การเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่

On Tuning Parameter Selection of Lasso Regression

จุฑาทิพย์ นันทสุวรรณ* และวิฐรา พึ่งพาพงศ์

ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

แขวงเมืองใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330

Jutatip Nuntasuwan* and Vitara Pungpapong

Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University,

Wangmai, Pathumwan, Bangkok 10330

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่ โดยใช้การตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอย และเปรียบเทียบผลที่ได้กับการเลือกพารามิเตอร์ปรับจาก 2 วิธี ที่ใช้กันอย่างแพร่หลายสำหรับการถดถอยแบบลาสโซ่ ได้แก่ วิธีการตรวจสอบไขว้ และวิธีการใช้เกณฑ์ข้อสนเทศของเบส์ โดยจำลองข้อมูลให้ครอบคลุมกับเหตุการณ์ที่อาจก่อให้เกิดปัญหาเกี่ยวกับข้อบังคับเบื้องต้นของการถดถอยทั้งหมด 6 กรณี เน้นไปที่การเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ สำหรับเกณฑ์ที่ใช้วัดประสิทธิ์ภาพของผลที่ได้จากการวิเคราะห์การถดถอยด้วยพารามิเตอร์ปรับจากวิธีต่าง ๆ ได้แก่ อัตราความผิดพลาดในการตรวจจับเชิงบวก อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าคลาดเคลื่อนจากการพยากรณ์ และค่าคลาดเคลื่อนของสัมประสิทธิ์การถดถอย ผลการศึกษาการจำลองข้อมูลพบว่าวิธีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยให้อัตราความผิดพลาดในการตรวจจับเชิงบวกต่ำที่สุด วิธีการตรวจสอบไขว้ให้อัตราความผิดพลาดในการตรวจจับเชิงลบต่ำกว่าอีก 2 วิธี นอกจากนี้วิธีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยและวิธีการตรวจสอบไขว้ ไม่มีวิธีใดวิธีหนึ่งที่เหมาะสมกว่าอย่างเด่นชัดกว่ากัน เมื่อพิจารณาจากค่าคลาดเคลื่อนของการพยากรณ์และสัมประสิทธิ์การถดถอยจากการวิเคราะห์กับข้อมูลจริง โดยใช้ข้อมูลไมโครอะเรย์ในโรคอัลไซเมอร์ ยังพบอีกว่าวิธีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยมีความเหมาะสมมากกว่าอีก 2 วิธี

คำสำคัญ : ข้อมูลที่มีมิติสูง; การถดถอยแบบลาสโซ่; พารามิเตอร์การปรับ; การตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอย; การตรวจสอบไขว้; เกณฑ์ข้อสนเทศของเบส์

Abstract

This research is aimed to propose a method to select a tuning parameter for lasso regression by using regression diagnostics. Here we compare the results with the two popular approaches in

***ผู้รับผิดชอบบทความ :** j.nuntasuwan@gmail.com

lasso tuning parameter selection including cross-validation and Bayesian Information Criteria. Simulation studies in 6 cases emphasizing on violation of the linearity and homoscedasticity assumptions were carried out. The performance of three methods are compared in terms of false positive rate, false negative rate, prediction error, and estimation error. Our simulation studies show that regression diagnostics approach yields the lowest false positive rates while cross-validation method provides the lower false negative rates. In addition, regression diagnostics and cross-validation methods are comparable in terms of prediction error and estimation error. For the real data analysis, we applied all three methods with the Alzheimer's disease microarray data set. The results show that regression diagnostics is the most appropriate methods.

Keywords: high-dimensional data; lasso regression; tuning parameter; regression diagnostics; cross-validation; Bayesian information criteria

1. บทนำ

การวิเคราะห์การถดถอยเชิงเส้นเป็นกระบวนการทางสถิติที่ใช้ในการวิเคราะห์ข้อมูล เพื่อหาความสัมพันธ์ของตัวแปร 2 ประเภท คือ ตัวแปรอิสระและตัวแปรตาม นอกจากนี้ยังสามารถใช้พยากรณ์ค่าของตัวแปรตามเมื่อทราบค่าความสัมพันธ์ของข้อมูลที่หาได้จากการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัวด้วยวิธีกำลังสองน้อยสุด (ordinary least squares method) การวิเคราะห์นี้มีข้อจำกัดอยู่หลายประการ และข้อจำกัดที่สำคัญประการหนึ่ง คือ สามารถวิเคราะห์ได้กับข้อมูลที่มีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระเท่านั้น หากข้อมูลที่นำมาวิเคราะห์มีขนาดตัวอย่างน้อยกว่าจำนวนตัวแปรอิสระจะเรียกข้อมูลประเภทนี้ว่าข้อมูลที่มีมิติสูง และจะไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีกำลังสองน้อยสุด นอกจากนี้ยังอาจเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์กันเองสูงซึ่งจะส่งผลให้การพยากรณ์เกิดความผิดพลาดสูง และปัญหาในการแปรผลลัพธ์ของตัวแบบที่ได้อีกด้วย [1]

การวิเคราะห์ข้อมูลที่มีมิติสูงจะนิยมใช้วิธี penalized regression ซึ่งเป็นวิธีที่เพิ่ม penalty term เข้าไปในสมการที่ใช้ประมาณค่าสัมประสิทธิ์การถดถอย ซึ่งจะอยู่ในรูปของการดำเนินการทางคณิตศาสตร์ต่อค่าสัมประสิทธิ์การถดถอยและถูกถ่วงน้ำหนักโดยค่าพารามิเตอร์ที่เรียกว่าพารามิเตอร์การปรับ (tuning parameter) วิธี penalized regression ที่เป็นที่รู้จักและนิยมใช้กันอย่างแพร่หลาย คือ วิธี least absolute shrinkage and selection operator หรือลาสโซ่ ที่เสนอโดย Tibshirani [2] เพื่อคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์การถดถอยในคราวเดียวกัน โดยการบีบค่าสัมประสิทธิ์บางตัวให้เป็นศูนย์ [1,2]

