

การประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวเมื่อเกิดปัญหาพหุสัมพันธ์ โดยใช้วิธีบูตสเตรป*

Liu Logistic Regression Coefficient Estimation with Multicollinearity Problem by Using the Bootstrapping Method

Received: October 25, 2018

Revised: March 18, 2019

Accepted: March 19, 2019

นฤมล สูดใจ (Narumol Sudjai)**

มณฑิรา ดวงสาพล (Monthira Duangsaphon)***

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอตัวประมาณค่าพารามิเตอร์ซริงค์เกจและประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวเมื่อเกิดปัญหาพหุสัมพันธ์โดยใช้วิธีบูตสเตรป นอกจากนั้นได้เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกเมื่อเกิดพหุสัมพันธ์ทั้งหมด 4 วิธี ได้แก่ วิธีการถดถอยลอจิสติกแบบหลิว วิธีการถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสเตรป วิธีการถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียงและวิธีการถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป โดยเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าคือค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจากการจำลองข้อมูล รวมถึงได้แสดงประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่ได้นำเสนอ

คำสำคัญ : วิธีบูตสเตรป ตัวประมาณแบบหลิว ตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียง พหุสัมพันธ์

* เพื่อเผยแพร่ผลงานวิจัย

** ภาควิชาศัลยศาสตร์ออร์โธปิดิกส์และกายภาพบำบัด คณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล

Department of Orthopaedic Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University,

narumol.sud@mahidol.ac.th, +66 (02) 4197529

*** สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University,

monthira.stat@gmail.com, +66 (02) 564-4440-59 ext. 2101-3 Press 315

Abstract

The objectives of this research are to propose the estimators of shrinkage parameter and estimate the Liu logistic regression coefficient with multicollinearity problem using the bootstrapping method. Moreover, we compare the performance of four methods for the logistic regression coefficient estimation with multicollinearity; Liu logistic regression, Liu logistic regression with bootstrapping method, almost unbiased Liu regression, and almost unbiased Liu regression with bootstrapping method. A simulation study was conducted to compare the appraise performance of four different estimation methods by mean square error (MSE). Also, we show the performance of proposed the estimators of shrinkage parameter.

Keywords: Bootstrapping method, Liu estimator, The almost unbiased Liu estimator, Multicollinearity

บทนำ

การนำวิธีการวิเคราะห์การถดถอยลอจิสติกมาใช้ในการวิจัยทางการแพทย์และสาธารณสุขมีเพิ่มมากขึ้นเพื่อให้งานวิจัยออกมาอย่างมีคุณภาพและมีความน่าเชื่อถือ โดยวัตถุประสงค์หลักที่สำคัญของการศึกษาคือเพื่อศึกษาอิทธิพลของตัวแปรอิสระที่มีผลต่อตัวแปรตาม ซึ่งตัวแปรตามเป็นตัวแปรเชิงกลุ่ม ยกตัวอย่างเช่น ผู้วิจัยต้องการศึกษาปัจจัยเสี่ยง ได้แก่ อายุ ขนาดของก้อนเนื้ออก ความรุนแรงของโรคและความเคยชินในการบริโภคอาหารที่ไม่ดี ที่มีผลต่อการกลับมาเป็นโรครั่ว (Recurrence) หลังการทำผ่าตัดรักษาในผู้ป่วยมะเร็งกระดุกและเนื้อเยื่ออ่อน สำหรับตัวแปรตามของการศึกษาคือการกลับมาเป็นโรครั่วอีก โดยกำหนดให้ 0 แทนผู้ป่วยที่ไม่กลับมาเป็นโรครั่วอีก และ 1 แทนผู้ป่วยที่กลับมาเป็นโรครั่วอีก วิธีการทางสถิติที่นิยมนำมาวิเคราะห์คือการวิเคราะห์การถดถอยลอจิสติกแบบสองกลุ่ม (Binary logistic regression analysis) ปัญหาที่พบบ่อยสำหรับงานวิจัยทางการแพทย์และสาธารณสุขคือตัวแปรอิสระมักจะมีความสัมพันธ์กันสูงหรือเรียกว่าเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) ส่งผลให้ความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกมีค่าสูงขึ้น เพื่อที่จะแก้ปัญหานี้อาจจะพิจารณาตัดตัวแปรอิสระบางตัวออกจากตัวแบบได้ แต่บางการศึกษาอาจเป็นเรื่องยากสำหรับนักวิจัยที่จะเลือกตัดตัวแปรใดตัวแปรหนึ่งออกจากตัวแบบ เนื่องจากตัวแปรอิสระทุกตัวมีความสำคัญและอาจจะมีอิทธิพลต่อตัวแปรตามของการศึกษานั้น การแก้ปัญหาเพื่อที่จะสามารถคงตัวแปรอิสระไว้ในตัวแบบได้ทุกตัวนั้นได้มีนักวิจัยเสนอตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก เช่น ตัวประมาณแบบสไตน์ เสนอโดย Stein ในปี ค.ศ. 1960 (Stein, 1960) ตัวประมาณแบบบริดจ์ เสนอโดย Schaefer และคณะ ในปี ค.ศ. 1984 (Schaefer, Roi, & Wolfe, 1984) และตัวประมาณแบบหลิว เสนอโดย Urgan และ Tez ในปี ค.ศ. 2008 (Urgan, & Tez, 2008) ซึ่งได้นำแนวคิดของหลิว (Liu, 1993) มาปรับใช้เพื่อสร้างตัวประมาณแบบหลิวสำหรับตัวแบบการถดถอยลอจิสติก โดยการนำเอาตัวประมาณแบบบริดจ์มารวมกับตัวประมาณแบบสไตน์เกิดเป็นตัวประมาณใหม่ที่เรียกว่าตัวประมาณแบบหลิวและมีค่าพารามิเตอร์ชริงค์เกจ (d) เป็นพารามิเตอร์ที่มีความสำคัญในการสร้างตัวประมาณแบบหลิว ด้วยเหตุนี้ในปี ค.ศ. 2012 Månsson และคณะ

(Månsson, Kibria, & Shukur, 2012) จึงได้เสนอตัวประมาณค่าพารามิเตอร์ซริงค์เกจ โดยพิจารณาค่าที่เหมาะสมของ d_j แต่ละตัว (The individual parameter: d_j) ขึ้นทั้งหมด 5 วิธี ถึงแม้ว่ารูปแบบของตัวประมาณค่าพารามิเตอร์ซริงค์เกจนี้ จะส่งผลให้ได้ตัวประมาณแบบหลิวที่มีประสิทธิภาพดีกว่าตัวประมาณภาวน่าจะเป็นสูงสุดก็ตาม แต่จะสังเกตได้ว่าค่าประมาณของพารามิเตอร์ซริงค์เกจอาจจะมีค่าเท่ากับศูนย์ ทำให้รูปแบบของตัวประมาณแบบหลิวเปลี่ยนไปเป็นรูปแบบของตัวประมาณแบบริดจ์ที่มีค่าประมาณของพารามิเตอร์ริดจ์เท่ากับ 1

