การเปรียบเทียบประสิทธิภาพของวิธีการถดถอยแบบพีนอลไลซ์ ในตัวแบบการถดถอยลอจิสติกภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง

Performance Comparison of Penalized Regression Methods in Logistic Regression Model under High-Dimensional Sparse Data with Muticollinerity

อภิสรา ศรีพานิช (Apisara Sripanich)* ดร.สุปราณี ลิสวัสดิ์ (Dr.Supranee Lisawadi)**
เบญจมาศ ตุลยนิติกุล (Benjamas Tulyanitikul)***

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์ในการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยและการคัดเลือกตัวแปรในตัวแบบการถดถอยลอจิสติก 5 วิธี ได้แก่วิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระในสมการถดถอยมีความสัมพันธ์เชิงเส้นสูง และเป็นการจำลองข้อมูลด้วยการจำลองมอนติคาร์โล ที่มีการทำซ้ำ 1,000 รอบ โดยเกณฑ์ที่ใช้ในการวัดประสิทธิภาพ ได้แก่ ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mPMSE) อัตราบวกเท็จ (FPR) และอัตราลบเท็จ (FNR) ผลการศึกษาพบว่าวิธีการถดถอยอิลาสติคเน็ตปรับได้ เป็นวิธีที่มีประสิทธิภาพดีที่สุด เนื่องจากค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย และอัตราบวกเท็จ มีค่าต่ำที่สุด และเมื่อประยุกต์ใช้กับชุดข้อมูลจริงพบว่าวิธีการถดถอยอิลาสติคเน็ตเป็นวิธีที่มีประสิทธิภาพดีที่สุดเช่นเดียวกับการจำลองข้อมูล

ABSTRACT

This research was aimed to compare the efficiency of regression coefficients estimation and variable selection of logistic regression model on five methods namely ridge regression, LASSO regression, elastic net regression, adaptive LASSO regression, and adaptive elastic net regression methods under high-dimensional sparse data and multicollinearity problem. The various situations and Monte Carlo simulation with 1,000 iterations were performed. The criteria for the performance measuring were False Positive Rate (FPR), False Negative Rate (FNR), and mean of prediction mean square error (mPMSE). The results showed that mean square error and False Positive Rate had the results in the same direction which the adaptive elastic net regression method performed the best. The results of real data show that the adaptive elastic net regression method performed the best as simulation.

คำสำคัญ: วิธีการถดถอยแบบพีนอลไลซ์ ข้อมูลมิติสูงแบบบางเบา การจำลองมอนติคาร์โล

Keywords: Penalized regression method, High-dimensional sparse data, Monte Carlo simulation

** นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*

*** ผู้ช่วยศาสตราจารย์ สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*

**** ผู้ช่วยศาสตราจารย์ สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*

บทนำ

ตัวแบบการถดถอยลอจิสติก (logistic regression model) เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ของตัวแปรอิสระ (independent variable) ที่มีต่อโอกาสที่จะเกิดเหตุการณ์ที่สนใจ และศึกษาการทำนาย (prediction) ความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ ในกรณีที่ตัวแปรตอบสนอง (response variable) เป็นตัวแปรทวิภาค (binary variable) เรียกว่าตัวแบบการถดถอยลอจิสติกทวิภาค (binary logistic regression) หรือเรียกว่า ตัวแบบลอจิต (logit model) และกรณีที่ตัวแปรตอบสนองเป็นแบบจำแนกประเภทแบบนามบัญญัติมากกว่า 2 ประเภท เรียกว่า ตัวแบบการถดถอยลอจิสติกพหุ (multinomial logistic regression model) และในกรณีที่ตัวแปรตอบสนองเป็นแบบจำแนกประเภทแบบอันดับมากกว่า 2 ประเภท เรียกว่า ตัวแบบการถดถอยลอจิสติกอันดับ (ordinal logistic regression model) และวิธีที่นิยมใช้ในการประมาณค่าของพารามิเตอร์สัมประสิทธิ์การถดถอย (coefficient of regression) $(\boldsymbol{\beta})$ คือวิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood method)

เนื่องจากในปัจจุบันมีข้อมูลจำนวนมากที่ถูกนำมาประยุกต์ใช้กับตัวแบบลอจิสติก เช่น ข้อมูลทางด้านการแพทย์ ข้อมูลทางด้านการเงิน เป็นต้น และในปัจจุบันได้มีข้อมูลที่มีมิติสูง (high-dimensional data) หรือข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง (sample size) อยู่จำนวนมากเช่นกัน ยกตัวอย่างเช่น ข้อมูลทางด้านชีวสารสนเทศ (bioinformatic) และทางด้านวิศวกรรมศาสตร์ เป็นต้น ซึ่งการที่มีจำนวนตัวแปรอิสระจำนวนมาก อาจส่งผลทำให้ตัวแปรอิสระเหล่านั้นมีแนวโน้มที่จะมีความสัมพันธ์เชิงเส้นซึ่งกันและกันเพิ่มสูงขึ้น หรือเรียกสถานการณ์ดังกล่าวว่า เกิดปัญหาความสัมพันธ์เชิงเส้นพหุ (multicollinearity)

