การเปรียบเทียบการคัดเลือกตัวแปรอิสระที่มีปัญหาสหสัมพันธ์เชิงเส้นพหุด้วยวิธีการถดถอยแบบริดจ์และการค้นหาแบบต้องห้าม

A Comparison of Variable Selection with Multicollinearity by Ridge Regression and Tabu Search

นิศาชล งามประเสริฐสิทธิ์* และ จิราวัลย์ จิตรถเวช
สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์
Nisachon Ngamprasertsit* and Jirawan Jitthavech
Department of Statistics, School of Applied Statistics, National Institute of Development Administration

บทคัดย่อ

การศึกษามีวัตถุประสงค์เพื่อเปรียบเทียบการคัดเลือกตัวแปรอิสระในการวิเคราะห์การถดถอยเชิงเส้นพหุที่ตัวแบบมีตัวแปรอิสระที่เกี่ยวข้องและไม่เกี่ยวข้องกับตัวแปรตาม โดยตัวแปรอิสระที่เกี่ยวข้อง 1 คู่ มีความสัมพันธ์กันสูง การคัดเลือกตัวแปรอิสระใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ โดยใช้วิธีการประมาณค่าคงตัวริดจ์ 4 วิธีคือ วิธีโฮเอิร์ล เคนนาร์ด และ บาลด์วิน (Hoerl, Kennard and Baldwin) วิธีลอว์เลสและแวง (Lawless and Wang) วิธีนอมูระ (Nomura) และวิธีคาลาฟและชูเกอร์ (Khalaf and Shukur) กับการคัดเลือกตัวแปรอิสระที่ประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธีการค้นหาแบบต้องห้าม (Tabu Search) ที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ (Penalty Function) เกณฑ์ที่ใช้ในการเปรียบเทียบการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ คือร้อยละของจำนวนครั้งที่แต่ละวิธีสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ตามตัวแบบจริง การศึกษาใช้วิธีการจำลองข้อมูล กำหนดขนาดตัวอย่างเท่ากับ 20, 60 และ 100 และกระทำซ้ำในแต่ละสถานการณ์ 500 ครั้ง เมื่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระเป็น 0.95, 0.99 และ 0.999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายทั้ง 2 ฟังก์ชัน มีร้อยละของการคัดเลือกได้ตัวแบบจริงมากกว่าวิธีอื่นๆ และค่อนข้างเสถียรในทุกขนาดตัวอย่าง ยกเว้นกรณีของริดจ์ที่มีการประมาณค่าคงตัวโดยวิธีคาลาฟและชูเกอร์เมื่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้นเป็น 0.9999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ มีร้อยละของการคัดเลือกได้ตัวแบบจริงสูงและค่อนข้างคงที่ โดยไม่ขึ้นกับขนาดตัวอย่างและค่าสัมประสิทธิ์สหสัมพันธ์ นอกจากนั้น ผลการศึกษาไม่พบตัวแบบ Underspecification และ Misspecification มีเพียงตัวแบบ Overspecification ซึ่งเป็นปัญหาที่มีความรุนแรงในการวิเคราะห์น้อยกว่าตัวแบบในสองกรณีแรก ในขณะที่วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และวิธีการถดถอยแบบขั้นตอนที่มีการประมาณค่าพารามิเตอร์กำลังสองน้อยที่สุดและแบบริดจ์ มีร้อยละของการคัดเลือกได้ตัวแบบจริงมีค่าต่ำ เมื่อขนาดตัวอย่างเท่ากับ 20 แต่จะเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น และมีร้อยละของตัวแบบ Underspecification ลดลงอย่างชัดเจน

คำสำคัญ : การคัดเลือกตัวแปร การค้นหาแบบต้องห้าม วิธีการถดถอยแบบขั้นตอน วิธีการถดถอยแบบริดจ์ สหสัมพันธ์เชิงเส้นพหุ

***Corresponding author.** E-mail: nisachon94@hotmail.com

Abstract

The purpose of this study is to compare variable selection methods for multiple linear regression models that have both relevant and irrelevant variables in full model when one pair of relevant variables has a high correlation coefficient. The variables are selected by the stepwise regression method with the multiple regression coefficients are estimated by the method of Ordinary Least Square (OLS) and Ridge Regression by Hoerl, Kennard and Baldwin, Lawless and Wang, Nomura and Khalaf and Shukur methods. The variables are again selected and the multiple regression coefficients are again estimated by the Tabu Search using two objective functions: mean squared error (MSE) and mean squared error augmented by a penalty function. The criterion of comparison is the percentage of selecting the true model. The comparisons, using simulation data, are performed with sample size 20, 60 and 100 and are repeated 500 times in each case of sample size. When the pairwise of correlation coefficientis 0.95, 0.99 and 0.999, the percentages of selecting the true model by Tabu Search using both objective functions are higher than those by other methods are rather stable for all cases of sample size except in the case of Ridge Regression using Khalaf and Shukar method. When the pairwise of correlation coefficient increases to 0.9999, the percentage of selecting the true model by Tabu Search using objective function of mean squared error augmented by a penalty function is high and quite stable, regardless of the sample size and correlation. Moreover, the Tabu Search using objective function of mean squared error augmented with a penalty function does not select any of the underspecified models and the misspecified models, only select a few overspecified models which its effects are less serious than those of the underspecified models. The percentages of selecting the true model by Tabu Search using objective function of mean squared error and by the stepwise method with OLS estimates and ridge estimates using all four methods are low when the sample size 20. But increase as the sample size increases and the percentages of selecting the underspecified models are clearly decreasing.

Keywords : Variable selection, Tabu search, Stepwise regression, Ridge regression, Multicollinearity

บทนำ

การวิเคราะห์การถดถอยเชิงเส้นพหุ เป็นตัวแบบที่ความสัมพันธ์ของตัวแปรตามขึ้นอยู่กับตัวแปรอิสระมากกว่า 1 ตัวแปร (Montgomery, Peck and Vining, 2006) เพื่ออธิบายหรือพยากรณ์ค่าของตัวแปรตาม แต่การใช้ตัวแปรอิสระมากเกินไปในสมการถดถอยทำให้ค่าพยากรณ์ที่ได้มีความคลาดเคลื่อนสูงและอาจเกิดปัญหาตัวแปรอิสระบางตัวมีพหุสัมพันธ์ (Multicollinearity) ต่อกันได้ ทำให้การประมาณค่าสัมประสิทธิ์การถดถอยที่ได้มีค่าไม่เสถียรและมีค่าความคลาดเคลื่อนสูง ซึ่งมักใช้วิธีการแก้ปัญหาโดยการตัดตัวแปรที่มีความสัมพันธ์ระหว่างกันสูงทิ้งไป หรือเก็บรวบรวมข้อมูลเพิ่มขึ้น เพื่อลดระดับความสัมพันธ์ระหว่างตัวแปรลง หรืออาจใช้วิธีการวิเคราะห์การถดถอยแบบริดจ์ (Ridge Regression) วิธีนี้ไม่จำเป็นต้องตัดตัวแปรอิสระที่มีความสัมพันธ์กันสูงออกจากตัวแบบ หลักการของการถดถอยแบบริดจ์คือ การทำให้ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นพหุมีค่าเพิ่มขึ้นเล็กน้อย เพื่อทำให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำลง เนื่องจากค่าประมาณความคลาดเคลื่อนของสัมประสิทธิ์การถดถอยเป็นฟังก์ชันของ $(X'X)^{-1}$ ดังนั้นการลดค่าความคลาดเคลื่อนดังกล่าวให้ต่ำลงจึงต้องพยายามลดค่า $(X'X)^{-1}$ ให้ต่ำลง ซึ่งจะทำได้โดยการบวกค่าคงตัวที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุมจะได้ตัวประมาณค่าสัมประสิทธิ์การถดถอยแบบริดจ์เท่ากับ $\hat{\beta}_r = (X'X + rI_p)^{-1} X'y$, $r \geq 0$ ซึ่งมีสมบัติที่มีความเอนเอียง (bias) นอกจากนี้ในการวิเคราะห์จะต้องมีการประมาณค่าคงตัว r เพื่อทำให้ค่า $(X'X)^{-1}$ มีความเสถียรมากขึ้น วิธีการประมาณค่าคงตัว r มีหลายวิธี เช่น วิธีของโฮเอิร์ล, เคนนาร์ด และ บาลด์วิน (Hoerl, Kennard and Baldwin, 1975) วิธีของแมคโดนัลและกาลาร์นัว (McDonald and Galarneau,1975) และวิธีของคาลาฟและชูเกอร์ (Khalaf and Shukur, 2005) เป็นต้น Dean W. Wichernและ Gilbert A. Churchill (1978) ได้ศึกษาเปรียบเทียบตัวประมาณค่าสัมประสิทธิ์การถดถอยแบบริดจ์ ด้วยวิธีของโฮเอิร์ลและเคนนาร์ด (Hoerl and Kennard), ลอว์เลส และแวง (Lawless and Wang, 1976), โฮเอิร์ล, เคนนาร์ดและบาลด์วิน แมคโดนัลและกาลาร์นัว และเคอร์รี่และมายเยอร์ (Khuriand Myers), นุสรา สถิตโพธิ์ศรี (2535) เสนอการเปรียบเทียบตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุ ในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระโดยวิธีการถดถอยแบบริดจ์ และวิธีการถดถอยแบบลาเท้นรูทธ์ (Latent Root Regression) และ ธันยากร ต้นชลขันธ์ (2538) เสนอการเปรียบเทียบการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุโดยใช้วิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบริดจ์และวิธีที่ใช้หลักการของริดจ์และสไตน์ ในกรณีเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ เป็นต้น ในทุกกรณีที่ทำการศึกษา การจำลองตัวแปรอิสระในตัวแบบมีเฉพาะตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามเท่านั้น