การวิเคราะห์การถดถอยลาสโซ่นั้น การหาค่าพารามิเตอร์การปรับ (tuning parameter) ที่เหมาะสมเป็นอีกหนึ่งประเด็นสำคัญที่ต้องคำนึงถึง เนื่องจากอาจส่งผลในเรื่องของการพยากรณ์ได้ โดยทั่วไปการหาค่าพารามิเตอร์การปรับจะนิยมใช้วิธีการตรวจสอบไขว้ (cross-validation, CV) เพื่อลดความผิดพลาดจากการทำนาย นอกจากนี้ยังมีการศึกษาที่พบว่าวิธีเกณฑ์ข้อสนเทศของเบส์ (Bayesian information criterion, BIC) เป็นอีกหนึ่งวิธีที่ใช้ในการประมาณค่าพารามิเตอร์การปรับได้อย่างมีประสิทธิ

ภาพเช่นกัน [5]

การทบทวนวรรณกรรมที่ผ่านมายังไม่พบว่ามีการตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยทั้ง 4 ข้อ ได้แก่ (1) การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น (2) การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ (3) การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ และ (4) การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน หลังจากการวิเคราะห์การถดถอยด้วยวิธีลาสโซ่ ผู้วิจัยจึงสนใจศึกษาเกี่ยวกับประเด็นนี้ โดยวิธีการที่ผู้วิจัยเสนอวิธีหาค่าพารามิเตอร์การปรับเพื่อหลีกเลี่ยงการเกิดการละเมิดข้อบังคับเบื้องต้น คือ การพิจารณาจากการตรวจสอบข้อบังคับเบื้องต้นทั้ง 4 ข้อ เป็นหลัก โดยเลือกช่วงของค่าพารามิเตอร์การปรับที่ก่อให้เกิดการละเมิดข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยน้อยที่สุด

2. วิธีการวิจัย

2.1 การตรวจสอบเงื่อนไขของการวิเคราะห์การถดถอย

การวิเคราะห์การถดถอยเป็นวิธีทางสถิติที่ใช้ในการศึกษาความสัมพันธ์ระหว่าง 2 ตัวแปร หรือมากกว่า 2 ตัวแปร โดยจะพิจารณาการพยากรณ์ตัวแปรที่สนใจ ซึ่งเรียกว่าตัวแปรตาม (dependent variable) จากตัวแปรอีกตัวหนึ่ง หรือตัวแปรอีกกลุ่มหนึ่ง หรือที่เรียกว่าตัวแปรอิสระ (independent variable) โดยมีตัวแบบการถดถอยอยู่ในรูป [4]

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i;\ i = 1,2,\ldots,n \quad (1)$$

เมื่อ Y_i และ X_i แทนค่าสังเกตที่ i ของตัวแปรตามและตัวแปรอิสระตามลำดับ เมื่อ $X_i = X_{i1}, X_{i2}, \ldots, X_{ip}$ โดยที่ X_i เป็นค่ามาตรฐาน (standardize) และ Y_i เป็นค่าศูนย์กลาง (centering); β_j แทนค่าสัมประสิทธิ์การถดถอยของตัวแบบเมื่อ $j = 1,2,\ldots,p$; ε_i แทนค่าความคลาดเคลื่อน โดย $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2 I_n)$

โดยทั่วไปการหาค่าประมาณของสัมประสิทธิ์การถดถอยจะหาได้จากวิธีกำลังสองน้อยสุด หาได้จาก

$$b = \underset{\beta}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 \quad (2)$$

ซึ่งจะมีการตรวจสอบเงื่อนไขทั้งหมด 4 ข้อ ได้แก่

(1) การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น ทำได้โดยพิจารณาจากแผนภาพการกระจายระหว่างตัวแปรอิสระและตัวแปรตามเพื่อดูแนวโน้มของข้อมูลเป็นเส้นตรงหรือไม่ หรือพิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือ (residual) และค่าพยากรณ์ (fitted value) ว่าเป็นไปอย่างสุ่มหรือไม่ หากมีการเพิ่มขึ้นหรือลดลงอย่างมีรูปแบบแสดงว่าฟังก์ชันการถดถอยไม่เป็นเชิงเส้น [4]

(2) การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ ทำได้โดยพิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือและค่าพยากรณ์ว่าเป็นไปอย่างสุ่มหรือไม่ หากมีลักษณะเป็นรูปคล้ายลำโพง แสดงว่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ หรือทดสอบ ได้แก่ Brown-Forsythe test, Breusch-Pagan test เป็นต้น [4-6]

(3) การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน ในกรณีที่เก็บรวบรวมข้อมูลตามลำดับเวลา ทำได้โดยพิจารณาจากแผนภาพการกระจายระหว่างค่าเศษเหลือและเวลาว่าเป็นไปอย่างสุ่มหรือมีการเพิ่มขึ้นหรือลดลงในลักษณะวัฏจักร หากเห็นความสัมพันธ์ระหว่างค่าส่วนเหลือกับเวลามีลักษณะเพิ่มขึ้นและลดลงเป็นวัฏจักร แสดงว่าเกิดปัญหาค่าความคลาดเคลื่อนไม่เป็นอิสระต่อกัน หรือทดสอบได้แก่ Runs test, Durbin-Watson test เป็นต้น [4-6]

(4) การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ ทำได้โดยพิจารณาจากแผนภาพ

การกระจายระหว่างค่าเศษเหลือและค่าคาดหวังของค่าส่วนเหลือโดยมีข้อสมมติว่า ถ้าการแจกแจงของค่าความคลาดเคลื่อนมีการแจกแจงปกติแล้ว กราฟจะเป็นลักษณะเส้นตรง หรือทดสอบ ได้แก่ Kolmogorov-mirnov test, Shapiro-wilk test, Lilliefors test เป็นต้น [4-6]

2.2 การถดถอยแบบลาสโซ่ (least absolute shrinkage and selection operator, Lasso)

