การเปรียบเทียบประสิทธิภาพของวิธีวิเคราะห์การถดถอยแบบริดจ์ แลซโซ และแลซโซแบบปรับปรุง ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันสูง

Performance Comparison of Ridge Regression, LASSO and Adaptive LASSO in Poisson Regression under High-Dimensional Sparse Data with Multicollinearity

ชุติกาญจน์ ชูสวัสดิ์ (Chutikarn Choosawat)* สุปราณี ลิสวัสดิ์ (Supranee Lisawadi)**

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์สัมประสิทธิ์การถดถอยปัวซง ด้วยวิธีการวิเคราะห์แบบริดจ์ วิธี LASSO และวิธี Adaptive LASSO ในกรณีข้อมูลมีมิติสูงแบบบางเบา และตัวแปรอิสระมีความสัมพันธ์กันค่อนข้างสูง ภายใต้สถานการณ์ต่างๆ โดยกำหนดให้พิจารณาตัวแปรอิสระเป็น 2 ลักษณะ หลังจากนั้นทำการจำลองข้อมูลด้วยโปรแกรม R และทำซ้ำ 1,000 ครั้ง เมื่อพิจารณาประสิทธิภาพของการวิเคราะห์ในแต่ละวิธี จากความถูกต้องในการพยากรณ์ พบว่าวิธี Adaptive LASSO จะให้ค่า PMSE ต่ำที่สุด เมื่อความสัมพันธ์ของตัวแปรอิสระเพิ่มสูง วิธีการวิเคราะห์การถดถอยแบบริดจ์ จะให้ค่า PMSE ต่ำลง และพิจารณาความผิดพลาดในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ พบว่า วิธี LASSO เกิดความผิดพลาดในการคัดเลือกตัวแปรอิสระแบบที่ 1 น้อยกว่า วิธี Adaptive LASSO แต่เกิดความผิดพลาดในการคัดเลือกตัวแปรอิสระแบบที่ 2 มากกว่าวิธี Adaptive LASSO

ABSTRACT

The purpose of this study was to compare the performance of Poisson regression among Ridge regression, LASSO, Adaptive LASSO under high-dimensional sparse data and independent variables which are highly correlated under various situations. We consider two types of independent variables. After performing 1,000 replications of simulation using R software, we found that the Adaptive LASSO gave the lowest PMSE. When the correlation is increased, PMSE for Ridge regression decreased. Considering a probability of incorrect selection, LASSO has a lower probability of incorrect selection type I than Adaptive LASSO. However, it has higher probability of incorrect selection type II than Adaptive LASSO.

คำสำคัญ: วิธีการวิเคราะห์แบบริดจ์ LASSO Adaptive LASSO

Keywords: Ridge regression, LASSO, Adaptive LASSO

** นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*

*** ผู้ช่วยศาสตราจารย์ สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์*

บทนำ

ตัวแบบการถดถอยปัวซง (Poisson Regression Model) เป็นวิธีการศึกษาลักษณะความสัมพันธ์ระหว่างตัวแปรที่ต้องการศึกษากับตัวแปรต่างๆ โดยเรียกตัวแปรที่ต้องการศึกษาว่า ตัวแปรตอบสนอง (Response Variable : $\mathbf{y}$) และเรียกตัวแปรต่างๆที่เกี่ยวข้องกับการศึกษาว่าตัวแปรอิสระ (Independent Variable : $\mathbf{X}$) โดยตัวแปรตอบสนองเป็นข้อมูลจำนวนนับ (Count Data)

โดยทั่วไป การวิเคราะห์การถดถอยสามารถแบ่งออกเป็น 2 ชนิด ถ้ามีตัวแปรอิสระในตัวแบบมีเพียงตัวเดียว เรียกว่า การถดถอยเชิงเส้นอย่างง่าย (Simple Regression) แต่ถ้ามีตัวแปรอิสระมีมากกว่าสองตัวขึ้นไป จะเรียกว่า การถดถอยพหุคูณ (Multiple Regression) โดยในงานวิจัยนี้สนใจในกรณีที่มีจำนวนตัวแปรอิสระเป็นจำนวนมาก ซึ่งอาจทำให้เกิดปัญหาความสัมพันธ์กันเองของตัวแปรอิสระ ซึ่งเป็นปัญหาที่พบ เมื่อจำนวนตัวแปรอิสระมีจำนวนมาก เรียกว่า ปัญหามัลติคอลิเนียร์ลิตี้ (Multicollinearity) หรือเกิดภาวะร่วมเชิงเส้น ในบางครั้งการที่มีตัวแปรอิสระมากเกินไป แต่ขนาดของข้อมูลที่เราศึกษาหรือขนาดตัวอย่างมีจำนวนน้อย นั่นคือ ขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระมากๆ $(n = p)$ ทำให้เกิดความไม่เพียงพอต่อการวิเคราะห์ข้อมูล จึงก่อให้เกิดปัญหาในการวิเคราะห์เช่นกัน โดยเรียกลักษณะข้อมูลนี้ว่า ข้อมูลที่มีมิติสูง (High-Dimensional)