ดังนั้นจากที่กล่าวมาข้างต้น ผู้วิจัยจึงสนใจศึกษาการวิเคราะห์การถดถอยในตัวแบบลอจิสติก กรณีที่ข้อมูลมีมิติสูงและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกัน ซึ่งภายใต้สถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกัน หรือสถานการณ์ที่ข้อมูลมีมิติสูง วิธีที่เรามักนิยมใช้ในการประมาณค่าสัมประสิทธิ์การถดถอย นั่นคือวิธีภาวะน่าจะเป็นสูงสุด อาจเป็นวิธีที่ไม่ดีนัก เนื่องจากเมื่อตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกันจะส่งผลทำให้ตัวประมาณ (estimator) ที่ได้มีประสิทธิภาพ (efficiency) ลดลง นั่นคือ ตัวประมาณจะมีส่วนเบี่ยงเบนมาตรฐาน (standard error) และความแปรปรวน (variance) ที่สูงขึ้น ซึ่งส่งผลทำให้การทำนายเกิดความคลาดเคลื่อนหรือไม่ถูกต้อง และความถูกต้องของการทำนายสามารถปรับปรุงให้ดีขึ้นได้โดยการลดขนาดของตัวประมาณ ดังนั้นจึงมีวิธีการหนึ่งที่นิยมใช้ในการวิเคราะห์ข้อมูลที่สามารถนำไปประยุกต์ใช้ได้กับข้อมูลที่มีมิติสูง และสามารถรับมือได้กับกรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกัน นั่นคือ วิธีการถดถอยแบบพีนอลไลซ์ (penalized regression method) ซึ่งเป็นการวิเคราะห์เพื่อหาค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอย $(\boldsymbol{\beta})$ ที่ทำให้ฟังก์ชันเป้าหมาย (objective function) ดังสมการ

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n} \left(1-y_i\right) \ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) \right] + P_\lambda\left(\boldsymbol{\beta}\right) \right\}$$

มีค่าต่ำที่สุด ภายใต้เงื่อนไข $P_\lambda\left(\boldsymbol{\beta}\right)$ ที่แตกต่างกัน ซึ่งฟังก์ชันดังกล่าวเรียกว่า penalty function โดยฟังก์ชันดังกล่าวจะมีลักษณะที่แตกต่างกันออกไป ซึ่งมีนักสถิติหลายๆ ท่านได้นำเสนอวิธีการถดถอยแบบพีนอลไลซ์ไว้ หนึ่งในนั้นคือ Hoerl, Kennard (1970) ได้เสนอวิธีการถดถอยแบบริดจ์ (ridge regression method) ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย เพื่อแก้ไขปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกัน โดยตัวประมาณที่ได้จะมีความเสถียร (stable) ในด้านความถูกต้องของการทำนาย และเป็นการปรับลดขนาดของตัวประมาณที่ได้ อีกทั้งยังช่วยลดความความแปรปรวนของตัวประมาณ แต่วิธีการถดถอยแบบริดจ์ยังขาดคุณสมบัติในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ Tibshirani (1996) จึงได้เสนอวิธีการถดถอยลาสโซ่ (least absolute shrinkage and selection operator regression method: lasso) ซึ่งมีคุณสมบัติในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยและคัดเลือกตัวแปรเข้าสู่ตัวแบบไปพร้อม ๆกัน แต่ Zou, Hastie (2005) พบว่า

วิธีการถดถอยลาสโซ่นั้นสามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้สูงสุดเท่ากับขนาดตัวอย่าง และเมื่อตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุซึ่งกันและกัน วิธีดังกล่าวจะมีประสิทธิภาพการคัดเลือกตัวแปรเข้าสู่ตัวแบบลดลง จึงได้นำเสนอวิธีการถดถอยอิลาสติคเน็ต (elastic net regression method) ซึ่งเป็นวิธีที่ถูกพัฒนาขึ้นมาเพื่อแก้ไขข้อจำกัดของวิธีการถดถอยลาสโซ่ และยังใช้ได้ดีในสถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นต่อกันสูงอีกด้วย ในเวลาต่อมา Zou (2006) พบว่าวิธีการถดถอยลาสโซ่นั้นยังขาดความคงเส้นคงวา (consistency) ในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ จึงเสนอค่าถ่วงน้ำหนัก (weight) เพิ่มเข้าไปใน penalty function ของวิธีการถดถอยลาสโซ่ ซึ่งเรียกวิธีดังกล่าวว่า วิธีการถดถอยลาสโซ่ปรับได้ (adaptive least absolute shrinkage and selection operator regression method: adaptive lasso) อีกทั้ง Zou, Zhang (2009) ยังได้นำเสนอวิธีการถดถอยอิลาสติคเน็ตปรับได้ (adaptive elastic net regression method) เพื่อเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยและการคัดเลือกของวิธีการถดถอยอิลาสติคเน็ตให้ดียิ่งขึ้น

ดังนั้นในงานวิจัยนี้ ผู้วิจัยจึงทำการเปรียบเทียบประสิทธิภาพของตัวประมาณด้วยวิธีการถดถอยแบบพีนอลไลซ์ 5 วิธีได้แก่ วิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ในตัวแบบลอจิสติก กรณีข้อมูลมีมิติสูง และตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุต่อกัน และสัมประสิทธิ์การถดถอยมีลักษณะบางเบา (sparse coefficient) นั่นคือค่าสัมประสิทธิ์การถดถอยส่วนมากเป็นศูนย์และส่วนน้อยที่ไม่เป็นศูนย์ หรือเรียกว่าตัวแบบบางเบา (sparse model) ภายใต้สถานการณ์ต่าง ๆ

วัตถุประสงค์การวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการวิเคราะห์การถดถอยแบบพีนอลไลซ์ 5 วิธี ได้แก่ วิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ในตัวแบบการถดถอยลอจิสติก กรณีข้อมูลมีมิติสูงแบบบางเบาและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ

วิธีการวิจัย

ตัวแบบการถดถอยลอจิสติก

กำหนดให้ $\mathbf{Y}=\left(y_1,y_2,y_3,...,y_n\right)'$ เป็นเวกเตอร์ตัวแปรตอบสนองที่เป็นตัวแปรทวิภาคหรือตัวแปรจำแนกประเภทที่มีค่าได้เพียง 2 ค่า เช่น สำเร็จ/ไม่สำเร็จ สินค้าชำรุด/สินค้าไม่ชำรุด เป็นต้น ที่มีขนาดตัวอย่างเท่ากับ n และมีตัวแปรอิสระ $\mathbf{X_i}=\left(x_{i1},x_{i2},x_{i3},...,x_{ip}\right)'$ ที่เป็นตัวแปรเชิงปริมาณ ดังนั้นฟังก์ชันมวลความน่าจะเป็น (probability mass function: p.m.f.) เขียนได้ดังนี้

$$f\left(y_i\mid\mathbf{X_i}\right)=\pi_i\left(\mathbf{X_i'\beta}\right)^{y_i}\left(1-\pi_i\left(\mathbf{X_i'\beta}\right)^{1-y_i}\right);i=1,2,3,...,n$$

