulative liquidity	10,000,000 ***IRR***
X_9	Private sector investment in new buildings	10,000,000 ***IRR***
X_{10}	The interest rate for loan in a time resolution	percentage (%)
X_{11}	The average construction cost of buildings by private sector at the time of completion of construction	10,000 ***IRR*** / m^2
X_{12}	The average of construction cost of buildings by private sector at the beginning of the construction	10,000 ***IRR*** / m^2
X_{13}	Official exchange rate with respect to dollars	***IRR***
X_{14}	Nonofficial (street market) exchange rate with respect to dollars	***IRR***
X_{15}	Population of the city	People
X_{16}	Gold price per ounce	***IRR***
y	Actual sales prices	10,000 ***IRR***

3.9 วิเคราะห์และสรุปผลทั้ง 5 วิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษภายใต้ข้อมูลที่มีมิติสูง

4. ผลการวิจัย

งานวิจัยนี้ได้ใช้มูลจากการจำลองในหัวข้อที่ 3 แสดงผลในหัวข้อ 4.1 และนำมาประยุกต์ใช้ข้อมูลจริง แสดงผลในหัวข้อ 4.2 ตามลำดับ

4.1 ผลจากการจำลองข้อมูล

การประมาณค่าสัมประสิทธิ์การถดถอยของวิธี ridge วิธี lasso วิธี elastic net วิธี Alasso และวิธี Aelastic net เมื่อค่าความคลาดเคลื่อนของตัวแบบการถดถอยเชิงเส้นพหุคูณมีการแจกแจงปรกติ การแจกแจงปรกติปลอมปน และการแจกแจงไวบูล ดังตารางที่ 2-4

Table 2 AMSE of the normal distribution $(N(\mu,\sigma^2))$, sample sizes (n), the number of independent variables (p) of ridge, lasso, elastic net, Alasso, and Aelastic net methods

The Residuals	p	n	Methods				
			Ridge	Lasso	Elastic net	Alasso	Aelastic net
$N(0,1)$	16	5	33.9256	26.9874	22.4558	7.7025	**3.3780**
		10	37.7946	16.6173	13.5643	6.5365	**3.6406**
		15	38.9704	7.5894	5.8060	4.1797	**2.7336**
	50	20	96.2498	70.4524	61.0480	16.5131	**7.6294**
		30	101.6025	44.2040	34.1240	11.1575	**5.6076**
		40	106.7399	22.5958	14.0320	7.2192	**3.7383**
	100	60	181.1241	86.9848	59.0713	12.8314	**7.6452**
		70	186.1973	49.0954	35.2171	8.4384	**4.8927**
		80	190.5646	28.7975	19.8804	6.0043	**3.3205**
$N(0,9)$	16	5	39.1396	35.3612	29.3117	9.4488	**4.1393**
		10	43.6695	22.1339	18.5876	8.0446	**4.6192**
		15	46.2458	12.8152	10.0233	6.0954	**4.0198**
	50	20	93.8416	70.0297	58.5286	16.8798	**7.6230**
		30	101.2664	45.8974	34.0036	10.8726	**6.0087**
		40	105.3065	21.4300	14.2404	6.0439	**3.5356**
	100	60	184.7577	93.3526	72.0972	14.2818	**8.2960**
		70	190.6009	57.6322	37.4456	9.2522	**5.4729**
		80	194.1427	39.8798	24.6886	6.7811	**5.1945**

Bold text meaning the lowest AMSE of penalized regression methods in each case

Table 3 AMSE of the contaminated normal distribution with percent and variance (σ_2^2) of contaminated data, sample sizes (n), the number of independent variables (p) of ridge, lasso, elastic net, Alasso, and Aelastic net methods

The Residuals	p	n	Methods				
			Ridge	Lasso	Elastic net	Alasso	Aelastic net
10 %, σ_2^2 = 25	16	5	35.0811	31.0671	24.7741	8.4406	**3.6405**
		10	38.9817	16.7417	13.8955	6.2428	**3.6534**
		15	41.0874	9.1498	6.9239	4.0533	**3.1888**
	50	20	96.8406	73.9534	64.6146	17.6593	**7.7842**
		30	102.9568	44.0060	34.0947	10.2773	**6.0346**
		40	105.9625	23.7546	17.1172	7.0021	**3.8908**
	100	60	182.8232	83.6694	61.6271	12.4524	**8.6602**
		70	187.1082	52.7917	38.6351	8.7435	**5.6113**
		80	190.7047	35.7055	22.9899	6.7027	**4.9833**
10 %, σ_2^2 = 100	16	5	38.8782	33.3775	27.6338	8.7207	**3.9104**
		10	44.2320	22.5650	19.9547	8.0482	**4.7798**
		15	47.3983	14.2904	11.8839	6.6052	**4.7059**
	50	20	96.2471	69.3985	59.6544	17.5066	**7.6667**
		30	102.8508	46.1801	33.6662	10.8649	**5.5765**
		40	106.1674	22.1053	14.9190	6.2030	**3.8285**
	100	60	186.8187	96.7751	71.6475	12.9031	**9.4370**
		70	191.2204	59.5943	42.4976	11.1551	**5.9500**
		80	195.7240	42.7192	26.4030	7.0348	**5.0815**