วิธีลาสโซ่ถูกเสนอโดย Tibshirani [2] ซึ่งพัฒนาขึ้นเพื่อใช้ในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง ($p > n$) โดยการบีบให้ค่าสัมประสิทธิ์ b_j ส่วนใหญ่เป็นศูนย์ และ b_j บางส่วนไม่เท่ากับศูนย์ (sparse estimator) ดังนั้นวิธีนี้จะสามารถเลือกตัวแปรเข้าสู่ตัวแบบและประมาณค่าสัมประสิทธิ์ b_j ได้ในคราวเดียวกัน ซึ่งจะสามารถหาค่าประมาณ b_j ได้ดังนี้ [1,2]

$$b = \underset{\beta}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3)$$

อย่างไรก็ตาม การวิเคราะห์การถดถอยด้วยวิธีนี้ยังมีข้อจำกัดอีกเล็กน้อย นั่นคือ วิธีลาสโซ่สามารถเลือกตัวแปรเข้าสู่ตัวแบบได้มากที่สุดเท่ากับจำนวนของค่าสังเกต n ตัว และหากตัวแปรอิสระมีความสัมพันธ์กันสูง วิธีลาสโซ่มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวเดียวจากกลุ่มตัวแปรนั้น ๆ ดังนั้นหากข้อมูลที่นำมาวิเคราะห์มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมาก ๆ หรือตัวแประอิสระมีความสัมพันธ์กันเองสูง ตัวแบบที่ได้จากการวิเคราะห์ด้วยวิธีลาสโซ่ก็อาจไม่มีความเหมาะสม [4]

วิธีการเลือกพารามิเตอร์การปรับสำหรับการถดถอยลาสโซ่ที่ใช้กันแพร่หลายมี 2 วิธี คือ วิธีการตรวจสอบไขว้ และวิธีเกณฑ์ข้อสนเทศของเบส์ ดังนี้

(1) วิธีการตรวจสอบไขว้ (cross-validation, CV) เป็นวิธีที่นิยมใช้กันโดยทั่วไปในการหาค่าพารามิเตอร์การปรับ โดยการแบ่งกลุ่มของชุดข้อมูลออกเป็นกลุ่มย่อย k จากนั้นนำข้อมูลเพียง $k-1$ กลุ่ม มาทำนายกลุ่มที่เหลืออีก 1 กลุ่ม ทำเช่นนี้ไป k ครั้ง แล้วพิจารณาค่าความผิดพลาดในการทำนาย โดยทั่วไปแล้วจะนิยมใช้ $k = 10$ มีขั้นตอนในการคำนวณดังนี้ [7,8]

(1.1) กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1, \dots, \lambda_m\}$

(1.2) แบ่งกลุ่มของข้อมูลทั้งหมด n ชุด ออกเป็น k กลุ่ม กลุ่มละเท่า ๆ กัน

(1.3) สำหรับ $k = 1,2, \dots, K$ ทำนายดังนี้

ครั้งที่ k ใช้ข้อมูลจำนวน $K-1$ ชุด คือ ชุดที่ 1 ถึงชุดที่ K ใด ๆ ที่ไม่ใช่ชุดที่ k

(ก) ในแต่ละ $\lambda_l;\ l = 1,2, \dots, m$ ประมาณค่าพารามิเตอร์ $\{b_j\}_l$ ในข้อมูลจำนวน $k-1$ ชุดที่เลือกมา เมื่อได้ค่าประมาณ $\{b_j\}_l$ แล้ว นำไปหาค่าพยากรณ์ $\{\hat{Y}_i\}_l$ ของข้อมูลชุดที่ k

(ข) คำนวณหาค่าความผิดพลาดจาก

$$e_k(\lambda_l) = \sum_{i=1}^{n_k} (Y_i - \{\hat{Y}_i\}_l)^2 \quad (4)$$

(1.4) ในแต่ละ $\lambda_l;\ l = 1,2, \dots, m$ คำนวณค่าความผิดพลาดเฉลี่ยของ λ_l จากทั้งหมด k ครั้ง ดังนี

$$CV(\lambda_l) = \frac{1}{n} \sum_{k=1}^{K} e_k(\lambda_l) \quad (5)$$

(1.5) ตัว λ_l ที่มีค่า $CV(\lambda_l)$ น้อยที่สุดจะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

(2) วิธีเกณฑ์ข้อสนเทศของเบส์ (Bayesian information criterion, BIC)

(2.1) กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซ็ตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1, \dots, \lambda_m\}$

(2.2) ในแต่ละ λ_l; $l = 1,2,\dots,m$ ประมาณค่าพารามิเตอร์ $\{b_j\}_l$ และประมาณค่า BIC จากข้อมูลทั้งหมด โดยคำนวณจาก [2]

$$BIC_l = \log(\hat{\sigma}_l^2) + |S_l| \times \frac{\log n}{n} \times C_n \quad (6)$$

เมื่อ $\hat{\sigma}_l^2 = SSE_l/n$; $C_n = \log\log p$; S_l คือ เซต $\{b_j\}_l$ ที่ไม่เท่ากับ 0

(2.3) ตัว λ_l ที่มีค่า BIC_l น้อยที่สุด จะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

2.3 วิธีการหาพารามิเตอร์การปรับโดยพิจารณาจากการตรวจสอบข้อบังคับเบื้องต้น (regression diagnostics, RD)

งานวิจัยนี้นำเสนอการใช้การตรวจสอบข้อบังคับเบื้องต้นของการวิเคราะห์การถดถอยในการเลือกพารามิเตอร์ปรับ ซึ่งมีรายละเอียดขั้นตอน ดังนี้

2.3.1 กำหนด λ ให้เป็นตัวประมาณแบบจุดมีค่าต่างกันไป จะได้ λ ที่เป็นไปได้เป็นเซตที่ประกอบไปด้วย λ ทั้งหมด m ตัว ดังนี้ $\Lambda = \{\lambda_1,\dots,\lambda_m\}$