ตัวประมาณที่กล่าวมาแล้วข้างต้นเป็นตัวประมาณเอนเอียง (Biased estimators) แต่ตัวประมาณเหล่านี้จะเป็นตัวประมาณที่เกือบจะไม่เอนเอียง (The almost unbiased estimators) เมื่อขนาดตัวอย่างมีมากเพียงพอ ด้วยเหตุนี้ในปี ค.ศ. 2015 Chang (Chang, 2015) จึงได้เสนอตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียง (The almost unbiased Liu estimator) ซึ่งตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงจะมีประสิทธิภาพดีกว่าตัวประมาณแบบหลิวในบางกรณีเมื่อพิจารณาจากทฤษฎีค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของตัวประมาณ แต่ Chang (Chang, 2015) ไม่ได้เสนอการเปรียบเทียบประสิทธิภาพของตัวประมาณแบบหลิวกับตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงจากการจำลองข้อมูลหรือประยุกต์ใช้กับข้อมูลจริงภายใต้สถานการณ์ต่าง ๆ นอกจากนั้น Farghali และ Abo-El-Hadid (Farghali, & Abo-El-Hadid, 2017) ได้ทำการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก 4 ตัว คือตัวประมาณภาวน่าจะเป็นสูงสุด ตัวประมาณแบบสไตน์ ตัวประมาณแบบริดจ์และตัวประมาณแบบหลิว โดยผลการวิจัยพบว่า ตัวประมาณแบบหลิวเป็นตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่มีประสิทธิภาพดีที่สุดเมื่อเปรียบเทียบกับตัวประมาณอื่น ๆ ที่ทำการศึกษาในกรณีที่เกิดปัญหาพหุสัมพันธ์สูงระหว่างตัวแปรอิสระ

จากการทบทวนวรรณกรรมข้างต้นที่เกี่ยวข้องกับตัวประมาณแบบหลิว ยังไม่มีการนำเสนอวิธีการประมาณค่าพารามิเตอร์ซริงค์เกจ (d) ที่หลากหลาย ผู้วิจัยจึงเห็นว่า สิ่งสำคัญที่สุดคือการประมาณค่าพารามิเตอร์ซริงค์เกจ (d) ที่เหมาะสมสำหรับแต่ละข้อมูล ซึ่งตัวประมาณค่าพารามิเตอร์ซริงค์เกจ (d) อาจจะขึ้นอยู่กับขนาดตัวอย่าง จำนวนตัวแปรอิสระและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ นอกจากปัญหาเรื่องพหุสัมพันธ์ระหว่างตัวแปรอิสระแล้ว อีกปัญหาหนึ่งที่พบมากสำหรับงานวิจัยทางด้านการแพทย์และสาธารณสุข คือ ขนาดตัวอย่างที่ทำการศึกษามีน้อย (Small sample size) ส่งผลให้การประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกได้ตัวประมาณค่าที่มีประสิทธิภาพลดลงและอาจจะส่งผลต่อความเที่ยง (Precision) และความแม่นยำ (Accuracy) ในการประมาณค่าดังกล่าวได้ วิธีการทางสถิติที่ไม่อิงพารามิเตอร์หนึ่งที่สามารถแก้ปัญหานี้คือ ใช้เทคนิคบูตสแตรปที่ขึ้นอยู่กับกลุ่มตัวอย่างซ้ำค่าสังเกต (The resampling observations)

ดังนั้นการวิจัยครั้งนี้ผู้วิจัยจึงเสนอตัวประมาณค่าพารามิเตอร์ซริงค์เกจ (d) ที่แตกต่างไปจากงานวิจัยของ Månsson และคณะ (Månsson, Kibria, & Shukur, 2012) และเนื่องจากตัวประมาณแบบหลิวเป็นฟังก์ชันของตัวประมาณภาวน่าจะเป็นสูงสุดซึ่งตัวประมาณภาวน่าจะเป็นสูงสุดจะมีประสิทธิภาพดีเมื่อขนาดตัวอย่างมีมากเพียงพอ เพื่อที่จะนำตัวประมาณแบบหลิวมาประยุกต์ใช้ในกรณีที่มีขนาดตัวอย่างน้อย ผู้วิจัยจึงนำเทคนิคบูตสแตรปที่ขึ้นอยู่กับกลุ่มตัวอย่างซ้ำค่าสังเกต มาใช้ร่วมกับการประมาณค่าด้วยวิธีการถดถอยลอจิสติกแบบหลิว พร้อมทั้งเปรียบเทียบประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่ได้จากวิธีการถดถอยลอจิสติกแบบหลิว วิธีการถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสแตรป วิธีการถดถอยลอจิสติกแบบหลิวที่

เกือบจะไม่เอนเอียงและวิธีการถดถอยลอจิสติกแบบหลิที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป ด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) นอกจากนั้นยังได้แสดงประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่ได้นำเสนอด้วยค่าประมาณพารามิเตอร์ซริงค์เกจที่เท่ากับศูนย์

วิธีวิจัย

การวิเคราะห์การถดถอยลอจิสติกแบบสองกลุ่ม (Binary logistic regression analysis)

ตัวแปรตอบสนองในตัวแบบการถดถอยลอจิสติกแบบสองกลุ่มมีการแจกแจงแบร์นูลลีที่มีพารามิเตอร์ π_i และความคลาดเคลื่อนสุ่ม (ε_i) มีค่าเฉลี่ยเท่ากับศูนย์และความแปรปรวนเท่ากับ $\pi_i(1-\pi_i)$ ตัวแบบการถดถอยลอจิสติกมีรูปแบบดังนี้

$$Y_i = \pi_i + \varepsilon_i \quad (1)$$

เมื่อ $\pi_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$; $i=1,2,3,...,n$ โดยที่ Y_i คือตัวแปรตอบสนอง และ x_i คือเวกเตอร์แถวขนาด $1 \times (p+1)$ ที่มีสมาชิกในแถวที่ i ของเมทริกซ์ X ส่วน X คือเมทริกซ์ของตัวแปรอิสระที่มีขนาด $n \times (p+1)$ และ β คือเวกเตอร์ของสัมประสิทธิ์การถดถอยลอจิสติกที่มีขนาด $(p+1) \times 1$ เมื่อ n คือขนาดตัวอย่าง และ p คือจำนวนตัวแปรอิสระ

เมื่อกำหนดให้ $y_i=0$ เมื่อเกิดเหตุการณ์ที่ไม่สนใจด้วยความน่าจะเป็นที่เกิดเหตุการณ์ที่ไม่สนใจคือ $1-\pi_i=P(Y_i=0|x_i)$ และ $y_i=1$ เมื่อเกิดเหตุการณ์ที่สนใจด้วยความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจคือ $\pi_i=P(Y_i=1|x_i)$ ดังนั้นเมื่อนำข้อมูลมาวิเคราะห์ซึ่งอยู่ในรูป (y_i, x_i) จะได้

$$P(Y_i=y_i)=\begin{cases} \pi_i^y(1-\pi_i)^{1-y_i} & ; y_i = 0,1 \\ 0 & ; y_i \text{ มีค่าอื่น ๆ} \end{cases}$$

