

การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์สำหรับตัวแบบ การถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีค่าผิดปกติในตัวแปรอิสระ

Comparison of Parameter Estimation Methods in Multiple Linear Regression Model

When Data Contain Outliers in Independent Variables

กัลยา บุญหล้า* และลักษณารัตน์ ศรีบัว

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

ตำบลท่าโพธิ์ อำเภอเมือง จังหวัดพิษณุโลก 65000

Kanlaya Boonlha* and Laksanarat Sribua

Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok,

Tha Pho, Muang, Phitsanulok 65000

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบการประมาณค่าพารามิเตอร์สำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธีการประมาณ 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ M-Andrews และวิธีตัวประมาณ GM-Huber กรณีตัวแปรอิสระมีค่าผิดปกติและความคลาดเคลื่อนมีการแจกแจงที่ เกณฑ์ในการเปรียบเทียบ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) โดยจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล แต่ละกรณีทำซ้ำ 1,000 ครั้ง ด้วยโปรแกรม R ผลการวิจัยพบว่าวิธีกำลังสองน้อยที่สุดให้ค่า MSE ต่ำที่สุด เมื่อองศาเสรีเท่ากับ 5 เมื่อตัวแปรอิสระไม่มีค่าผิดปกติ ที่ขนาดตัวอย่างเท่ากับ 50 วิธีตัวประมาณ M-Andrews ให้ค่า MSE ต่ำที่สุด เมื่อองศาเสรีเท่ากับ 5 และวิธีตัวประมาณ GM-Huber ให้ค่า MSE ต่ำที่สุด เมื่อองศาเสรีเท่ากับ 3 ทุกขนาดตัวอย่างและทุกร้อยละค่าผิดปกติ

คำสำคัญ : การถดถอยเชิงเส้นพหุคูณ; ค่าผิดปกติ; ตัวประมาณ M-Andrews; วิธีตัวประมาณ GM-Huber; วิธีกำลังสองน้อยที่สุด

Abstract

The purpose of this study was to compare the methods of regression coefficient estimation in the multiple linear regression model for three methods, the ordinary least squares (OLS),

M-Andrews and GM-Huber methods, in case of outliers on independent variable and error is t -distribution. The approach of estimator comparison was mean square error (MSE). The dataset was generated by the Monte Carlo simulation technique, repeated 1,000 times for each situation with R programming. The study results were as follows: the ordinary least squares methods indicated the lowest MSE when the degree of freedom equal to 5 in case of no outliers on independent. While the case of outliers on the independent variable, the GM-Andrews methods indicated the lowest MSE when the degree of freedom equal to 3 for all sample sizes and all percentage of outliers. However, the degree of freedom equal to 5, the M-Andrews methods indicated the lowest MSE in most cases.

Keywords: multiple linear regression; outlier; M-Andrews estimator; GM-Huber estimator; ordinary least square

1. บทนำ

การวิเคราะห์การถดถอย (regression analysis) เป็นวิธีการทางสถิติที่ศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระ (independent variable) กับตัวแปรตาม (dependent variable) เป็นการศึกษาความสัมพันธ์เชิงเส้น (linearity) ถ้าตัวแปรอิสระมีมากกว่าหนึ่งตัวกับตัวแปรตามหนึ่งตัว เรียกว่า การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (multiple linear regression analysis) ซึ่งมีข้อตกลงเบื้องต้น คือ ความคลาดเคลื่อนสุ่มเป็นอิสระต่อกันมีการแจกแจงปกติ ค่าเฉลี่ยของความคลาดเคลื่อนมีค่าเท่ากับศูนย์ ความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ ความคลาดเคลื่อนไม่มีความสัมพันธ์กัน และตัวแปรอิสระไม่มีความสัมพันธ์กัน การสร้างสมการถดถอยเริ่มจากการหาค่าประมาณ β_j ของพารามิเตอร์ β_j ส่วนใหญ่การประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณจะใช้วิธีกำลังสองน้อยที่สุด (ordinary least square method, OLS) โดยมีสมบัติเป็นตัวประมาณเชิงเส้นไม่เอนเอียงที่ดีที่สุด (best linear unbiased estimator, BLUE)

ในทางปฏิบัติการรวบรวมข้อมูลบางครั้งอาจเกิดปัญหาข้อมูลบางค่าแตกต่างจากข้อมูลอื่นมาก บางค่ามีค่าสูงมาก บางค่ามีค่าต่ำมาก เรียกว่า ค่าผิดปกติ (outlier) สาเหตุของการเกิดค่าผิดปกติบางครั้งอาจเกิดจากความผิดพลาดที่เป็นเหตุการณ์ที่อธิบายได้ เช่น วัดผิด จดข้อมูลผิด หรือลงรหัสผิด เมื่อเกิดค่าผิดปกติขึ้นจะมีผลทำให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดเกิดความผิดพลาดขึ้นได้ ถ้าข้อมูลเกิดจากการบันทึกข้อมูลผิดพลาด วิธีแก้ไขอาจตัดข้อมูลที่มีค่าผิดปกติออกหรืออาจปรับแก้ค่าให้เหมาะสม แต่ถ้าข้อมูลที่มีค่าผิดปกติได้รับการบันทึกมาอย่างถูกต้องและผู้วิเคราะห์ไม่สามารถอธิบายถึงสาเหตุของการเกิดค่าผิดปกติ ควรนำค่าผิดปกติมาวิเคราะห์ด้วยแต่ไม่ควรประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด ดังนั้นการวิเคราะห์การถดถอยที่มีความแกร่ง (robust regression) เป็นทางเลือกหนึ่งที่ใช้ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยนักสถิติหลายท่านได้คิดวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งไว้หลายวิธี เช่น Huber [1] ได้เสนอวิธี M (M method) เป็นวิธีที่นำฟังก์ชันที่ถูก