2.3.2 ในแต่ละ λ_l; $l = 1,2,\dots,m$ วิเคราะห์การถดถอยลาสโซ่

2.3.3 ในแต่ละ λ_l; $l = 1,2,\dots,m$ ตรวจสอบข้อบังคับเบื้องต้นของการถดถอย โดยมีเกณฑ์ในการตัดสินใจดังนี้ (1) การตรวจสอบฟังก์ชันการถดถอยเชิงเส้น พิจารณาจาก residuals plot และค่าสหสัมพันธ์ระหว่างค่าเศษเหลือและค่าพยากรณ์ (r) (2) การตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ ใช้ Breusch-Pagan test (3) การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน ใช้ Runs test และ (4) การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ ใช้ Shapiro-wilk test ที่ระดับนัยสำคัญ 0.05

2.3.4 ตัว λ_l ที่ไม่ปฏิเสธสมมติฐานหลักของการตรวจสอบความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ การตรวจสอบค่าความคลาดเคลื่อนเป็นอิสระต่อกัน การตรวจสอบค่าความคลาดเคลื่อนมีการแจกแจงปกติ และมีค่าสหสัมพันธ์ระหว่างค่าเศษเหลือและค่าพยากรณ์ (r) น้อยที่สุดจะเป็นพารามิเตอร์การปรับที่เหมาะสมที่สุด

3. ผลวิจัยจากการจำลองข้อมูล

จำลองข้อมูลแบบตัดขวาง (cross-sectional data) ให้มีสถานการณ์ที่ก่อให้เกิดปัญหาเกี่ยวกับเงื่อนไขการถดถอย โดยเน้นไปที่การเกิดปัญหาฟังก์ชันการถดถอยไม่เป็นเชิงเส้นและปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ โดยกำหนดขนาดตัวอย่าง (n) เท่ากับ 100 และจำนวนตัวแปรอิสระ (p) เท่ากับ 1,000

3.1 กรณีที่ 1 : Normal ให้ $Y_i = X_i\beta_j + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $\beta_j = \begin{cases} 1.5\,; j = 1,..,25 \\ 0\,; \text{อื่นๆ} \end{cases}$ และ $X_i \overset{iid}{\sim} N(0, I_p)$ โดยที่ $X_i = X_{i1}, X_{i2}, \dots, X_{ip}$ สำหรับ $i = 1,2,\dots,n\,; j = 1,2,\dots,p$

สำหรับกรณีที่ 2-4 เป็นการจำลองข้อมูลให้เกิดปัญหาค่าความแปรปรวนของค่าความคลาดเคลื่อนมีค่าไม่คงที่ โดยอ้างอิงการจำลองข้อมูลมาจาก Dezeure และคณะ [9] ดังนี้

3.2 กรณีที่ 2 : Non-constant variance (equal correlation) ให้ $Y_i = X_i\beta_j + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $\beta_j = \begin{cases} 1.5\,; j = 1,..,25 \\ 0\,; \text{อื่นๆ} \end{cases}$ และ $X_i \sim N(0, \Sigma_p)$; $\Sigma_{j,k} = \begin{cases} 0.8\,; j \neq k \\ 1\,; \text{อื่นๆ} \end{cases}$ โดยที่ $X_i = X_{i1}, X_{i2}, \dots, X_{ip}$ สำหรับ $i = 1,2,\dots,n\,; j = 1,2,\dots,p$

3.3 กรณีที่ 3 : Non-constant variance (toeplitz) ให้ $Y_i = X_i\beta_j + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $\beta_j = \begin{cases} 1.5\,; j = 1,..,25 \\ 0\,; \text{อื่นๆ} \end{cases}$ และ $X_i \sim N(0, \Sigma_p)$ โดยที่ $\Sigma_{j,k} = 0.9^{|j-k|}$, $X_i = X_{i1}, X_{i2}, \dots, X_{ip}$ สำหรับ $i = 1,2,\dots,n\,; j = 1,2,\dots,p$

3.4 กรณีที่ 4 : Non-constant variance (exponential decay) ให้ $Y_i = X_i\beta_j + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $\beta_j = \begin{cases} 1.5 ; j = 1,..,25 \\ 0 \ ; \text{อื่นๆ} \end{cases}$ และ $X_i \sim N(0, \Sigma_p)$ โดยที่ $(\Sigma^{-1})_{j,k} = 0.4^{|j-k|/5}$, $X_i = X_{i1}, X_{i2}, \dots, X_{ip}$ สำหรับ $i = 1,2,\dots,n ; j = 1,2,\dots,p$

3.5 กรณีที่ 5 : Errors in predictors ให้ $Y_i = Z_i\beta_j + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $\beta_j = \begin{cases} 1.5 ; j = 1,..,25 \\ 0 \ ; \text{อื่นๆ} \end{cases}$ และ $X_i = Z_i + \xi_i$ โดยที่ $Z_i \overset{iid}{\sim} N(0, I_p)$, $\xi_i \overset{iid}{\sim} N(0, I_p)$, $X_i = X_{i1}, X_{i2}, \dots, X_{ip}$; $Z_i = Z_{i1}, Z_{i2}, \dots, Z_{ip}$ และ $\xi_i = \xi_{i1}, \xi_{i2}, \dots, \xi_{ip}$ สำหรับ $i = 1,2,\dots,n ; j = 1,2,\dots,p$

3.6 กรณีที่ 6 : Latent-variable model ให้ $Y_i = 1.5Z_1 + 1.5Z_2 + 1.5Z_3 + \varepsilon_i$ เมื่อ $\varepsilon_i \overset{iid}{\sim} N(0,1)$, $X_j = sign(5.5 - j)Z_1 1_{\{j \le 10\}} + sign(15.5 - j)Z_2 1_{\{11 \le j \le 20\}} + Z_3 1_{\{21 \le j \le 25\}} + \xi_j$ โดยที่ $Z_1, Z_2, Z_3 \overset{iid}{\sim} N(0,1)$, $\xi_j \overset{iid}{\sim} N(0,1)$, $X_j = X_{1j}, X_{2j}, \dots, X_{nj}$; $Z_1 = Z_{11}, Z_{21}, \dots, Z_{n1}$; $Z_2 = Z_{12}, Z_{22}, \dots, Z_{n2}$; $Z_3 = Z_{13}, Z_{23}, \dots, Z_{n3}$ และ $\xi_j = \xi_{1j}, \xi_{2j}, \dots, \xi_{nj}$ สำหรับ $i = 1,2,\dots,n ; j = 1,2,\dots,p$

