การเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso + MLE และวิธี Lasso + Partial Ridge

EFFICIENCY COMPARISON ON METHODS TO CONSTRUCT CONFIDENCE INTERVALS FOR PARAMETERS IN HIGH-DIMENSIONAL LOGISTIC REGRESSION MODELS BETWEEN A BOOTSTRAP LASSO + MLE AND A BOOTSTRAP LASSO + PARTIAL RIDGE

ณิชากร ไทยวงษ์[1*] และวิฐรา พึ่งพาพงศ์[2]

Nichagorn Thaiwong[1*] and Vitara Pungpapong[2]

[1] นักศึกษาระดับปริญญาโท, ภาควิชาสถิติ, คณะพาณิชย์ศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย

[1] Graduate student, Department of Statistics, Chulalongkorn Business School, Chulalongkorn University.

[2] ผู้ช่วยศาสตราจารย์, ภาควิชาสถิติ, คณะพาณิชย์ศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย

[2] Assistant Professor, Department of Statistics, Chulalongkorn Business School, Chulalongkorn University.

[*]**Corresponding author, E-mail:** 6380116126@student.chula.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+Partial Ridge ซึ่งในการศึกษานี้จะจำลองข้อมูลทั้งหมด 4 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความผิดพลาดในการตรวจจับเชิงบวก และค่าความผิดพลาดในการตรวจจับเชิงลบ จากการศึกษาภายใต้ขอบเขตจากการจำลองข้อมูล ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอน ด้วยวิธี Lasso+Partial Ridge จะให้ความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Lasso + MLE และพบว่าทั้งวิธี Lasso+Partial Ridge และวิธี Lasso + MLE มีค่า

ความน่าจะเป็นครอบคลุมใกล้เคียงกัน นอกจากนี้ผลการศึกษายังพบว่า วิธี Lasso+Partial Ridge ให้ค่าความผิดพลาดในการตรวจจับเชิงบวกและค่าความผิดพลาดในการตรวจจับเชิงลบต่ำกว่าวิธี Lasso+MLE ทั้ง 4 ชุดข้อมูล

คำสำคัญ: การถดถอยลอจิสติกทวิภาค, การถดถอยแบบริดจ์, การถดถอยลาสโซ่, การสุ่มตัวอย่างบูตสแทร็ป, การประมาณช่วงความเชื่อมั่น

Abstract

This research is aimed to compare the efficiency of methods to construct confidence intervals for parameters in high-dimensional logistic regression models between a bootstrap Lasso + MLE and a bootstrap Lasso + Partial Ridge. In this study, there are 4 simulation data sets. Also, the confidence intervals are constructed by 4 methods: (i) Parametric Bootstrap Lasso+MLE (ii) Parametric Bootstrap Lasso+Partial Ridge (iii) Paired Bootstrap Lasso+MLE, and (iv) Paired Bootstrap Lasso+Partial Ridge. The performance of all 4 methods is compared in terms of average width value, coverage probability value, false positive value, and false negative value. From our simulation studies, they show that a bootstrap Lasso + Partial Ridge method gives the smallest average width and both a bootstrap Lasso + MLE method and a bootstrap Lasso + Partial Ridge method have similar coverage probability values. In addition, we found that the values of false positive and false negative are the smallest in a bootstrap Lasso + Partial Ridge method for all data sets.

Keywords: BINARY LOGISTIC REGRESSION, RIDGE REGRESSION, LASSO REGRESSION, BOOTSTRAP SAMPLING, CONFIDENCE INTERVALS ESTIMATION

บทนำ

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นการวิเคราะห์ที่ถูกนำมาใช้อย่างแพร่หลายและมีวัตถุประสงค์เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งการวิเคราะห์การถดถอยลอจิสติกที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่า การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression) สำหรับการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกจะใช้วิธีภาวะน่าจะเป็นสูงสุดหรือ Maximum Likelihood Estimator (MLE) อันเป็นการคำนวณทวนซ้ำ (Iterative Algorithm) แต่มีข้อจำกัดว่าตัวแปรอิสระต้องไม่มีความสัมพันธ์กันเองสูง หรือไม่มีปัญหาเรื่องความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) และจะคำนวณได้ในกรณีที่ข้อมูลมีขนาดตัวอย่างมากกว่าจำนวนตัวแปรอิสระ