เลือกอย่างเหมาะสมมาแทนที่ความคลาดเคลื่อนกำลังสองในวิธีกำลังสองน้อยที่สุด และ Simpona และ Montgomery [2] ได้พัฒนาจากวิธี M มาเป็นวิธี GM (generalized M method) โดยปรับฟังก์ชันของความคลาดเคลื่อนที่ถูกเลือกอย่างเหมาะสม

Tantrakul [3] ได้ศึกษาเรื่องการเปรียบเทียบวิธีการถดถอยที่มีความแข็งแกร่งสำหรับตัวแบบการถดถอยเชิงเส้นแบบพหุคูณ โดยใช้ตัวแปรอิสระ 2 ตัวแปร และศึกษาวิธีการประมาณ 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS วิธีตัวประมาณ GM ทั้งนี้เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ซึ่งการเปรียบเทียบดังกล่าวทำภายใต้เงื่อนไขของข้อมูลที่มีค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ในตัวแปรอิสระและตัวแปรตาม พบว่าวิธีตัวประมาณ GM มีประสิทธิภาพมากที่สุด Thitacharee [4] ได้ศึกษาเรื่องการเปรียบเทียบความแข็งแกร่งของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีค่านอกเกณฑ์ระดับไม่รุนแรงในตัวแปรอิสระและตัวแปรตาม โดยศึกษาวิธีการประมาณ 4 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธี LTS วิธี M-Andrews และวิธี GM-Huber ทั้งนี้เกณฑ์ในการเปรียบเทียบความแกร่ง คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ซึ่งการเปรียบเทียบดังกล่าวทำภายใต้เงื่อนไขของข้อมูลที่มีค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ในตัวแปรอิสระและตัวแปรตาม ผลการวิจัยสรุปได้ว่าวิธีตัวประมาณ M-Andrews มีประสิทธิภาพมากที่สุด

ผู้วิจัยมีความสนใจที่จะศึกษาวิธีการประมาณค่าพารามิเตอร์สำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณ เมื่อมีค่าผิดปกติในตัวแปรอิสระ โดยใช้วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ M-Andrews และวิธีตัวประมาณ GM-Huber ภายใต้สถานการณ์ที่ข้อมูลมีค่าผิดปกติแบบต่าง ๆ แล้วเปรียบเทียบตัวประมาณของ

ทุกวิธีด้วยค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของตัวประมาณค่าสัมประสิทธิ์การถดถอย การศึกษาครั้งนี้ผู้วิจัยสนใจศึกษาในกรณีที่มีตัวแปรอิสระ 3 ตัวแปร กำหนดให้มีร้อยละค่าผิดปกติ 0, 10, 20 และ 30 ภายใต้สถานการณ์ที่ความคลาดเคลื่อนมีการแจกแจงแบบที่

2. วิธีการวิจัย

การวิจัยครั้งนี้ ผู้วิจัยได้ใช้เทคนิคการจำลองข้อมูลมอนติคาร์โล (Monte Carlo simulation technique) มาสร้างสถานการณ์ต่าง ๆ ด้วยการเขียนฟังก์ชันในโปรแกรม R เพื่อสร้างข้อมูลให้เป็นไปตามการวิจัย ทำซ้ำ 1,000 ครั้ง โดยดำเนินการวิจัยดังต่อไปนี้

2.1 สร้างข้อมูลตัวแปรตาม (Y) ตัวแปรอิสระ X_1 , X_2 และ X_3 และความคลาดเคลื่อน (ε) เมื่อกำหนดขนาดตัวอย่าง (n) ที่ศึกษาเท่ากับ 20, 30, 50 และ 100 ซึ่งมีขั้นตอนการสร้าง ดังนี้

ขั้นตอนที่ 1 สร้างตัวแปรอิสระ X_1 , X_2 และ X_3 มีการแจกแจงเอกรูป (uniform distribution) $U(-1,1)$, $U(-3,3)$ และ $U(-5,5)$ ตามลำดับ

ขั้นตอนที่ 2 สร้างความคลาดเคลื่อน (ε) มีการแจกแจงแบบที่ 3 องศาเสรีเท่ากับ 3 และ 5

ขั้นตอนที่ 3 สร้างข้อมูลตัวแปรตาม (Y) ให้มีลักษณะอยู่ในรูปแบบเชิงเส้น (linear model) กำหนดให้ค่าพารามิเตอร์หรือค่าสัมประสิทธิ์การถดถอย $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$ และ $\beta_3 = 1$