โดยวิธีการที่ใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก คือ วิธีภาวะน่าจะเป็นสูงสุด (Maximum likelihood method) ซึ่งตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum likelihood estimator) มีรูปแบบดังนี้

$$\hat{\beta}_{ML} = (X^T \hat{W} X)^{-1} (X^T \hat{W} \hat{z}) \quad (2)$$

เมื่อ $\hat{\beta}_{ML}$ คือ เวกเตอร์ของตัวประมาณภาวะน่าจะเป็นสูงสุดที่มีขนาด $(p+1) \times 1$ และ X คือ เมทริกซ์ของตัวแปรอิสระที่มีขนาด $n \times (p+1)$ ส่วน $\hat{W}$ คือ เมทริกซ์ทแยงมุมที่มีขนาด $n \times n$ ซึ่งมีสมาชิกในแนวเส้นทแยงมุมเป็น $\hat{\pi}_i(1-\hat{\pi}_i)$; $i=1,2,3,...,n$ และ $\hat{z}$ คือ เวกเตอร์คอลัมน์ขนาด $n \times 1$ ที่มีสมาชิกเป็น $\hat{z}_i = \log(\hat{\pi}_i) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$; $i=1,2,3,...,n$ โดยที่ y_i คือ ค่าของตัวแปรตอบสนองหน่วยที่ i ; $i=1,2,3,...,n$ การวัดประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกจะวัดจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean square error: MSE) ซึ่งค่า MSE ของตัวประมาณภาวะน่าจะเป็นสูงสุดมีรูปแบบดังนี้

$$\text{MSE}(\hat{\beta}_{ML}) = E \left[(\hat{\beta}_{ML} - \beta)' (\hat{\beta}_{ML} - \beta) \right] = \text{tr} \left[(X'WX)^{-1} \right] = \sum_{j=1}^{p+1} \frac{1}{\lambda_j} \quad (3)$$

เมื่อ λ_j คือค่าเฉพาะ (Eigen value) ที่ j ของเมทริกซ์ $X'WX$ โดยที่ $j=1,2,3,...,p+1$

ตัวประมาณแบบหลิ

ตัวประมาณแบบหลิถูกเสนอขึ้นเป็นครั้งแรกโดย Liu ในปี ค.ศ. 1993 (Liu, 1993) เพื่อใช้แก้ปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระสำหรับตัวแบบการถดถอยเชิงเส้นโดยตัวประมาณแบบหลินี้เกิดจากการนำตัวประมาณแบบบริดจ์และตัวประมาณแบบสไตน์มารวมเข้าด้วยกันและต่อมาในปี ค.ศ. 2008 Urgan และ Tez (Urgan, & Tez, 2008) ได้นำแนวคิดของ Liu (Liu, 1993) มาปรับใช้เพื่อสร้างตัวประมาณแบบหลิสำหรับตัวแบบการถดถอยลอจิสติกเมื่อเกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระโดยมีรูปแบบดังนี้

$$\hat{\beta}_{LE} = (X'WX + I)^{-1} (X'WX + dI) \hat{\beta}_{ML} \quad (4)$$

เมื่อ d คือค่าพารามิเตอร์ซริงค์เกจ (shrinkage parameter) มีค่าอยู่ระหว่าง 0 ถึง 1 และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean square error: MSE) ของตัวประมาณแบบหลิมีรูปแบบดังนี้

$$\text{MSE}(\hat{\beta}_{LE}) = \sum_{j=1}^{p+1} \frac{(\lambda_j + d)^2}{\lambda_j (\lambda_j + 1)^2} + (d-1)^2 \sum_{j=1}^{p+1} \frac{\alpha_j^2}{(\lambda_j + 1)^2} \quad (5)$$

เมื่อ α_j^2 คือ สมาชิกตัวที่ j ของเวกเตอร์ของผลคูณระหว่างเวกเตอร์เฉพาะ (Eigenvector) ของเมทริกซ์ $X'WX$ กับเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่ได้จากวิธีภาวะน่าจะเป็นสูงสุด และ λ_j คือค่าเฉพาะ (Eigen value) ตัวที่ j ของเมทริกซ์ $X'WX$ เมื่อ $j=1,2,3,...,p+1$

ในปี ค.ศ. 2012 Månsson และคณะ (Månsson, Kibria, & Shukur, 2012) ได้เสนอตัวประมาณค่าพารามิเตอร์ซริงค์เกจทั้งหมด 5 ตัว ซึ่งพิจารณาจากตัวประมาณค่าพารามิเตอร์ซริงค์เกจ d_j แต่ละตัว (The individual parameter) โดย $d_j = (\alpha_j^2 - 1) / ((1/\lambda_j) + \alpha_j^2)$ และใช้แนวคิดของ Hoerl และ Kennard (Hoerl, & Kennard, 1970) Kibria (Kibria, 2003) และ Khalaf และ Shukur (Khalaf & Shukur, 2005) จากผลการศึกษาพบว่า ตัวประมาณแบบหลิที่ใช้ตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่ได้จากการพิจารณาดำแหน่งของควอนไทล์ต่ำสุดของตัวประมาณค่าพารามิเตอร์ซริงค์เกจแต่ละตัว มีประสิทธิภาพดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ ซึ่งตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่เป็นตำแหน่งของควอนไทล์ต่ำสุดของตัวประมาณค่าพารามิเตอร์ซริงค์เกจแต่ละตัวมีรูปแบบดังนี้

$$\hat{d}_1 = \max \left(0, \min \left(\frac{\hat{\alpha}_j^2 - 1}{\frac{1}{\hat{\lambda}_j} + \hat{\alpha}_j^2} \right) \right), j=1,2,3,...,p+1$$

ตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียง

แม้ว่าตัวประมาณแบบหลิวจะเหมาะสมสำหรับการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกเมื่อเกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระ แต่ตัวประมาณแบบหลิวก็ยังคงเป็นตัวประมาณเอนเอียง (Biased estimator) ดังนั้นในปี ค.ศ. 2015 Chang (Chang, 2015) ได้เสนอตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงขึ้น โดยมีรูปแบบดังนี้

$$\hat{\beta}_{AULE} = [I - (1-d)^2(X'WX + I)^{-2}] \hat{\beta}_{ML} \quad (6)$$

เมื่อ d คือ ค่าพารามิเตอร์ชริงค์เกจ (shrinkage parameter) มีค่าอยู่ระหว่าง 0 ถึง 1 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean square error: MSE) ของตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงมีรูปแบบดังนี้

$$MSE(\hat{\beta}_{AULE}) = \sum_{j=1}^{p+1} \frac{(\lambda_j + d)^2 (2 + \lambda_j - d)^2}{\lambda_j (\lambda_j + 1)^4} + \sum_{j=1}^{p+1} \frac{(1-d)^4 \alpha_j^2}{(\lambda_j + 1)^4} \quad (7)$$