ข้อมูลในปัจจุบันมีขนาดใหญ่และซับซ้อนมากขึ้น เนื่องจากความสามารถในการจัดเก็บข้อมูลที่มีความทันสมัย ทำให้เกิดข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง เรียกว่า ข้อมูลที่มีมิติสูง (High Dimensional Data) ซึ่งพบได้มากในข้อมูลด้านการแพทย์ วิทยาศาสตร์อวกาศและเทคโนโลยี โดยในการวิเคราะห์ข้อมูลที่มีมิติสูงจะเกิดปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ ทำให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดไม่มีประสิทธิภาพ ซึ่งมีอีกวิธีการหนึ่งที่นิยมใช้ในการวิเคราะห์ข้อมูลที่มีมิติสูง คือ วิธี penalized regression อันเป็นการปรับตัวแบบภายใต้เงื่อนไขที่เรียกว่า penalty function โดยการถดถอยที่ปรับด้วย penalty function ที่มีการใช้งานอย่างแพร่หลาย คือ การถดถอยแบบริดจ์ (Ridge Regression) และการถดถอยลาสโซ่ (Lasso regression) โดยในการคัดเลือกตัวแปรเข้าหรือออกจากตัวแบบสำหรับการถดถอยแบบริดจ์ สามารถใช้การทดสอบสมมุติฐานสำหรับสัมประสิทธิ์การถดถอยแต่ละตัวได้ ในขณะที่วิธีการถดถอยแบบลาสโซ่เป็นวิธีที่ใช้อย่างแพร่หลายมากที่สุด เพราะสามารถประมาณค่าและคัดเลือกตัวแปรเข้าสู่ตัวแบบได้ในคราวเดียวกัน แต่จะทำให้ได้ค่าประมาณสัมประสิทธิ์การถดถอยส่วนใหญ่เท่ากับศูนย์ เรียกว่า sparse estimator อย่างไรก็ตาม ตัวประมาณจากวิธีลาสโซ่ไม่สามารถหาการแจกแจงค่าตัวอย่าง (Sampling Distribution) ได้อย่างแน่ชัด ในการทดสอบสมมติฐานดังกล่าว จึงสามารถใช้การสุ่มตัวอย่างบูตสแทร็ป (Bootstrap Sampling) เพื่อสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอย ซึ่งหากช่วงความเชื่อมั่นไม่ครอบคลุมค่าศูนย์ จะสามารถแปลผลได้ว่าสัมประสิทธิ์การถดถอยนั้นมีค่าแตกต่างจากศูนย์แบบมีนัยสำคัญทางสถิติ

จากการศึกษางานวิจัย Lui and Yu (2013) ได้นำเสนอการวิเคราะห์ข้อมูลที่มีมิติสูงด้วยการสร้างความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยการสุ่มตัวอย่างบูตสแทร็ป และการประมาณค่าสัมประสิทธิ์การถดถอย 2 ขั้นตอน โดยเริ่มจากการสร้างตัวแบบการถดถอยเชิงเส้น (Linear Regression) ที่ปรับด้วย penalty function แบบลาสโซ่ และประมาณค่าสัมประสิทธิ์การถดถอยอีกครั้งด้วยวิธีกำลังสองน้อยที่สุด (Lasso+OLS) ซึ่งพบว่าสัมประสิทธิ์การถดถอยที่คำนวณได้จากวิธีนี้ประสบปัญหา Beta-min condition หรือก็คือ สัมประสิทธิ์การถดถอยมีค่าน้อยและเข้าใกล้ศูนย์จำนวนมาก ต่อมา Dezeure et al. (2014) ได้ศึกษาประสิทธิภาพของการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยวิธี Lasso+OLS กับข้อมูลจำลอง พบว่า วิธีนี้ยังไม่มีประสิทธิภาพเพียงพอในการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้น เนื่องจากสัมประสิทธิ์การถดถอยที่คำนวณได้มีค่าน้อยและเข้าใกล้ศูนย์จำนวนมาก ทำให้ความน่าจะเป็นครอบคลุม (Coverage Probability) ต่ำกว่า 50 เปอร์เซ็นต์

Liu et al. (2020) ได้นำเสนอการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นโดยใช้ตัวอย่างบูตสแทร็ป สำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่ Liu et al. (2020) นำเสนอ คือ วิธี Lasso+Partial Ridge ซึ่งเป็นการประมาณ 2 ขั้นตอน เริ่มจากการสร้างตัวแบบการถดถอยเชิงเส้นที่ปรับด้วย penalty function แบบลาสโซ่ จากนั้นจึงใช้วิธีการถดถอยแบบริดจ์ในการหาสัมประสิทธิ์การถดถอยในขั้นตอนที่ 2 ซึ่งในขั้นตอนนี้จะใช้ penalty function แบบริดจ์กับเฉพาะสัมประสิทธิ์การถดถอยที่เท่ากับศูนย์จากวิธีลาสโซ่เท่านั้น สำหรับสัมประสิทธิ์ที่ไม่เท่ากับศูนย์จากวิธีลาสโซ่ ในขั้นตอนที่ 2 นี้จะไม่

มีการปรับด้วย penalty function ใด ๆ Liu et al. (2020) พบว่าช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นที่ได้จากวิธี Lasso+Partial Ridge แคบกว่าของวิธี Lasso+OLS และยังได้ค่าความน่าจะเป็นครอบคลุมสูงกว่าวิธี Lasso+OLS อีกด้วย ดังนั้นการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยเชิงเส้นด้วยวิธี Lasso+Partial Ridge จึงเป็นเทคนิคที่น่าสนใจ เพราะสามารถใช้วิเคราะห์ข้อมูลที่มีมิติสูงและสามารถนำมาปรับใช้กับการวิเคราะห์การถดถอยเชิงเส้นได้ดี

แม้ว่าจะมีงานวิจัยได้นำเสนอการประมาณ 2 ขั้นตอนในการสร้างความเชื่อมั่นของสัมประสิทธิ์การถดถอยสำหรับข้อมูลที่มีมิติสูง อย่างไรก็ตาม งานวิจัยที่กล่าวมาทั้งหมดจะดำเนินการศึกษาเฉพาะกรณีของตัวแบบการถดถอยเชิงเส้นกรณีที่ตัวแปรตามมีการแจกแจงแบบปรกติเท่านั้น ในงานวิจัยนี้จึงสนใจที่จะศึกษาการประมาณ 2 ขั้นตอนสำหรับตัวแบบการถดถอยลอจิสติกทวิภาค โดยเปรียบเทียบประสิทธิภาพการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกทวิภาคสำหรับข้อมูลที่มีมิติสูง โดยใช้การประมาณ 2 ขั้นตอนได้แก่ วิธี Lasso+MLE และวิธี Lasso+Partial Ridge