ขั้นตอนที่ 4 นำค่าของตัวแปรอิสระ X_1 , X_2 และ X_3 มาตรวจสอบค่าผิดปกติ โดยใช้แผนภาพกล่อง (Box and Whisker plot) และถ้าพบว่าค่าของตัวแปรอิสระ X_1 , X_2 และ X_3 ทุกค่าอยู่ในช่วง $(Q_1 - 1.5(IQR), Q_3 + 1.5(IQR))$ จะถือว่าค่าข้อมูลของตัวแปรมีค่าปกติอ้างอิงจาก Tantrakul [3]

2.2 ปรับข้อมูลตัวแปรอิสระ X_1 , X_2 และ X_3 จะปรับตามระดับค่าผิดปกติตามเงื่อนไขของการตรวจสอบค่าผิดปกติโดยใช้แผนภาพกล่อง (Box and Whisker plot) โดยในระดับค่าผิดปกติจะกำหนดให้มีร้อยละค่าผิดปกติเท่ากับ 0, 10, 20 และ 30 และขั้นตอนของการสร้างข้อมูลที่มีค่าผิดปกติในตัวแปรอิสระอ้างอิงจาก Tantrakul [3] มีดังนี้

ขั้นตอนที่ 1 นำข้อมูลของตัวแปรอิสระที่ต้องการสร้างให้มีค่าผิดปกติมาเรียงลำดับจากน้อยไปมาก และกำหนดให้ $A_1, A_2, A_3, \dots, A_{n-2}, A_{n-1}, A_n$ คือ ค่าผิดปกติที่จำลองขึ้นเรียงลำดับจากน้อยไปมาก

ขั้นตอนที่ 2 แทนค่าของตัวแปรอิสระที่มากที่สุดด้วยค่าหนึ่ง ซึ่งให้มีค่ามากขึ้นกว่าเดิม หรือแทนค่าตัวแปรอิสระที่น้อยที่สุดด้วยค่าหนึ่ง ซึ่งทำให้มีค่าน้อยลงกว่าเดิม เพื่อให้ได้ค่าผิดปกติในระดับที่กำหนด และมีจำนวนตามที่ต้องการ และ L มีค่าเท่ากับ 2 เมื่อต้องการค่าผิดปกติระดับไม่รุนแรง เนื่องจากจะได้ค่าที่อยู่กึ่งกลางของช่วง $(Q_1 - (IQR), Q_1 - 1.5(IQR))$ หรือ $(Q_3 + 1.5(IQR), Q_3 + 3(IQR))$ ซึ่งค่าที่อยู่ในช่วงเหล่านี้เป็นค่าผิดปกติระดับไม่รุนแรงตามเงื่อนไขการตรวจสอบค่าผิดปกติของวิธีแผนภาพกล่อง (Box and Whisker plot) หรือ L จะมีค่าเท่ากับ 5 เมื่อต้องการค่าผิดปกติระดับรุนแรง เนื่องจากจะได้ค่าที่น้อยกว่าค่าของ $Q_1 - 3(IQR)$ หรือมากกว่าค่าของ $Q_3 + 3(IQR)$ ซึ่งจะเป็นค่าผิดปกติระดับรุนแรงตามเงื่อนไขการตรวจสอบค่าผิดปกติของวิธีแผนภาพกล่อง (Box and Whisker plot)

2.3 หาตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ด้วยวิธีการประมาณค่า 3 วิธี คือ

2.3.1 วิธีกำลังสองน้อยที่สุด (ordinary least square method, OLS)

$$\beta = (XX)^{-1}XY \quad (1)$$

เมื่อ Y แทนเมทริกซ์ของตัวแปรตาม ขนาด $n \times 1$; X แทนเมทริกซ์ของตัวแปรอิสระ ขนาด $n \times p$; $p = k + 1$

2.3.2 วิธีตัวประมาณ M (M method) เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews [5]

ขั้นตอนที่ 1 คำนวณ $\beta^{(0)}$ เริ่มต้นจากวิธีกำลังสองน้อยที่สุด

ขั้นตอนที่ 2 หาค่า $e_i^{(1)}$ และ $s^{(1)}$ จาก

$$e_i^{(1)} = Y - X\beta^{(0)} \quad (2)$$

โดยที่ $i = 1, 2, \dots, n$

$$s^{(1)} = \frac{\text{median}[e_i^{(1)} - \text{median}(e^{(1)})]}{0.6745} \quad (3)$$

ขั้นตอนที่ 3 คำนวณ z ในรอบที่ 1

กล่าวคือ $z^{(1)} = \frac{e_i^{(1)}}{s^{(1)}}$

ขั้นตอนที่ 4 คำนวณ $\beta^{(1)}$ ในรอบที่ 1

$$\beta^{(1)} = (XW^{(1)}X)^{-1}XW^{(1)}Y \quad (4)$$

W แทนเมทริกซ์ทแยงมุม ขนาด $n \times n$

$$W^{(1)} = \begin{bmatrix} w_1^{(1)}(z) & 0 & L & 0 \\ 0 & w_2^{(1)}(z) & L & 0 \\ M & M & O & M \\ 0 & 0 & L & w_n^{(1)}(z) \end{bmatrix}_{n \times n}$$