นอกจากปัญหาการใช้ตัวแปรอิสระมากเกินไปในตัวแบบแล้ว ในทางตรงกันข้าม หากตัวแบบขาดตัวแปรอิสระที่สำคัญไปจะทำให้ค่าพยากรณ์ที่ได้มีความคลาดเคลื่อนสูงและเป็นปัญหาที่รุนแรงมากกว่าการมีตัวแปรที่ไม่สำคัญอยู่ในตัวแบบ โดยทั่วไปความรู้ในการบ่งชี้ตัวแปรอิสระในตัวแบบมีความจำกัดเนื่องจากความไม่สมบูรณ์และครบถ้วนในความรู้และข้อมูลที่นำมาศึกษา ดังนั้นวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบจึงมีความสำคัญมากในการอธิบายหรือพยากรณ์ค่าของตัวแปรตามในตัวแบบ ซึ่งวิธีที่มีอยู่ในโปรแกรมสำเร็จรูปทั่วไปและในโปรแกรมสำเร็จรูป SAS (Games and Lerner, 1981) มีหลายวิธี กล่าวคือ วิธีพิจารณาทุกตัวแบบที่เป็นไปได้ (All Possible Subsets) วิธีเลือกตัวแปรอิสระแบบไปข้างหน้า (Forward Selection) วิธีตัดตัวแปรอิสระออกแบบถอยหลัง (Backward Elimination) วิธีการถดถอยแบบขั้นตอน (Stepwise Regression) (Montgomery, Peck and Vining, 2006) วิธีการปรับปรุงค่า R^2 สูงสุด (MAXR: Maximum R^2 Improvement) วิธีการปรับปรุงค่า R^2 ต่ำสุด (MINR: Minimum R^2 Improvement) วิธีการคัดเลือกเหล่านี้เป็นวิธีการที่มีอยู่ในโปรแกรมสำเร็จรูปทางสถิติ นอกจากนี้ยังมีวิธีการหาค่าอุตตมะโดยการจัดกลุ่ม (Combinatorial) ที่สามารถใช้คัดเลือกตัวแปรอิสระที่เหมาะสมได้เช่นกัน แต่ยังไม่มีให้ใช้ในโปรแกรมสำเร็จรูปทางสถิติที่ใช้ในการวิเคราะห์การถดถอย เช่น วิธีการค้นหาแบบต้องห้าม (Tabu Search) เป็นวิธีแก้ปัญหาเพื่อหาค่าเหมาะสมที่สุด โดยใช้ข้อจำกัดและเวลาในการประมวลผลน้อยที่สุด (Glover, 1990) เกณฑ์ที่ใช้หยุดกระบวนการทำงาน คือ จำนวนรอบสูงสุดและการลดลงของค่าฟังก์ชันเป้าหมาย (Objective Function) วิธีนี้ใช้ระยะเวลาน้อยในการคัดเลือกตัวแปรเข้าสู่ตัวแบบและได้ตัวแบบที่มีความคลาดเคลื่อนในการพยากรณ์ต่ำ หรือกล่าวได้ว่าเป็นวิธีที่ช่วยในการค้นหาตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามในตัวแบบเพื่อลดปัญหาการมีจำนวนตัวแปรอิสระในตัวแบบมากเกินไป (Overspecification) หรือน้อยเกินไป (Underspecification) วิธีการค้นหาแบบต้องห้าม มีผู้สนใจนำมาใช้ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ เช่น Drezner and George (1999) ได้เสนอวิธีการคัดเลือกตัวแปรอิสระในการวิเคราะห์การถดถอยเชิงเส้นพหุโดยวิธีการค้นหาแบบต้องห้าม เปรียบเทียบกับวิธีการถดถอยแบบขั้นตอนและวิธีการปรับปรุงค่า R^2 สูงสุด บุษยา ปภาพจน์ (2548) ได้ทำการศึกษาเปรียบเทียบวิธีการคัดเลือกตัวแบบที่เหมาะสมที่สุด

ในการวิเคราะห์การถดถอยเชิงเส้นพหุโดยเปรียบเทียบการคัดเลือกตัวแบบ 4 วิธีคือ วิธีการค้นหาแบบต้องห้าม วิธีการถดถอยแบบขั้นตอน วิธีเลือกตัวแปรอิสระแบบไปข้างหน้า และวิธีการปรับปรุงค่า R^2 สูงสุด โดยทั้ง Drezner and George และบุษยา ศึกษาเฉพาะกรณีที่ไม่มีสหสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ และกานต์ณัฐ ณ บางช้าง (2554) ได้ทำการศึกษาเปรียบเทียบการคัดเลือกตัวแปรอิสระ โดยใช้วิธีการค้นหาแบบต้องห้ามที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ (MAE) กับวิธีการถดถอยแบบขั้นตอน โดยแบ่งตัวแปรอิสระที่ทำการคัดเลือกออกเป็น 2 กลุ่มคือ ตัวแปรที่เกี่ยวข้องและไม่เกี่ยวข้องกับตัวแปรตาม การประมาณค่าพารามิเตอร์ใช้วิธีกำลังสองน้อยที่สุดในกรณีที่มีและไม่มีสหสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระในกลุ่มที่ 1 ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระมีค่าระหว่าง 0.95-0.99

วัตถุประสงค์

1) เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยเชิงเส้นพหุโดยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์

2) เพื่อเปรียบเทียบความถูกต้องในการคัดเลือกตัวแปรอิสระที่มีพหุสัมพันธ์กันสูงเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุโดยใช้วิธีการค้นหาแบบต้องห้าม (Tabu Search) ที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษกับวิธีการถดถอยแบบขั้นตอนที่มีการประมาณค่าพารามิเตอร์แบบกำลังสองน้อยที่สุดและแบบริดจ์

ทฤษฎีที่เกี่ยวข้อง

ตัวแบบและฐานคติของการถดถอยเชิงเส้นพหุ

ตัวแบบการถดถอยเชิงเส้นพหุ $y = X\beta + \varepsilon$ (1)

โดยที่ y เป็น เวกเตอร์ค่าสังเกตของตัวแปรตาม y ขนาด $n \times 1$

X เป็น เมทริกซ์ของตัวแปรอิสระขนาด $n \times (k + 1)$

β เป็น เวกเตอร์ของพารามิเตอร์ของตัวแบบขนาด $(k + 1) \times 1$

ε เป็น เวกเตอร์ของความคลาดเคลื่อนสุ่มขนาด $n \times 1$

ฐานคติของตัวแบบการถดถอยเชิงเส้นพหุ

ความคลาดเคลื่อนสุ่มมีข้อกำหนดดังนี้ $\varepsilon \sim N(0, \sigma^2 I)$ เมื่อ I เป็นเมทริกซ์เอกลักษณ์ขนาด $n \times n$

n เป็นขนาดตัวอย่างและ k เป็นจำนวนตัวแปรอิสระในตัวแบบ

ตัวแบบการถดถอยสามารถเขียนในรูปของคาโนนิคอล (Canonical Form) ได้ดังนี้

$$y = Z\alpha + \varepsilon \quad (2)$$

เมื่อ $Z = XT$, $\alpha = T'\beta$, T เป็นเมทริกซ์เชิงตั้งฉากขนาด $p \times p$ สมาชิกในสดมภ์ของเมทริกซ์เป็นเวกเตอร์เฉพาะของ $X'X$ ที่สอดคล้องกับค่าเฉพาะ λ และ $Z'Z = \Lambda$ และ Λ เป็นเมทริกซ์ทแยงมุมขนาด $p \times p$ ที่มีสมาชิกเป็นค่าเฉพาะของ $X'X$