โดยที่ $\pi_i\left(\mathbf{X_i'\beta}\right)$ คือความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งคำนวณได้จาก $\pi_i\left(\mathbf{X_i'\beta}\right)=\dfrac{\exp\left(\mathbf{X_i'\beta}\right)}{1+\exp\left(\mathbf{X_i'\beta}\right)}$ $;i=1,2,3,...,n$ โดยที่ $\boldsymbol{\beta}$ คือพารามิเตอร์สัมประสิทธิ์การถดถอยที่ไม่ทราบค่า และ n คือขนาดตัวอย่าง

ดังนั้น ตัวแบบการถดถอยลอจิสติก เขียนได้ดังนี้ $\pi_i\left(\mathbf{X_i'\beta}\right)=\dfrac{\exp\left(\mathbf{X_i'\beta}\right)}{1+\exp\left(\mathbf{X_i'\beta}\right)}$ $;i=1,2,3,...,n$

วิธีการถดถอยแบบพีนอลไลซ์

1) วิธีการถดถอยแบบริดจ์

Hoerl, Kennard (1970) ได้เสนอวิธีการถดถอยแบบริดจ์ขึ้นมา ซึ่งวิธีการดังกล่าวเป็นอีกหนึ่งวิธีที่นิยมใช้เพื่อแก้ไขปัญหาความสัมพันธ์เชิงเส้นพหุ โดยจะเป็นการปรับค่าสัมประสิทธิ์การถดถอยให้ลู่เข้าสู่ศูนย์แต่ไม่เท่ากับศูนย์ ดังนั้นตัวประมาณ $\hat{\boldsymbol{\beta}}_{Ridge}$ ทุกตัวที่ได้จะมีขนาดเล็ก และตัวประมาณที่ได้จะมีความเสถียรในความถูกต้องของการทำนาย แต่วิธีการถดถอยแบบริดจ์ยังขาดคุณสมบัติการคัดเลือกตัวแปรเข้าสู่ตัวแบบ ซึ่งส่งผลทำให้ยากต่อการอธิบายผลลัพธ์ โดยตัวประมาณสัมประสิทธิ์การถดถอยแบบริดจ์ เป็นดังนี้

$$\hat{\boldsymbol{\beta}}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n}\left(1-y_i\right)\ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right)\right] + \lambda\sum_{j=1}^{p}\beta_j^2 \right\}$$

โดยที่ $\lambda \geq 0$ คือพารามิเตอร์ปรับแต่ง (tuning parameter) หรือค่าคงตัวบวก ทำหน้าที่ควบคุมการหดตัว (shrinkage) ของตัวประมาณ $\hat{\boldsymbol{\beta}}_{Ridge}$ โดยทั่วไปจะใช้วิธี k-fold cross validation ในการหาค่าที่เหมาะสม

2) วิธีการถดถอยลาสโซ่

Tibshirani (1996) ได้นำเสนอวิธีการถดถอยลาสโซ่ขึ้นมา โดยวิธีการถดถอยดังกล่าวมีคุณสมบัติที่เป็นได้ทั้งการประมาณค่าสัมประสิทธิ์การถดถอยและการคัดเลือกตัวแปรเข้าสู่ตัวแบบได้ในคราวเดียวกัน นั่นคือตัวประมาณที่ได้จากวิธีการถดถอยลาสโซ่ส่วนใหญ่จะมีค่าเท่ากับศูนย์และบางส่วนที่ไม่เท่ากับศูนย์ แต่วิธีการถดถอยลาสโซ่นั้นยังมีข้อจำกัดในการคัดเลือกตัวแปรอิสระบางประการ นั่นคือ วิธีดังกล่าวสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้สูงสุดเท่ากับขนาดตัวอย่าง และในกรณีที่เกิดปัญหาความสัมพันธ์เชิงเส้นพหุ วิธีดังกล่าวจะมีแนวโน้มที่จะคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเพียงแค่ตัวเดียวจากกลุ่มตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นต่อกัน โดยตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่ เป็นดังนี้

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n}\left(1-y_i\right)\ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right)\right] + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \right\}$$

โดยที่ $\lambda \geq 0$ คือพารามิเตอร์ปรับแต่งหรือค่าคงตัวบวก ทำหน้าที่ควบคุมการหดตัวของตัวประมาณ $\hat{\boldsymbol{\beta}}_{LASSO}$ โดยทั่วไปจะใช้วิธี k-fold cross validation ในการหาค่าที่เหมาะสม

3) วิธีการถดถอยอิลาสติคเน็ต

Zou, Hastie (2005) ได้นำเสนอวิธีการถดถอยอิลาสติคเน็ตขึ้นมา โดยที่วิธีดังกล่าวมีคุณสมบัติเช่นเดียวกับวิธีการถดถอยลาสโซ่ นั่นคือ สามารถประมาณค่าสัมประสิทธิ์การถดถอยและคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ในคราวเดียวกัน ซึ่งวิธีการถดถอยอิลาสติคเน็ตเป็นการรวมกันระหว่างวิธีการถดถอยแบบริดจ์และวิธีการถดถอยลาสโซ่ อีกทั้งวิธีนี้ยังเหมาะสมสำหรับการวิเคราะห์ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง และเหมาะสำหรับการวิเคราะห์ที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมาก ๆ $\left(p >> n\right)$ อีกด้วย โดยตัวประมาณสัมประสิทธิ์การถดถอยแบบอิลาสติคเน็ต เป็นดังนี้

$$\hat{\boldsymbol{\beta}}_{Elastic\ \ net} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n}\left(1-y_i\right)\ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right)\right] + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2 \right\}$$