เปรียบเทียบประสิทธิภาพของค่าพารามิเตอร์การปรับที่ได้จาก 3 วิธี คือ CV, BIC และ RD โดยดูถึง 4 เงื่อนไข ดังนี้

(1) อัตราความผิดพลาดในการตรวจจับเชิงบวก (false positive rate, FPR)

$$FPR = \frac{\sum_{j=1}^{p} 1_{\{b_j \neq 0 \ and \ \beta_j = 0\}}}{\sum_{j=1}^{p} 1_{\{b_j \neq 0\}}}$$

(2) อัตราความผิดพลาดในการตรวจจับเป็นลบ (false negative rate, FNR)

$$FNR = \frac{\sum_{j=1}^{p} 1_{\{b_j = 0 \ and \ \beta_j \neq 0\}}}{\sum_{j=1}^{p} 1_{\{b_j = 0\}}}$$

(3) ค่าความผิดพลาดในการพยากรณ์ (prediction error, PE)

$$PE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

(4) ค่าความผิดพลาดของค่าสัมประสิทธิ์การถดถอย (estimation error, EE)

$$BE = \sum_{j=1}^{p} |b_j - \beta_j|$$

ตารางที่ 1 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าพารามิเตอร์การปรับสำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1	170.6 (3.507)	1.000 (0.000)	244.7 (5.606)
กรณีที่ 2	862.9 (9.004)	1.000 (0.000)	7.004 (0.657)
กรณีที่ 3	25.71 (0.419)	27.97 (0.523)	30.47 (0.639)
กรณีที่ 4	80.33 (1.109)	107.9 (2.209)	75.67 (2.201)
กรณีที่ 5	194.0 (3.492)	1.000 (0.000)	218.3 (3.846)
กรณีที่ 6	40.71 (0.515)	112.6 (3.261)	116.1 (3.410)

จากตารางที่ 1 พบว่าค่าพารามิเตอร์การปรับจากวิธี BIC ในข้อมูลจำลองกรณีที่ 1, 2 และ 5 จะได้ค่าพารามิเตอร์การปรับเท่ากับ 1 เสมอ ทำให้มีค่าสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์มีจำนวนเท่ากับจำนวนขนาดตัวอย่าง จึงจะไม่พิจารณาประสิทธิภาพของการหาค่าพารามิเตอร์การปรับด้วยวิธีนี้ ในกรณีที่ที่ 1, 2 และ 5

ตารางที่ 2 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดในการตรวจจับเชิงบวก (FPR) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	0.444 (0.020)	-	**0.333 (0.034)**
กรณีที่ 2‡	0.850 (0.008)	-	**0.780 (0.004)**
กรณีที่ 3*	0.074 (0.007)	0.038 (0.005)	**0.000 (0.018)**
กรณีที่ 4*	0.839 (0.013)	**0.500 (0.044)**	0.889 (0.017)
กรณีที่ 5‡	0.720 (0.024)	-	**0.683 (0.030)**
กรณีที่ 6*	0.250 (0.012)	**0.000 (0.013)**	**0.000 (0.019)**

ตัวหนา คือ วิธีที่เหมาะสมที่สุด; * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman rank est); ‡ มีความแตกต่างทางสถิติที่ระดับนัยสำคัญ 0.05 (Wilcoxon rank sum test)

จากตารางที่ 2 พบว่าข้อมูลจำลองทุกกรณี ยกเว้นกรณีที่ 4 การหาค่าพารามิเตอร์การปรับด้วยวิธี RD มีความเหมาะสมมากที่สุด ในขณะที่ข้อมูลจำลองกรณีที่ 4 การหาค่าพารามิเตอร์การปรับด้วยวิธี BIC จะมีความเหมาะสมมากที่สุด

ตารางที่ 3 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของอัตราความผิดพลาดในการตรวจจับเชิงลบ (FNR) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**0.020 (<0.01)**	-	0.023 (<0.01)
กรณีที่ 2‡	0.020 (<0.01)	-	**0.005 (<0.01)**
กรณีที่ 3	0	0	0
กรณีที่ 4	0.024 (<0.01)	0.024 (<0.001)	0.024 (0.001)
กรณีที่ 5‡	**0.023 (<0.01)**	-	0.024 (<0.01)
กรณีที่ 6*	**0.011 (<0.01)**	0.024 (<0.001)	0.023 (<0.01)

ตัวหนา คือ วิธีที่เหมาะสมที่สุด; * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman rank est); ‡ มีความแตกต่างทางสถิติที่ระดับนัยสำคัญ 0.05 (Wilcoxon rank sum test)

ตารางที่ 4 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าความผิดพลาดในการพยากรณ์ (PE) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**5190 (103.1)**	-	5503 (96.31)
กรณีที่ 2‡	9155 (173.8)	-	**4.661 (0.880)**
กรณีที่ 3*	**159.5 (4.347)**	165.7 (4.855)	178.5 (5.391)
กรณีที่ 4*	**998.7 (20.29)**	1031 (23.64)	1011 (21.02)
กรณีที่ 5‡	5871 (103.7)	-	**5860 (102.5)**
กรณีที่ 6*	**340.9 (9.297)**	722.6 (22.88)	762.8 (20.54)

ตัวหนา คือ วิธีที่เหมาะสมที่สุด; * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman rank est); ‡ มีความแตกต่างทางสถิติที่ระดับนัยสำคัญ 0.05 (Wilcoxon rank sum test)

จากตารางที่ 3 พบว่าในข้อมูลจำลองกรณีที่ 1, 5, และ 6 การหาค่าพารามิเตอร์การปรับด้วยวิธี CV จะมีความเหมาะสมมากที่สุด ในขณะที่ข้อมูลจำลองกรณีที่ 2 การหาค่าพารามิเตอร์การปรับด้วย RD จะมีความเหมาะสมมากที่สุด แต่ข้อมูลจำลองกรณีที่ 3 และ 4 สามารถใช้การหาค่าพารามิเตอร์การปรับได้ทั้ง 3 วิธี เนื่องจาก FNR ไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