ที่มี w_i เป็นสมาชิกในแนวทแยงมุม โดยที่ w_i เป็นฟังก์ชันถ่วงน้ำหนักของ Andrews [6]

$$w(z) = \begin{cases} \frac{\sin(z/c)}{z/c} & ; |z| \leq c\pi \\ 0 & ; |z| > c\pi \end{cases} \quad (5)$$

c มีค่าเท่ากับ 1.339 และ π เป็นค่าคงที่

ขั้นตอนที่ 5 คำนวณค่า $|\beta^{(0)} - \beta^{(1)}|$

ขั้นตอนที่ 6 ถ้า $|\beta^{(0)} - \beta^{(1)}| > 0.001$

ให้ไปทำซ้ำในรอบที่ m แต่ถ้า $|\beta^{(0)} - \beta^{(1)}| \leq 0.001$ พบว่าค่าประมาณสัมประสิทธิ์การถดถอยทุกค่าไม่มากกว่า 0.001 นั่นคือ จะได้ค่าประมาณสัมประสิทธิ์การถดถอยจากรอบที่ 1 และสิ้นสุดการคำนวณ

2.3.3 วิธีตัวประมาณ GM (generalized M method) เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Huber

ขั้นตอนที่ 1 คำนวณ $\beta^{(0)}$ เริ่มต้นจากวิธีกำลังสองน้อยที่สุด

ขั้นตอนที่ 2 หาค่า $e_i^{(1)}$ และ $s^{(1)}$

ขั้นตอนที่ 3 คำนวณ π_i ในรอบที่ 1

$$\text{กล่าวคือ } \pi_i = \frac{\text{median}|z_i|}{z_i}$$

ขั้นตอนที่ 4 คำนวณ $\beta^{(1)}$ ในรอบที่ 1

$$\beta^{(1)} = (XW^{(1)}X)^{-1}XW^{(1)}Y \quad (6)$$

W แทนเมทริกซ์ทแยงมุม ขนาด $n \times n$ ที่มี w_i เป็นสมาชิกในแนวทแยงมุม โดยที่ w_i เป็นฟังก์ชันถ่วงน้ำหนักของ Huber

$$w(z) = \begin{cases} \frac{-c}{u} & ; \quad u < -c \\ 1 & ; \quad -c \leq u \leq c \\ \frac{c}{u} & ; \quad u > c \end{cases} \quad (7)$$

c มีค่าเท่ากับ 1.345

ขั้นตอนที่ 5 คำนวณค่า $|\beta^{(0)} - \beta^{(1)}|$

ขั้นตอนที่ 6 ถ้า $|\beta^{(0)} - \beta^{(1)}| > 0.001$

ให้ไปทำซ้ำในรอบที่ m แต่ถ้า $|\beta^{(0)} - \beta^{(1)}| \leq 0.001$ พบว่าค่าประมาณสัมประสิทธิ์การถดถอยทุกค่าไม่มากกว่า 0.001 นั่นคือ จะได้ค่าประมาณสัมประสิทธิ์การถดถอยจากรอบที่ 1 และสิ้นสุดการคำนวณ

2.4 เกณฑ์การเปรียบเทียบประสิทธิภาพแต่ละวิธีด้วยค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของการประมาณค่าสัมประสิทธิ์การถดถอย ภายใต้สถานการณ์ต่าง ๆ ที่กำหนดขึ้น วิธีใดให้ค่า MSE ที่ต่ำกว่าจะเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีประสิทธิภาพมากกว่า

$$MSE = \frac{\sum_{j=0}^k \left[\sum_{i=1}^{1,000} \frac{(\beta_{ij} - \hat{\beta}_{ij})^2}{1,000} \right]}{k+1} \quad (8)$$

เมื่อ β_{ij} แทนค่าสัมประสิทธิ์การถดถอยที่กำหนด ตัวที่ j ของการทำซ้ำรอบที่ i ; $\hat{\beta}_{ij}$ แทนค่าประมาณสัมประสิทธิ์การถดถอย ตัวที่ j ของการทำซ้ำรอบที่ i ; k แทนจำนวนตัวแปรอิสระในสมการถดถอย; i แทนจำนวนรอบในการทำซ้ำ โดย $i = 1, 2, \dots, 1000$

3. ผลการวิจัยและวิจารณ์

การศึกษานี้เปรียบเทียบการประมาณค่าพารามิเตอร์สำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธีการประมาณ 3 วิธี คือ วิธีการถดถอยน้อยที่สุด (OLS)

วิธีตัวประมาณ M-Andrews และวิธีตัวประมาณ GM-Huber โดยจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล ทำซ้ำ 1,000 ครั้ง ในแต่ละกรณี ด้วยโปรแกรม R ซึ่งได้ผลการวิจัยดังนี้