จากตัวแบบการถดถอยพหุใน (1) ตัวประมาณค่าพารามิเตอร์เท่ากับ

$$\hat{\beta} = (X'X)^{-1}X'y \quad (3)$$

ตัวประมาณค่าพารามิเตอร์ของตัวแบบในรูปคาโนนิคอลเขียนได้เป็น

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \quad (4)$$

ตัวประมาณค่าการถดถอยเชิงเส้นพหุจากวิธีการถดถอยแบบริดจ์ เขียนได้เป็น

$$\hat{\beta}_r = (X'X + rI)^{-1}X'y \quad (5)$$

โดยใช้วิธีการประมาณค่าคงตัว r 4 วิธีดังนี้

วิธีโฮเอิร์ล, เคนนาร์ด และ บาลด์วิน

$$r_{HKB} = \frac{p\sigma^2}{\hat{\alpha}'\hat{\alpha}} \quad (6)$$

เมื่อ $\hat{\sigma}^2 = \dfrac{y'y - \hat{\alpha}'T'X'y}{n - p}$, $\hat{\alpha} = T'\hat{\beta}$, $\hat{\alpha}_i$ เป็นสมาชิกตัวที่ i ของเวกเตอร์ $\hat{\alpha}$, λ_i เป็นสมาชิกตัวที่ i ของเมทริกซ์ Λ, t_m เป็นค่าลักษณะเฉพาะที่มีค่ามากที่สุด (The Largest Eigenvalue) ของเมทริกซ์ $X'X$, $\hat{\alpha}_{max} = \max(\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_p)$ และ p เป็นจำนวนพารามิเตอร์ที่ต้องประมาณค่าในตัวแบบ

วิธีลอว์เลสและแวง

$$r_{LW} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2} \quad (7)$$

วิธีนอมูระ (Nomura, 1988)

$$r_N = p\hat{\sigma}^2 / \sum_{i=1}^{p}[\hat{\alpha}_i^2/\{1 + (1 + \lambda_i(\hat{\alpha}_i^2/\hat{\sigma}^2)^{1/2})\}] \quad (8)$$

วิธีคาลาฟและชูเกอร์

$$r_{KS} = \frac{t_m\hat{\sigma}^2}{t_m\hat{\alpha}_{max}^2 + (n - p)\hat{\sigma}^2} \quad (9)$$

วิธีการคัดเลือกตัวแปรอิสระ

วิธีการถดถอยแบบขั้นตอน กล่าวคือมีการเลือกตัวแปรอิสระเข้าในตัวแบบการถดถอยครั้งละหนึ่งตัวแปร ตัวแปรอิสระใดที่ถูกเลือกเข้าอยู่ในตัวแบบการถดถอยแล้ว อาจถูกนำออกไปได้ภายหลัง หากพบว่าตัวแปรอิสระนั้นไม่มีนัยสำคัญ

การค้นหาแบบต้องห้ามเป็นวิธีการหาคำตอบแบบมีเหตุผล (Metaheuristic) วิธีหนึ่ง โดยวิธีดังกล่าวถูกดัดแปลงมาจากวิธีการทางปัญญาประดิษฐ์ (Artificial Intelligence) ให้มีความเหมาะสมในการใช้คำตอบเดิมในการตอบปัญหาเดิมที่มีวัตถุประสงค์เปลี่ยนแปลงไป และมีการนำมาใช้ในการหาคำตอบที่ใกล้เคียงกับค่าอุตมะ (Optimum)(Glover, 1990) ซึ่งนิยมใช้กับปัญหาในการตัดสินใจ เช่น ปัญหาการหาเส้นทางเดินของพนักงานขาย (Travelling Salesman Problem) ปัญหาการจัดช่องทางในการให้บริการในร้านสะดวกซื้อ ปัญหาการจัดตารางเวลาการทำงานในโรงงาน (Job Shop Scheduling) เป็นต้น ขั้นตอนของวิธีการนี้ไม่ซับซ้อนและผลลัพธ์ที่ได้มีประสิทธิภาพสูง แนวคิดของวิธีการนี้ ใช้การจดจำจากรอบการทำซ้ำที่ผ่านมา โดยใช้หน่วยความจำในเครื่องคอมพิวเตอร์มาช่วยในการประมวลผล นอกจากนี้ ยังสามารถใช้ในการหาคำตอบของปัญหาที่มีความซับซ้อนและมีตัวแปรที่ใช้พิจารณาเป็นจำนวนมาก ช่วยป้องกันการเกิดปัญหาคำตอบที่มีค่าต่ำสุดเฉพาะที่ (Local Minimum) โดยแก้ไขการย้อนกลับไปหาคำตอบเดิมที่อาจจะมีค่ามากกว่าคำตอบที่มีค่าต่ำสุดเฉพาะที่และดำเนินการค้นหาคำตอบต่อไปจนกระทั่งได้คำตอบที่ใกล้เคียงกับค่าต่ำสุดทั่วไป (Global Minimum) หรือค้นหาคำตอบต่อไปอีกระยะหนึ่ง (อาทิเช่น Fred Glover กำหนดจำนวนรอบการค้นหาเป็น 30 รอบ) และไม่พบคำตอบที่มีค่าต่ำกว่าค่าต่ำสุดทั่วไปที่ได้ค้นพบแล้ว จึงจะหยุดกระบวนการการค้นหา รูปแบบการค้นหาคำตอบประกอบด้วยการค้นหาที่สำคัญ 2 รูปแบบ คือ การค้นหาคำตอบโดยใช้หน่วยความจำระยะสั้น (Short Term Memory) ซึ่งเป็นหน่วยความจำตามเวลา (Recency Base Memory) หรือการค้นหาที่จดจำอดีตในการค้นหาที่ผ่านมาในระยะสั้น และการค้นหาคำตอบโดยการใช้หน่วยความจำระยะยาว (Long Term Memory) ซึ่งถือเป็นหน่วยความจำตามความถี่ (Frequency Base Memory) หมายถึง การค้นหาที่จดจำอดีตเพื่อช่วยให้การค้นหาคำตอบเป็นไปอย่างมีประสิทธิภาพมากขึ้น

ขั้นตอนในการทำงานของวิธีการค้นหาแบบต้องห้ามสรุปได้ดังนี้

1) เริ่มจากการหาคำตอบเริ่มต้น โดยมีการจัดทำโครงสร้างหน่วยความจำมาใช้ในการเก็บค่า

2) สร้างคำตอบและเลือกคำตอบจากคำตอบที่สร้างขึ้นที่มีความเหมาะสมที่สุดโดยจะหยุดการค้นหาคำตอบตามเงื่อนไขที่กำหนดไว้

3) ปรับปรุง (Update) คำตอบโดยการแทนที่คำตอบที่ดีที่สุดในปัจจุบันด้วยคำตอบที่ดีกว่า จากนั้นจึงเก็บคำตอบบันทึกไว้

วิธีการค้นหาแบบต้องห้าม (Tabu Search) เป็นวิธีการแก้ปัญหาที่ไม่ซับซ้อน แต่ความสำคัญอยู่ที่การออกแบบโครงสร้างและการจัดการข้อมูลในหน่วยความจำ รวมถึงการกำหนดลักษณะ (Attribute) ของคำตอบ การคัดเลือกตัวแปรอิสระโดยวิธีการค้นหาแบบต้องห้ามมีขั้นตอนดังนี้

1. กำหนดค่าพิสัยเริ่มต้น โดยใช้การพิจารณาจากตัวประมาณค่า OLS

2. สร้างคำตอบเริ่มต้นโดยกำหนดค่าตัวแปรอิสระต่างๆ จากการสุ่มค่าในพิสัยที่กำหนด พิสัยของตัวแปรอิสระแต่ละตัวอาจแตกต่างกันได้ตามการประมาณการเริ่มต้น

3. สร้างเซตคำตอบข้างเคียง (Neighborhood) ของค่าคำตอบปัจจุบันที่กำหนดขึ้น

4. การตรวจสอบรายการต้องห้าม (Tabu List) ซึ่งเป็นรายการที่บันทึกข้อมูลกระบวนการค้นหาในอดีต ถ้าคำตอบข้างเคียงนั้นเป็นรายการต้องห้ามเกินระยะเวลาที่กำหนด ให้ยกเลิกการเป็นรายการต้องห้ามของคำตอบข้างเคียงนั้น

5. ดำเนินการค้นหาคำตอบจากเซตคำตอบข้างเคียง ซึ่งต้องไม่เป็นรายการต้องห้ามและเป็นคำตอบที่ดีกว่าคำตอบที่มีค่าฟังก์ชันเป้าหมายดีที่สุดจากเซตคำตอบข้างเคียง

6. กำหนดเกณฑ์ความปรารถนา (Aspiration Criteria) ในการตรวจสอบคำตอบข้างเคียง ถ้าคำตอบข้างเคียงนั้นเป็นรายการต้องห้าม แต่สามารถให้คำตอบที่มีค่าฟังก์ชันเป้าหมายที่ดีกว่าก็สามารถเลือกมาเป็นคำตอบได้