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพของวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอน ด้วยวิธี Lasso+MLE และวิธี Lasso+Partial Ridge

แนวคิด ทฤษฎี กรอบแนวคิด

1. การวิเคราะห์การถดถอยลอจิสติก

การวิเคราะห์การถดถอยลอจิสติก (Logistic Regression Analysis) เป็นการวิเคราะห์ที่มีวัตถุประสงค์เพื่อประมาณหรือทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยที่ตัวแปรตามจะเป็นตัวแปรจำแนกประเภท (Categorical Variable) และอาจแบ่งออกได้เป็นข้อมูลทวิภาค (Dichotomous Data) ซึ่งก็คือข้อมูล 2 กลุ่ม หรือมากกว่า 2 กลุ่ม สำหรับการวิเคราะห์การถดถอยลอจิสติกกรณีที่ตัวแปรตามแบ่งออกเป็น 2 กลุ่ม จะเรียกว่าการวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression)

ในงานวิจัยนี้เราสนใจการวิเคราะห์การถดถอยลอจิสติกทวิภาค ซึ่งจะสร้างตัวแบบตามสมการ

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \qquad (1)$$

โดยในการประมาณค่าสัมประสิทธิ์การถดถอยจะใช้วิธีภาวะน่าจะเป็นสูงสุด หรือ Maximum Likelihood Estimator (MLE) ซึ่งหาได้จากการทำให้ Log Likelihood function มีค่ามากที่สุด ดังสมการ

$$argmax(l(\beta)) = argmax(\sum_{i=1}^{n} y_i \log\frac{\pi_i}{1-\pi_i} + (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)) \qquad (2)$$

เมื่อกำหนดให้ y_i คือ จำนวนผลสำเร็จของข้อมูลสังเกต i และ $i = 1,2,\ldots,n$

นอกจากการประมาณค่าสัมประสิทธิ์การถดถอยแล้ว เรามักสนใจที่จะตรวจสอบว่าตัวแปรอิสระใดบ้างที่มีความสัมพันธ์กับโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งสามารถเขียนได้อยู่ในรูปสมมติฐาน ดังนี้

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0 \qquad (3)$$

เมื่อ $j = 1,2,\ldots,p$ ในการใช้ตัวประมาณ MLE ในการประมาณค่าสัมประสิทธิ์การถดถอย ทำให้ได้ว่าค่าประมาณสัมประสิทธิ์ของการถดถอยลอจิสติกแต่ละตัว มีการกระจายตัวของค่าตัวอย่างโดยประมาณแบบปรกติ โดยที่ $\hat{\beta}_j \sim N(\beta_j, \hat{\sigma}_j^2)$ ดังนั้นจะได้ว่าตัวสถิติทดสอบสำหรับทดสอบสมมติฐาน คือ

$$z = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim N(0,1) \qquad (4)$$

และสามารถหาช่วงความเชื่อมั่น $(1-\alpha)100\%$ ของ $\hat{\beta}_j$ ได้ โดยมีขีดจำกัดล่างของช่วงความเชื่อมั่น คือ $\left(\hat{\beta}_j - Z_{1-\frac{\alpha}{2}}\hat{\sigma}_j\right)$ และขีดจำกัดบนของช่วงความเชื่อมั่น คือ $\left(\hat{\beta}_j + Z_{1-\frac{\alpha}{2}}\hat{\sigma}_j\right)$

2. การประมาณค่าสัมประสิทธิ์ด้วยการวิเคราะห์การถดถอยลอจิสติกส์ที่ปรับด้วย penalty function

วิธีการหนึ่งที่ใช้กันอย่างแพร่หลายในการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง คือ penalized regression ซึ่งจะช่วยลดค่าสัมประสิทธิ์ของตัวแปรที่ส่งผลต่อตัวแบบน้อยให้เป็นศูนย์ และคัดออกมา โดยปรับให้สมการ (5) มีค่าน้อยที่สุด

$$l_\lambda(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda P_\lambda(\beta) \qquad (5)$$

เมื่อ $P_\lambda(\beta)$ คือ penalty function และ λ คือ พารามิเตอร์ที่มีการปรับค่าแล้ว (Tuning Parameter) ซึ่ง $\lambda \geq 0$

ต่อไปจะเสนอ penalty function 2 วิธีที่มีการใช้งานอย่างแพร่หลายได้แก่ การถดถอยแบบริดจ์ (Ridge Regression) และการถดถอยลาสโซ่ (Lasso Regression)

2.1 การถดถอยแบบริดจ์ (Ridge Regression)

Hoerl และ Kennard (1962) ได้พัฒนาและนำเสนอคุณสมบัติทางสถิติของการวิเคราะห์การถดถอยแบบริดจ์ซึ่งเป็นวิธีที่นิยมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความสัมพันธ์กันสูง (Multicollinearity) หรือเกิดภาวะร่วมเชิงเส้น โดยวิธีการถดถอยแบบริดจ์จะประมาณค่าสัมประสิทธิ์การถดถอยจากการทำให้สมการที่ (6) มีค่าน้อยที่สุด

$$l_\lambda^R(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda_2 \sum_{j=1}^{p}\beta_j^2 \qquad (6)$$