จากตารางที่ 4 พบว่าข้อมูลจำลองกรณีที่ 1, 3, 4 และ 6 วิธีการหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือวิธี CV ในขณะที่ข้อมูลจำลองกรณีที่ 2 และ 4 ที่การหาค่าพารามิเตอร์การปรับด้วย RD จะมีความเหมาะสมมากที่สุด

ตารางที่ 5 ค่ามัธยฐานและค่าความคลาดเคลื่อนมาตรฐานของค่าความผิดพลาดค่าสัมประสิทธิ์การถดถอย (EE) สำหรับข้อมูลจำลอง 6 กรณี

	CV	BIC	RD
กรณีที่ 1‡	**36.23 (0.195)**	-	37.19 (0.076)
กรณีที่ 2‡	54.36 (0.513)	-	**44.41 (0.811)**
กรณีที่ 3*	**7.598 (0.195)**	7.659 (0.197)	7.800 (0.205)
กรณีที่ 4*	37.67 (0.036)	**37.44 (0.027)**	37.75 (0.047)
กรณีที่ 5‡	37.87 (0.109)	-	**37.60 (0.067)**
กรณีที่ 6*	**37.07 (0.065)**	37.43 (0.037)	37.46 (0.029)

ตัวหนา คือ วิธีที่เหมาะสมที่สุด; * มีความแตกต่างทางสถิติ ที่ระดับนัยสำคัญ 0.05 (Friedman rank est); ‡ มีความแตกต่างทางสถิติที่ระดับนัยสำคัญ 0.05 (Wilcoxon rank sum test)

จากตารางที่ 5 พบว่าในข้อมูลจำลองกรณีที่ 1, 3 และ 6 การหาค่าพารามิเตอร์การปรับด้วยวิธี CV จะมีความเหมาะสมมากที่สุด ในขณะที่ข้อมูลจำลองกรณีที่ 2 และ 5 การหาค่าพารามิเตอร์การปรับที่เหมาะสมที่สุดคือ RD และสำหรับข้อมูลจำลองกรณีที่ 4 การหา

ค่าพารามิเตอร์การปรับด้วยวิธี BIC จะมีความเหมาะสมมากที่สุด

4. การปรับใช้กับข้อมูลจริง : การวิเคราะห์ข้อมูลไมโครอะเรย์ในโรคอัลไซเมอร์

ข้อมูลงานวิจัยของ Blalock และคณะ [10] ซึ่งประกอบไปด้วยข้อมูลจากผู้ป่วยที่ได้รับวินิจฉัยว่าเป็นโรคอัลไซเมอร์ระยะแรก (incipient Alzheimer's disease) จำนวน 31 ราย โดยนำเนื้อสมองจากการชันสูตรศพของผู้ป่วยมาสกัดการแสดงออกของยีนในสมองส่วนฮิปโปแคมปัสซึ่งเป็นตัวแปรอิสระในข้อมูลชุดนี้ สำหรับตัวแปรตามในการศึกษานี้ ผู้วิจัยใช้คะแนนจากการทดสอบสภาพสมองอย่างย่อ (mini mental h status examination, MMSE) ซึ่งเป็นคะแนนตั้งแต่ 2 (most severe : หนักที่สุด) ถึง 30 (normal : ปกติ) [11] ในการวิเคราะห์นี้ ผู้วิจัยได้ใช้การแสดงออกของยีน จำนวน 3,413 ตำแหน่ง จากทั้งหมด 9,921 ตำแหน่ง ที่ผ่านการคัดกรองด้วยการวิเคราะห์สหสัมพันธ์ (correlation analysis) โดย Pearson's test ว่ายีนตัวดังกล่าวมีความสัมพันธ์กับคะแนน MMSE และคะแนนของโปรตีนลำเลียงสารสู่เซลล์ประสาท (neurofibrillary tangle, NFT) ซึ่งมีความเกี่ยวข้องกับคะแนน MMSE [10]

ผลจากการวิเคราะห์การถดถอยลาสโซ่โดยใช้วิธีการหาพารามิเตอร์การปรับทั้ง 3 วิธี นั่นคือ CV, BIC และ RD ได้ค่าพารามิเตอร์การปรับเท่ากับ 132.8409, 1 และ 18.4825 ตามลำดับ ในกรณีที่ใช้วิธี BIC ไม่นำมาพิจารณาเพราะค่าสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์มีจำนวนเท่ากับจำนวนขนาดตัวอย่าง สำหรับวิธี CV และ RD ได้จำนวนสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์เท่ากับ 5 และ 25 ตำแหน่ง ตามลำดับ

เมื่อวิเคราะห์การถดถอยลาสโซ่ด้วยวิธี RD พบว่าค่า r ที่เท่ากับ 0.7085 ทำให้ได้ค่าพารามิเตอร์การปรับที่เหมาะสมที่สุด ซึ่งเท่ากับ 18.4825 ส่งผลให้มียีนในตัวแบบการถดถอยทั้งหมด 25 ตำแหน่ง โดยมีรายงานว่ายีนดังกล่าวเกี่ยวข้องกับการเกิดโรคอัลไซเมอร์ในคนจำนวน 13 ตำแหน่ง ดังตารางที่ 6 นอกจากนี้ยังพบอีกว่ายีนทั้ง 5 ตำแหน่งในตัวแบบการถดถอยลาสโซ่ด้วยวิธี CV อยู่ในตัวแบบการถดถอยลาสโซ่ด้วยวิธี RD เช่นกัน (ตัวหนาในตารางที่ 6) ทั้งนี้ยีน 2 ใน 5 ตำแหน่งจากวิธี CV มีรายงานว่าเกี่ยวข้องกับการเกิดโรคอัลไซเมอร์ในคน