7. ถ้าไม่สามารถหาคำตอบที่มีค่าฟังก์ชันเป้าหมายดีกว่าปัจจุบันได้ ให้เรียกใช้หน่วยความจำระยะยาว (Long Term Memory) มาใช้ในการหาคำตอบเริ่มต้นใหม่

8. ย้ายตำแหน่ง (Move) ไปยังคำตอบปัจจุบันที่ถูกเลือกแล้วปรับปรุงรายการต้องห้าม (Tabu List) ให้เป็นปัจจุบันและเพิ่มจำนวนครั้งในการเลือกข้อมูลความถี่ (Frequency Move) เป็นการใช้หน่วยความจำระยะยาวบันทึกการค้นหาทุกค่าตลอดช่วงของการค้นหา โดยใช้หลักการสร้างความหลากหลาย (Diversification) เพื่อค้นหาคำตอบในบริเวณที่แตกต่างจากที่ค้นพบมาแล้ว

9. ทำซ้ำจนกระทั่งได้ค่าฟังก์ชันเป้าหมายอยู่ในเกณฑ์ที่กำหนดขึ้นหรือครบตามจำนวนครั้งที่กำหนด

อนึ่งการค้นหาแบบต้องห้ามไม่ได้ใช้การประมาณค่าสัมประสิทธิ์การถดถอย (β) แบบ OLS หรือ ฟังก์ชันภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) แต่การค้นหาแบบต้องห้ามจะค้นหาค่าของสัมประสิทธิ์การถดถอย (β) ที่ทำให้ฟังก์ชันเป้าหมายมีค่าต่ำสุด โดยไม่ต้องคำนวณค่า $(X'X)^{-1}$ ดังนั้นการค้นหาแบบต้องห้ามจึงสามารถใช้ได้กับข้อมูลทั้งที่ตัวแปรอิสระไม่มีและมีพหุสัมพันธ์ (Multicollinearity) โดยทั่วไปฟังก์ชันเป้าหมายคือค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (F_1)

$$F_1 = \frac{(y - \hat{y})'(y - \hat{y})}{n - k - 1} \qquad (10)$$

โดยค้นหาค่า $\hat{\beta}$ ที่ทำให้ F_1 มีค่าต่ำสุด

อีกแนวคิดหนึ่งที่นำมาใช้ในการศึกษานี้คือนำ $\hat{\beta}'\hat{\beta}$ มาพิจารณาด้วยในการหาค่าต่ำสุดของ F_1 กล่าวคือต้องการให้ F_1 มีค่าน้อยในขณะเดียวกันกับค่า $\hat{\beta}'\hat{\beta}$ ต้องไม่มากด้วย จึงนำมาสู่ฟังก์ชันเป้าหมาย F_2 ซึ่งเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ

$$F_2 = \frac{(y - \hat{y})'(y - \hat{y})}{n - k - 1} + \frac{c\hat{\beta}'\hat{\beta}}{n - p} \qquad (11)$$

เมื่อ $0 < c \leq 1$ เป็นค่าคงตัวที่ต้องค้นหาค่าที่เหมาะสมพร้อมๆ กับค่า $\hat{\beta}$ เพื่อทำให้ F_2 มีค่าต่ำสุด

y เป็นเวกเตอร์ค่าสังเกต, $\hat{\beta}$ เป็นเวกเตอร์ของค่าสัมประสิทธิ์การถดถอยที่ประมาณจากวิธีการค้นหาแบบต้องห้าม, $\hat{y}$ เป็นเวกเตอร์ค่าพยากรณ์ที่ได้มาจากสมการถดถอยที่ประมาณค่าพารามิเตอร์ของตัวแบบโดยวิธีการค้นหาแบบต้องห้าม, n เป็นจำนวนค่าสังเกต และ $p = k + 1$ เป็นจำนวนพารามิเตอร์ที่ต้องประมาณค่าในตัวแบบ

การคัดเลือกตัวแปรอิสระโดยใช้วิธีการค้นหาแบบต้องห้ามใช้วิธีประมาณค่าพารามิเตอร์แบบต้องห้ามและค้นหาตัวแบบเพื่อทำให้ฟังก์ชัน F_1 และ F_2 มีค่าต่ำสุด

นิยามคำศัพท์ที่เกี่ยวข้องดังนี้

ตัวแบบเต็มรูป (Full Model) หมายถึง ตัวแบบการถดถอยเชิงเส้นพหุที่ประกอบด้วยตัวแปรอิสระทั้งหมดที่พิจารณา ประกอบด้วยตัวแปรอิสระจำนวนหนึ่งที่เกี่ยวข้องกับตัวแปรตามโดยมีทฤษฎีที่เกี่ยวข้องสนับสนุนและหรือการศึกษาในอดีตที่ผ่านมาและตัวแปรที่ไม่เกี่ยวข้องกับตัวแปรตามจำนวนหนึ่ง

ตัวแบบ Overspecification หมายถึง ตัวแบบที่ประกอบด้วยตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามครบทุกตัว และยังมีตัวแปรอิสระบางตัวที่ไม่เกี่ยวข้องกับตัวแปรตามรวมอยู่ในตัวแบบ

ตัวแบบ Underspecification หมายถึง ตัวแบบที่ไม่มีตัวแปรอิสระที่ไม่เกี่ยวข้องอยู่ในตัวแบบ แต่มีตัวแปรอิสระบางตัวที่เกี่ยวข้องกับตัวแปรตามขาดหายไปจากตัวแบบ

ตัวแบบ Misspecification หมายถึง ตัวแบบที่มีตัวแปรอิสระบางตัวแปรที่เกี่ยวข้องกับตัวแปรตามขาดหายไปจากตัวแบบและมีตัวแปรอิสระบางตัวแปรที่ไม่เกี่ยวข้องกับตัวแปรตามรวมอยู่ในตัวแบบ

ตัวแบบจริง (True Model) หมายถึง ตัวแบบการถดถอยเชิงเส้นพหุที่มีตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามครบทุกตัว

วิธีการวิจัย

ขั้นตอนในการดำเนินงานวิจัย

งานวิจัยนี้ ศึกษากรณีตัวแปรอิสระมีการแจกแจงเอกรูป (Uniform Distribution) และความคลาดเคลื่อนมีการแจกแจงปรกติ (Normal Distribution) กำหนดจำนวนตัวแปรอิสระมีการแจกแจงเอกรูป 7 ตัว ซึ่งแยกเป็น 2 กลุ่ม กลุ่มที่ 1 คือตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามจำนวน 5 ตัว ตัวแปรเหล่านี้มีทฤษฎีบทหรือผลการศึกษาในอดีตสนับสนุนว่ามีความสำคัญกับตัวแปรตามและมีตัวแปรอิสระ 1 คู่ ในกลุ่มนี้มีความสัมพันธ์กันสูงในระดับ 0.95-0.9999 ตัวแปรอิสระกลุ่มที่ 2 เป็นตัวแปรที่ไม่มีทฤษฎีบทหรือผลการศึกษาในอดีตสนับสนุนและไม่แน่ใจว่าตัวแปรอิสระเหล่านี้ควรอยู่ในตัวแบบหรือไม่มี จำนวน 2 ตัว การวิจัยใช้การจำลองแบบมอนติคาร์โล (Monte Carlo Simulation) มีขนาดตัวอย่างเท่ากับ 20, 60 และ 100 โดยทำซ้ำ จำนวน 500 ครั้งในแต่ละสถานการณ์ การจำลองข้อมูล วิเคราะห์ข้อมูลและคัดเลือกตัวแบบใช้โปรแกรม Matlab Version 2011a และการคัดเลือกตัวแบบโดยวิธีการค้นหาแบบต้องห้ามดัดแปลงมาจาก “Ts Directed by direct search method for nonlinear global optimization” (Hedarand Fukushima, 2003: 329-349) ขั้นตอนการดำเนินงานดังนี้

1) สร้างประชากร N = 200,000 จำนวน 2 กลุ่ม แต่ละกลุ่มประกอบด้วย $X_1, X_2, X_3, X_4, X_5, X_6, X_7$

2) การสร้างตัวแปรอิสระ X_1 และ X_3 ที่มีความสัมพันธ์กันโดยใช้วิธีรากที่สอง (square root method) โดยจำลองตัวแปรช่วย U_1 และ U_2 ที่มีการแจกแจงเอกรูป ที่เป็นอิสระต่อกันโดย $U_i \sim (a,b)$ และ X_i เป็นตัวแปรอิสระที่ $E(X_i) = \frac{a+b}{2}$, $\sigma_i^2 = Var(X_i) = \frac{(b+a)^2}{12}$