3.1 กรณีที่ข้อมูลไม่มีค่าผิดปกติในตัวแปรอิสระและความคลาดเคลื่อนมีการแจกแจงแบบที่ เมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE ของการประมาณค่าสัมประสิทธิ์การถดถอยแต่ละวิธีมีค่าลดลง ที่องศาเสรีเท่ากับ 3 วิธี GM-Huber ให้ค่า MSE ต่ำที่สุดในทุกขนาดตัวอย่าง แต่ที่องศาเสรีเท่ากับ 5 เมื่อขนาดตัวอย่างเท่ากับ 20 วิธี OLS และ M-Andrews ให้ค่า MSE ต่ำที่สุดเท่ากัน เมื่อแสดงผลลัพธ์ด้วยทศนิยมตำแหน่งที่ 4 แต่เมื่อพิจารณาที่ทศนิยมตำแหน่งที่ 6 พบว่าวิธี OLS ให้ค่า MSE ต่ำกว่าวิธี M-Andrews ขณะที่ขนาดตัวอย่างเท่ากับ 30 วิธี OLS ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ดังตารางที่ 1

3.2 กรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรอิสระ 1 ตัว และความคลาดเคลื่อนแจกแจงที่ พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE ของการประมาณค่าสัมประสิทธิ์การถดถอยแต่ละวิธีมีค่าลดลง และองศาเสรีที่แตกต่างกันของการแจกแจงแบบที่มีผลต่อค่า MSE โดยค่า MSE มีแนวโน้มลดลงเมื่อองศาเสรีเพิ่มขึ้น กรณีที่องศาเสรีเท่ากับ 3 สำหรับทุกขนาดตัวอย่างและทุกร้อยละการเกิดค่าผิดปกติทั้งค่าผิดปกติระดับไม่รุนแรงและรุนแรง วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด กรณีที่องศาเสรีเท่ากับ 5 ค่าผิดปกติระดับไม่รุนแรงของขนาดตัวอย่างเท่ากับ 20 และ 30 ร้อยละ 10 ของการเกิดค่าผิดปกติ และขนาดตัวอย่างเท่ากับ 30 ร้อยละ 30 ของการเกิดค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำ

ที่สุด เมื่อขนาดตัวอย่างเท่ากับ 20 ร้อยละ 20 และ 30 ของการเกิดค่าผิดปกติ และขนาดตัวอย่างเท่ากับ 30 ร้อยละ 20 ของการเกิดค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 ทุกร้อยละการเกิดค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ในขณะที่ค่าผิดปกติระดับรุนแรงของขนาดตัวอย่างเท่ากับ 20 และ 30 ทุกร้อยละการเกิดค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นขนาดตัวอย่างเท่ากับ 30 ร้อยละ 10 ของการเกิดค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 ทุกร้อยละการเกิดค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ดังตารางที่ 2

3.3 กรณีที่ข้อมูลมีค่าผิดปกติตัวแปรอิสระ 2 ตัว และความคลาดเคลื่อนแจกแจงที่ กรณีที่องศาเสรีเท่ากับ 3 สำหรับทุกขนาดตัวอย่างและทุกร้อยละการเกิดค่าผิดปกติทั้งค่าผิดปกติระดับไม่รุนแรงและรุนแรง วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด กรณีที่องศาเสรีเท่ากับ 5 ค่าผิดปกติระดับไม่รุนแรงของขนาดตัวอย่างเท่ากับ 20, 30 และ 50 ทุกร้อยละของค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นขนาดตัวอย่างเท่ากับ 50 ร้อยละ 20 ของค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 100 ร้อยละ 10 ของค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 100 ร้อยละ 20

และ 30 ของค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ขณะที่ค่าผิดปกติระดับรุนแรงของขนาดตัวอย่างเท่ากับ 20, 30 และ 50 ทุกร้อยละของค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นขนาดตัวอย่างเท่ากับ 50 ร้อยละ 20 ของค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 100 เกิดค่าผิดปกติร้อยละ 10 วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด และเมื่อเกิดค่าผิดปกติร้อยละ 20 และ 30 วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ดังตารางที่ 3

3.4 กรณีที่ข้อมูลมีค่าผิดปกติตัวแปรอิสระ 3 ตัว และความคลาดเคลื่อนแจกแจงที่ กรณีที่องศาเสรีเท่ากับ 3 สำหรับทุกขนาดตัวอย่างและทุกร้อยละการเกิดค่าผิดปกติทั้งค่าผิดปกติระดับไม่รุนแรงและรุนแรง วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด กรณีที่องศาเสรีเท่ากับ 5 ค่าผิดปกติระดับไม่รุนแรงของขนาดตัวอย่างเท่ากับ 20 และ 30 ร้อยละ 20 ของการเกิดค่าผิดปกติ และขนาดตัวอย่างเท่ากับ 30 ร้อยละ 10 ของการเกิดค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 20 ร้อยละ 10 และ 30 ของการเกิดค่าผิดปกติ ขนาดตัวอย่างเท่ากับ 30 ร้อยละ 30 ของการเกิดค่าผิดปกติ ขนาดตัวอย่างเท่ากับ 50 ทุกร้อยละการเกิดค่าผิดปกติ และขนาดตัวอย่างเท่ากับ 100 ร้อยละ 10 ของการเกิดค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 100 ร้อยละ 20 และ 30 ของการเกิดค่าผิดปกติ วิธี GM-Huber