สำหรับการปรับใช้กับข้อมูลงานวิจัยของ Blalock และคณะ พบว่าการหาพารามิเตอร์การปรับด้วยวิธี RD ให้ผลการวิเคราะห์ที่ดีกว่า เนื่องจากวิธี RD มียีนในตัวแบบที่มีรายงานว่าเกี่ยวข้องกับโรคอัลไซเมอร์ในคนจำนวน 13 ตำแหน่ง ตามรายการอ้างอิงในตารางที่ 6 ในขณะที่วิธี CV นำยีนตัวดังกล่าวเข้ามาในตัวแบบเพียง 2 ตำแหน่ง ทั้งนี้ยีนในตัวแบบ RD อีก 12 ตำแหน่ง ที่ยังไม่มีรายงานว่ามีความเกี่ยวข้องกับการเกิดโรคอัลไซเมอร์ในคนนั้นไม่ได้หมายความว่ายีนเหล่านี้จะไม่มีความสัมพันธ์กับการเกิดโรคอัลไซเมอร์ หากแต่ต้องอาศัยผู้เชี่ยวชาญทางด้านการแพทย์ในการตรวจสอบและสรุปผลต่อไป

5. สรุปและอภิปรายผล

งานวิจัยชิ้นนี้นำเสนอวิธีการคัดเลือกพารามิเตอร์การปรับสำหรับการถดถอยแบบลาสโซ่ด้วยวิธีการตรวจสอบข้อบังคับเบื้องต้นของการถดถอย (RD) และเปรียบเทียบวิธีการหาค่าพารามิเตอร์การปรับสำหรับการถดถอยแบบลาสโซ่ด้วยวิธี CV, BIC และ RD โดยวิเคราะห์ข้อมูลจำลองทั้งหมด 6 กรณี และเปรียบเทียบประสิทธิภาพของพารามิเตอร์การปรับโดยอัตราความผิดพลาดในการตรวจจับเชิงบวก อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าความ

ผิดพลาดในการพยากรณ์ และค่าความผิดพลาดค่าสัมประสิทธิ์การถดถอย

การวิจัยพบว่าวิธี BIC ไม่เหมาะสมในการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ เนื่องจากได้ค่าพารามิเตอร์การปรับเท่ากับ 1 ในหลายกรณี ทำให้เลือกพารามิเตอร์เข้าตัวแบบเท่ากับจำนวนตัวอย่างตามข้อจำกัดของการถดถอยแบบลาสโซ่

ตารางที่ 6 รายชื่อยีนที่มีค่าสัมประสิทธิ์การถดถอยที่ไม่เท่ากับศูนย์ จากการหาค่าพารามิเตอร์การปรับด้วยวิธี RD

ลำดับ	ชื่อยีน	โครโมโซม	ตำแหน่ง	รายการอ้างอิง
1	CASP8AP2	6	6q15	
2	CCDC59	12	12q21.31	
3	CHD2	15	15q26	Shen, Ji, Yuan *et al.* [11]
4	CLDN10	13	13q31-q34	
5	CNAP1, NCAPD2	12	12p13.3	Li *et al*. [12]
6	CREB3	9	9p13.3	
7	CSNK1A1	5	5q32	
8	GNA14	9	9q21.2	Lazarczyk *et al.* [13]
9	GRM6	5	5q35	
10	HOMER-3	19	19p13.11	Kyratzi and Efthimiopoulos [14]
11	IFNGR2	21	21q22.11	Wilcock and Griffin [15]
12	KIAA0749	12	12q13.12	Schneider *et al*. [16]
13	**MAB21L2**	**4**	**4q31.3**	
14	MGC2840, ALG8	11	11q14.1	
15	**MYO1F**	**19**	**19p13.3-p13.2**	**Orre *et al.* [17]**
16	OR7E2P	11	11q14.2	
17	PER2	2	2q37.3	Ma, Jiang, and Zhang [18]
18	PRPF4B	6	6p25.2	Wong [19]
19	RI58, IFIT5	10	10q23.31	Soler-López *et al.* [20]
20	RSU1	10	10p13	
21	SSBP1	7	7q34	Wu *et al.* [21]
22	**SEPT6**	**X**	**Xq24**	**Hu *et al.* [22]**
23	**TIX1**	**20**	**20q12**	
24	TPD52	8	8q21	Yokoyama *et al*. [23]
25	**TUBG2**	**17**	**17q21**	

ตัวหนา คือ ยีนที่พบในตัวแบบการถดถอยลาสโซ่ทั้งวิธี CV และ RD

จากที่ได้จากการจำลองข้อมูล พบว่าเมื่อพิจารณาเกณฑ์อัตราความผิดพลาดในการตรวจจับเชิงบวกแล้ว วิธี RD เป็นวิธีที่เหมาะสมในการหาค่าพารามิเตอร์การปรับสำหรับการถดถอยแบบลาสโซ่ที่สุด ในขณะที่เกณฑ์อัตราความผิดพลาดในการตรวจจับเชิงลบ ค่าความผิดพลาดในการพยากรณ์ และค่าความผิดพลาดค่าสัมประสิทธิ์การถดถอยพบว่าวิธี CV และวิธี RD ไม่มีวิธีใดวิธีหนึ่งที่เหมาะสมกว่าอย่างเด่นชัด

ผลจากการปรับใช้การวิเคราะห์การถดถอยลาสโซ่โดยใช้วิธีการหาพารามิเตอร์การปรับทั้ง 3 วิธีกับข้อมูลไมโครอะเรย์ในโรคอัลไซเมอร์ของ Blalock และคณะ [10] พบว่าได้ผลลัพธ์ในทางเดียวกับการวิเคราะห์ข้อมูลจำลอง นั่นคือ วิธี BIC เป็นวิธีที่ไม่เหมาะสมในการหาค่าพารามิเตอร์การปรับสำหรับการวิเคราะห์การถดถอยลาสโซ่ เนื่องจากได้ค่าพารามิเตอร์การปรับเท่ากับ 1 รวมถึงวิธี RD เป็นวิธีที่เหมาะสมกว่าวิธี CV เนื่องจากพบยีนที่เกี่ยวข้องกับการเกิดโรคอัลไซเมอร์ในคนมากกว่า