เมื่อ α_j^2 คือ สมาชิกตัวที่ j ของเวกเตอร์ของผลคูณระหว่างเวกเตอร์เฉพาะ (Eigenvector) ของเมทริกซ์ $X'WX$ กับเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่ได้จากวิธีภาวะน่าจะเป็นสูงสุด และ λ_j คือค่าเฉพาะ (Eigen value) ตัวที่ j ของเมทริกซ์ $X'WX$ เมื่อ $j=1,2,3,\dots,p+1$ นอกจากนั้น Chang (Chang, 2015) ได้แสดงให้เห็นว่าตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงจะเป็นตัวประมาณที่มีประสิทธิภาพดีกว่าตัวประมาณแบบหลิวในบางกรณี

เสนอตัวประมาณค่าพารามิเตอร์ชริงค์เกจ

งานวิจัยนี้ได้เสนอตัวประมาณค่าพารามิเตอร์ชริงค์เกจขึ้น 2 ตัว โดยตัวแรกประยุกต์จากวิธีของ Hoerl และ Kennard (Hoerl, & Kennard, 1970) ซึ่งเป็นการแทนค่าพารามิเตอร์ที่ไม่ทราบค่าด้วย $\hat{\alpha}_{\max}^2$ และ $\hat{\lambda}_{\max}$ แต่ในที่นี้จะเลือกแทนด้วยค่า $\hat{\alpha}_{\min}^2$ และ $\hat{\lambda}_{\min}$ เพื่อให้ค่าประมาณพารามิเตอร์ชริงค์เกจมีค่าต่ำ ๆ ซึ่งจะส่งผลให้ค่าประมาณสัมประสิทธิ์การถดถอยลอจิสติกที่ได้ อาจจะเป็นตัวประมาณที่มีประสิทธิภาพดีกว่ากรณีที่ค่าประมาณพารามิเตอร์ชริงค์เกจมีค่าสูง ส่วนตัวที่ 2 จะประมาณค่าพารามิเตอร์ชริงค์เกจโดยการนำ $MSE(\hat{\beta}_{LE})$ มาหาอนุพันธ์อันดับที่ 1 เทียบกับค่า d แล้วให้เท่ากับ 0 หลังจากนั้นแก้สมการหาค่า d เพื่อให้ค่า MSE มีค่าต่ำสุด ตัวประมาณค่าพารามิเตอร์ชริงค์เกจที่ผู้วิจัยนำเสนอมีรูปแบบตามลำดับดังนี้

$$\hat{d}_2 = \max \left(0, \frac{\hat{\alpha}_{\min}^2 - 1}{\frac{1}{\hat{\lambda}_{\min}} + \hat{\alpha}_{\min}^2} \right) \quad \text{และ} \quad \hat{d}_3 = \max \left(0, 1 - \frac{\sum_{j=1}^{p+1} \left\{ \frac{1}{\lambda_j (\lambda_j + 1)} \right\}}{\sum_{j=1}^{p+1} \left\{ \frac{1 + \lambda_j \alpha_j^2}{\lambda_j (\lambda_j + 1)^2} \right\}} \right)$$

เมื่อ $\hat{\alpha}_{\min}^2$ คือ α_j^2 ที่มีค่าน้อยที่สุด และ $\hat{\lambda}_{\min}$ คือ λ_j ที่มีค่าน้อยที่สุด โดยที่ α_j^2 คือสมาชิกตัวที่ j ของเวกเตอร์ของผลคูณระหว่างเวกเตอร์เฉพาะ (Eigenvector) ของเมทริกซ์ $X'WX$ กับเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่ได้จากวิธีภาวะน่าจะเป็นสูงสุด และ λ_j คือค่าเฉพาะ (Eigen value) ตัวที่ j ของเมทริกซ์ $X'WX$ เมื่อ $j=1,2,3,...,p+1$

ตัวประมาณแบบหลิวด้วยวิธีบูตสเตรปและตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป

การเกิดปัญหาหุ้มนัยสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) และขนาดตัวอย่างที่ทำการศึกษามีน้อย (Small sample size) ส่งผลให้การประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกนั้นได้ตัวประมาณค่าที่มีประสิทธิภาพต่ำลง ในงานวิจัยนี้ได้แก้ปัญหาดังกล่าวโดยนำเทคนิคบูตสเตรปที่ขึ้นอยู่กับ การสุ่มตัวอย่างซ้ำค่าสังเกต (The resampling observations) มาใช้ร่วมกับการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวและการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียง โดยมีขั้นตอนดังนี้

ขั้นตอนที่ 1 สร้างตัวอย่างบูตสเตรปขนาด n ด้วยการสุ่มตัวอย่างซ้ำค่าสังเกตจากข้อมูลตั้งต้น $z_i = (y_i, x_i)$, $i=1,2,...,n$ โดยทำการสุ่มแบบคืนที่ด้วยความน่าจะเป็น $\frac{1}{n}$ จะได้ค่าสังเกตใหม่ $z_i^* = (y_i^*, x_i^*)$, $i=1,2,...,n$

ขั้นตอนที่ 2 ประมาณค่าพารามิเตอร์ β ด้วยตัวประมาณแบบหลิวและตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียง โดยใช้วิธีการประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$

ขั้นตอนที่ 3 ทำซ้ำขั้นตอนที่ 1-2 เท่ากับจำนวนรอบบูตสเตรป โดยกำหนดให้จำนวนรอบบูตสเตรปเป็น B ครั้ง จะได้ตัวประมาณแบบหลิวและตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงในแต่ละรอบบูตสเตรป

ขั้นตอนที่ 4 นำค่าประมาณพารามิเตอร์ β ที่ได้ในแต่ละรอบบูตสเตรปมาหาค่าเฉลี่ย จะได้ตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสเตรปและตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป ดังนี้

$$\bar{\beta}_{LE(\hat{d}_1)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{LE(\hat{d}_1)}^{*(b)}, \quad \bar{\beta}_{LE(\hat{d}_2)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{LE(\hat{d}_2)}^{*(b)} \quad \text{และ} \quad \bar{\beta}_{LE(\hat{d}_3)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{LE(\hat{d}_3)}^{*(b)}$$

$$\bar{\beta}_{AULE(\hat{d}_1)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{AULE(\hat{d}_1)}^{*(b)}, \quad \bar{\beta}_{AULE(\hat{d}_2)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{AULE(\hat{d}_2)}^{*(b)} \quad \text{และ} \quad \bar{\beta}_{AULE(\hat{d}_3)}^* = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{AULE(\hat{d}_3)}^{*(b)}$$

เมื่อ $\hat{\beta}_{LE(\hat{d}_1)}^{*(b)}$, $\hat{\beta}_{LE(\hat{d}_2)}^{*(b)}$ และ $\hat{\beta}_{LE(\hat{d}_3)}^{*(b)}$ คือ ตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิว สำหรับ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ ตามลำดับ ในรอบบูตสเตรปที่ b ; $b=1,2,...,B$