ดังนั้น งานวิจัยนี้จึงได้ศึกษาการวิเคราะห์การถดถอยพหุคูณในตัวแบบปัวซง ในลักษณะข้อมูลที่มีมิติสูงและตัวแปรอิสระมีความสัมพันธ์กันสูงอีกด้วย เมื่อข้อมูลอยู่ในลักษณะนี้ วิธีภาวะน่าจะเป็นสูงสุด ซึ่งเป็นวิธีที่นิยมใช้ในการวิเคราะห์การถดถอยปัวซง อาจจะไม่เป็นวิธีที่ดีนัก เพราะเมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง ส่งผลให้ตัวประมาณด้วยวิธีนี้มีความแปรปรวนสูงขึ้น และทำให้การประมาณค่าไม่ถูกต้อง และไม่มีประสิทธิภาพ นอกจากนี้การอธิบายผลของตัวแบบมีความยากและซับซ้อนมากขึ้น ดังนั้น วิธีภาวะน่าจะเป็นสูงสุด อาจเป็นวิธีที่ไม่เหมาะสมกับสถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์กันสูง หรือข้อมูลที่มีมิติสูง จึงมีวิธีการหนึ่งที่นิยมใช้ในการวิเคราะห์ข้อมูลที่มีลักษณะดังกล่าว คือ วิธีวิเคราะห์การถดถอยแบบพีนอลไลซ์ (Penalized Regression) ซึ่งเป็นการวิเคราะห์เพื่อหาค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอย $(\boldsymbol{\beta})$ ที่ทำให้ฟังก์ชันเป้าหมาย (Objective Function) ดังสมการนี้

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

ให้มีค่าน้อยที่สุด ภายใต้เงื่อนไขที่แตกต่างกัน ที่เรียกว่า ฟังก์ชันพีนอลตี้ (Penalty Function) โดยฟังก์ชันนี้จะมีด้วยกันหลายรูปแบบ ซึ่งจะให้ค่าประมาณพารามิเตอร์สัมประสิทธิ์การถดถอยแตกต่างกัน ในงานวิจัยนี้จะศึกษาวิธี Penalized Regression 3 วิธี คือ วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge Regression) วิธีการวิเคราะห์การถดถอยแบบแลซโซ (LASSO) และวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง (Adaptive LASSO)

Hoerl, Kennard (1970) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ ในตัวแบบเชิงเส้น และต่อมา Kristofer, Ghazi (2011) ได้นำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์สำหรับตัวแบบการถดถอยปัวซง ซึ่งเป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง สามารถคำนวณหาค่าสัมประสิทธิ์การถดถอย จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood จากวิธีประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุด แต่เนื่องจากวิธีของริดจ์ สามารถใช้ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยทุกตัว ให้มีค่าเข้าใกล้ศูนย์ หรือทำให้มีขนาดเล็กลง (Shrink) และตัวประมาณที่ได้จะมีความเสถียร แต่วิธีนี้ยังขาดคุณสมบัติในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ

Tibshirani (1996) ได้เสนอวิธีการวิเคราะห์การถดถอยแบบแลซโซ ในตัวแบบเชิงเส้น และต่อมา Park, Hastie (2007) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซโซ สำหรับตัวแบบการถดถอย

ปัวซง วิธีนี้ไม่เพียงแต่ประมาณค่าสัมประสิทธิ์การถดถอย ยังมีคุณสมบัติคัดเลือกตัวแปรเข้าสู่ตัวแบบ ถึงแม้ว่าจะสามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้นั้น แต่ถ้ากรณีที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง หรือเกิดภาวะร่วมเชิงเส้น วิธีแลซโซ จะเลือกตัวแปรอิสระเพียงตัวเดียวจากตัวแปรที่มีความสัมพันธ์กันในกลุ่มนั้น โดยไม่คำนึงว่าตัวแปรอิสระนั้นมีความสัมพันธ์กับตัวแปรสนองมากที่สุดหรือไม่ ดังนั้น วิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีแลซโซ ยังมีข้อจำกัดบางประการในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ที่ไม่มีความคงเส้นคงวา (Consistency) ต่อมา Park, Hastie (2007) ได้ศึกษาต่อจากงานของ Fan, Li (2001) จากข้อจำกัดที่ว่า วิธี LASSO ยังขาดคุณสมบัติความคงเส้นคงวา ในตัวแบบการถดถอยปัวซง ส่งผลทำให้การประมาณค่าด้วยวิธีนี้มีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบมากขึ้น ทำให้ช่วยลดเอนเอียงในการประมาณค่าดีกว่าวิธีของแลซโซแบบเดิม และเรียกว่าวิธีการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีแลซโซแบบปรับปรุง

Hossain, Lee (2012) ได้ศึกษาการเปรียบเทียบประสิทธิภาพการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีพีนอลไลซ์และวิธีการหดค่าลงของการประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอย (Shrinkage) ในตัวแบบการถดถอยปัวซง โดยศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยหลายวิธี เทียบกับ วิธีการวิเคราะห์สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด โดยพบว่า ในการจำลองข้อมูลที่มีมิติต่ำ $(n < p)$ วิธีการวิเคราะห์แบบแลซโซและวิธีการวิเคราะห์แลซโซแบบปรับปรุง จะมีประสิทธิภาพ เมื่อตัวแปรอิสระที่ไม่มีผล $(\beta = 0)$ มีเป็นจำนวนมากอยู่ในตัวแบบ

Algamal, Lee (2015) ได้ศึกษาในข้อมูลที่มีมิติต่ำ ในตัวแบบการถดถอยปัวซง นอกจากนี้ตัวแปรอิสระมีความสัมพันธ์กันสูงอีกด้วย โดยใช้วิธีการวิเคราะห์แบบพีนอลไลซ์เพื่อแก้ปัญหาที่เกิดขึ้น นอกจากนี้ได้ศึกษาการคัดเลือกตัวแปรของวิธีการวิเคราะห์การถดถอยของทั้ง 3 วิธี ได้แก่ วิธีการวิเคราะห์แบบแลซโซ, วิธีการวิเคราะห์แลซโซแบบปรับปรุง, วิธีการวิเคราะห์แบบลาแลซโซ (RALASSO) ซึ่งเป็นการรวมวิธีการวิเคราะห์แบบแลซโซและแลซโซแบบปรับปรุงเข้าด้วยกัน เพื่อเป็นการแก้ปัญหาการเกิดข้อมูลที่มีมิติสูง และให้ผลสรุปว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุงมีประสิทธิภาพสูงสุด ในตัวแบบการถดถอยปัวซง กรณีข้อมูลมีมิติต่ำ แต่วิธีการวิเคราะห์แบบลาแลซโซ จะแสดงให้เห็นถึงประสิทธิภาพของการคัดเลือกตัวแปรได้อย่างมีประสิทธิภาพ และมีความแม่นยำมากขึ้น

Ivanoff et al. (2016) ได้ศึกษาข้อมูลที่มิติสูง ในตัวแบบการถดถอยปัวซง โดยในงานวิจัยนี้จะประมาณค่าสัมประสิทธิ์การถดถอย ด้วยวิธีการวิเคราะห์แบบแลซโซ และวิธีการวิเคราะห์แบบกรุ๊ปแลซโซ (Group LASSO) ซึ่งจะเป็นการรวมวิธีการวิเคราะห์ทั้ง 2 วิธีดังกล่าว เพื่อใช้ในการวิเคราะห์การถดถอยหรือประมาณค่าสัมประสิทธิ์การถดถอย ในส่วนของการคัดเลือกตัวแปรอิสระ จะขึ้นอยู่กับการถ่วงน้ำหนักในฟังก์ชันพีนอลตี้ ที่ต้องใช้วิธีการนี้ เนื่องจากวิธีพื้นฐานสามารถใช้ในตัวแบบเกาส์เซี่ยน แต่ไม่สามารถนำมาใช้ได้โดยตรงกับตัวแบบปัวซง เพราะชนิดของข้อมูลที่ใช้แตกต่างกัน ดังนั้นในงานวิจัยนี้จะแสดงเห็นถึงประสิทธิภาพในการวิเคราะห์การถดถอยด้วยวิธีการวิเคราะห์แบบ LASSO และ วิธี Group LASSO เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ในตัวแบบปัวซง

ดังนั้นในงานวิจัยนี้ ผู้วิจัยจึงได้ศึกษาการเปรียบเทียบประสิทธิภาพของตัวประมาณด้วยวิธีพีนอลไลซ์ 3 วิธี ได้แก่ Ridge regression, LASSO และ Adaptive LASSO ในตัวแบบการถดถอยปัวซง กรณีที่ข้อมูลมีมิติสูง และตัวแปรอิสระมีความสัมพันธ์กันสูง และมีค่าสัมประสิทธิ์การถดถอยแบบบางเบา เรียกว่า ตัวแบบบางเบา (Sparse Model) หรือกล่าวอีกอย่างหนึ่งว่าค่าสัมประสิทธิ์การถดถอยส่วนน้อยไม่เป็นศูนย์ และส่วนมากเป็นศูนย์ อยู่ภายใต้สถานการณ์ที่แตกต่างกันหลายเงื่อนไข โดยเงื่อนไขที่กำหนดจะครอบคลุมคุณสมบัติของตัวประมาณในแต่ละวิธี