เมื่อคำนวณค่าของตัวแปร $X_1 = \mu_1 + \sigma_1 U_1$ โดย μ_1 และ σ_1 คือ ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของตัวแปร X_1 ตามลำดับ คำนวณ $X_3 = \mu_3 + \sigma_3(\rho U_1 + \sqrt{1 - \rho_{13}^2} U_2)$ โดย μ_3 และ σ_3 คือ ค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของตัวแปร ประชากรที่ 1 กำหนดค่าสัมประสิทธิ์สหสัมพันธ์ $\rho_{13} = 0.999$ ประชากรที่ 2 กำหนดค่าสัมประสิทธิ์สหสัมพันธ์ $\rho_{13} = 0.9999$

3) กำหนดให้ตัวแปรอิสระมีการแจกแจงเอกรูป โดย $X_1 \sim U(20,100)$, $X_2 \sim U(20,120)$, $X_3 \sim U(5,90)$, $X_4 \sim U(-30,30)$, $X_5 \sim U(-60,60)$, $X_6 \sim U(-20,40)$, และ $X_7 \sim U(-50,50)$

4) สร้างความคลาดเคลื่อนที่เป็นอิสระต่อกันที่มีการแจกแจงปรกติ $\varepsilon \sim N(0,30)$

5) ตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตาม y คือ X_1, X_2, X_3, X_4, และ X_5 และตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตาม y คือ X_6, X_7 กำหนดค่าสัมประสิทธิ์การถดถอยดังนี้

$\beta' = [30\ 20\ 10 - 5\ 15\ 4\ 0\ 0]$

6) สร้างตัวแปรตาม y ภายใต้ตัวแบบเต็มรูป

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$

7) เมื่อได้ประชากรตามที่ต้องการ ดำเนินการสุ่มตัวอย่างขนาด 20, 60 และ 100 แล้วคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_1 และ X_3 ในแต่ละตัวอย่างที่สุ่มได้

8) เนื่องจากปัญหาสหสัมพันธ์เชิงเส้นพหุ เป็นปัญหาที่พบบ่อยในตัวอย่างมากกว่าในประชากร เพื่อให้แน่ใจว่าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าสูงจริง เนื่องจากตัวอย่างสุ่มที่ได้มักมีค่าสัมประสิทธิ์สหสัมพันธ์ต่ำกว่าในประชากร ดังนั้นหากพบว่าตัวอย่างสุ่มชุดใดมีค่า $r_{13} < 0.945$ จะไม่นำมาพิจารณาและตัวอย่างที่ได้ถ้าค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_1 และ X_3 มีค่าอยู่ระหว่าง 0.945-0.954 จัดอยู่ในกลุ่ม $r_{13} = 0.95$ ถ้า r_{13} มีค่าอยู่ระหว่าง 0.985-0.994 จัดอยู่ในกลุ่ม $r_{13} = 0.99$ ถ้า r_{13} มีค่าอยู่ระหว่าง 0.9985-0.9994 จัดอยู่ในกลุ่ม $r_{13} = 0.999$ ถ้า r_{13} มีค่าอยู่ระหว่าง 0.99985-0.99994 จัดอยู่ในกลุ่ม $r_{13} = 0.9999$ ตามลำดับ

9) ทำการคัดเลือกตัวแบบจริง เพื่อให้ได้ตัวแบบที่มีตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตามครบทุกตัว ในการศึกษาครั้งนี้ตัวแบบจริงคือ $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ โดยใช้วิธีการคัดเลือกตัวแบบ 2 วิธี คือ วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด และวิธีการถดถอยแบบริดจ์ ที่ใช้วิธีการประมาณค่าคงตัว r 4 วิธีคือ วิธีโฮเอิร์ล เคนนาร์ด และ บาลด์วิน วิธีลอว์เลสและแวง วิธีนอมูระและวิธีคาลาฟและชูเกอร์ โดยกำหนดระดับนัยสำคัญในการเลือกตัวแปรอิสระเข้าและออกจากตัวแบบเท่ากับ 0.05 กำหนดพิสัยของค่าพารามิเตอร์เริ่มต้นจากวิธีกำลังสองน้อยที่สุดโดยกำหนดดังนี้

$b_0 = (10,50)$, $b_1 = (0,40)$, $b_2 = (0,20)$, $b_3 = (-10,0)$, $b_4 = (0,50)$, $b_5 = (0,10)$, $b_6 = (-5,5)$, $b_7 = (-10,10)$ และกำหนดระยะที่อยู่ในรายการต้องห้ามเท่ากับ 10 รอบ การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ โดยการทดสอบสมมติฐาน $H_0 : \beta_i = 0$ เทียบกับ $H_1 : \beta_i \neq 0$; $i = 1,2,...,7$ ใช้สถิติทดสอบทีดังนี้

$t_0 = \dfrac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$ โดยที่ $SE(\beta_i) = \sqrt{\dfrac{\sum_{i=1}^{500}(\beta_i - \bar{\beta}_i)^2}{500}}$ จะปฏิเสธสมมติฐาน H_0 เมื่อ $|t_0| > t_{\frac{\alpha}{2},n-k-1}$ ถ้าไม่ปฏิเสธสมมติฐาน H_0 แสดงว่าตัวแปรอิสระนั้นสามารถนำออกจากตัวแบบได้ จะต้องประมาณค่าพารามิเตอร์ใหม่ เพื่อหาสมการถดถอยที่เหมาะสมต่อไป แต่หากปฏิเสธ H_0 แสดงว่าตัวแปรอิสระนั้นควรอยู่ในตัวแบบทำต่อไปจนกระทั่งไม่สามารถนำตัวแปรอิสระออกจากตัวแบบได้อีก กระบวนการคัดเลือกจะหยุด

10) คำนวณเกณฑ์การตัดสินใจในการเปรียบเทียบการคัดเลือกตัวแปรอิสระ คือร้อยละของจำนวนครั้งที่แต่ละวิธีสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ตามตัวแบบจริง ตัวแบบ Overspecification ตัวแบบ Underspecification และตัวแบบ Misspecification จากการทำซ้ำจำนวน 500 ครั้งในแต่ละวิธีในแต่ละสถานการณ์

ผลการวิจัย

เมื่อขนาดตัวอย่างเท่ากับ 20 ตัวแปรอิสระ X_1 และ X_3 มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.95, 0.99 และ 0.999 วิธีการค้นหาแบบต้องห้าม มีร้อยละการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องมากที่สุด โดยทั้งสองฟังก์ชันเป้าหมายมีร้อยละการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องใกล้เคียงกันและไม่มีตัวแบบ Underspecification และ Misspecification มีเพียงตัวแบบ Overspecification เพียงเล็กน้อย ซึ่งน้อยกว่าวิธีอื่นโดยวิธีการถดถอยแบบขั้นตอน ที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด และวิธีการถดถอยแบบริดจ์ที่ประมาณค่าคงตัว r 4 วิธีคือ วิธีโฮเอิร์ล เคนนาร์ด และ บาลด์วิน วิธีลอว์เลสและแวง วิธีนอมูระ และวิธีคาลาฟและชูเกอร์ มีร้อยละการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องน้อยกว่าวิธีการค้นหาแบบต้องห้าม

และมีตัวแบบ Underspecification Overspecification และตัวแบบ Misspecification โดยมีร้อยละของตัวแบบ Underspecification มากที่สุดในทุกวิธี

กรณีตัวแปรอิสระ X_1 และ X_3 มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.9999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ มีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบถูกต้องมากที่สุดร้อยละ 98.4 และไม่มีตัวแบบ Underspecification และ Misspecification มีเพียงตัวแบบ Overspecification เพียงร้อยละ 1.6 ในขณะที่วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย มีการคัดเลือกตัวแปรอิสระได้ตัวแบบจริงเพียงร้อยละ 14.8 และวิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ที่ประมาณค่าคงตัว r 4 วิธีคือ วิธีโฮเอิร์ล เคนนาร์ด และบาลด์วิน วิธีลอว์เลสและแวง วิธีนอมูระและวิธีคาลาฟและชูเกอร์ มีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบถูกต้องเพียงร้อยละ 18.0,19.6, 18.0, 18.0 และ 2.4 ตามลำดับ และมีตัวแบบ Underspecification Overspecification และตัวแบบ Misspecification โดยมีร้อยละของตัวแบบ Underspecification มากที่สุดในทุกวิธีดูรายละเอียดได้จากตารางที่ 1 และภาพที่ 1

ในกรณีขนาดตัวอย่างเป็น 60 และ 100 และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ X_1 และ X_3 เป็น 0.95, 0.99 และ 0.999 ได้ผลการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเป็นไปในทำนองเดียวกันกับกรณีขนาดตัวอย่างเป็น 20 แต่ในกรณีที่ตัวแปรอิสระ X_1 และ X_3 มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบถูกต้องมากที่สุด ซึ่งได้แสดงรายละเอียดไว้ในตารางที่ 2-3 และ ภาพที่ 2-3