Table 1 The MSE of the coefficient estimations when the error is t-distribution without outliers

n	d.f = 3			d.f = 5		
	OLS	M	GM	OLS	M	GM
20	0.2008	0.1817	<u>0.1509</u>	<u>0.1147</u>	<u>0.1147</u>	0.1353
30	0.1165	0.1160	<u>0.0839</u>	<u>0.0719</u>	0.0720	0.0820
50	0.0664	0.0657	<u>0.0428</u>	0.0401	0.0400	<u>0.0397</u>
100	0.0333	0.0329	<u>0.0203</u>	0.0197	0.0197	<u>0.0191</u>

ให้ค่า MSE ต่ำที่สุด ในขณะที่ค่าผิดปกติระดับรุนแรงของขนาดตัวอย่างเท่ากับ 20, 30 และ 50 ทุกร้อยละการเกิดค่าผิดปกติ วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นขนาดตัวอย่างเท่ากับ 30 ร้อยละ 10 ของการเกิด

ค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 100 ทุกร้อยละการเกิดค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด ดังตารางที่ 4

Table 2 The MSE of the coefficient estimations with outliers in 1 independent variable

d.f	n	Percentages of outliers	Mild Outliers			Extreme Outliers		
			OLS	M	GM	OLS	M	GM
3	20	10	0.2040	0.2035	<u>0.1663</u>	0.2015	0.2014	<u>0.1578</u>
		20	0.2025	0.2011	<u>0.1651</u>	0.1973	0.1964	<u>0.1633</u>
		30	0.2011	0.2000	<u>0.1645</u>	0.1989	0.1966	<u>0.1616</u>
	30	10	0.1204	0.1197	<u>0.0826</u>	0.1206	0.1199	<u>0.0830</u>
		20	0.1199	0.1194	<u>0.0823</u>	0.1200	0.1194	<u>0.0830</u>
		30	0.1181	0.1172	<u>0.0797</u>	0.1195	0.1187	<u>0.0816</u>
	50	10	0.0699	0.0692	<u>0.0439</u>	0.0691	0.0686	<u>0.0438</u>
		20	0.0695	0.0688	<u>0.0443</u>	0.0687	0.0680	<u>0.0441</u>
		30	0.0693	0.0686	<u>0.0437</u>	0.0660	0.0646	<u>0.0401</u>
	100	10	0.0330	0.0329	<u>0.0201</u>	0.0328	0.0327	<u>0.0200</u>
		20	0.0346	0.0344	<u>0.0194</u>	0.0325	0.0323	<u>0.0197</u>
		30	0.0326	0.0325	<u>0.0199</u>	0.0324	0.0323	<u>0.0198</u>
5	20	10	<u>0.1138</u>	0.1139	0.1438	0.1112	<u>0.1111</u>	0.1350
		20	0.1108	<u>0.1107</u>	0.1337	0.1095	<u>0.1094</u>	0.1323
		30	0.1120	<u>0.1119</u>	0.1321	0.1071	<u>0.1070</u>	0.1310
	30	10	<u>0.0715</u>	0.0716	0.0765	<u>0.0705</u>	0.0706	0.0742
		20	0.0712	<u>0.0709</u>	0.0757	0.0704	<u>0.0699</u>	0.0737
		30	<u>0.0685</u>	0.0686	0.0743	0.0701	<u>0.0698</u>	0.0751
	50	10	0.0397	0.0396	<u>0.0390</u>	0.0392	0.0392	<u>0.0384</u>
		20	0.0394	0.0393	<u>0.0383</u>	0.0390	0.0390	<u>0.0379</u>
		30	0.0392	0.0392	<u>0.0388</u>	0.0390	0.0389	<u>0.0387</u>
	100	10	0.0195	0.0195	<u>0.0189</u>	0.0193	0.0194	<u>0.0188</u>
		20	0.0193	0.0193	<u>0.0188</u>	0.0192	0.0192	<u>0.0186</u>
		30	0.0192	0.0192	<u>0.0187</u>	0.0191	0.0191	<u>0.0185</u>

Table 3 The MSE of the coefficient estimations with outliers in 2 independent variables