2.2 การถดถอยลาสโซ่ (Lasso Regression)

Tibshirani (1996) ได้เสนอวิธีการวิเคราะห์การถดถอยแบบลาสโซ่ (Least Absolute Shrinkage and Selection Operator Regression: Lasso Regression) โดยจะประมาณค่าสัมประสิทธิ์การถดถอยจากการทำให้สมการที่ (7) มีค่าน้อยที่สุด

$$l_\lambda^L(\beta) = \sum_{i=1}^{n}\left(-y_i \log\frac{\pi_i}{1-\pi_i} - (1-y_i)\log\left(1-\frac{\pi_i}{1-\pi_i}\right)\right) + \lambda_1 \sum_{j=1}^{p}|\beta_j| \;\; ; \lambda > 0 \qquad (7)$$

3. วิธีบูตสแทร็ปสำหรับการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติก

3.1 วิธี Parametric Bootstrap

Efron (1979) กล่าวว่า วิธีบูตสแทร็ปเป็นวิธีการประมาณค่าโดยใช้การสุ่มตัวอย่างจากประชากรแบบใส่คืน (replacement) นั่นคือ มีโอกาสที่ตัวอย่างจะสุ่มได้ซ้ำกัน โดยที่แต่ละหน่วยตัวอย่างมีโอกาสในการถูกสุ่มเท่ากัน ทำการสุ่มตัวอย่างจำนวนด้วยจำนวนครั้งที่มากพอ เพื่อสร้างการแจกแจงของตัวสถิติตัวอย่างแล้วนำมาใช้ในการประมาณค่าพารามิเตอร์ที่สนใจ มีขั้นตอนดังต่อไปนี้

1) นำข้อมูล $(\boldsymbol{X},\boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคตามสมการที่ (1) เพื่อประมาณค่า $\hat{\pi}$

2) สุ่มตัวอย่างซ้ำแบบใส่คืน จำนวน n ตัว จาก $Y^{*}{\sim}Bin(1,\hat{\pi})$ จะได้ตัวอย่างบูตสแทร็ป คือ $\boldsymbol{Y}^{*}$ โดยที่เวกเตอร์ $\boldsymbol{Y}^{*}=(Y_1^{*},Y_2^{*},\ldots,Y_n^{*})^{T}$

3) คำนวณค่าสัมประสิทธิ์การถดถอยลอจิสติกทวิภาค ตามสมการ (2) โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_p\right)^{T}$

4) ทำตามขั้นตอน 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}=\left(\widehat{\boldsymbol{\beta}}^{(1)},\widehat{\boldsymbol{\beta}}^{(2)},\ldots,\widehat{\boldsymbol{\beta}}^{(B)}\right)^{T}$

5) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

6) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j,u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

3.2 วิธี Paired Bootstrap

เรียกอีกชื่อว่าวิธี vector bootstrap เป็นวิธีการประมาณค่าโดยใช้การสุ่มตัวอย่างจากประชากรแบบใส่คืนคล้ายกับวิธี Parametric Bootstrap ซึ่งจับสุ่มเป็นคู่ และมีขั้นตอนดังต่อไปนี้

1) กำหนดให้ข้อมูล คือ $(\boldsymbol{X},\boldsymbol{Y})$ แล้วนำมาสุ่มตัวอย่างแบบใส่คืน จำนวน n ตัว จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^{*},\boldsymbol{Y}^{*})$ โดยที่ $(\boldsymbol{X}^{*},\boldsymbol{Y}^{*})=(X_1^{*},Y_1^{*}),\ldots,(X_n^{*},Y_n^{*})$

2) นำตัวอย่างที่สุ่มมาจากข้อ 1) มาประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกทวิภาค ตามสมการ (2) โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}=\left(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_p\right)^{T}$

3) ทำตามขั้นตอน 1) และ 2) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}=\left(\widehat{\boldsymbol{\beta}}^{(1)},\widehat{\boldsymbol{\beta}}^{(2)},\ldots,\widehat{\boldsymbol{\beta}}^{(B)}\right)^{T}$

4) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

5) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j,u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

4. วิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยโลจิสติกโดยใช้การประมาณสองขั้นตอน

4.1 วิธี Parametric Bootstrap Lasso+MLE

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X},\boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำบูตสแทร็ปเท่ากับ B ครั้ง

เป้าหมาย : หาช่วงความเชื่อมั่น $[l_j,u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติก เมื่อ $j=1,2,\ldots,p$

ขั้นตอน :

1) นำข้อมูล $(\boldsymbol{X},\boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคด้วยวิธีลาสโซ่

2) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ และ $\hat{\beta}_j=\begin{cases}0 & ;\hat{\beta}_{Lasso,j}=0\\ \hat{\beta}_{MLE,j} & ;\hat{\beta}_{Lasso,j}\neq 0\end{cases}$ เมื่อ $\hat{\boldsymbol{\beta}}_{\boldsymbol{Lasso,j}}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j

3) นำค่า $\widehat{\boldsymbol{\beta}}$ ที่ได้ในข้อ 2) ประมาณค่า $\hat{\pi}$

4) สุ่มตัวอย่างซ้ำแบบใส่คืน จำนวน n ตัว จาก $Y^*\sim Bin(1,\hat{\pi})$ จะได้ตัวอย่างบูตสแทร็ป คือ $\boldsymbol{Y}^*$ โดยที่เวกเตอร์ $\boldsymbol{Y}^*=(Y_1^*,Y_2^*,\ldots,Y_n^*)^T$

5) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (7) มีค่าน้อยที่สุด

6) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ และ $\hat{\beta}_j=\begin{cases}0 & ;\hat{\beta}_{Lasso,j}=0\\ \hat{\beta}_{MLE,j} & ;\hat{\beta}_{Lasso,j}\neq 0\end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j ที่ได้ในขั้นตอนที่ 5) และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j

7) ทำตามขั้นตอนที่ 4), 5) และ 6) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}=\left(\widehat{\boldsymbol{\beta}}^{(1)},\widehat{\boldsymbol{\beta}}^{(2)},\ldots,\widehat{\boldsymbol{\beta}}^{(B)}\right)^T$

8) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

9) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j,u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

4.2 วิธี Parametric Bootstrap Lasso+Partial Ridge

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X},\boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำบูตสแทร็ปเท่ากับ B ครั้ง

เป้าหมาย : หาช่วงความเชื่อมั่น $[l_j,u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติก เมื่อ $j=1,2,\ldots,p$

ขั้นตอน :

1) นำข้อมูล $(\boldsymbol{X},\boldsymbol{Y})$ มาสร้างตัวแบบการถดถอยลอจิสติกทวิภาคด้วยวิธีลาสโซ่

2) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ;\hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ;\hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j

3) นำค่า $\widehat{\boldsymbol{\beta}}$ ที่ได้ในข้อ 2) ประมาณค่า $\hat{\pi}$

4) สุ่มตัวอย่างซ้ำแบบใส่คืน จำนวน n ตัว จาก $Y^* \sim Bin(1,\hat{\pi})$ จะได้ตัวอย่างบูตสแทร็ป คือ $\boldsymbol{Y}^*$ โดยที่เวกเตอร์ $\boldsymbol{Y}^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^T$t

5) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (7) มีค่าน้อยที่สุด

6) คำนวณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ โดยการใช้ penalty function แบบริดจ์สำหรับสัมประสิทธิ์ตัวที่เท่ากับศูนย์จากวิธีการลาสโซ่เท่านั้น กล่าวคือ เมื่อให้ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และให้ $\hat{S}_{Lasso} = \{j: \hat{\beta}_{Lasso,j} \neq 0\}$ แล้วสัมประสิทธิ์ที่ได้จากวิธีการริดจ์ในขั้นตอนที่ 5) จะเขียนแทนด้วยเวกเตอร์ $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ โดยหาได้จาก

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{argmin}\left\{\frac{1}{2n}\|Y^* - X^*\beta\|_2^2 + \frac{\lambda_2}{2}\sum_{j \notin \hat{S}_{Lasso}}\beta_j^2\right\} \qquad (8)$$

7) ทำตามขั้นตอนที่ 4), 5) และ 6) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = (\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)})^T$

8) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1-\frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

9) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j, u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

4.3 วิธี Paired Bootstrap Lasso+MLE

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X},\boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำบูตสแทร็ปเท่ากับ B ครั้ง
เป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติก เมื่อ $j = 1,2,\ldots,p$

ขั้นตอน :

1) สุ่มตัวอย่างบูตสแทร็ปจำนวน n ตัว จากข้อมูล $(\boldsymbol{X},\boldsymbol{Y})$ แบบใส่คืน จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^*,\boldsymbol{Y}^*)$ โดยที่ $(\boldsymbol{X}^*,\boldsymbol{Y}^*) = (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$

2) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (7) มีค่าน้อยที่สุด

3) เลือกเฉพาะตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยลาสโซ่ไม่เท่ากับศูนย์ จากนั้นนำไปคำนวณค่าสัมประสิทธิ์การถดถอย โดยใช้วิธี MLE โดยให้สัมประสิทธิ์ที่ได้เขียนแทนด้วย $\widehat{\boldsymbol{\beta}}$ โดยที่เวกเตอร์

$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ และ $\hat{\beta}_j = \begin{cases} 0 & ; \hat{\beta}_{Lasso,j} = 0 \\ \hat{\beta}_{MLE,j} & ; \hat{\beta}_{Lasso,j} \neq 0 \end{cases}$ เมื่อ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และ $\hat{\beta}_{MLE,j}$ คือ สัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดตัวที่ j

4) ทำตามขั้นตอนที่ 1), 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$

5) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1 - \frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

6) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j, u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

4.4 วิธี Paired Bootstrap Lasso+Partial Ridge

กำหนดให้ : ข้อมูล คือ $(\boldsymbol{X}, \boldsymbol{Y})$ ที่ระดับความเชื่อมั่น $1-\alpha$ และให้จำนวนทำซ้ำบูตสแทร็ปเท่ากับ B ครั้ง

เป้าหมาย : หาช่วงความเชื่อมั่น $[l_j, u_j]$ สำหรับสัมประสิทธิ์การถดถอยโลจิสติก เมื่อ $j = 1,2,\ldots,p$

ขั้นตอน :

1) สุ่มตัวอย่างบูตสแทร็ปจำนวน n ตัว จากข้อมูล $(\boldsymbol{X}, \boldsymbol{Y})$ แบบใส่คืน จะได้ตัวอย่างสุ่มชุดใหม่ คือ $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$ โดยที่ $(\boldsymbol{X}^*, \boldsymbol{Y}^*) = (X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$