วัตถุประสงค์การวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเครื่องมือในการวิเคราะห์การถดถอยแบบ penalized regression 3 วิธี ได้แก่ Ridge regression , LASSO และ Adaptive LASSO ในตัวแบบถดถอยปัวซง (Poisson Regression) สำหรับข้อมูลที่มีมิติสูงแบบบางเบา ในกรณีที่ตัวแปรอิสระเกิดภาวะร่วมเชิงเส้น โดยจำลองข้อมูลด้วยโปรแกรม R ภายใต้สถานการณ์ต่างๆ

วิธีการวิจัย

ตัวแบบการถดถอยปัวซง (Poisson Regression Model)

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

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

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

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

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

ดังนั้น ตัวแบบการถดถอยปัวซง $\qquad Y_i=E[Y_i]+\varepsilon_i \qquad , \quad i=1,2,3,\dots,n$

โดยที่ ค่าเฉลี่ยของตัวแปรตอบสนองในแต่ละ i หรือ $E[Y_i]$ จะมีค่าเท่ากับ μ_i จะได้

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

การหาตัวประมาณสัมประสิทธิ์การถดถอยในตัวแบบปัวซง

วิธีการเคราะห์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum likelihood)

การประมาณค่าพารามิเตอร์สัมประสิทธิ์การถดถอยด้วยวิธีนี้ จะอาศัยหลักการของความน่าจะเป็น โดยหาฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของตัวแปรสุ่ม แล้วจึงหาค่าสูงสุดของฟังก์ชันภาวะน่าจะเป็นนี้ โดยเทียบกับตัวประมาณที่ยังไม่ทราบค่า ดังนี้

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

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

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

และตัวแบบการถดถอยปัวซง จะได้ฟังก์ชันภาวะน่าจะเป็น (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

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

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

เนื่องจากสมการข้างต้น ไม่อยู่ในรูปเชิงเส้น เพื่อประมาณค่า $\boldsymbol{\beta}$ ต่อไป โดยใช้วิธีการ iterative weighted least square (IWLS) อัลกอริทึม ในการแก้หาคำตอบ จากบทความของ Mansson, Shukur (2011) ได้นำเสนอตัวประมาณสัมประสิทธิ์ถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด ดังนี้

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

โดยที่ $\hat{\mathbf{W}} = diag\left[\hat{\mu}_i\right]$ และ $\hat{\mathbf{z}}$ คือ เวกเตอร์ของ $\hat{z}_i = \log\left(\hat{\mu}_i\right) + \dfrac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$

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

วิธีการเคราะห์การถดถอยด้วยวิธีพีนอลไลซ์ (Penalized Regression)

1) วิธีการวิเคราะห์การถดถอยแบบริดจ์

Kristofer, Ghazi (2011) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ โดยตัวประมาณริดจ์สำหรับตัวแบบการถดถอยปัวซง เป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง (Multicollinearity) หรือเกิดภาวะร่วมเชิงเส้น สามารถคำนวณหาค่าสัมประสิทธิ์การถดถอย จากการหาค่าต่ำสุด ของฟังก์ชันลบของลอการิทึมของฟังก์ชันน่าจะเป็น (Log Likelihood) จากวิธีภาวะน่าจะเป็นสูงสุด ภายใต้ L_2 penalty บน $\boldsymbol{\beta}$ โดยที่ $L_2 = \|\boldsymbol{\beta}\|_2 = \sum_{j=1}^{p}\beta_j^{\,2}$

ตัวประมาณสัมประสิทธิ์การถดถอยแบบริดจ์ ของ $\boldsymbol{\beta}$ สามารถนิยามได้โดย

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

เมื่อ λ คือพารามิเตอร์ปรับแต่ง(Tuning Parameter) ซึ่งควบคุมขนาดการหดตัว (Shrinkage) ของ $\hat{\boldsymbol{\beta}}_{ridge}$

2) วิธีการวิเคราะห์การถดถอยแบบแลซโซ

Park, Hastie (2007) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบแลซ-โซ สำหรับตัวแบบการถดถอยปัวซง โดยวิธีการวิเคราะห์นี้เป็นวิธีที่นิยมสำหรับการประมาณค่าและคัดเลือกตัวแปรในคราวเดียวกัน สามารถคำนวณหาค่าสัมประสิทธ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด

ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 penalty บน $\boldsymbol{\beta}$ โดยที่ $L_1 = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$