โดยที่ $\lambda_1 \geq 0, \lambda_2 \geq 0$ คือพารามิเตอร์ปรับแต่งหรือค่าคงตัวบวก ทำหน้าที่ควบคุมการหดตัวของตัวประมาณ $\hat{\boldsymbol{\beta}}_{Elastic\ \ net}$ โดยทั่วไปจะใช้วิธี k-fold cross validation ในการหาค่าที่เหมาะสม

4) วิธีการถดถอยลาสโซ่ปรับได้

Zou (2006) พัฒนาวิธีการถดถอยลาสโซ่ปรับได้ขึ้นมาเพื่อแก้ไขข้อจำกัดของวิธีการถดถอยลาสโซ่ ในเรื่องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ โดยเพิ่มค่าถ่วงน้ำหนักเข้าไปใน penalty function ของวิธีการถดถอยลาสโซ่ เพื่อให้มีคุณสมบัติในการคัดเลือกตัวแปรอิสระที่มีความแม่นและมีประสิทธิภาพมากยิ่งขึ้น ซึ่งแนวคิดคือกำหนดให้ค่าถ่วงน้ำหนักมากกับสัมประสิทธิ์การถดถอยที่มีค่าน้อยและให้ค่าถ่วงน้ำหนักน้อยกับสัมประสิทธิ์การถดถอยที่มีค่ามาก และเมื่อขนาดตัวอย่างลู่เข้าสู่อนันต์วิธีการถดถอยลาสโซ่ปรับได้จะมีความสามารถในการคัดเลือกตัวแปรอิสระเสมือนกับว่าทราบตัวแบบที่แท้จริง (true model) (เบญจมาส, อัชฌา, 2562) โดยตัวประมาณสัมประสิทธิ์การถดถอยลาสโซ่ปรับได้ เป็นดังนี้

$$\hat{\boldsymbol{\beta}}_{AL} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n} \left(1-y_i\right) \ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) \right] + \lambda \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right| \right\}$$

โดยที่ $\hat{w}_j = \dfrac{1}{\left|\hat{\beta}_{j(\text{Ridge})}\right|^{\gamma}}$; $j = 1, 2, \ldots, p, \gamma > 0$ โดยทั่วไปกำหนดให้ $\gamma = 1$ และ $\lambda \geq 0$ คือพารามิเตอร์ปรับแต่งหรือค่าคงตัวบวก ทำหน้าที่ควบคุมการหดตัวของตัวประมาณ $\hat{\boldsymbol{\beta}}_{AL}$ โดยทั่วไปจะใช้วิธี k-fold cross validation ในการหาค่าที่เหมาะสม

5) วิธีการถดถอยอิลาสติคเน็ตปรับได้

Zou, Zhang (2009) ได้ศึกษาและนำเสนอวิธีการถดถอยอิลาสติคเน็ตปรับได้ขึ้นมา เพื่อพัฒนาวิธีการถดถอยอิลาสติคให้มีประสิทธิภาพในการประมาณค่าสัมประสิทธิ์การถดถอยและการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเพิ่มยิ่งขึ้น ซึ่งเป็นการรวมกันระหว่างวิธีการถดถอยอิลาสติคเน็ตและวิธีการถดถอยแลซโซ่ปรับได้เข้าด้วยกัน อีกทั้งเมื่อขนาดตัวอย่างลู่เข้าสู่อนันต์ วิธีการถดถอยอิลาสติคเน็ตปรับได้จะมีความสามารถในการคัดเลือกตัวแปรเสมือนว่าทราบตัวแบบที่แท้จริง ซึ่งคุณสมบัติดังกล่าวไม่ปรากฏในวิธีการถดถอยอิลาสติคเน็ต และค่าประมาณของสัมประสิทธิ์การถดถอยที่ได้จากวิธีการถดถอยอิลาสติคเน็ตปรับได้จะมีลักษณะบางเบา โดยตัวประมาณสัมประสิทธิ์การถดถอยแบบอิลาสติคเน็ตปรับได้ เป็นดังนี้

$$\hat{\boldsymbol{\beta}}_{AE} = \arg\min_{\boldsymbol{\beta}} \left\{ -\left[\sum_{i=1}^{n} y_i \ln\left(\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) + \sum_{i=1}^{n} \left(1-y_i\right) \ln\left(1-\pi_i\left(\mathbf{X}_i'\boldsymbol{\beta}\right)\right) \right] + \lambda_1 \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

โดยที่ $\hat{w}_j = \dfrac{1}{\left|\hat{\beta}_{j(\text{Ridge})}\right|^{\gamma}}$; $j = 1, 2, \ldots, p, \gamma > 0$ โดยทั่วไปกำหนดให้ $\gamma = 1$ และ $\lambda_1 \geq 0, \lambda_2 \geq 0$ คือพารามิเตอร์ปรับแต่งหรือค่าคงตัวบวก ทำหน้าที่ควบคุมการหดตัวของตัวประมาณ $\hat{\boldsymbol{\beta}}_{AE}$ โดยทั่วไปจะใช้วิธี k-fold cross validation ในการหาค่าที่เหมาะสม

เกณฑ์ที่ใช้พิจารณา

1) ประสิทธิภาพในการทำนาย โดยจะวัดจากค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mean of Prediction Mean Square Error: mPMSE) ภายใต้การจำลอง (simulation) ข้อมูล t ครั้ง โดยวิธีที่ให้ค่าดังกล่าวต่ำที่เข้าใกล้ศูนย์ยิ่งแสดงถึงความแม่น (accuracy)

$$mPMSE = mean\left(PMSE_1, PMSE_2, \ldots, PMSE_t\right)$$

โดยที่ $PMSE_j = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$; $j = 1, 2, 3, ..., t$ เมื่อ t คือ จำนวนครั้งของการจำลองข้อมูล

2) ประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ โดยวัดจากอัตราบวกเท็จ (False Positive Rate: FPR) และอัตราลบเท็จ (False Negative Rate: FNR) โดยเกณฑ์การวัดประสิทธิภาพความแม่นในการคัดเลือกตัวแปรอิสระนี้จะทำการเปรียบเทียบเพียงแค่ 4 วิธี โดยยกเว้นวิธีการถดถอยแบบริดจ์ เนื่องจากวิธีดังกล่าวไม่มีคุณสมบัติในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ

2.1) อัตราลบเท็จ เป็นการวัดความผิดพลาดกรณีที่ค่าพารามิเตอร์ที่แท้จริงไม่เท่ากับศูนย์ แต่ตัวแปรอิสระไม่ถูกคัดเลือกเข้าสู่ตัวแบบ (Identify Criterion 1: IC1) ซึ่งสามารถคำนวณได้ดังนี้

$$FNR = \frac{IC1}{(p-q)\times r}$$

โดยที่ $IC1 = \left\{ j = 0, ..., p ; \beta_j \neq 0, \hat{\beta}_j = 0 \right\}$ เมื่อ $p-q$ คือ จำนวนสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดให้มีค่าเท่ากับศูนย์

2.2) อัตราบวกเท็จ เป็นการวัดความผิดพลาดกรณีที่ค่าพารามิเตอร์ที่แท้จริงเท่ากับศูนย์ แต่ตัวแปรอิสระถูกคัดเลือกเข้าสู่ตัวแบบ (Identify Criterion 2: IC2) ซึ่งสามารถคำนวณได้ดังนี้

$$FPR = \frac{IC2}{q\times r}$$

โดยที่ $IC2 = \left\{ j = 0, ..., p \ ; \beta_j = 0, \hat{\beta}_j \neq 0 \right\}$ เมื่อ q คือ จำนวนสัมประสิทธิ์การถดถอยเริ่มต้นกำหนดให้มีค่าไม่เท่ากับศูนย์

โดยในงานวิจัยครั้งนี้ เมื่อพิจารณาในส่วนของการจำลองข้อมูล ผู้วิจัยจะพิจารณาค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mPMSE) หรือเกณฑ์การวัดประสิทธิภาพการทำนายเป็นเกณฑ์ที่ใช้ตัดสินใจเลือกวิธีการถดถอยที่ดีที่สุดเป็นหลัก และจะนำเกณฑ์การวัดประสิทธิภาพในการคัดเลือกตัวแปร คือค่าอัตราลบเท็จ (FNR) และอัตราบวกเท็จ (FPR) มาพิจารณาร่วม นั่นคือเกณฑ์ที่ใช้ในการตัดสินใจเลือกวิธีการถดถอยที่ดีที่สุด คือ วิธีการถดถอยที่ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mPMSE) ต่ำที่สุด และให้ค่าดังกล่าวต่ำที่สุดสอดคล้องกับอัตราลบเท็จ (FNR) หรืออัตราบวกเท็จ (FPR) และเมื่อพิจารณาการประยุกต์ใช้กับชุดข้อมูลจริง เกณฑ์ที่ใช้ในการตัดสินใจเลือกวิธีการถดถอยที่ดีที่สุดจะพิจารณาเพียงแค่ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mPMSE) หรือพิจารณาเพียงแค่เกณฑ์การวัดประสิทธิภาพการทำนาย

การดำเนินการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงทดลอง โดยเป็นการจำลองด้วยวิธีมอนติคาร์โล (Monte Carlo method) ซึ่งการวิจัยครั้งนี้ได้จำลองข้อมูลด้วยโปรแกรม Rstudio เวอร์ชัน 3.6.3 และทำซ้ำ 1,000 ครั้งเพื่อให้ผลลัพธ์มีความคงที่หรือมีความเสถียร

ขอบเขตของการวิจัย

ผู้วิจัยจำลองข้อมูลที่ใช้ในการทดลองภายใต้สถานการณ์ต่าง ๆ ดังนี้

1. รูปแบบความสัมพันธ์ของการถดถอยลอจิสติก คือ

$$P\left(y_i=1\mid \mathbf{X_i}\right)=\pi_i\left(\mathbf{X_i'\beta}\right)=\frac{\exp\left(\mathbf{X_i'\beta}\right)}{1+\exp\left(\mathbf{X_i'\beta}\right)};i=1,2,3,...,n$$

โดยที่ $\pi_i\left(\mathbf{x_i'\beta}\right)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ

$\mathbf{X}_i=\left(x_{i1},x_{i2},...,x_{ip}\right)'$ คือ ตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณ

$\boldsymbol{\beta}=\left(\beta_1,\beta_2,\beta_3,...,\beta_p\right)'$ คือ พารามิเตอร์สัมประสิทธิ์การถดถอยที่ไม่ทราบค่า

n คือ ขนาดตัวอย่าง

p คือ จำนวนตัวแปรอิสระ

2. กำหนดให้ตัวแปรอิสระมีการแจกแจงปรกติหลายตัวแปร (multivariate normal distribution) ที่มีค่าเฉลี่ยเท่ากับศูนย์ และเมทริกซ์ความแปรปรวนร่วมเท่ากับ Σ ขนาด $p\times p$ หรือเขียนได้ว่า $N\left(0,\Sigma_{p\times p}\right)$

3. กำหนดความสัมพันธ์ระหว่างตัวแปรอิสระที่ i และ j ให้มีรูปแบบความสัมพันธ์แบบคงที่ (constant correlation) $\rho_k=r$ โดยที่พิจารณาให้ค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ระหว่างตัวแปรอิสระที่ i และ j เท่ากับ $r=0.5,0.6,0.7,0.8,0.9$ นั่นคือ

$$\Sigma_{p\times p}=\begin{bmatrix}1 & \rho_k & \cdots & \rho_k\\ \rho_k & 1 & \cdots & \rho_k\\ \vdots & \vdots & \ddots & \vdots\\ \rho_k & \rho_k & \cdots & 1\end{bmatrix}$$