และ $\hat{\beta}_{AULE(\hat{d}_1)}^{*(b)}$, $\hat{\beta}_{AULE(\hat{d}_2)}^{*(b)}$ และ $\hat{\beta}_{AULE(\hat{d}_3)}^{*(b)}$ คือ ตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียง สำหรับ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ ตามลำดับ ในรอบบูตสเตรปที่ b ; $b=1,2,...,B$

การวิจัยครั้งนี้ทำการจำลองข้อมูลภายใต้สถานการณ์ที่จำนวนตัวแปรอิสระเท่ากับ 23 และ 4 ขนาดตัวอย่างเท่ากับ 20 40 80 และ 200 โดยขนาดตัวอย่างเล็กคือ ขนาดตัวอย่างน้อยกว่า 30 เท่าของจำนวนตัวแปรอิสระและขนาดตัวอย่างใหญ่คือ ขนาดตัวอย่างเท่ากับ 200 (กัลยา, 2549) และกำหนดระดับความสัมพันธ์

ระหว่างตัวแปรอิสระ (ρ) เท่ากับ 0.85 0.95 และ 0.99 ซึ่งสูตรที่ใช้ในการสร้างตัวแปรอิสระให้เกิดพหุสัมพันธ์คือ $x_{ij} = (1 - \rho^2)^{1/2} z_{ij} + \rho z_{ip}$, $i = 1, 2, \dots, n$ และ $j = 1, 2, \dots, p$ เมื่อ z_{ij} คือเลขสุ่มเทียมที่สร้างมาจากการแจกแจงปรกติมาตรฐาน ส่วนตัวแปรตอบสนองสร้างจากการแจกแจงแบร์นูลลีที่มีพารามิเตอร์ π_i เมื่อ $\pi_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$ ภายใต้ $\beta' \beta = 1$ ผู้วิจัยจึงกำหนดให้ $\beta_0 = 0$ และ $\beta_1, \beta_2, \dots, \beta_p = 1/\sqrt{p}$ การจำลองข้อมูลตามแผนการดำเนินการวิจัยนี้ใช้เทคนิคมอนติคาร์โลบนโปรแกรม R เวอร์ชัน 3.4.3 โดยในการจำลองจะกระทำซ้ำทั้งหมด 1,000 ครั้ง และการทำบูตสเตรปจะกระทำซ้ำ 500 ครั้ง โดยมีสูตรที่ใช้ในการคำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) คือ

$$MSE = \frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{\beta}_i - \beta)' (\hat{\beta}_i - \beta) \quad (8)$$

เมื่อ $\hat{\beta}_i$ คือเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่มีขนาด $(p+1) \times 1$ ในการจำลองรอบที่ i , $i = 1, 2, \dots, 1,000$

ผู้วิจัยเปรียบเทียบประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) โดยตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกที่มีประสิทธิภาพดีที่สุดจะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด นอกจากนั้นผู้วิจัยได้นำจำนวนรอบที่ค่าประมาณพารามิเตอร์ซริงค์เกจเท่ากับศูนย์มาเป็นเกณฑ์ที่ใช้ในการประกอบการตัดสินใจเพื่อวัดประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ โดยตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่ทำให้การประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิwmมีประสิทธิภาพดีจะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) น้อยและจำนวนรอบที่ค่าประมาณพารามิเตอร์ซริงค์เกจเท่ากับศูนย์มีค่าน้อยด้วย

ผลการวิจัย

จากตารางที่ 1 พบว่า ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.85 กรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 2 ภายใต้ขนาดตัวอย่างที่เท่ากัน เมื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ เห็นได้ชัดว่า ในทุก ๆ วิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่มีประสิทธิภาพดีที่สุดคือ $\hat{d}_2$ เนื่องจากให้ค่า MSE ต่ำที่สุด รองลงมาคือ $\hat{d}_1$ และ $\hat{d}_3$ ตามลำดับ นอกจากนั้นสังเกตได้ว่า ที่ขนาดตัวอย่างเท่ากับ 20 40 และ 80 ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$ และ $\hat{d}_2$ ให้ค่า MSE ใกล้เคียงกัน ส่วนที่ขนาดตัวอย่างเท่ากับ 200 ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$ และ $\hat{d}_2$ ส่งผลให้ได้ค่า MSE ใกล้เคียงกันมาก นอกจากนั้นเมื่อเปรียบเทียบวิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก ในทุก ๆ ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ เห็นได้ชัดว่า ที่ขนาดตัวอย่างเท่ากับ 20 40 และ 80 วิธีถดถอยลอจิสติกแบบหลิwmด้วยวิธีบูตสเตรป (BLE) ให้ค่า MSE ต่ำที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 200 วิธีถดถอยลอจิสติกแบบหลิwmที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป (BAULE) ให้ค่า MSE ต่ำที่สุด ส่วนผลลัพธ์ในกรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 3 และ 4 เป็นไปในทิศทางเดียวกันกับจำนวนตัวแปรอิสระเท่ากับ 2 เกือบทุกสถานการณ์

จากตารางที่ 2 พบว่า ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.95 กรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 2 ภายใต้ขนาดตัวอย่างที่เท่ากัน เมื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ ชริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ เห็นได้ชัดว่า ในทุก ๆ วิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก ตัวประมาณค่าพารามิเตอร์ชริงค์เกจที่มีประสิทธิภาพดีที่สุดคือ $\hat{d}_2$ เนื่องจากให้ค่า MSE ต่ำที่สุด รองลงมาคือ $\hat{d}_1$ และ $\hat{d}_3$ ตามลำดับ และสังเกตได้ว่าที่ขนาดตัวอย่างเท่ากับ 40 และ 80 ตัวประมาณค่าพารามิเตอร์ชริงค์เกจ $\hat{d}_1$ และ $\hat{d}_2$ ส่งผลให้ได้ค่า MSE ใกล้เคียงกัน ส่วนที่ขนาดตัวอย่างเท่ากับ 200 ตัวประมาณค่าพารามิเตอร์ชริงค์เกจ $\hat{d}_1$ และ $\hat{d}_2$ ส่งผลให้ได้ค่า MSE ใกล้เคียงกันมาก นอกจากนั้นเมื่อเปรียบเทียบวิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก เห็นได้ชัดว่า ในทุก ๆ ตัวประมาณค่าพารามิเตอร์ชริงค์เกจ วิธีการถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสเตรป (BLE) ให้ค่า MSE ต่ำที่สุด ส่วนผลลัพธ์ในกรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 3 และ 4 เป็นไปในทิศทางเดียวกันกับจำนวนตัวแปรอิสระเท่ากับ 2 เกือบทุกสถานการณ์