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

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

ซึ่ง λ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดของการหดตัว (Shrinkage) ของ $\hat{\boldsymbol{\beta}}_{lasso}$

3) วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง

Park, Hastie (2007) ได้ศึกษาต่อจากงานของ Fan, Li (2001) โดยใช้วิธีการวิเคราะห์ด้วยวิธีแลซโซแบบปรับปรุงในตัวแบบถดถอยปัวซง เพื่อเพิ่มประสิทธิภาพในการการประมาณค่าพารามิเตอร์และการคัดเลือกตัวแปรมากขึ้น โดยมีการให้ความสำคัญกับตัวแปร โดยการถ่วงน้ำหนักให้กับตัวแปร สามารถคำนวณหาค่าสัมประสิทธิ์ถดถอย (Coefficient) จากการหาค่าต่ำสุด ของฟังก์ชันลบของ log likelihood ภายใต้ L_1 penalty บน $\boldsymbol{\beta}$ โดยที่ $L_2 = \|\boldsymbol{\beta}\|_2 = \sum_{j=1}^{p} \beta_j^{\,2}$

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

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

ซึ่ง λ คือ พารามิเตอร์ปรับแต่ง (Tuning Parameter) ควบคุมขนาดของการหดตัว (Shrinkage) ของ $\hat{\boldsymbol{\beta}}_{Adaplasso}$

เกณฑ์ที่ใช้พิจารณา

1. ประสิทธิภาพในการคัดเลือกตัวแปรของตัวประมาณ เมื่อกำหนด $\boldsymbol{\beta}$ ตามสถานการณ์ต่างๆ โดยค่าสัมประสิทธิการถดถอยส่วนใหญ่มีค่าเป็นศูนย์และบางส่วนไม่ได้มีค่าเป็นศูนย์ ทำการนับการจำนวนตัวแปรที่ถูกคัดเลือกผิดพลาดไปจากตัวแบบโดยความผิดพลาดในการคัดเลือกตัวแปร 2 แบบ คือ กรณีที่ค่าพารามิเตอร์ไม่เท่ากับ 0 แต่ตัวแปรอิสระถูกคัดเลือกเข้าไปในตัวแบบเท่ากับ 0 (Identify Criterion 1: IC1) และกรณีที่ค่าพารามิเตอร์เท่ากับ 0 แต่ตัวแปรอิสระถูกคัดเลือกในตัวแบบไม่เท่ากับ 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

$$P(\text{IC1}) = \frac{\text{IC1}}{15 \times m} \quad \text{และ} \quad P(\text{IC2}) = \frac{\text{IC2}}{(p-15) \times m}$$

เมื่อ m คือ จำนวนครั้งของการจำลองข้อมูล และ 15 คือ จำนวนพารามิเตอร์ที่กำหนดให้ไม่เท่ากับ 0

2. ประสิทธิภาพของการพยากรณ์โดยวัดจากค่ามัธยฐานของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของค่าพยากรณ์ (Prediction Mean Square Error : PMSE) มีค่าน้อยที่สุด

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

การดำเนินการวิจัย

การวิจัยนี้เป็นการวิจัยเชิงทดลอง โดยวิธีการจำลอง (Simulation Study) ซึ่งการวิจัยครั้งนี้ได้จำลองข้อมูลด้วยโปรแกรม R เวอร์ชั่น 3.1.2 โดยทำซ้ำ 1,000 ครั้ง ขอบเขตของงานวิจัยและขั้นตอนในการดำเนินการทดลอง มีดังนี้

ขอบเขตของการวิจัย

ผู้วิจัยจำลองข้อมูลที่ใช้ในการทดลองภายใต้สถานการณ์ต่างๆ ดังนี้

1. ขนาดตัวอย่าง $n = 25, 50$
2. จำนวนตัวแปรอิสระ $p = 50, 100, 200$
3. กำหนดให้ $\mathbf{X} : N(\mathbf{0}, \mathbf{\Sigma})$, $\boldsymbol{\mu} = \exp(\mathbf{X}_i^T \boldsymbol{\beta})$ และ $\boldsymbol{\varepsilon} : Poi(1)$ เพื่อสร้างเวกเตอร์ตัวแปรตาม $(\mathbf{y})$ ที่มีการแจกแจงปัวซง โดยมีเวกเตอร์ค่าเฉลี่ยเท่ากับ $\boldsymbol{\mu}$ จากตัวแบบ $y_i = \exp(\mathbf{X}_i^T \boldsymbol{\beta}) + \varepsilon_i$
4. กำหนดสถานการณ์ของการทดลองสำหรับแต่ละค่า n และ p โดยกำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระ $\mathbf{X}$ และค่าพารามิเตอร์ $\boldsymbol{\beta}$ ในแต่ละกรณี ดังนี้