2) คำนวณค่าสัมประสิทธิ์การถดถอยลาสโซ่ จากการทำให้สมการที่ (7) มีค่าน้อยที่สุด

3) คำนวณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ โดยการใช้ penalty function แบบริดจ์สำหรับสัมประสิทธิ์ตัวที่เท่ากับศูนย์จากวิธีการลาสโซ่เท่านั้น กล่าวคือ เมื่อให้ $\hat{\beta}_{Lasso,j}$ คือ สัมประสิทธิ์การถดถอยลาสโซ่ตัวที่ j และให้ $\hat{S}_{Lasso} = \{j: \hat{\beta}_{Lasso,j} \neq 0\}$ แล้วสัมประสิทธิ์ที่ได้จากวิธีการริดจ์ในขั้นตอนที่ 3) จะเขียนแทนด้วยเวกเตอร์ $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ โดยหาได้จาก

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{argmin} \left\{ \frac{1}{2n} \|Y^* - X^*\beta\|_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}_{Lasso}} \beta_j^2 \right\} \qquad (9)$$

4) ทำตามขั้นตอนที่ 1), 2) และ 3) ซ้ำทั้งหมด B ครั้ง จะได้ตัวประมาณ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})} = \left(\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(B)}\right)^T$

5) คำนวณขีดจำกัดล่าง (l_j) และขีดจำกัดบน (u_j) ของช่วงความเชื่อมั่น โดยใช้ควอนไทล์ที่ $\frac{\alpha}{2}$ และ $1 - \frac{\alpha}{2}$ ของค่าประมาณที่ได้จากบูตสแทร็ป $\widehat{\boldsymbol{\beta}}^{(\boldsymbol{b})}$

6) จะได้ช่วงความเชื่อมั่นที่ระดับ คือ $[l_j, u_j]$ ที่ระดับความเชื่อมั่น $(1-\alpha)100\%$

วิธีดำเนินการวิจัย

1. สร้างข้อมูลจำลองที่มีการกำหนดตัวแปร ดังต่อไปนี้

1.1 กำหนดให้ขนาดตัวอย่าง $n = 200$ จำนวนตัวแปรอิสระ $p = 500$

1.2 สร้างตัวแปรอิสระ $\boldsymbol{X}$ จาก $\boldsymbol{X} \sim N(0, \sum)$ โดยที่เวกเตอร์ $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$

กำหนด $\sum$ จาก 2 วิธี ได้แก่ Toeplitz และ Equal Correlation ดังสมการต่อไปนี้

- วิธี Toeplitz: $\sum_{ij} = \rho^{|i-j|}$ ด้วย $\rho = 0.5$
- วิธี Equal Correlation: $\sum_{ij} = \rho$ ด้วย $\rho = 0.5$

1.3 ประมาณค่า $\widehat{\boldsymbol{\beta}}^0$ โดยที่เวกเตอร์ $\widehat{\boldsymbol{\beta}}^0 = \left(\hat{\beta}_0^0 = 0, \hat{\beta}_1^0, \ldots, \hat{\beta}_p^0\right)^T$ จาก 2 วิธี ดังนี้

- วิธีที่ 1 (Hard Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim Unif\left(\frac{1}{3}, 1\right)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าเท่ากับ 0 เมื่อ $j = 1,2,\ldots,p$
- วิธีที่ 2 (Weak Sparsity) สุ่ม $\hat{\beta}_j^0$ แบบไม่ใส่คืนจำนวน 10 ตัว จากเวกเตอร์ $\widehat{\boldsymbol{\beta}}^0$ จากนั้นกำหนดค่าเป็น $\hat{\beta}_j^0 \sim N(0,0.001)$ และให้ $\hat{\beta}_j^0$ ที่เหลือมีค่าลดลงตามสมการ $\hat{\beta}_j^0 = \frac{1}{(j+3)^2}$ เมื่อ $j = 1,2,\ldots,p$

1.4 สร้าง $\boldsymbol{Y}$ จาก $\boldsymbol{Y} \sim Bin(1, \hat{\pi})$ โดยที่เวกเตอร์ $\boldsymbol{Y} = (Y_1, Y_2, .., Y_n)^T$ และ $\hat{\pi} = \frac{exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}{1+exp(\widehat{\boldsymbol{\beta}}^0 \boldsymbol{X}^T)}$

2. สร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติก โดยใช้การประมาณ 2 ขั้นตอน ทั้ง 4 วิธี ได้แก่ ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge

3. เปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากทั้ง 4 วิธี โดยใช้ความกว้างเฉลี่ยของช่วงความเชื่อมั่น (average width: AW), ค่าความน่าจะเป็นครอบคลุม (coverage probability: CP), ค่าความผิดพลาดในการตรวจจับเชิงบวก (false positive: FP) และค่าความผิดพลาดในการตรวจจับเชิงลบ (false negative: FN) ดังสมการต่อไปนี้

$$\mathbf{AW} = \frac{\sum_{j=1}^{p}(u_j - l_j)}{p} \quad (8)$$

$$\boldsymbol{CP} = \frac{\text{จำนวนช่วงความเชื่อมั่นที่ครอบคลุมค่าพารามิเตอร์จริง}}{p} \quad (9)$$

$$FP = \left|\hat{S} \cap S^c\right| \quad (10)$$

$$FN = \left|\hat{S}^c \cap S\right| \quad (11)$$

<u>กำหนดให้</u>

S คือ เซตของ j ที่ค่าสัมประสิทธิ์การถดถอยที่แท้จริงที่มีค่าไม่เท่ากับ 0 หรือ $S = \{j : \beta_j \neq 0\}$