4. กำหนดให้ขนาดตัวอย่าง $\left(n\right)$ ที่ศึกษา คือ 50 และ 100

5. กำหนดให้จำนวนตัวแปรอิสระ $\left(p\right)$ มีจำนวนมากกว่าขนาดตัวอย่าง $\left(p>n\right)$ โดยที่กำหนดให้จำนวนตัวแปรอิสระที่ศึกษา คือ 200 500 และ 1,000

6. กำหนดค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่ 15 ตัวแรกมีขนาดเล็กที่ไม่เท่ากับศูนย์ และ $p-15$ ตัว มีค่าเป็นศูนย์ ดังนี้ $\beta_1=1,\quad \beta_2,\beta_3=-0.5,\quad \beta_4,\beta_5,\beta_6=0.1,\quad \beta_7,...,\beta_{10}=0.05,\quad \beta_{11},...,\beta_{15}=0.01$ และ $\beta_{16},...,\beta_{p-15}=0$ ในทุกสถานการณ์

7. นำชุดข้อมูลจริงที่เป็นข้อมูลการจำแนกผู้ป่วยมะเร็งเม็ดเลือดขาวฉับพลันชนิดมัยอีมอยด์ และ ผู้ป่วยมะเร็งเม็ดเลือดขาวเฉียบพลันชนิดลิมฟอยด์ จำนวน 72 คน ผ่านการแสดงออกของยีนจำนวน 3,571 ชนิด (Golub et al., 1999) โดยที่ตัวแปรอิสระที่เป็นข้อมูลเชิงปริมาณ

ผลการวิจัย

ผลการวิจัยจากการจำลองข้อมูล

จากการจำลองข้อมูลภายใต้สถานการณ์ต่าง ๆ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา โดยมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ 2 ตัวใด ๆ เท่ากับ 0.5 0.6 0.7 0.8 และ 0.9 ผลลัพธ์ที่ได้เป็นดังตารางที่ 1 และตางรางที่ 2 ดังนี้

ตารางที่ 1 เปรียบเทียบค่าเฉลี่ยของ PMSE ของแต่ละวิธี

r	n	p	mPMSE				
			ridge	LASSO	elastic net	adaptive LASSO	adaptive elastic net
0.5	50	200	0.21271	0.20976	0.20855	0.17707	**0.16797**
		500	0.21271	0.22091	0.21690	0.17833	**0.16515**
		1000	0.21600	0.20677	0.21326	0.17183	**0.15983**
	100	200	0.20499	0.19765	0.19368	0.16602	**0.15706**
		500	0.20032	0.19495	0.19575	0.15945	**0.14875**
		1000	0.21088	0.20092	0.20856	0.16269	**0.15016**
0.6	50	200	0.20949	0.20930	0.20542	0.18092	**0.17173**
		500	0.20161	0.19774	0.19745	0.16534	**0.15348**
		1000	0.21709	0.20568	0.21509	0.16978	**0.15736**
	100	200	0.22030	0.21958	0.21633	0.18617	**0.17738**
		500	0.21121	0.20771	0.20742	0.17233	**0.15963**
		1000	0.21414	0.20430	0.20961	0.16785	**0.15469**
0.7	50	200	0.22309	0.22091	0.21737	0.19097	**0.18163**
		500	0.21283	0.20612	0.20754	0.17853	**0.16853**
		1000	0.21670	0.20739	0.21328	0.17281	**0.15990**
	100	200	0.21586	0.21242	0.20735	0.18532	**0.17592**
		500	0.21484	0.21130	0.20937	0.17755	**0.16733**
		1000	0.22527	0.21363	0.22213	0.17715	**0.16278**
0.8	50	200	0.22270	0.21907	0.21565	0.19264	**0.18500**
		500	0.22043	0.21736	0.21771	0.18691	**0.17591**
		1000	0.20940	0.20578	0.20729	0.17543	**0.16569**
	100	200	0.22485	0.22042	0.21853	0.19710	**0.18908**
		500	0.22591	0.22232	0.22326	0.19086	**0.18089**
		1000	0.21982	0.21289	0.21675	0.17717	**0.16504**
0.9	50	200	0.22742	0.21839	0.21532	0.20190	**0.19612**
		500	0.22051	0.21974	0.21704	0.19355	**0.18670**
		1000	0.21035	0.20834	0.20570	0.18489	**0.17808**
	100	200	0.22771	0.21930	0.21532	0.20521	**0.19983**
		500	0.22213	0.22044	0.21813	0.19647	**0.18909**
		1000	0.21476	0.21480	0.18620	0.21111	**0.17927**

หมายเหตุ : ตัวหนา แทน วิธีที่ให้ค่าเฉลี่ยของ PMSE ต่ำที่สุด

จากตารางที่ 1 พบว่าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยอิลาสติคเน็ตปรับได้ ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายต่ำที่สุดทุกกรณี ในทางตรงกันข้ามวิธีการถดถอยแบบริดจ์ให้ค่าดังกล่าวสูงที่สุดทุกกรณีเช่นกัน และพบว่าค่า mPMSE จะมีแนวโน้มลดลง เมื่อขนาดตัวอย่างเพิ่มขึ้นและมีค่าสัมประสิทธิ์สหสัมพันธ์ $r = 0.5$ แต่เมื่อขนาดตัวอย่างเพิ่มขึ้นแต่ค่าสัมประสิทธิ์สหสัมพันธ์เพิ่มสูงขึ้นเป็น $r = 0.6, 0.7, 0.8, 0.9$ ผลโดยส่วนมากค่า

mPMSE ก็จะมีแนวโน้มเพิ่มสูงขึ้นด้วยเช่นกัน อีกทั้งยังพบว่าเมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น ถ้าพิจารณาวิธีการถดถอยลาสโซ่ปรับได้และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ค่า mPMSE จะมีแนวโน้มลดลง ในทำนองเดียวกันวิธีการถดถอยแบบริดจ์เมื่อค่าสัมประสิทธิ์สหสัมพันธ์มีค่าสูงมาก เช่น $r = 0.8, 0.9$ ค่า mPMSE จะมีแนวโน้มลดลงเมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น