อนึ่ง ด้วยเทคโนโลยีทางด้านวิทยาศาสตร์ชีว-การแพทย์ที่ก้าวหน้า ส่งผลให้ข้อมูลที่มีมิติสูงมีมากขึ้น เนื่องจากสามารถสกัดข้อมูลจากตำแหน่งของยีนหรือดีเอ็นเอที่ละเอียดขึ้น โดยส่วนใหญ่ผู้วิจัยทางวิทยาศาสตร์ชีวการแพทย์ได้ให้ความสำคัญกับอัตราความผิดพลาดในการตรวจจับเชิงบวกเป็นอย่างมาก ทั้งนี้เพราะหากมีความผิดพลาดในการตรวจจับเชิงบวกเกิดขึ้น นั่นหมายความว่ายีนที่ไม่เกี่ยวข้องกับโรคจริงนั้นถูกระบุว่ามีความสัมพันธ์กับโรคนั้น ๆ ซึ่งผู้ป่วยที่มียีนดังกล่าวอาจได้รับการรักษาที่ไม่เหมาะสมและส่งผลกระทบที่ร้ายแรงต่อผู้ป่วยก็เป็นได้ ดังนั้นอัตราความผิดพลาดเชิงบวกจึงส่งผลกระทบที่ร้ายแรงกว่าอัตราความผิดพลาดเชิงลบ ซึ่งหากพิจารณาถึงอัตราความผิดพลาดเชิงบวก จะเห็นได้ชัดเจนว่าวิธี RD นั้นให้ค่าอัตราความผิดพลาดเชิงบวกที่ต่ำกว่าอีกสองวิธี จึงเป็นวิธีที่น่าสนใจอีกวิธีหนึ่งในการเลือกพารามิเตอร์ปรับสำหรับการถดถอยลาสโซ่

6. รายการอ้างอิง

[1] วิฐรา พึ่งพาพงศ์, 2558, บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง, ว. วิทยาศาสตร์และเทคโนโลยี 23: 212-223.

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

[3] Chand, S., 2012, On tuning parameter selection of lasso-type methods: A Monte Carlo study, 9th International Bhurban Conference on Applied Sciences & Technology, National Centre for Physics (NCP), Islamabad.

[4] พิษณุ เจียวคุณ, 2550, การวิเคราะห์การถดถอย, สถานบริการวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเชียงใหม่, เชียงใหม่,

[5] สุพล ดุรงค์วัฒนา, 2558, Regression Models: Analytics-based Approach, บริษัท แดเน็กซ์ อินเตอร์คอร์ปอเรชั่น จำกัด, กรุงเทพฯ.

[6] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W., 2005, Applied Linear Statistical Models, 5th Ed., The McGraw-Hill Companies, Inc., Singapore.

[7] กัลยา วานิชย์บัญชา, 2552, การวิเคราะห์ข้อมูลหลายตัวแปร, สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 589 น.

[8] Syed, A.R., 2011, A Review of Cross Validation and Adaptive Model Selection, M.S. Thesis, Department of Mathematics and Statistics, College of Arts and Sciences,

Georgia State University,

[9] Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N., 2015, High-dimensional Inference: Confidence intervals, p-values and R-Software hdi, Stat. Sci. 30: 533-558.

[10] Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R. and Landfield, P.W., 2004, Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses, PNAS. 101: 2173-2178.

[11] Shen, T., Ji, F., Yuan, Z. and Jiao, J., 2015, CHD2 is Required for embryonic neurogenesis in the developing cerebral cortex, Stem Cells 33: 1794-806.

[12] Li, Y., Chu, L.W., d, Li, Z., Yik, P.Y. and Song, Y.Q., 2009, A study on the association of the chromosome 12p13 locus with sporadic late-onset Alzheimer's disease in Chinese, Dement Geriatr. Cogn. Disord. 27: 508-512.

[13] Lazarczyk, M.J., Haller, S., Savioz, A., Gimelli, S., Bena, F. and Giannakopoulos, P., 2017, Heterozygous deletion of Chorein exons 70-73 and GNA14 exons 3-7 in a Brazilian patient presenting with probable Tau-negative early-onset Alzheimer disease, Alzheimer Dis. Assoc. Disord. 31: 82-85.

[14] Kyratzi, E. and Efthimiopoulos, S., 2014, Calcium regulates the interaction of amyloid precursor protein with Homer3 protein, Neurobiol. Aging. 35: 2053-2063.

[15] Wilcock, D.M. and Griffin, W.S.T., 2013, Down's syndrome, neuroinflammation, and Alzheimer neuropathogenesis, J. Neuroinflamm. 10: 84.

[16] Schneider, A., Huentelman, M.J., Kremerskothen, J., Duning, K., Spoelgen, R. and Nikolich, K., 2010, KIBRA: A new gateway to learning and memory?, Front. Aging Neurosci. 2: 4.

[17] Orre, M., Kamphuis, W., Osborn, L.M., Jansen, A.H.P., Kooijman, L., Bossers, K. and Hol, E.M., 2014, Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction, Neurobiol. Aging. 35: 2746-2760.

[18] Ma, Z., Jiang, W. and Zhang, E.E., 2016, Orexin signaling regulates both the hippocampal clock and the circadian oscillation of Alzheimer's disease-risk genes, Sci. Rep. 6: 36035.

[19] Wong, J., 2013, Altered expression of RNA splicing proteins in Alzheimer's disease patients: Evidence from two microarray studies, Dement Geriatr. Cogn. Disord. Extra. 3: 74-85.

[20] Soler-López, M., Zanzoni, A., Lluís, R., Stelzl, U. and Aloy, P., 2011, Interactome mapping suggests new mechanistic details underlying Alzheimer's disease, Genome Res. 21: 364-376.

[21] Wu, Y., Zhang, S., Xu, Q., Zou, H., Zhou, W., Cai, F., Li, T. and Song, W., 2015,

Regulation of global gene expression and cell proliferation by APP, Sci. Rep. 6: 22460.

[22] Hu, Y.S., Xin, J., Hu, Y., Zhang, L. and Wang, J., 2017, Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach, AZRT, 9: 29.

[23] Yokoyama, J.S., Bonham†, L.W., Sears†, R.L., Klein, E., Karydas, A., Kramer, J.H., Miller, B.L. and Coppola, G., 2015, Decision tree analysis of genetic risk for clinically heterogeneous Alzheimer's disease, BMC Neurol. 15: 47.