4.1 ตัวแปรอิสระที่สัมพันธ์กันแบ่งออกเป็น 2 กลุ่ม

โดยกลุ่มแรก คือ ตัวแปรอิสระ 15 ตัวที่สัมพันธ์กัน $\left(x_{i(1)}, \dots, x_{i(15)}\right)$

และกลุ่มที่สอง คือ ตัวแปรอิสระตัวที่เหลือที่สัมพันธ์กัน $\left(x_{i(16)}, \dots, x_{i(p-15)}\right)$ โดยที่ตัวแปรอิสระในกลุ่ม 1 และกลุ่มที่ 2 เป็นอิสระต่อกัน ซึ่งค่าสหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)}, \dots, x_{i(15)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r = 0.5, \quad \mathrm{i}, j = 1, \dots, 15$

และค่าสหสัมพันธ์(Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ$\left(x_{i(16)}, \dots, x_{i(p-15)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r = 0.5, \quad \mathrm{i}, j = 1, \dots, 15$ เมื่อ

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

และ $\beta_{16}, \dots, \beta_{p-15}$

4.2 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.1 ยกเว้น $r = 0.9$

4.3 ตัวแปรอิสระทั้งหมด p ตัวที่สัมพันธ์กัน $\left(x_{i(1)}, \dots, x_{i(p)}\right)$ซึ่งค่าสหสัมพันธ์ (Pairwise Correlation) ระหว่างตัวแปรอิสระที่ i และ j ของ $\left(x_{i(1)}, \dots, x_{i(p)}\right)$ คือ $r^{|i-j|}$ เมื่อ $r = 0.5, \quad \mathrm{i}, j = 1, \dots, p$

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

และ $\beta_{16}, \dots, \beta_{p-15}$

4.4 กำหนดรูปแบบความสัมพันธ์ของตัวแปรอิสระเหมือน 3.3 ยกเว้น $r = 0.9$

ผลการวิจัย

จากการจำลองข้อมูลภายใต้สถานการณ์ต่างๆเพื่อเปรียบเทียบประสิทธิภาพด้วยวิธีการวิเคราะห์แบบริดจ์ (Ridge-regression) วิธีการวิเคราะห์แบบแลซโซ (LASSO) และวิธีการวิเคราะห์แลซโซแบบปรับปรุง (Adaptive LASSO) ในกรณีที่ข้อมูลมีมิติสูง $(n = p)$ โดยในกรณีที่ 1 และ 2 เป็นการแบ่งตัวแปรอิสระออกเป็น 2 กลุ่ม ที่อิสระกัน โดยภายในกลุ่มมีสหสัมพันธ์ระหว่างตัวแปรอิสระ 2 ตัวใดๆ ขนาดปานกลาง $(r = 0.5)$ และสูง $(r = 0.9)$ ตามลำดับ และกรณีที่ 3 และ 4 เป็นกรณีที่ตัวแปรอิสระมีกลุ่ม และมีความสัมพันธ์ระหว่างตัวแปรอิสระ ปานกลาง และสูง เช่นกัน

ตางรางที่ 1 เปรียบเทียบค่ามัธยฐานของ PMSE ของแต่ละวิธี

		median PMSE					
		ตัวแปรอิสระ 2 กลุ่ม			**ตัวแปรอิสระ 1 กลุ่ม**		
n	**p**	**Ridge**	**Lasso**	**Adaptive LASSO**	**Ridge**	**Lasso**	**Adaptive LASSO**
		$r = 0.5$					
		กรณี 1			กรณี 3		
25	50	11.993590	6.204300	4.707064	3.286708	3.066254	2.57964
25	100	3.308921	3.399607	4.134070	5.600545	5.760652	4.005341
25	200	5.689742	6.201270	5.423309	4.683963	5.654731	3.687385
50	50	6.878677	3.800490	3.128080	5.718789	3.115685	3.134062
50	100	2.727958	2.819876	2.145643	3.337378	3.115685	2.41614
50	200	2.810461	3.043582	2.143438	3.579415	3.650788	2.246439
		$r = 0.9$					
		กรณี 2			กรณี 4		
25	50	2.515508	2.457489	2.330324	2.192461	2.189538	2.065781
25	100	2.180289	2.190415	2.110024	2.091885	2.100496	2.33808
25	200	1.809780	1.917122	2.113250	2.372414	2.463096	2.129494
50	50	2.348484	2.241956	2.090518	2.236713	2.100975	1.981776
50	100	2.368332	2.344496	2.098472	2.209969	2.24266	2.128063
50	200	1.870907	1.930425	1.651600	1.906751	1.945714	1.678539