Bold text meaning the lowest AMSE of penalized regression methods in each case

ตารางที่ 2-4 พบว่าวิธี ridge เมื่อขนาดตัวอย่าง (n) เพิ่มขึ้น จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) เพิ่มขึ้นด้วย สอดคล้องกับงานวิจัยของ Thongteeraparp [16] ส่วนวิธี lasso วิธี elastic net วิธี Alasso และวิธี Aelastic net เมื่อขนาดตัวอย่าง (n) เพิ่มขึ้น จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) ต่ำลง

วิธี Aelastic net ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) ต่ำที่สุด ในทุกขนาดตัวอย่าง (n) ทุกจำนวนตัวแปรอิสระ (p) และทุกค่าความคลาดเคลื่อนของตัวแบบการถดถอยเชิงเส้นพหุคูณที่มีการแจกแจงปรกติ การแจกแจงปรกติปลอมปน และการแจกแจงไวบูล

4.2 ผลจากข้อมูลจริง

Table 4 AMSE of the Weibull distribution $(W(\alpha,\beta))$, sample sizes (n), the number of independent variables (p) of ridge, lasso, elastic net, Alasso, and Aelastic net methods

The Residuals	p	n	Methods				
			Ridge	Lasso	Elastic Net	ALasso	AElastic Net
$W(1,1.5)$	16	5	33.0417	28.2946	22.9671	8.0242	**3.5182**
		10	37.0135	15.6834	12.7651	5.7262	**3.4748**
		15	39.122	7.0851	5.1782	3.713	**2.5124**
	50	20	95.5383	71.9751	61.5638	18.3407	**7.788**
		30	100.6384	46.5271	35.2927	11.0467	**5.9334**
		40	106.0123	25.3405	15.4053	6.4533	**4.2603**
	100	60	181.1241	86.9848	59.0713	12.8314	**7.6452**
		70	186.1973	49.0954	35.2171	8.4384	**4.8927**
		80	190.5646	28.7975	19.8804	6.0043	**3.3205**
$W(1,3)$	16	5	33.6305	30.1172	24.1374	8.5887	**3.6626**
		10	37.0765	16.2172	12.9472	6.0041	**3.4910**
		15	38.1400	7.3388	4.8485	3.5092	**2.1785**
	50	20	95.9776	73.5361	58.7985	18.3416	**7.6663**
		30	101.5208	43.7559	33.5337	10.2491	**5.9679**
		40	106.2578	21.6254	15.1425	5.8569	**3.4669**
	100	60	185.3214	91.2636	68.1079	13.8273	**8.1367**
		70	190.7263	64.2791	47.5868	10.5184	**6.3380**
		80	194.3859	35.3487	23.3838	7.1841	**4.4780**

Bold text meaning the lowest AMSE of penalized regression methods in each case

Table 5 MSE of building residential apartments data set in Tehran, Iran (2018) with sample sizes (n) of ridge, lasso, elastic net, Alasso, and Aelastic net methods

Sample sizes (n)	Methods				
	Ridge	Lasso	Elastic net	Alasso	Aelastic net
$n=5$	104,678.60	105,560.00	105,560.00	39,850.79	**4,053.28**
$n=10$	194,735.60	152,793.90	135,290.40	51,930.01	**3,492.19**
$n=15$	162,579.30	200,074.90	207,856.10	59,951.53	**5,283.75**

Bold text meaning the lowest AMSE of penalized regression methods in each case

ผลการวิเคราะห์ของวิธี ridge วิธี lasso วิธี elastic net วิธี Alasso และวิธี Aelastic net แสดงค่าความเคลื่อนกำลังสองเฉลี่ย (MSE) ของชุดข้อมูลการสร้างอาคารที่อยู่อาศัย ดังตารางที่ 5

ตารางที่ 5 พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) จะเพิ่มขึ้นทุกวิธี เนื่องจากการใช้ข้อมูลจริงมีข้อจำกัดหลายด้าน เช่น การแจกแจงของค่าความคลาดเคลื่อนที่ไม่ทราบของข้อมูล แต่เมื่อมองโดยภาพรวม วิธี Aelastic net ให้ค่าความเคลื่อนกำลังสองเฉลี่ย (MSE) ต่ำที่สุดทุกขนาดตัวอย่าง ทั้งยังมีผลลัพธ์สอดคล้องกับผลการวิเคราะห์ของชุดข้อมูลที่จำลองขึ้นในงานวิจัยครั้งนี้ที่ว่าวิธี Aelastic net ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (AMSE) ต่ำที่สุด โดยสมการถดถอยของชุดข้อมูลจริงวิธี Aelastic net เมื่อขนาดตัวอย่าง (n) เป็น 5 10 และ 15 ดังนี้