ตารางที่ 1 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามค่าสัมประสิทธิ์สหสัมพันธ์และวิธีการคัดเลือกตัวแบบ เมื่อ n=20

n = 20, r_{13} = 0.95				
วิธี	Correct.	Under.	Over.	Miss.
OLS	89.8	1.0	9.2	0.0
HKB	89.6	4.8	5.4	0.2
LW	89.8	1.0	9.2	0.0
N	87.0	4.6	8.2	0.2
KS	39.4	56.2	4.4	0.0
TABU(F1)	98.2	0.0	1.8	0.0
TABU(F2)	97.0	0.0	3.0	0.0

n = 20, r_{13} = 0.99				
วิธี	Correct.	Under.	Over.	Miss.
OLS	87.2	1.6	10.8	0.4
HKB	88.8	4.6	6.6	0.0
LW	87.2	1.6	10.8	0.4
N	84.8	4.6	10.6	0.0
KS	36.8	59.6	3.6	0.0
TABU(F1)	98.0	0.0	2.0	0.0
TABU(F2)	97.6	0.0	2.4	0.0

n = 20, r_{13} = 0.999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	81.0	7.2	10.4	1.4
HKB	86.4	7.6	5.4	0.6
LW	81.0	7.2	10.4	1.4
N	79.8	9.2	9.8	1.2
KS	31.2	64.2	3.8	0.8
TABU(F1)	94.0	4.0	1.8	0.2
TABU(F2)	97.8	0.0	2.2	0.0

n = 20, r_{13} = 0.9999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	18.0	71.6	2.6	7.8
HKB	19.6	74.0	2.8	3.6
LW	18.0	71.6	2.6	7.8
N	18.0	72.0	2.6	7.4
KS	2.4	94.4	0.6	2.6
TABU(F1)	14.8	83.8	1.2	0.2
TABU(F2)	98.4	0.0	1.6	0.0

ตารางที่ 2 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามค่าสัมประสิทธิ์สหสัมพันธ์และวิธีการคัดเลือกตัวแบบ เมื่อ n=60

n = 60, r_{13} = 0.95				
วิธี	Correct.	Under.	Over.	Miss.
OLS	89.6	0.0	10.4	0.0
HKB	93.6	0.0	6.4	0.0
LW	89.6	0.0	10.4	0.0
N	89.6	0.0	10.4	0.0
KS	56.2	37.2	6.6	0.0
TABU(F1)	97.2	0.0	2.8	0.0
TABU(F2)	97.2	0.0	2.8	0.0

n = 60, r_{13} = 0.99				
วิธี	Correct.	Under.	Over.	Miss.
OLS	89.6	0.0	10.4	0.0
HKB	93.0	0.0	7.0	0.0
LW	89.6	0.0	10.4	0.0
N	89.6	0.0	10.4	0.0
KS	59.4	33.2	7.4	0.0
TABU(F1)	96.6	0.0	3.4	0.0
TABU(F2)	96.0	0.0	4.0	0.0

n = 60, r_{13} = 0.999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	90.4	0.0	9.6	0.0
HKB	94.2	0.0	5.8	0.0
LW	90.4	0.0	9.6	0.0
N	90.4	0.0	9.6	0.0
KS	53.6	37.4	9.0	0.0
TABU(F1)	98.0	0.0	2.0	0.0
TABU(F2)	97.6	0.0	2.4	0.0

n = 60, r_{13} = 0.9999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	52.4	36.2	6.4	5.0
HKB	72.0	19.2	6.8	2.0
LW	52.4	36.2	6.4	5.0
N	52.4	36.2	6.4	5.0
KS	24.6	67.6	4.4	3.4
TABU(F1)	55.2	44.2	0.6	0.0
TABU(F2)	97.6	0.0	2.4	0.0

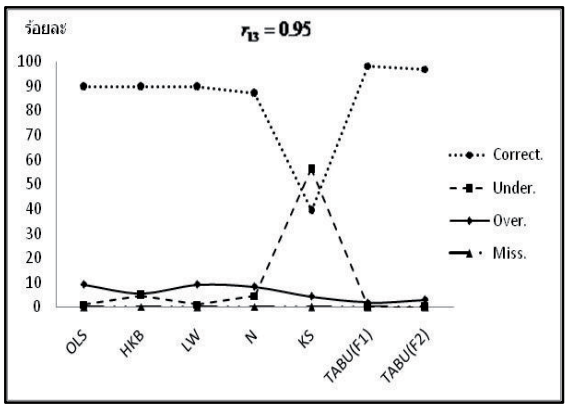
ตารางที่ 3 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามค่าสัมประสิทธิ์สหสัมพันธ์และวิธีการคัดเลือกตัวแบบ เมื่อ n=100

n = 100, r_{13} = 0.95				
วิธี	Correct.	Under.	Over.	Miss.
OLS	91.2	0.0	8.8	0.0
HKB	94.4	0.0	5.6	0.0
LW	91.2	0.0	8.8	0.0
N	91.2	0.0	8.8	0.0
KS	67.0	24.6	8.4	0.0
TABU(F1)	97.4	0.0	2.6	0.0
TABU(F2)	96.6	0.0	3.4	0.0

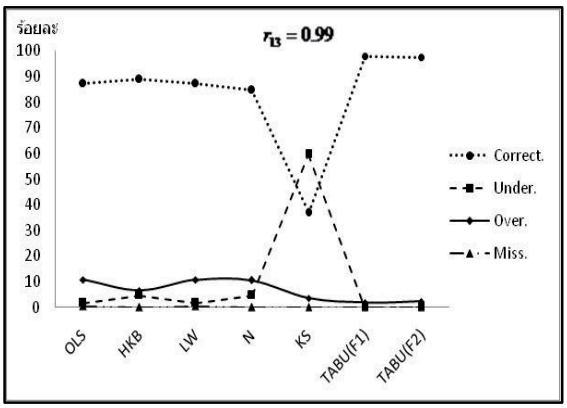
n = 100, r_{13} = 0.99				
วิธี	Correct.	Under.	Over.	Miss.
OLS	90.8	0.0	9.2	0.0
HKB	94.8	0.0	5.2	0.0
LW	90.8	0.0	9.2	0.0
N	90.8	0.0	9.2	0.0
KS	62.4	30.4	7.2	0.0
TABU(F1)	97.0	0.0	3.0	0.0
TABU(F2)	96.8	0.0	3.2	0.0

n = 100, r_{13} = 0.999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	89.8	0.0	10.2	0.0
HKB	93.4	0.0	6.6	0.0
LW	89.8	0.0	10.2	0.0
N	89.8	0.0	10.2	0.0
KS	67.4	23.4	9.2	0.0
TABU(F1)	96.6	0.0	3.4	0.0
TABU(F2)	96.4	0.0	3.6	0.0

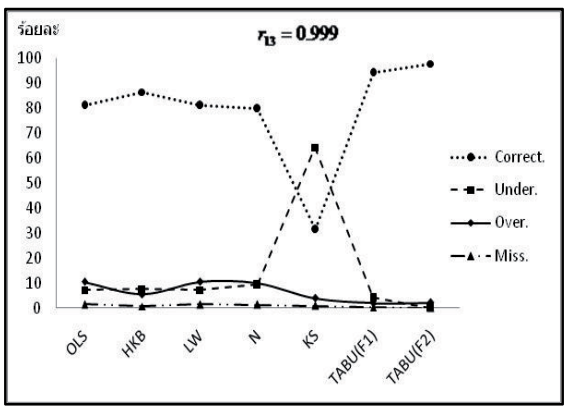
n = 100, r_{13} = 0.9999				
วิธี	Correct.	Under.	Over.	Miss.
OLS	76.0	13.6	8.4	2.0
HKB	84.0	8.0	7.4	0.6
LW	76.0	13.6	8.4	2.0
N	76.0	13.6	8.4	2.0
KS	49.6	41.4	7.2	1.8
TABU(F1)	80.0	17.8	2.0	0.2
TABU(F2)	97.4	0.0	2.6	0.0



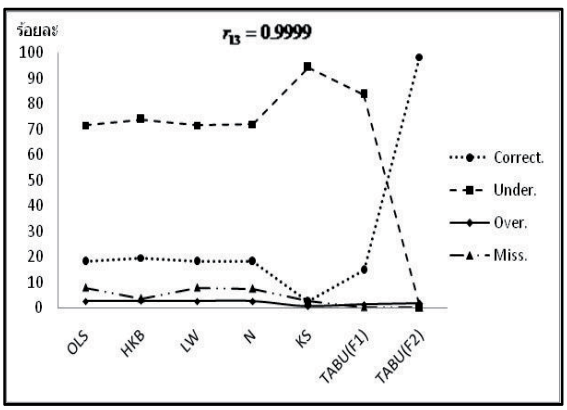
ภาพ 1ก.