จากตารางที่ 1 พบว่า กรณีที่ 1 วิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของ PMSE ต่ำสุด ในทุกกรณี ยกเว้นกรณีที่ $p = 100$ วิธีการวิเคราะห์การถดถอยแบบริดจ์ จะให้ค่ามัธยฐานของ PMSE ต่ำสุด เมื่อ $p = 100, 200$ ค่ามัธยฐานของ PMSE ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ จะให้ค่าต่ำกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซ แต่สูงกว่า วิธีการวิเคราะห์การถดถอยแบบริดจ์ และเมื่อขนาดตัวอย่างเพิ่มมากขึ้น จะส่งผลทำให้ค่ามัธยฐานของ PMSE ของทั้ง 3 วิธี มีค่าลดลง

กรณีที่ 2 วิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของ PMSE ต่ำกว่าวิธีกว่าแบบแลซโซ และวิธีการวิเคราะห์แบบริดจ์ ตามลำดับ ในทุกกรณี ยกเว้น เมื่อขนาดตัวอย่าง $n = 25$ พบว่า เมื่อจำนวนตัวแปรอิสระ $p = 200$ ค่ามัธยฐานของ PMSE ของวิธีการวิเคราะห์การถดถอยแบบริดจ์ ต่ำที่สุด และเมื่อขนาดตัวอย่าง $n = 50, p = 200$ ค่ามัธยฐานของ PMSE ของวิธีการวิเคราะห์การถดถอยแบบแลซโซ จะมีค่าสูงที่สุด และเมื่อขนาดตัวอย่างเพิ่มมากขึ้น จะส่งผลทำให้ค่ามัธยฐานของ PMSE ของทั้ง 3 วิธี มีค่าลดลงเช่นเดียวกับกรณีที่ 1

เมื่อเทียบกรณีที่ 1 และ 2 จะเห็นว่า เมื่อความสัมพันธ์เพิ่มขึ้น จาก $r = 0.5$ เป็น $r = 0.9$ วิธีการวิเคราะห์การถดถอยแบบริดจ์จะมีค่ามัธยฐานของ PMSE ต่ำลง

กรณี 3 และ กรณี 4 ซึ่งเป็นกรณีที่มีตัวแปรอิสระกลุ่มเดียว และตัวแปรอิสระ 2 ตัวใดๆ มีความสัมพันธ์ระดับปานกลาง และสูง ตามลำดับ ยังคงให้ผลไปในทิศทางเดียวกับกรณีที่ 1 และ 2 ตามลำดับ และวิธีการวิเคราะห์การถดถอยแบบแลซโซแบบปรับปรุง จะให้ค่ามัธยฐานของ PMSE ต่ำสุด ในทุกกรณี

ตารางที่ 2 เปรียบเทียบความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรของวิธี LASSO และ Adaptive LASSO

		Probability							
		ตัวแปรอิสระ 2 กลุ่ม				**ตัวแปรอิสระ 1 กลุ่ม**			
n	**p**	**LASSO**		**Adaptive LASSO**		**LASSO**		**Adaptive LASSO**	
		IC1	**IC2**	**IC1**	**IC2**	**IC1**	**IC2**	**IC1**	**IC2**
		$r = 0.5$							
		กรณีที่ 1				กรณีที่ 3			
25	50	0.238800	0.358314	0.239933	0.350400	0.162733	0.392371	0.239666	0.383342
25	100	0.234000	0.167529	0.325933	0.163800	0.279866	0.158058	0.366200	0.155552
25	200	0.417933	0.077464	0.508400	0.076600	0.252133	0.078772	0.424200	0.077200
50	50	0.352066	0.332571	0.367733	0.330000	0.322733	0.328371	0.293533	0.325942
50	100	0.281533	0.163376	0.469266	0.155400	0.393266	0.153764	0.490666	0.148541
50	200	0.425800	0.075145	0.698533	0.071700	0.453533	0.074043	0.733000	0.071908
		$r = 0.9$							
		กรณีที่ 2				กรณีที่ 4			
25	50	0.086600	0.403285	0.105600	0.395942	0.104266	0.403085	0.136400	0.397942
25	100	0.129933	0.167635	0.179066	0.164811	0.122200	0.169552	0.178066	0.165752
25	200	0.123466	0.079124	0.235066	0.077475	0.126600	0.079021	0.253333	0.077702
50	50	0.140466	0.391000	0.168666	0.384685	0.129466	0.386142	0.135333	0.377085
50	100	0.184133	0.163717	0.248066	0.160494	0.175733	0.157588	0.217333	0.153647
50	200	0.157466	0.078945	0.339933	0.077113	0.162933	0.078335	0.333266	0.076567