ตารางที่ 2 เปรียบเทียบค่าความผิดพลาดในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของแต่ละวิธี

r	n	P	LASSO		elastic net		adaptive LASSO		adaptive elastic net	
			FNR	FPR	FNR	FPR	FNR	FPR	FNR	FPR
0.5	50	200	0.16253	0.07919	0.38686	0.07741	0.32933	0.07800	0.65113	**0.07518**
		500	0.15586	0.03064	0.39066	0.03021	0.38526	0.03031	0.85026	**0.02958**
		1000	0.16866	0.015139	0.45706	0.01501	0.41926	0.01502	0.97920	**0.01482**
	100	200	0.35046	0.07369	0.73153	0.07045	0.56813	0.07096	1.02113	**0.06749**
		500	0.36740	0.02879	0.84753	0.02787	0.71786	0.02783	1.40246	**0.02670**
		1000	0.33480	0.01461	0.80006	0.01425	0.72560	0.01442	1.61066	**0.01396**
0.6	50	200	0.18873	0.07949	0.47620	0.07758	0.30060	0.07878	0.64140	**0.07628**
		500	0.22160	0.03021	0.61600	0.02943	0.37140	0.02982	0.90886	**0.02878**
		1000	0.18266	0.01511	0.48960	0.01498	0.45046	0.01500	1.10593	**0.01475**
	100	200	0.24133	0.07665	0.52586	0.07307	0.51046	0.07427	0.95213	**0.06968**
		500	0.31420	0.02931	0.71133	0.02859	0.66220	0.02869	1.35966	**0.02757**
		1000	0.34673	0.01484	0.83246	0.01453	0.81553	0.01458	1.70760	**0.01410**
0.7	50	200	0.14313	0.07997	0.35380	0.07820	0.28333	0.07870	0.60340	**0.07603**
		500	0.18526	0.03066	0.57186	0.03022	0.35926	0.03047	0.87500	**0.02978**
		1000	0.17140	0.01506	0.47900	0.01493	0.40820	0.01497	1.03933	**0.01468**
	100	200	0.27766	0.07660	0.64313	0.07274	0.49626	0.07441	0.93433	**0.07036**
		500	0.30413	0.02987	0.75526	0.02907	0.66080	0.02924	1.32040	**0.02823**
		1000	0.25946	0.01498	0.78380	0.01465	0.60273	0.01479	1.65973	**0.01431**
0.8	50	200	0.14506	0.07975	0.38066	0.07766	0.26820	0.07879	0.58306	**0.07637**
		500	0.14913	0.03067	0.42780	0.03028	0.32113	0.03040	0.80926	**0.02979**
		1000	0.19900	0.01516	0.65306	0.01505	0.36300	0.01512	1.04193	**0.01493**
	100	200	0.24013	0.07782	0.51206	0.07565	0.42140	0.07589	0.83193	**0.07257**
		500	0.22940	0.03031	0.55806	0.02956	0.57773	0.02980	1.24900	**0.02865**
		1000	0.28326	0.01495	0.74193	0.01469	0.73366	0.01475	1.69026	**0.01435**
0.9	50	200	0.14073	0.07940	0.43393	0.07701	0.20593	0.07882	0.51533	**0.07626**
		500	0.15806	0.03078	0.53040	0.03038	0.27926	0.03061	0.78760	**0.03008**
		1000	0.18946	0.01518	0.78380	0.01504	0.29373	0.01517	0.97860	**0.01499**
	100	200	0.21246	0.07845	0.61906	0.07509	0.31320	0.07770	0.70260	**0.07444**
		500	0.25066	0.03061	0.69226	0.03007	0.48460	0.03039	1.15880	**0.02965**
		1000	0.27833	0.01504	0.98573	0.01476	0.56800	0.01492	1.53286	**0.01458**

หมายเหตุ : ตัวหนา แทน วิธีที่ให้ค่า FPR ต่ำที่สุด

จากตารางที่ 2 พบว่าค่าความผิดพลาดกรณีที่ค่าพารามิเตอร์ที่แท้จริงเท่ากับศูนย์ แต่ตัวประมาณสัมประสิทธิ์ถดถอยที่ได้มีค่าไม่เท่ากับศูนย์หรืออัตราบวกเท็จ วิธีการถดถอยอิลาสติคเน็ตปรับได้มีค่าต่ำที่สุด โดยที่วิธีการถดถอยลาสโซ่ให้ค่าดังกล่าวสูงที่สุด แต่เมื่อพิจารณาค่าความผิดพลาดกรณีที่ค่าพารามิเตอร์ที่แท้จริงไม่เท่ากับศูนย์ แต่ตัวประมาณสัมประสิทธิ์ถดถอยที่ได้มีค่าเท่ากับศูนย์หรืออัตราลบเท็จ วิธีการถดถอยลาสโซ่ให้ค่าดังกล่าวต่ำที่สุด ซึ่งวิธีการถดถอยอิลาสติคเน็ตปรับได้ให้ค่าดังกล่าวสูงที่สุด และเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าอัตราบวกเท็จมีแนวโน้มลดต่ำลง ในขณะที่อัตราลบเท็จมีแนวโมเพิ่มสูงขึ้น

ผลการวิจัยจากชุดข้อมูลจริง

จากการนำชุดข้อมูลจริงมาวิเคราะห์ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา ผลลัพธ์ที่ได้เป็นดังตารางที่ 3 ดังนี้

ตารางที่ 3 เปรียบเทียบค่าเฉลี่ยของ PMSE ของแต่ละวิธี

n	p	mPMSE				
		ridge	adaptive LASSO	LASSO	elastic net	adaptive elastic net
72	3,571	0.01544	0.00972	0.00950	0.00850	**0.00826**