ภาพ 1ข.

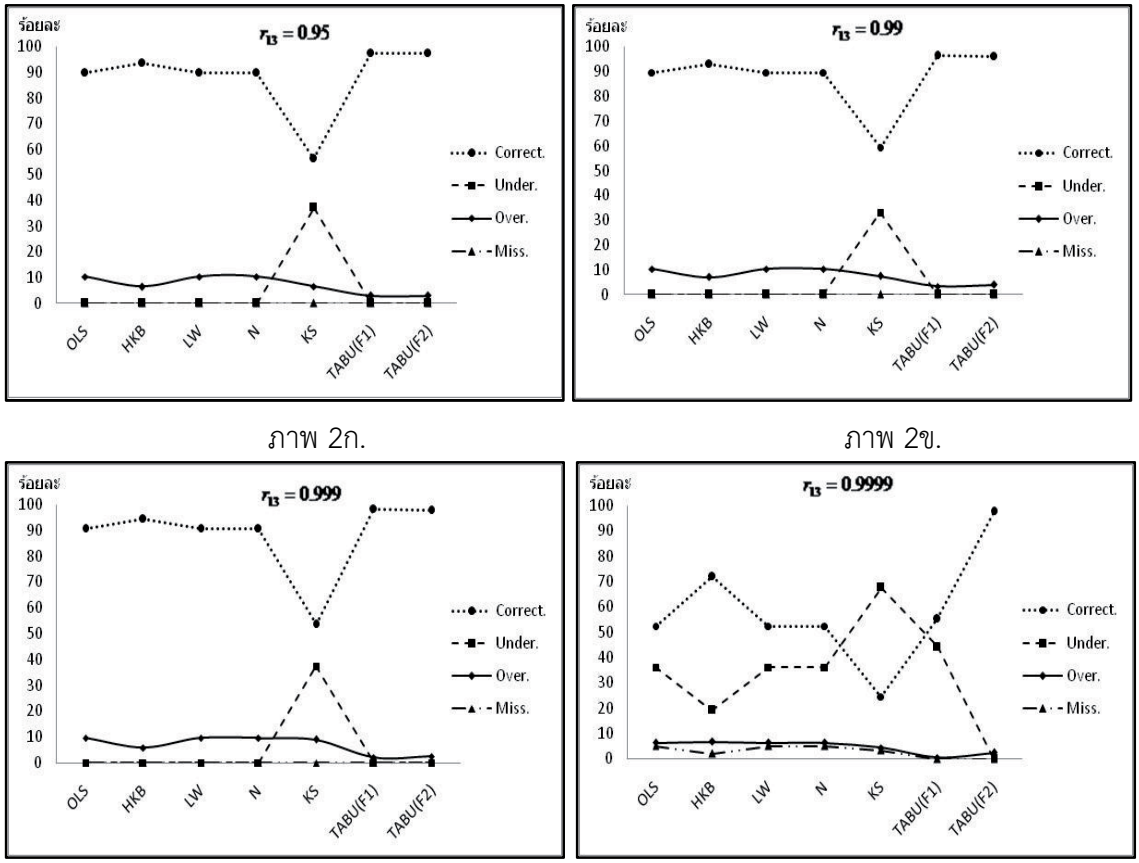


ภาพ 1ค.



ภาพ 1ง.

ภาพที่ 1 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามวิธีการคัดเลือกตัวแบบและค่าสัมประสิทธิ์สหสัมพันธ์ เมื่อ n=20



ภาพ 2ก. ภาพ 2ข.

ภาพ 2ค. ภาพ 2ง.

ภาพที่ 2 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามวิธีการคัดเลือกตัวแบบและค่าสัมประสิทธิ์สหสัมพันธ์ เมื่อ n=60

จากตารางที่ 1-3

r_{13} เป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ X_1 และ X_3

OLS เป็นการคัดเลือกตัวแปรอิสระโดยใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์ด้วยวิธีกำลังสองน้อยที่สุด

HKB เป็นการคัดเลือกตัวแปรอิสระโดยใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์ด้วยวิธีการถดถอยแบบริดจ์ โดยใช้วิธีโฮเอิร์ล เคนนาร์ด และ บาลด์วิน

LW เป็นการคัดเลือกตัวแปรอิสระโดยใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์ด้วยวิธีการถดถอยแบบริดจ์ โดยใช้วิธีลอว์เลสและแวง

N เป็นการคัดเลือกตัวแปรอิสระโดยใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์ด้วยวิธีการถดถอยแบบริดจ์ โดยใช้วิธีนอมูระ

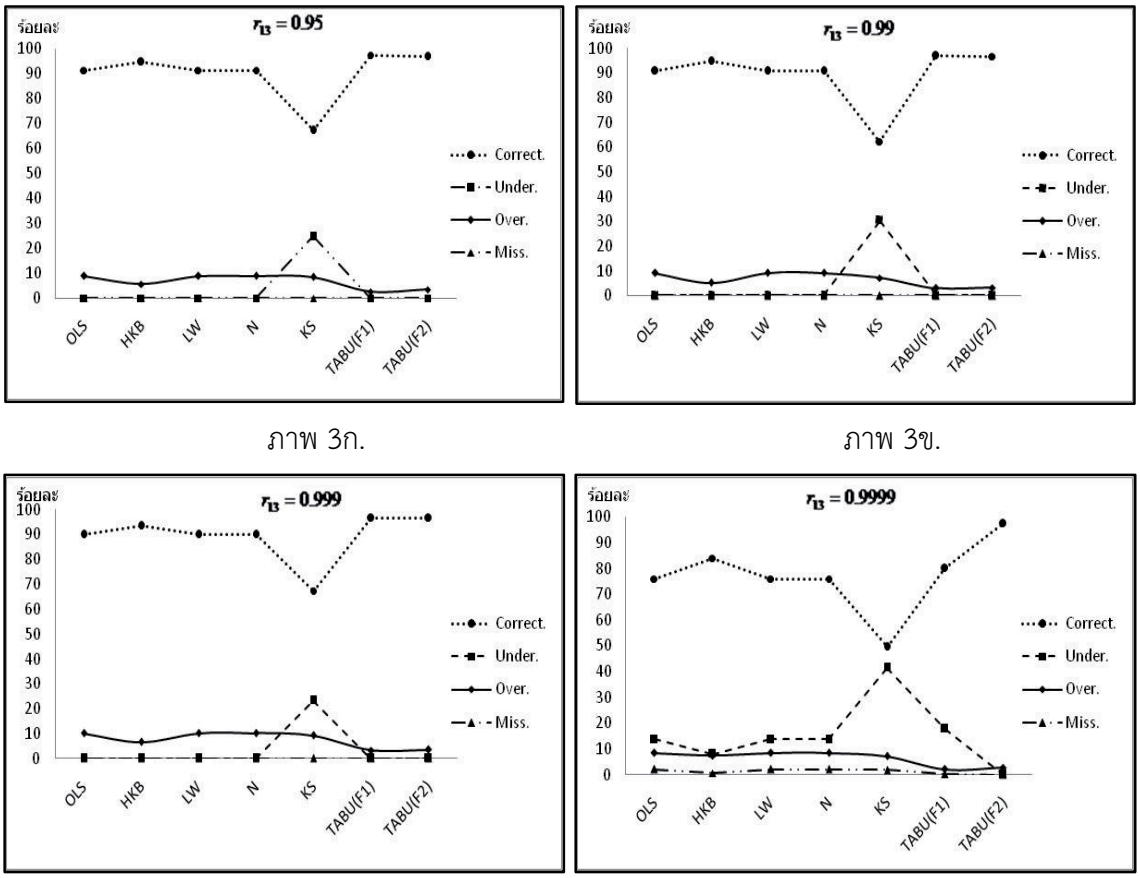
KS เป็นการคัดเลือกตัวแปรอิสระโดยใช้วิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์ด้วยวิธีการถดถอยแบบริดจ์ โดยใช้วิธีคาลาฟและชูเกอร์

TABU(F1) เป็นการคัดเลือกตัวแปรอิสระที่ประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีการค้นหาแบบต้องห้ามที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

TABU(F2) เป็นการคัดเลือกตัวแปรอิสระที่ประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีการค้นหาแบบต้องห้ามที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ

สรุปและอภิปรายผลการวิจัย

เนื่องจากการคำนวณค่าประมาณพารามิเตอร์โดยวิธีการค้นหาแบบต้องห้ามไม่ได้ใช้วิธีการหาเมทริกซ์ผกผันของ X'X จึง



ภาพ 3ก. ภาพ 3ข.

ภาพ 3ค. ภาพ 3ง.