df	n	Percentages of outliers	Mild Outliers			Extreme Outliers		
			OLS	M	GM	OLS	M	GM
3	20	10	0.1131	0.1123	<u>0.0882</u>	0.0645	0.0644	<u>0.0515</u>
		20	0.0881	0.0874	<u>0.0659</u>	0.0556	0.0551	<u>0.0430</u>
		30	0.0696	0.0687	<u>0.0553</u>	0.0521	0.0516	<u>0.0418</u>
	30	10	0.0658	0.0656	<u>0.0476</u>	0.0393	0.0392	<u>0.0276</u>
		20	0.0517	0.0514	<u>0.0361</u>	0.0343	0.0342	<u>0.0233</u>
		30	0.0423	0.0419	<u>0.0316</u>	0.0327	0.0324	<u>0.0221</u>
	50	10	0.0380	0.0376	<u>0.0251</u>	0.0232	0.0227	<u>0.0150</u>
		20	0.0302	0.0299	<u>0.0196</u>	0.0206	0.0204	<u>0.0132</u>
		30	0.0270	0.0267	<u>0.0171</u>	0.0197	0.0195	<u>0.0126</u>
	100	10	0.0189	0.0188	<u>0.0113</u>	0.0113	0.0112	<u>0.0068</u>
		20	0.0147	0.0146	<u>0.0089</u>	0.0098	0.0098	<u>0.0059</u>
		30	0.0128	0.0127	<u>0.0077</u>	0.0093	0.0093	<u>0.0056</u>
5	20	10	0.0615	<u>0.0614</u>	0.0755	0.0354	<u>0.0353</u>	0.0436
		20	0.0489	<u>0.0488</u>	0.0600	0.0312	<u>0.0311</u>	0.0380
		30	0.0420	<u>0.0419</u>	0.0511	0.0293	<u>0.0292</u>	0.0354
	30	10	0.0379	<u>0.0378</u>	0.0418	0.0221	<u>0.0220</u>	0.0240
		20	0.0295	<u>0.0294</u>	0.0332	0.0191	<u>0.0190</u>	0.0218
		30	0.0266	<u>0.0265</u>	0.0288	0.0185	<u>0.0184</u>	0.0204
	50	10	0.0215	<u>0.0214</u>	0.0220	0.0132	<u>0.0131</u>	0.0133
		20	<u>0.0170</u>	0.0171	0.0172	0.0117	<u>0.0116</u>	0.0115
		30	0.0153	<u>0.0152</u>	0.0156	0.0113	<u>0.0112</u>	0.0113
	100	10	0.0106	<u>0.0105</u>	0.0107	0.0065	<u>0.0064</u>	0.0064
		20	0.0086	0.0086	<u>0.0083</u>	0.0058	0.0058	<u>0.0056</u>
		30	0.0076	0.0075	<u>0.0072</u>	0.0055	0.0055	<u>0.0051</u>

3.5 สำหรับกรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรอิสระ ซึ่งมีค่าผิดปกติระดับไม่รุนแรงและระดับรุนแรง มีร้อยละการเกิดค่าผิดปกติเท่ากับ 10, 20 และ 30 และมีขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100 เมื่อ

ความคลาดเคลื่อนมีการแจกแจงแบบที่ สรุปลผลการวิจัยข้อมูลได้ว่าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ดีที่สุด เมื่อองศาเสรีเท่ากับ 3 คือ วิธี GM-Huber ดังตารางที่ 6 และเมื่อองศาเสรีเท่ากับ 5 ส่วน

ใหญ่วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นกรณี
ขนาดตัวอย่างเท่ากับ 30 ร้อยละ 10 ของค่าผิดปกติ
วิธี OLS ให้ค่า MSE ต่ำที่สุด และกรณีขนาดตัวอย่าง

เท่ากับ 50 ร้อยละ 20 ของค่าผิดปกติ กรณีขนาด
ตัวอย่างเท่ากับ 100 ทุกร้อยละค่าผิดปกติ วิธี GM-
Huber ให้ค่า MSE ต่ำที่สุด ดังตารางที่ 5

Table 4 The MSE of the coefficient estimations with outliers in 3 independent variables

d.f	n	Percentages of outliers	Mild Outliers			Extreme Outliers		
			OLS	M	GM	OLS	M	GM
3	20	10	0.1091	0.1087	<u>0.0852</u>	0.0572	0.0570	<u>0.0466</u>
		20	0.0825	0.0820	<u>0.0629</u>	0.0477	0.0471	<u>0.0365</u>
		30	0.0674	0.0665	<u>0.0516</u>	0.0437	0.0431	<u>0.0319</u>
	30	10	0.0640	0.0634	<u>0.0469</u>	0.0388	0.0389	<u>0.0250</u>
		20	0.0527	0.0541	<u>0.0341</u>	0.0306	0.0300	<u>0.0204</u>
		30	0.0429	0.0423	<u>0.0295</u>	0.0289	0.0284	<u>0.0197</u>
	50	10	0.0370	0.0366	<u>0.0247</u>	0.0213	0.0210	<u>0.0140</u>
		20	0.0287	0.0284	<u>0.0191</u>	0.0184	0.0181	<u>0.0121</u>
		30	0.0253	0.0250	<u>0.0160</u>	0.0175	0.0173	<u>0.0112</u>
	100	10	0.0186	0.0186	<u>0.0110</u>	0.0105	0.0105	<u>0.0062</u>
		20	0.0141	0.0140	<u>0.0085</u>	0.0090	0.0089	<u>0.0053</u>
		30	0.0121	0.0120	<u>0.0073</u>	0.0084	0.0084	<u>0.0050</u>
5	20	10	0.0589	<u>0.0588</u>	0.0730	0.0317	<u>0.0316</u>	0.0394
		20	<u>0.0457</u>	0.0458	0.0554	0.0270	<u>0.0269</u>	0.0327
		30	0.0386	<u>0.0385</u>	0.0469	0.0250	<u>0.0249</u>	0.0318
	30	10	<u>0.0365</u>	0.0366	0.0411	<u>0.0200</u>	0.0201	0.0221
		20	<u>0.0278</u>	0.0279	0.0313	0.0168	<u>0.0167</u>	0.0194
		30	0.0248	<u>0.0247</u>	0.0268	0.0161	<u>0.0160</u>	0.0173
	50	10	0.0210	<u>0.0209</u>	0.0214	0.0122	<u>0.0121</u>	0.0125
		20	0.0163	<u>0.0162</u>	0.0164	0.0105	<u>0.0104</u>	0.0106
		30	0.0144	<u>0.0143</u>	0.0148	0.0101	<u>0.0100</u>	0.0104
	100	10	0.0103	<u>0.0102</u>	0.0105	0.0060	0.0060	<u>0.0059</u>
		20	0.0082	0.0082	<u>0.0080</u>	0.0054	0.0054	<u>0.0052</u>
		30	0.0071	0.0071	<u>0.0068</u>	0.0050	0.0050	<u>0.0048</u>