จากตารางที่ 3 พบว่า ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.99 กรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 2 ภายใต้ขนาดตัวอย่างที่เท่ากัน เมื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ ชริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ เห็นได้ชัดว่า ในทุก ๆ วิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก ตัวประมาณค่าพารามิเตอร์ชริงค์เกจที่มีประสิทธิภาพดีที่สุดคือ $\hat{d}_2$ เนื่องจากส่งผลให้ได้ค่า MSE ต่ำที่สุด รองลงมาคือ $\hat{d}_1$ และ $\hat{d}_3$ ตามลำดับ นอกจากนั้นเมื่อเปรียบเทียบวิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก เห็นได้ชัดว่า ในทุก ๆ ตัวประมาณค่าพารามิเตอร์ชริงค์เกจ วิธีการถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสเตรป (BLE) ให้ค่า MSE ต่ำที่สุด ส่วนผลลัพธ์ในกรณีที่กำหนดจำนวนตัวแปรอิสระเท่ากับ 3 และ 4 เป็นไปในทิศทางเดียวกันกับจำนวนตัวแปรอิสระเท่ากับ 2 เกือบทุกสถานการณ์

นอกจากนั้นในแต่ละระดับความสัมพันธ์ระหว่างตัวแปรอิสระ เมื่อขนาดตัวอย่างเพิ่มขึ้นจะส่งผลให้ค่า MSE มีแนวโน้มลดลง ส่วนเมื่อจำนวนตัวแปรอิสระเพิ่มขึ้นจะส่งผลให้ค่า MSE มีแนวโน้มเพิ่มขึ้นด้วย

ตารางที่ 1 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.85

วิธีการ ประมาณ	จำนวนตัวแปรอิสระเท่ากับ 2			จำนวนตัวแปรอิสระเท่ากับ 3			จำนวนตัวแปรอิสระเท่ากับ 4		
	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$
n=20									
LE	1.5277	1.4861	1.8777	1.8241	1.7509	2.8322	2.0635	1.9667	4.2888
BLE	1.4587	1.4519	1.6722	1.6813	1.6451	2.5858	1.7922	1.7375	3.9832
AULE	2.3266	2.2768	2.7716	3.0827	2.9794	4.6190	3.7923	3.6130	7.3653
BAULE	2.0096	2.0037	2.1867	2.6252	2.6076	4.1718	3.5862	3.5529	4.9862
n=40									
LE	1.1949	1.1919	1.2752	1.5228	1.5186	1.7682	1.8550	1.8191	2.2922
BLE	1.0866	1.0863	1.2532	1.4939	1.4932	1.5682	1.5540	1.5378	2.2810
AULE	1.6504	1.6475	1.7259	2.3415	2.3364	2.6394	3.0054	3.0047	3.5856
BAULE	1.3590	1.3586	1.3724	2.0775	2.0748	2.6112	2.3646	2.3632	2.7238
n=80									
LE	1.0275	1.0275	1.0413	1.2965	1.2965	1.3298	1.5788	1.5788	1.6491
BLE	0.8972	0.8970	0.9344	1.1477	1.1477	1.1486	1.3746	1.3743	1.4958
AULE	1.2298	1.2298	1.2376	1.6961	1.6961	1.7223	2.1696	2.1696	2.2323
BAULE	1.1195	1.1193	1.1398	1.2623	1.2623	1.2644	1.9095	1.9093	1.9955
n=200									
LE	0.8141	0.8141	0.8144	0.9598	0.9598	0.9600	1.1851	1.1851	1.1859
BLE	0.7354	0.7354	0.7354	0.8493	0.8493	0.8578	0.9103	0.9103	1.0245
AULE	0.8623	0.8623	0.8624	1.0599	1.0599	1.0600	1.3606	1.3606	1.3610
BAULE	0.7336	0.7336	0.7336	0.7808	0.7808	0.7808	0.9985	0.9985	0.9995

LE = วิธีการถดถอยโลจิสติกแบบหลิ่ว

BLE = วิธีการถดถอยโลจิสติกแบบหลิ่วด้วยวิธีบูตสเตรป

AULE = วิธีการถดถอยโลจิสติกแบบหลิ่วที่เกือบจะไม่เอนเอียง

BAULE = วิธีการถดถอยโลจิสติกแบบหลิ่วที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป

ตารางที่ 2 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.95

วิธีการ ประมาณ	จำนวนตัวแปรอิสระเท่ากับ 2			จำนวนตัวแปรอิสระเท่ากับ 3			จำนวนตัวแปรอิสระเท่ากับ 4		
	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$
n=20									
LE	2.1367	1.7573	3.2354	2.3609	1.7624	6.3006	3.3606	2.0302	12.3205
BLE	1.6242	1.5416	2.9969	1.8455	1.6748	5.7209	2.8920	1.8766	9.9082
AULE	3.5009	2.9642	5.2415	4.4639	3.3242	11.2944	6.5583	4.0932	21.8962
BAULE	2.2796	2.1329	3.3915	2.9977	2.3373	8.7392	5.8543	4.6031	18.7326
n=40									
LE	1.2952	1.2248	1.7986	1.6871	1.5604	3.4564	2.1259	1.8577	5.3898
BLE	1.0878	1.0844	1.5913	1.5490	1.5444	2.2875	1.6057	1.5481	3.3098
AULE	2.0521	1.9521	2.7112	2.8658	2.4976	5.5037	3.5271	3.1996	7.8546
BAULE	1.8153	1.7810	1.9270	2.7608	2.4932	5.4978	2.9743	2.8787	6.6342
n=80									
LE	1.1095	1.0962	1.2228	1.3762	1.3724	1.8464	1.6151	1.6115	2.5785
BLE	0.9225	0.9095	0.9606	1.1493	1.1493	1.3949	1.5216	1.4747	2.5081
AULE	1.6130	1.5996	1.7433	2.2488	2.2431	2.9164	2.8363	2.8292	4.2333
BAULE	1.2867	1.2857	1.3170	2.0050	2.0005	2.3640	2.0386	2.0334	2.4719
n=200									
LE	0.9575	0.9573	0.9734	1.2283	1.2283	1.2801	1.5325	1.5324	1.7028
BLE	0.7615	0.7614	0.7685	0.8593	0.8593	0.8838	1.2594	1.2592	1.4628
AULE	1.1958	1.1956	1.2065	1.7080	1.7080	1.7560	2.2519	2.2518	2.4254
BAULE	0.9127	0.9127	0.9169	1.1042	1.1042	1.1168	1.1528	1.1527	1.2184

LE = วิธีการถดถอยลอจิสติกแบบหลิ่ว

BLE = วิธีการถดถอยลอจิสติกแบบหลิ่วด้วยวิธีบูตสตรอป

AULE = วิธีการถดถอยลอจิสติกแบบหลิ่วที่เกือบจะไม่เอนเอียง

BAULE = วิธีการถดถอยลอจิสติกแบบหลิ่วที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสตรอป

ตารางที่ 3 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.99

วิธีการ ประมาณ	จำนวนตัวแปรอิสระเท่ากับ 2			จำนวนตัวแปรอิสระเท่ากับ 3			จำนวนตัวแปรอิสระเท่ากับ 4		
	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$
n=20									
LE	7.1434	2.6073	13.5578	10.2585	2.9031	32.1549	16.4569	3.5329	64.0972
BLE	4.5871	1.7146	6.0089	4.9981	1.7940	27.7499	5.0320	2.1844	46.5194
AULE	12.7460	5.5311	23.9183	23.822	8.1201	61.6426	33.4733	8.7201	114.6395
BAULE	8.3673	3.0765	14.5181	17.9211	4.7934	54.9004	24.9987	5.3072	93.3313
n=40									
LE	3.6756	1.7972	6.2171	3.7397	1.8613	13.5757	5.8580	1.9953	24.9862
BLE	2.7994	1.5049	3.9704	2.9412	1.6924	14.8504	2.7178	1.9248	23.8773
AULE	5.8756	3.4471	10.3189	7.8227	3.5896	24.8028	12.2182	4.3298	44.3420
BAULE	4.9380	3.0394	7.0956	5.5144	4.1599	12.2721	6.4848	3.0403	41.3489
n=80									
LE	1.9228	1.3688	2.6361	2.1728	1.4965	6.7670	3.2231	1.7548	12.6375
BLE	1.5706	1.2762	1.7260	1.5857	1.4587	4.1688	1.9117	1.5162	8.1998
AULE	3.4061	2.7743	4.9221	3.8629	2.8561	10.8478	5.2259	3.1899	20.3433
BAULE	2.6601	2.2563	2.8754	2.8601	2.3925	5.2338	4.2954	2.4957	16.9158
n=200									
LE	1.2485	1.1549	1.5060	1.3342	1.2499	2.8588	1.6955	1.5924	4.7519
BLE	0.7677	0.7649	0.8193	0.8104	0.8083	1.2966	1.5538	0.8576	3.7051
AULE	1.9987	1.8931	2.3638	2.4402	2.3056	4.9278	3.1246	2.9428	8.1884
BAULE	0.8117	0.8068	0.8974	1.3812	1.3717	1.5926	1.7322	1.6393	5.2451

LE = วิธีการถดถอยลอจิสติกแบบหลิ

BLE = วิธีการถดถอยลอจิสติกแบบหลิด้วยวิธีบูตสเตรป

AULE = วิธีการถดถอยลอจิสติกแบบหลิที่เกือบจะไม่เอนเอียง

BAULE = วิธีการถดถอยลอจิสติกแบบหลิที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป

จากผลการวิจัยที่ได้จากการจำลองข้อมูลข้างต้น ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_2$ เป็นตัวประมาณที่ส่งผลให้ค่า MSE ของทุก ๆ ตัวประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกมีค่าต่ำที่สุด รองลงมาคือ $\hat{d}_1$ และ $\hat{d}_3$ ตามลำดับ แต่จากการพิจารณารูปแบบของตัวประมาณค่าพารามิเตอร์ซริงค์เกจทั้ง 3 ตัว แล้วพบว่า มีโอกาสที่จะมีค่าเท่ากับศูนย์ได้ ซึ่งจะทำให้รูปแบบของตัวประมาณแบบหลิเปลี่ยนไปเป็นตัวประมาณแบบริดจ์ที่มีค่าประมาณของพารามิเตอร์ริดจ์เท่ากับ 1 ดังนั้นเพื่อให้เห็นค่าประมาณของ $\hat{d}_1, \hat{d}_2$ และ $\hat{d}_3$ ที่ได้จากข้อมูลนั้น ๆ ผู้วิจัยจึงได้นำเสนอผลลัพธ์ในบางสถานการณ์ที่ศึกษาที่จำนวนรอบของค่าประมาณพารามิเตอร์ซริงค์เกจมีค่าเท่ากับศูนย์มาเป็นเกณฑ์ที่ใช้ในการประกอบการตัดสินใจร่วมกับค่า MSE เพื่อวัดประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ โดยสถานการณ์ที่นำเสนอคือจำนวนตัวแปรอิสระเท่ากับ 3 ขนาดตัวอย่างเท่ากับ 80 และระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.95 และ 0.99 ดังแสดงในตารางที่ 4

ตารางที่ 4 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และจำนวนรอบที่ค่าประมาณพารามิเตอร์ซริงค์เกจมีค่าเท่ากับศูนย์ เมื่อจำนวนตัวแปรอิสระเท่ากับ 3 ขนาดตัวอย่างเท่ากับ 80 และระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.95 และ 0.99

วิธีการถดถอยลอจิสติกแบบหลิว	ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ		
	$\hat{d}_1$	$\hat{d}_2$	$\hat{d}_3$
$\rho = 0.95$			
MSE	1.38	1.37	1.85
จำนวนรอบที่ $d=0$	986	986	692
$\rho = 0.99$			
MSE	1.83	1.29	5.93
จำนวนรอบที่ $d=0$	833	833	256

จากตารางที่ 4 พบว่า ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.95 และ 0.99 จำนวนตัวแปรอิสระเท่ากับ 3 และขนาดตัวอย่างเท่ากับ 80 เมื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ เห็นได้ชัดว่า ตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่มีประสิทธิภาพดีที่สุดคือ $\hat{d}_2$ เพราะส่งผลให้ได้ตัวประมาณแบบหลิวที่ให้ค่า MSE ต่ำที่สุด แต่จะสังเกตได้ว่าการทำซ้ำในการจำลอง 1,000 รอบ เมื่อนับจำนวนรอบที่ค่าประมาณของพารามิเตอร์ซริงค์เกจมีค่าเท่ากับศูนย์นั้นสูงที่สุดเมื่อเปรียบเทียบกับตัวประมาณค่าพารามิเตอร์ซริงค์เกจอื่น ๆ จากผลการวิจัยนี้ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_3$ เป็นอีกตัวประมาณหนึ่งที่มีประสิทธิภาพสำหรับนำไปประยุกต์ใช้กับข้อมูลจริงเนื่องจากการพิจารณาจำนวนรอบที่ค่าประมาณของพารามิเตอร์ซริงค์เกจมีค่าเท่ากับศูนย์รวมด้วย แม้ว่าตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_3$ นี้จะส่งผลให้ได้ตัวประมาณแบบหลิวที่ให้ค่า MSE สูงกว่าตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_2$ ก็ตาม แต่จะสังเกตได้ว่าการทำซ้ำในการจำลอง 1,000 รอบ เมื่อนับจำนวนรอบที่ค่าประมาณของพารามิเตอร์ซริงค์เกจมีค่าเท่ากับศูนย์นั้นน้อยกว่าตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_2$ มาก นอกจากนี้ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_3$ ไม่ได้เป็นตัวประมาณค่าพารามิเตอร์แต่ละตัว (The individual parameter: d_j) เหมือนกับ $\hat{d}_1$ และ $\hat{d}_2$ แต่เป็นตัวประมาณค่าพารามิเตอร์ซริงค์เกจเป็นค่าเดียว (A single value: $\hat{d}$) อย่างแท้จริง

สรุปผลและอภิปรายผล

จากการจำลองข้อมูลและวิเคราะห์ผล กรณีที่ตัวแปรอิสระเท่ากับ 2 3 และ 4 ขนาดตัวอย่างเท่ากับ 20 40 80 และ 200 และระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.85 0.95 และ 0.99 สามารถสรุปผลได้ดังนี้ เมื่อเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_1$, $\hat{d}_2$ และ $\hat{d}_3$ ในทุก ๆ วิธีการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก ภายใต้ขนาดตัวอย่างที่เท่ากัน ตัวประมาณค่าพารามิเตอร์ซริงค์เกจที่มีประสิทธิภาพดีที่สุดคือ $\hat{d}_2$ และเมื่อเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกในทุก ๆ ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ กรณีที่ขนาดตัวอย่างเล็ก วิธีการถดถอยลอจิสติกแบบหลิวด้วย