$\hat{S}$ คือ เซตของ j ที่ค่าประมาณของสัมประสิทธิ์การถดถอยปฏิเสธสมมติฐานว่างว่าค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับ 0 หรือ $\hat{S} = \{j : \text{ปฏิเสธ } H_0\}$ เมื่อ $j = 1,2,\ldots,p$

4. สรุปผล โดยนำเสนอข้อมูลในรูปกราฟและตาราง เพื่อตรวจสอบดูว่าวิธีการแบบใดให้ผลดีกว่ากัน

ผลการวิจัย

สำหรับงานวิจัยนี้ จะนำเสนอผลการเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกเป็น 4 ส่วน คือ ส่วนที่ 1 เปรียบเทียบค่าความกว้างเฉลี่ยของช่วงความเชื่อมั่น ส่วนที่ 2 เปรียบเทียบค่าความน่าจะเป็นครอบคลุม ส่วนที่ 3 เปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก และส่วนที่ 4 เปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงลบ

จากการศึกษาในตารางที่ 1 ผลปรากฏว่า การสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอน ด้วยวิธี Lasso+Partial Ridge จะให้ความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Lasso + MLE ทุกชุดข้อมูล

จากตารางที่ 2 พบว่าค่าความน่าจะเป็นครอบคลุมของทั้งวิธี Lasso+Partial Ridge และวิธี Lasso + MLE มีค่าใกล้เคียงกัน

จากตารางที่ 3 และ 4 พบว่าค่าความผิดพลาดในการตรวจจับเชิงบวกและค่าความผิดพลาดในการตรวจจับเชิงลบของวิธี Lasso+Partial Ridge มีค่าต่ำกว่าของวิธี Lasso+MLE ทุกชุดข้อมูล

ตารางที่ 1 ความกว้างเฉลี่ยของช่วงความเชื่อมั่นของแต่ละวิธีจากข้อมูลจำลอง

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ความกว้างเฉลี่ยของช่วงความเชื่อมั่น	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 1	วิธี Parametric Bootstrap Lasso+MLE	7.084e+12	1.168e+13
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.003**	**0.001**
	วิธี Paired Bootstrap Lasso+MLE	10.630	0.803
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.094	0.002
ชุดที่ 2	วิธี Parametric Bootstrap Lasso+MLE	3.441	4.602
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.003**	**0.001**
	วิธี Paired Bootstrap Lasso+MLE	3.119e+12	7.173e+12
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.105	0.002
ชุดที่ 3	วิธี Parametric Bootstrap Lasso+MLE	1.215e+13	1.838e+13
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.006**	**0.001**
	วิธี Paired Bootstrap Lasso+MLE	3.321e+11	1.669e+12
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.166	0.020

ตารางที่ 1 (ต่อ)

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ความกว้างเฉลี่ยของช่วงความเชื่อมั่น	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 4	วิธี Parametric Bootstrap Lasso+MLE	1.103e+13	1.464e+13
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.006**	**0.001**
	วิธี Paired Bootstrap Lasso+MLE	2.069e+12	6.941e+12
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.157	0.020

หมายเหตุ: ตัวพิมพ์หนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละชุดข้อมูล

ตารางที่ 2 ค่าความน่าจะเป็นครอบคลุมของแต่ละวิธีจากข้อมูลจำลอง

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ค่าความน่าจะเป็นครอบคลุม	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 1	วิธี Parametric Bootstrap Lasso+MLE	0.954	0.012
	วิธี Parametric Bootstrap Lasso+Partial Ridge	0.980	0.000
	วิธี Paired Bootstrap Lasso+MLE	**0.994**	**0.004**
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.980	0.000
ชุดที่ 2	วิธี Parametric Bootstrap Lasso+MLE	0.904	0.019
	วิธี Parametric Bootstrap Lasso+Partial Ridge	0.980	0.000
	วิธี Paired Bootstrap Lasso+MLE	**0.999**	**0.001**
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.980	0.000
ชุดที่ 3	วิธี Parametric Bootstrap Lasso+MLE	0.956	0.014
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.980**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	0.838	0.039
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.979	0.001
ชุดที่ 4	วิธี Parametric Bootstrap Lasso+MLE	0.957	0.014
	วิธี Parametric Bootstrap Lasso+Partial Ridge	0.980	0.000
	วิธี Paired Bootstrap Lasso+MLE	**0.999**	**0.001**
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.979	0.001

หมายเหตุ: ตัวพิมพ์หนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลลัพธ์ดีที่สุดในแต่ละชุดข้อมูล

ตารางที่ 3 ค่าความผิดพลาดในการตรวจจับเชิงบวกของแต่ละวิธีจากข้อมูลจำลอง

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ค่าความผิดพลาดในการตรวจจับเชิงบวก	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 1	วิธี Parametric Bootstrap Lasso+MLE	18.320	6.001
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	2.860	1.750
	วิธี Paired Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
ชุดที่ 2	วิธี Parametric Bootstrap Lasso+MLE	40.280	9.192
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	0.160	0.468
	วิธี Paired Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
ชุดที่ 3	วิธี Parametric Bootstrap Lasso+MLE	19.340	6.847
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	80.620	19.513
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.040	0.198
ชุดที่ 4	วิธี Parametric Bootstrap Lasso+MLE	19.640	6.724
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	0.440	0.611
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.020	0.141