หมายเหตุ : ตัวหนา แทน วิธีที่ให้ค่าเฉลี่ยของ PMSE ต่ำที่สุด

จากตารางที่ 3 พบว่าจากการใช้ชุดข้อมูลจริง วิธีการประมาณค่าสัมประสิทธิ์การถดถอยอิลาสติคเน็ตปรับได้ ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายต่ำที่สุด และในทางตรงกันข้ามพบว่า วิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนายสูงที่สุด

อภิปรายและสรุปผลการวิจัย

การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอย 5 วิธี ได้แก่ วิธีการถดถอยแบบริดจ์ วิธีการถดถอยลาสโซ่ วิธีการถดถอยอิลาสติคเน็ต วิธีการถดถอยลาสโซ่ปรับได้ และวิธีการถดถอยอิลาสติคเน็ตปรับได้ ในกรณีที่ข้อมูลมีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์เชิงเส้นสูงหรือเกิดปัญหาความสัมพันธ์เชิงเส้นพหุ เมื่อพิจารณาในส่วนของการจำลองข้อมูล จะพิจารณาเปรียบเทียบจากความถูกต้องของการทำนายและความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ผลการวิจัย เป็นดังนี้

ถ้าพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 5 วิธี โดยพิจารณาความถูกต้องของการทำนาย พบว่า วิธีการถดถอยอิลาสติคเน็ตปรับได้มีประสิทธิภาพดีที่สุด เนื่องจากให้ค่า mPMSE ต่ำที่สุด

ถ้าพิจารณาเปรียบเทียบความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ พบว่า วิธีการถดถอยอิลาสติคเน็ตปรับได้มีค่าอัตราบวกเท็จต่ำที่สุด นั่นคือวิธีการถดถอยอิลาสติคเน็ตปรับได้มีโอกาสที่จะคัดเลือกตัวแปรอิสระนั้นเข้าสู่ตัวแบบทั้ง ๆ ที่ไม่ได้กำหนดให้ตัวแปรอิสระนั้นอยู่ในตัวแบบต่ำที่สุด แต่วิธีการถดถอยดังกล่าว จะมีโอกาสที่จะไม่คัดเลือกตัวแปรอิสระนั้นเข้าสู่ตัวแบบทั้ง ๆ ที่กำหนดให้ตัวแปรอิสระนั้นอยู่ในตัวแบบตั้งแต่แรก หรืออัตราลบเท็จสูงที่สุด

จะได้ว่า ถ้าพิจารณาทั้งประสิทธิภาพความถูกต้องของการทำนายและความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ในส่วนของการจำลองข้อมูล พบว่า วิธีการถดถอยอิลาสติคเน็ตปรับได้จะมีประสิทธิภาพดีที่สุด เมื่อข้อมูลมีมิติสูงแบบบาง

เบา และตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง เนื่องจากให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการทำนาย (mPMSE) และอัตราบวกเท็จ (FPR) มีค่าต่ำที่สุด

และเมื่อพิจารณาในส่วนของการประยุกต์ใช้กับชุดข้อมูลจริง ในกรณีที่ข้อมูลมีมิติสูง เมื่อพิจารณาจากความถูกต้องของการทำนาย พบว่า วิธีการถดถอยอิลาสติคเน็ตปรับได้มีประสิทธิภาพดีที่สุด เนื่องจากให้ค่า mPMSE ต่ำที่สุด

ดังนั้น จากที่กล่าวมาข้างต้น จะได้ข้อสรุปที่ว่าวิธีการถดถอยอิลาสติคเน็ตปรับได้มีประสิทธิภาพดีที่สุด โดยให้ผลลัพธ์ไปในทิศทางเดียวกันทั้งในการจำลองข้อมูลและการประยุกต์ใช้กับชุดข้อมูลจริง

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สุปราณี ลิสวัสดิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งกรุณาสละเวลาให้คำปรึกษาให้ทุก ๆ ด้านและคอยติดตามความคืบหน้าในทุกขั้นตอนการทำวิทยานิพนธ์ และขอขอบพระคุณ คณาจารย์สาขาสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่ให้ความรู้แก่ผู้วิจัย เพื่อมาประยุกต์ใช้ในการทำวิทยานิพนธ์

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา. การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. พิมพ์ครั้งที่ 12. กรุงเทพฯ: โรงพิมพ์สามลดา: 2560.

ฑิฆัมพร สาระกอ และนัท กุลวานิช. การเปรียบเทียบประสิทธิภาพการพยากรณ์และการคัดเลือกตัวแปรของวิธีเพิ่มลดตัวแปรแบบขั้นตอน วิธีแลสโซ่ วิธีอีลาสติคเน็ต และวิธีแลสโซ่ปรับปรุง สำหรับผลกระทบขนาดเล็ก และมีค่าสัมประสิทธิ์บางตัวเป็นศูนย์. การประชุมสัมมนาทางวิชาการ มทร.ตะวันออก มรภ.กลุ่มศรีอยุธยา และราชนครินทร์วิชาการและวิจัย; 14-16 พฤษภาคม 2557; ชลบุรี.

เบญจมาศ รุ่งศรานนท์ และอัชฌา อระวีพร. การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษภายใต้ข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี 2562; 28(8): 1346-1358.

Choosawat O, Reangsephet O, Srisuradetchai P. and Lisawadi S. Performance Comparison of Penalized Regression Methods in Poisson Regression under High-Dimensional Sparse Data with Multicollinearity. Thailand Statistician 2020; 18(3): 306-318.

Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970; 12(1): 55-67.

Golub TR, Slonim DK, Tamayo,..., Coller H, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286(5439): 531–537.

Myers RH, Montgomery DC, Vining GG. and Robinson TJ. Generalized Linear Models with Application in Engineering and the Sciences. Second Edition. United States of America: A John wiley & sons; 2010.

Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of Royal Statistical Society B (Methodological) 1996; 58(1): 267-288.

Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 2005; 67: 301-320.

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

Zou H, Zhang T. On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics 2009; 37(4): 1733-1751.