วิธีบูตสเตรป (BLE) มีประสิทธิภาพมากที่สุด และเมื่อขนาดตัวอย่างใหญ่ วิธีการถดถอยลอจิสติกแบบหลิวด้วยวิธีบูตสเตรป (BLE) ให้ค่า MSE ต่ำที่สุดเกือบจะทุกสถานการณ์ ยกเว้นที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.85 วิธีการถดถอยลอจิสติกแบบหลิวที่เกือบจะไม่เอนเอียงด้วยวิธีบูตสเตรป (BAULE) ให้ค่า MSE ต่ำที่สุด

จากผลการวิจัยข้างต้น จะเห็นได้ชัดว่าปัจจัยที่ส่งผลต่อการเปลี่ยนแปลงค่า MSE ประกอบด้วย ขนาดตัวอย่าง ระดับความสัมพันธ์ระหว่างตัวแปรอิสระและจำนวนตัวแปรอิสระ โดยการเพิ่มขนาดตัวอย่างส่งผลให้ค่า MSE มีแนวโน้มลดลงเพราะขนาดตัวอย่างที่เพิ่มขึ้นจะส่งผลให้ค่าความคลาดเคลื่อนในการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกลดลง ดังนั้นตัวประมาณค่าที่ได้จะมีความเที่ยง (Precision) และความแม่นยำ (Accuracy) เมื่อขนาดตัวอย่างที่ทำการศึกษามีมากเพียงพอ ส่วนการเกิดปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระนั้นจะเห็นได้ว่าเมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้นส่งผลให้ค่า MSE มีแนวโน้มเพิ่มขึ้นด้วย นอกจากนี้การเพิ่มจำนวนตัวแปรอิสระก็ส่งผลให้ค่า MSE มีแนวโน้มเพิ่มขึ้นด้วยเช่นกันเพราะจำนวนตัวแปรอิสระเพิ่มขึ้นทำให้เกิดความสัมพันธ์ระหว่างตัวแปรอิสระมากขึ้นตามไปด้วย

จากการจำลองสถานการณ์ในงานวิจัยครั้งนี้ แม้ว่ารูปแบบของตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_2$ ส่งผลให้ได้ตัวประมาณแบบหลิวที่มีประสิทธิภาพดีกว่าเมื่อใช้ตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_3$ ก็ตาม แต่ค่าประมาณของพารามิเตอร์ซริงค์เกจอาจจะมีค่าเท่ากับศูนย์ ซึ่งส่งผลให้รูปแบบของตัวประมาณแบบหลิวเปลี่ยนไปเป็นรูปแบบของตัวประมาณแบบริดจ์ที่มีค่าประมาณของพารามิเตอร์ริดจ์เท่ากับ 1 การจำลองสถานการณ์ในงานวิจัยนี้ใช้เทคนิคมอนติคาร์โล กระทำซ้ำทั้งหมด 1,000 รอบ เมื่อนับจำนวนรอบที่ค่าประมาณของพารามิเตอร์ซริงค์เกจด้วยวิธี $\hat{d}_3$ มีค่าเท่ากับศูนย์นั้นน้อยมากเมื่อเปรียบเทียบกับวิธี $\hat{d}_2$ ดังนั้นตัวประมาณค่าพารามิเตอร์ซริงค์เกจ $\hat{d}_3$ อาจจะเป็นทางเลือกหนึ่งเพื่อนำมาใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกแบบหลิว

ข้อเสนอแนะ

1. วิธีการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติก เมื่อเกิดปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระด้วยวิธีการถดถอยลอจิสติกแบบหลิวเป็นวิธีที่มีประสิทธิภาพดี ดังนั้นจึงควรศึกษาและพัฒนาวิธีการประมาณค่าพารามิเตอร์ซริงค์เกจให้มีค่าใกล้เคียงศูนย์
2. ควรเพิ่มขนาดตัวอย่างให้มากกว่า 200 เพื่อจะได้เห็นประสิทธิภาพของตัวประมาณแบบหลิวที่เกือบจะไม่เอนเอียงว่าดีกว่าตัวประมาณแบบหลิวหรือไม่
3. แม้ว่าตัวประมาณแบบหลิวจะมีประสิทธิภาพดีแต่ยังมีข้อด้อยในเรื่องของการประมาณค่าพารามิเตอร์ซริงค์เกจที่อาจจะมีค่าเท่ากับศูนย์ ดังนั้นจึงควรทำการศึกษาวิธีประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกด้วยตัวประมาณแบบหลิว-ไทป์ (Liu-Type estimator) เพิ่มเติม ซึ่งตัวประมาณนี้อาจจะมีประสิทธิภาพที่ดีกว่าตัวประมาณแบบหลิว

References

- Chang, X. (2015). On the almost unbiased Ridge and Liu estimator in the logistic regression model. Paper presented at the International Conference on Social Science, Education Management and Sports Education (SSEMSE 2015), Beijing.
- Farghali, R. A., & Abo-El-Hadid, S. M. (2017). Evaluating the Performance of Liu Logistic Regression Estimator. *Research Journal of Mathematics and Statistics*, 9(2), 11-19.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1), 55-67.
- Khalaf, G., & Shukur, G. (2005). Choosing Ridge Parameter for Regression Problems. *Communications in Statistics - Theory and Methods*, 34(5), 1177-1182.
- Kibria, B. M. G. (2003). Performance of Some New Ridge Regression Estimators. *Communications in Statistics - Simulation and Computation*, 32(2), 419-435.
- Liu, K. (1993). A new class of biased estimate in linear regression. *Commun Statist Theor Meth*, 22, 393-402.
- Månsson, K., Kibria, B. M. G., & Shukur, G. (2012). On Liu estimators for the logit regression model. *Economic Modelling*, 29(4), 1483-1488.
- Schaefer, R.L., Roi, L.D., & Wolfe, R.A. (1984). A ridge logistic estimator. *Commun Statist Theor Meth*, 13(1), 99-113.
- Stein, C. (1960). Multiple regression. In I. Olkin (Ed.), *Contributions to probability and statistics: essays in honor of Harold Hotelling* (pp. 424-443). Palo Alto: Stanford University Press.
- Urgan, N. N., & Tez, M. (2008). Liu estimator in logistic regression when the data are collinear. Paper presented at the 20th EURO Mini Conference "Continuous Optimization and Knowledge Based Technologies", Vilnius. Retrieved August 25, 2017, from http://leidykla.vgtu.lt/conferences/europt_2008/files/056.html
- Vanichbuncha, K. (2006). "kān wikhroḥ sathiti khan sūṅg dūai SPSS for windows" [Advanced statistical analysis with SPSS for windows]. 5th edition. Bangkok: Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University.