ภาพที่ 3 ร้อยละของจำนวนครั้งที่คัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จำแนกตามวิธีการคัดเลือกตัวแบบและค่าสัมประสิทธิ์สหสัมพันธ์ เมื่อ n=100

ไม่มีปัญหาเหมือนกับการประมาณค่าพารามิเตอร์โดยใช้เมทริกซ์ผกผันของ X'X อย่างเช่นในวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ เมื่อตัวแปรอิสระมีสหสัมพันธ์เชิงเส้นพหุสูง เพราะเมื่อข้อมูลเปลี่ยนแปลงไปเล็กน้อย จะทำให้ค่าเมทริกซ์ผกผันของ X'X เปลี่ยนไปมากส่งผลให้ค่าประมาณพารามิเตอร์มีค่าไม่เสถียร จึงทำให้ค่าประมาณพารามิเตอร์เชื่อถือไม่ได้และส่งผลถึงการทดสอบสมมุติฐานด้วย จึงอาจสรุปได้ว่า ในกรณีที่ข้อมูลมีสหสัมพันธ์เชิงเส้นพหุกันสูง วิธีการค้นหาแบบต้องห้ามช่วยในการประมาณค่าสัมประสิทธิ์การถดถอยได้ใกล้เคียงกับค่าพารามิเตอร์ของตัวแบบมากกว่าวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์

กรณีตัวแปรอิสระ X_1 และ X_3 มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.95, 0.99, 0.999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษมีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องมากกว่าวิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด และวิธีการถดถอยแบบริดจ์ในทุกขนาดตัวอย่าง จากผลการวิเคราะห์ไม่พบตัวแบบ Underspecification และ Misspecification มีเพียงตัวแบบ Overspecification ซึ่งเป็นปัญหาที่รุนแรงน้อยกว่าการมีตัวแปรอิสระที่สำคัญขาดหายไปจากตัวแบบ ยกเว้นกรณีที่ขนาดตัวอย่างเป็น 20 และค่าสัมประสิทธิ์สหสัมพันธ์ 0.999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย มีตัวแบบ Underspecification ร้อยละ 4.0 และตัวแบบ Misspecification ร้อยละ 0.2 ในขณะที่การคัดเลือกตัวแบบด้วยวิธีการถดถอยแบบขั้นตอนมีร้อยละของตัวแบบ Underspecification ในทุกขนาดตัวอย่าง ซึ่งผลการศึกษาในส่วนนี้สอดคล้องกับผลการศึกษาของ กานต์ณัฐ ณ บางช้าง (2554)

ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันในระดับ 0.95-0.99 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องมากกว่าวิธีการถดถอยแบบขั้นตอน

กรณีตัวแปรอิสระ X_1 และ X_3 มีค่าสัมประสิทธิ์สหสัมพันธ์ 0.9999 วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย และวิธีการถดถอยแบบขั้นตอนที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด และวิธีการถดถอยแบบริดจ์ มีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องลดลงมาก และเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น คือมีร้อยละของตัวแบบ Underspecification ค่อนข้างสูงและน้อยลงเมื่อขนาดตัวอย่างเพิ่มขึ้น แต่วิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ มีร้อยละของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องมากที่สุด และค่อนข้างคงที่ในทุกขนาดตัวอย่างและไม่มีตัวแบบ Underspecification และ Misspecification เลย

เนื่องจาก เมื่อตัวแปรอิสระ (X_i) มีความสัมพันธ์กันสูงมาก (r_{ij} = 0.9999) ค่าประมาณพารามิเตอร์ ($\hat{\beta}_i$) โดยวิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยจะให้ค่าที่ต่างกันมาก ดังนั้นเมื่อตัวอย่างมีขนาดเล็ก (n=20) จะทำให้ค่าความคลาดเคลื่อนมาตรฐานของค่าประมาณพารามิเตอร์ ($SE(\hat{\beta}_i)$) มีค่าสูง ซึ่งส่งผลต่อการทดสอบสมมุติฐานในการคัดเลือกตัวแปรอิสระทำให้มีร้อยละของตัวแบบ Underspecification ค่อนข้างสูง แต่เมื่อขนาดตัวอย่างใหญ่ขึ้นก็ส่งผลให้การทดสอบสมมติฐานมีความถูกต้องมากขึ้น ในส่วนของวิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษ สมการ (11) มีแนวคิดที่นำ $\hat{\beta}'\hat{\beta}$ มาพิจารณาด้วยในการหาค่าต่ำสุดของฟังก์ชันเป้าหมาย กล่าวคือ ต้องการให้ F_1 สมการที่ (10) มีค่าน้อยในขณะเดียวกันกับค่า $\hat{\beta}'\hat{\beta}$ ต้องไม่มากด้วย ซึ่งเป็นการจำกัดขนาดของ $\hat{\beta}_i$ ทำให้ค่า $\hat{\beta}_i$ ค่อนข้างเสถียร ไม่สูงหรือต่ำจนเกินไป

ดังนั้น สามารถสรุปได้ว่าวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบโดยใช้วิธีการค้นหาแบบต้องห้ามที่ใช้ฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยปรับด้วยฟังก์ชันการลงโทษมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระได้ดีที่สุด กล่าวคือ สามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้ถูกต้องมากที่สุดและค่อนข้างคงที่ในทุกกรณี โดยความถูกต้องในการคัดเลือกไม่ได้ขึ้นอยู่กับขนาดตัวอย่างและระดับสหสัมพันธ์ระหว่างตัวแปรอิสระอย่างเช่นในวิธีการถดถอยแบบขั้นตอน และวิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ย แต่ข้อจำกัดของวิธีการค้นหาแบบต้องห้ามคือ เมื่อมีข้อมูลเพียงชุดเดียว การประมาณค่าความเบี่ยงเบนมาตรฐานของตัวประมาณค่าสัมประสิทธิ์การถดถอยต้องใช้วิธีการสุ่มซ้ำ (Resampling) เช่น วิธี Bootstrap Resampling วิธี Jackknife Resampling (Wu, 1986: 1261-1295) เป็นต้น

กิตติกรรมประกาศ

ขอขอบคุณ รศ. ดร. วิชิต หล่อจีระชุณห์กุล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ ที่ช่วยให้คำปรึกษาตลอดจนแนวคิดที่เป็นประโยชน์อย่างมากสำหรับการทำวิจัยเรื่องนี้

เอกสารอ้างอิง

กานต์ณัฐ ณ บางช้าง. (2554). *การคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นพหุด้วยวิธีการค้นหาแบบต้องห้าม*. วิทยานิพนธ์ปริญญามหาบัณฑิต, สาขาสถิติ, คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์.

ธันยากร ต้นชลขันธ์. (2538). *การเปรียบเทียบการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุโดยใช้วิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบริดจ์ และวิธีที่ใช้หลักการของริดจ์และสไตน์ ในกรณีเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ*. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ, คณะพาณิชยศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย.

นุสรา สถิตโพธิ์ศรี. (2535). *การเปรียบเทียบตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุโดยวิธีการถดถอยแบบริดจ์และวิธีการถดถอยแบบลาเท้นรูทธ์ ในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ*. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ, คณะพาณิชยศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย.

บุษยา ปภาพจน์. (2548). *การเปรียบเทียบวิธีการคัดเลือกตัวแบบที่ดีที่สุดในการวิเคราะห์การถดถอยเชิงเส้นพหุ*. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ, คณะพาณิชยศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย.

Drezner, Z., & George, A. (1999). Tabu Search Model Selection in Multiple Regression Analysis. *Communication in Statistics - Simulation and Computation, 28*(2), 349-367.

Games, P.A., & Lerner, J.V. (1981). *Maximum R^2 Improvement and Stepwise Multiple Regression as Related to Over-fitting*. Boston: Department of Counseling Development and Education Psychology, Boston College.

Glover, F. (1990). Tabu Search: a Tutorial. *Interfaces, 20*, 74-94.

Hedar, A., & Fukushima, M. (2003). TS Directed by Direct Search Methods for Nonlinear Global Optimization. *European Journal of Operational Research, 170*(2), 329-349.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge Regression: Some Simulations. *Communications in Statistics, 4*(2), 105-123.

Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression Problem. *Communications in Statistics –Theory and Methods, 34*, 1177-1182.

Lawless, J., & Wang, P. (1976). A simulation study of ridge and other regression Estimators. *Communication Statistics – Theory and Methods, 5*(4), 307-323.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to Linear Regression Analysis* (4th ed.). New Jersey: John Willey & Sons.

Nomura, M. (1988). On The Almost Unbiased Ridge Regression Estimation. *Communication Statistics – Simulations, 17*(3), 729-743.

Wichern, D., & Churchill, G. (1978). A Comparison of Ridge Estimators. *Technometrics, 20*(3), 301-311.

Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals ofStatistics, 14*(4), 1261-1295.