จากตารางที่ 2 พบว่า วิธีการวิเคราะห์การถดถอยแบบแลซโซ มีความน่าจะเป็นที่เกิดความผิดพลาดในการคัดเลือกตัวแปรซึ่งเมื่อกำหนดให้พารามิเตอร์มีค่าไม่เท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าเป็น 0 (IC1) น้อยกว่าวิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง แต่เมื่อพิจารณา IC2 ซึ่งเป็นความผิดพลาดที่เมื่อกาหนดให้พารามิเตอร์มีค่าเท่ากับ 0 แต่ตัวประมาณสัมประสิทธิ์การถดถอยที่ได้มีค่าไม่เป็น 0 พบว่า วิธีการวิเคราะห์การถดถอยแลซโซแบบปรับปรุง มีโอกาสเกิดความผิดพลาดในการคัดเลือกตัวแปรน้อยกว่าวิธีการวิเคราะห์การถดถอยแบบแลซโซ ในทุกกรณี และความน่าจะเป็นที่จะเกิด IC2 ลดลงเมื่อจานวนตัวแปรอิสระ p เพิ่มขึ้น แต่เมื่อจานวนตัวแปรอิสระ p เท่ากับขนาดตัวอย่าง n จะเห็นว่า โอกาสที่จะเกิด IC2 มีค่าสูงที่สุด

อภิปรายและสรุปผลการวิจัย

ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอย ด้วยวิธีการวิเคราะห์แบบริดจ์ (Ridge regression) วิธีการวิเคราะห์แบบแลซโซ (LASSO) และวิธีการวิเคราะห์แลซโซแบบปรับปรุง (Adaptive LASSO) ในกรณีที่ข้อมูลมีมิติสูง $(n = p)$ และตัวแปรอิสระมีความสัมพันธ์กันค่อนข้างสูง นั่นคือ $r = 0.5, 0.9$ โดยจะพิจารณาจากความถูกต้องในการพยากรณ์ และความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ผลการวิจัยเป็นดังนี้

ถ้าพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี โดยพิจารณาจากความถูกต้องในการพยากรณ์ จะพบว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุง (Adaptive LASSO) ดีที่สุดในกรณีที่ข้อมูลมีมิติสูง ถึงแม้ว่าในกรณีนี้ที่ขนาดตัวอย่างเท่ากับจำนวนตัวแปรอิสระ ก็ยังให้ผลเช่นเดียวกัน

ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันเพิ่มมากขึ้น วิธีการวิเคราะห์แบบริดจ์ (Ridge regression) ก็จะมีประสิทธิภาพเพิ่มมากขึ้นด้วยเช่นกัน โดยพิจารณาจากความถูกต้องในการพยากรณ์ ซึ่งสอดคล้องกับผลงานวิจัยของ Hoerl, Kennard (1970) ที่ว่า วิธีการวิเคราะห์แบบริดจ์ สามารถแก้ไขปัญหาภาวะร่วมเชิงเส้นได้

ถ้าพิจารณาเปรียบเทียบความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จะพบว่า วิธีการวิเคราะห์แลซโซแบบปรับปรุง มีโอกาสในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆที่ตัวแปรอิสระนั้นไม่ควรอยู่ในตัวแบบมากกว่าวิธีการวิเคราะห์แบบแลซโซแต่มีโอกาสที่จะไม่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งๆ ที่ตัวแปรอิสระนั้นควรอยู่ในตัวแบบน้อยกว่าวิธีการวิเคราะห์แบบแลซโซ

ดังนั้น ถ้าพิจารณาทั้งประสิทธิภาพในการพยากรณ์และความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ วิธีการวิเคราะห์แลซโซแบบปรับปรุง จะให้ประสิทธิภาพดีที่สุด เมื่อข้อมูลมีมิติสูงแบบบางเบา

ผลการวิจัยในตัวแปรอิสระ 1 กลุ่ม และ 2 กลุ่ม ให้ผลไปในทิศทางเดียวกัน

กิตติกรรมประกาศ

ขอขอบพระคุณผศ.ดร.สุปราณี ลิสวัสดิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งกรุณาสละเวลาให้คำปรึกษา ความรู้และคำแนะนำตลอดระยะการทำวิทยานิพนธ์

ขอขอบพระคุณ คณาจารย์สาขาวิชาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่ให้ความรู้แก่ผู้วิจัย เพื่อมาประยุกต์ใช้ในวิทยานิพนธ์ และขอขอบพระคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ สำหรับทุนการศึกษาตลอดการศึกษาระดับปริญญาโทของผู้วิจัย

เอกสารอ้างอิง

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

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 2001; 96(456): 1348-1360.

Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society 2007; 69(4): 659-677.

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

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

Algamal ZY, Lee MH. Adjusted adaptive lasso in high-dimensional poisson regression model. Modern Applied Science 2015; 9(4): 170.

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