$\hat{y}_i = 6,798.6316 - 3.3362X_1 - 9.3918X_2$; $i = 1,2,...,5$

$\hat{y}_i = 6,793.8757 - 3.3222X_1 - 9.4017X_2$; $i = 1,2,...,10$

$\hat{y}_i = 6,771.5421 - 3.3412X_1 - 9.1724X_2$; $i = 1,2,...,15$

5. สรุปผลการวิจัย

งานวิจัยนี้ผู้วิจัยต้องการเปรียบเทียบวิธีการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยวิธีการถดถอยแบบปรับด้วยฟังก์ชันการลงโทษ 5 วิธี คือ วิธีการถดถอยริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่แบบปรับปรุง และวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง กรณีที่จำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างในลักษณะข้อมูลที่มีมิติสูง เมื่อค่าความคลาดเคลื่อนของตัวแบบการถดถอยเชิงเส้นพหุคูณมีการแจกแจงปรกติ การแจกแจงปรกติปลอมปน และการแจกแจงไวบูล พบว่าวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุงให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในทุกการแจกแจงของค่าความคลาดเคลื่อนของตัวแบบการถดถอยเชิงเส้นพหุคูณ ทุกขนาดตัวอย่าง และทุกจำนวนตัวแปรอิสระ และเมื่อนำมาประยุกต์ใช้กับข้อมูลจริงก็ให้ผลเช่นเดียวกับการจำลองข้อมูล เนื่องจากวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุงมีถึง 3 พารามิเตอร์ การปรับได้แก่ λ_1, λ_2 และ γ และผลการวิเคราะห์ยังแสดงให้เห็นว่าวิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่แบบปรับปรุง และวิธีการถดถอยอิลาสติคเน็ตแบบปรับปรุง ประมาณค่าสัมประสิทธิ์การถดถอยดีกว่าวิธีการถดถอยริดจ์ เนื่องจากวิธีการถดถอยริดจ์ไม่ได้ตัดตัวแปรอิสระออกจากตัวแบบของการถดถอยเชิงเส้นพหุคูณ และงานวิจัยครั้งต่อไปผู้วิจัยควรศึกษากรณีที่ตัวแปรอิสระเกิดพหุสัมพันธ์กัน

6. References

[1] Hoerl, A.E. and Kennard, R.W., 1970, Ridge regression: Biased estimation for nonorthogonal problems, J. Am. Stat. Assoc. 12: 55-67.

[2] Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, J. Royal Stat. Soc. B. 58: 267-288.

[3] Zou, H. and Hastie, T., 2005, Regularization and variable selection via the elastic net, J. Royal Stat. Soc. B 67: 301-320.

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

[5] Zou, H. and Zhang, T., 2009, On the

adaptive elastic net with a diverging number of parameters, Ann. Stat. 37: 1733-1751.

[6] Phakdee, N. , 2009, Comparisons of Estimation of Multiple Regression Coefficients with Existent Multicollinearity among Independent Variables by Ridge Regression Method, Master Thesis, King Mongkut' s Institute of Technology Ladkrabang, Bangkok, 88 p. (in Thai)

[7] Choosawat, C. and Lisawadi, S. , 2018, Performance comparison of ridge regression, LASSO and adaptive LASSO in poisson regression under high-dimensional sparse data with multicollinearity, pp. 305- 314, 19th National Graduate Research Conference, Khon Kaen University, Khon Kaen. (in Thai)

[8] Algamal, Z. Y. and Lee, M. H. , 2015, Regu larized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification, Comput. Biol. Med. 67: 136-145.

[9] Sinsomboonthong, S. , 2017, Regression Analysis, Jamjuree Product, Bangkok, 494 p. (in Thai)

[10] Boonstra, P.S., Mukherjee, B. and Taylar, J. M. , 2015, A small- sample choice of the tuning parameter in ridge regression, Stat. Sin. 25: 1185-1206.

[11] Efron, B. , Hastie, T. , Johnstone, I. and Tibshirani, R., 2004, Least angle regression, Ann. Stat. 32: 407-499.

[12] Hastie, T. , Tibshirani, R. and Friedman, J. , 2009, The Elements of Statistical Learning: Data Mining Inference and Prediction, 2nd Ed., Springer, California, 527 p.

[13] Zou, H., Hastie, T. and Tibshirani, R., 2007, On the degrees of freedom of lasso, Ann. Stat. 35: 2173-2192.

[14] Phuenaree, B. , 2007, An Estimation of Variance Components for Randomized Complete Block design by Bootstrap Method, Master Thesis, Chulalongkorn University, Bangkok, 249 p. (in Thai)

[15] Rafiei, M. H. and Adeli, H. , Residential Building Data Set, Available Source: http://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set#, February 19, 2018.

[16] Thongteeraparp, W. , 1994, Development of a Statistical Package for Ridge Regression Analysis, Master Thesis, Kasetsart University, Bangkok, 171 p. (in Thai)