หมายเหตุ: ตัวพิมพ์หนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลดีที่สุดในแต่ละชุดข้อมูล

ตารางที่ 4 ค่าความผิดพลาดในการตรวจจับเชิงลบของแต่ละวิธีจากข้อมูลจำลอง

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ค่าความผิดพลาดในการตรวจจับเชิงลบ	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 1	วิธี Parametric Bootstrap Lasso+MLE	9.280	1.485
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	8.960	3.023
	วิธี Paired Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**

ตารางที่ 4 (ต่อ)

ชุดข้อมูล	วิธีการสร้างช่วงความเชื่อมั่น	ค่าความผิดพลาดในการตรวจจับเชิงลบ	
		ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
ชุดที่ 2	วิธี Parametric Bootstrap Lasso+MLE	9.340	0.688
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	1.380	3.457
	วิธี Paired Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
ชุดที่ 3	วิธี Parametric Bootstrap Lasso+MLE	9.180	1.480
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	9.540	0.646
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.580	2.322
ชุดที่ 4	วิธี Parametric Bootstrap Lasso+MLE	9.460	1.487
	วิธี Parametric Bootstrap Lasso+Partial Ridge	**0.000**	**0.000**
	วิธี Paired Bootstrap Lasso+MLE	3.980	4.926
	วิธี Paired Bootstrap Lasso+Partial Ridge	0.560	2.242

หมายเหตุ: ตัวพิมพ์หนา หมายถึง วิธีการสร้างช่วงความเชื่อมั่นที่ได้ผลดีที่สุดในแต่ละชุดข้อมูล

สรุปและอภิปรายผล

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกในข้อมูลที่มีมิติสูง โดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+MLE และวิธี Lasso+Partial Ridge ซึ่งในการศึกษานี้จะจำลองข้อมูลทั้งหมด 4 ชุด และเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่นที่ได้จากการสร้างช่วงความเชื่อมั่นทั้งหมด 4 วิธี ได้แก่ วิธี Parametric Bootstrap Lasso+MLE, วิธี Parametric Bootstrap Lasso+Partial Ridge, วิธี Paired Bootstrap Lasso+MLE และวิธี Paired Bootstrap Lasso+Partial Ridge โดยใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของช่วงความเชื่อมั่น คือ ความกว้างเฉลี่ยของช่วงความเชื่อมั่น ค่าความน่าจะเป็นครอบคลุม ค่าความผิดพลาดในการตรวจจับเชิงบวก และค่าความผิดพลาดในการตรวจจับเชิงลบ ซึ่งผลที่ได้พบว่าการสร้างช่วงความเชื่อมั่นด้วยวิธี Lasso+Partial Ridge จะให้ความกว้างเฉลี่ยของช่วงความเชื่อมั่นแคบกว่าวิธี Lasso + MLE และพบว่าทั้งวิธี Lasso+Partial Ridge และวิธี Lasso+MLE มีค่าความน่าจะเป็นครอบคลุมใกล้เคียงกัน นอกจากนี้ผลการศึกษายังพบว่า วิธี Lasso+Partial Ridge ให้ค่าความผิดพลาดในการ

ตรวจจับเชิงบวกและค่าความผิดพลาดในการตรวจจับเชิงลบต่ำกว่าวิธี Lasso+MLE จึงสรุปได้ว่าการสร้างช่วงความเชื่อมั่นสำหรับสัมประสิทธิ์การถดถอยลอจิสติกโดยใช้การประมาณสองขั้นตอนด้วยวิธี Lasso+Partial Ridge มีประสิทธิภาพมากกว่าวิธี Lasso+MLE

เอกสารอ้างอิง

Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1), 226–248.

Algamal, Z. Y. (2015). High Dimensional Logistic Regression Model using Adjusted Elastic Net Penalty. *Peer Reviewed Articles*, 11(4), 667-676.

Algamal, Z. Y. and Lee, M. H. (2015). Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. *Computers in Biology and Medicine*, 67, 136–145.

Buhlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212–1242.

Dezeure, R., Buhlmann, P., Meier, L. and Meinshausen, N. (2014). High-dimensional Inference: Confidence intervals, p-values and R-Software hdi. *Statistical Science*, 30(4), 533–558.

Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2017). Robust and sparse estimation methods for high dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 12, 211-222.

Liu, H. and Yu, B. (2013). Asymptotic properties of lasso+mLS and lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7, 3124–3169.

Liu, H., Xu, X., and Li, J. J. (2020). A bootstrap lasso+ partial ridge method to construct confidence intervals for parameters in high-dimensional sparse linear models. *Statistica Sinica,* 30, 1333-1355.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4), 1567-1594.

จุฑาทิพย์ นันทสุวรรณ. (2561). การเลือกพารามิเตอร์การปรับสำหรับวิธีการถดถอยแบบลาสโซ่. *ว. วิทยาศาสตร์และเทคโนโลยี*, 26(3), 393-404.

วิฐรา พึ่งพาพงศ์. (2558). บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. *ว. วิทยาศาสตร์และเทคโนโลยี*, 23(2), 212-223.