Table 5 The estimation method with the lowest mean square error when the outliers in independent variables

d.f	n	Mild Outliers			Extreme Outliers		
		Percentages of outliers			Percentages of outliers		
		10	20	30	10	20	30
3	20	GM	GM	GM	GM	GM	GM
	30	GM	GM	GM	GM	GM	GM
	50	GM	GM	GM	GM	GM	GM
	100	GM	GM	GM	GM	GM	GM
5	20	M	M	M	M	M	M
	30	OLS	M	M	M	M	M
	50	M	M	M	M	GM	M
	100	M	GM	GM	GM	GM	GM

การศึกษาการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ และการจำลองข้อมูลในสถานการณ์ต่าง ๆ กรณีข้อมูลมีค่าผิดปกติในตัวแปรอิสระ และความคลาดเคลื่อนมีการแจกแจงที่ ถ้าองศาเสรีเท่ากับ 3 วิธีตัวประมาณ GM-Huber สามารถประมาณค่าสัมประสิทธิ์การถดถอยเหมาะสมที่สุด และองศาเสรีเท่ากับ 5 วิธีตัวประมาณ M-Andrews สามารถประมาณค่าสัมประสิทธิ์การถดถอยเหมาะสมที่สุด ซึ่งสอดคล้องกับงานวิจัยของ Tantrakul [3] ที่เปรียบเทียบวิธีการถดถอยที่มีความแข็งแกร่งสำหรับตัวแบบการถดถอยเชิงเส้นแบบพหุคูณ ซึ่งผลการศึกษากรณีข้อมูลมีค่าผิดปกติในตัวแปรอิสระและความคลาดเคลื่อนมีการแจกแจงที่ พบว่าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ดีที่สุดสำหรับองศาเสรีเท่ากับ 3 คือ วิธีตัวประมาณ GM-Huber และที่องศาเสรีเท่ากับ 5 เมื่อข้อมูลไม่เป็นไปตามข้อตกลงเบื้องต้นโดยทั่วไปของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ กรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรอิสระ วิธีตัวประมาณ M-Andrews

สามารถประมาณค่าสัมประสิทธิ์การถดถอยเหมาะสมที่สุด ซึ่งสอดคล้องกับงานวิจัยของ Tantrakul [4] ที่เปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์การถดถอยในสมการถดถอยเชิงเส้นพหุคูณที่มีตัวแปรอิสระ 2 ตัว เมื่อข้อมูลค่าผิดปกติ ซึ่งผลการศึกษากรณีที่มีข้อมูลมีค่าผิดปกติในตัวแปรอิสระ พบว่าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ดีที่สุดคือ วิธีตัวประมาณ M-Andrews

4. สรุป

ผลการเปรียบเทียบค่าคลาดเคลื่อนกำลังสองเฉลี่ยวิธีการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี พบว่าเมื่อข้อมูลที่มีค่าผิดปกติในตัวแปรอิสระ เมื่อความคลาดเคลื่อนมีการแจกแจงแบบที่ ค่าผิดปกติระดับไม่รุนแรง ระดับรุนแรง และร้อยละการเกิดค่าผิดปกติคงที่ พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่า MSE ของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี มีค่าลดลง ขณะที่ขนาดตัวอย่างและ

ร้อยละการเกิดค่าผิดปกติคงที่ พบว่าค่า MSE เมื่อระดับค่าผิดปกติไม่รุนแรงมีค่ามากกว่าระดับค่าผิดปกติรุนแรง ภายใต้ความคลาดเคลื่อนมีการแจกแจงแบบปกติ และที่ความคลาดเคลื่อนมีการแจกแจงแบบที่และเมื่อขนาดตัวอย่างคงที่มีค่าผิดปกติระดับไม่รุนแรงและระดับรุนแรง พบว่าเมื่อร้อยละการเกิดค่าผิดปกติในตัวแปรอิสระเพิ่มขึ้น ค่า MSE ของวิธีการประมาณสัมประสิทธิ์การถดถอยทั้ง 3 วิธี มีค่าลดลงสำหรับกรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรอิสระ ซึ่งมีค่าผิดปกติระดับไม่รุนแรงและระดับรุนแรง มีร้อยละการเกิดค่าผิดปกติเท่ากับ 10, 20 และ 30 และมีขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100 เมื่อความคลาดเคลื่อนมีการแจกแจงแบบที่สรุปผลการวิจัยข้อมูลได้ว่าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ดีที่สุดเมื่อองศาเสรีเท่ากับ 3 คือ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุดและเมื่อองศาเสรีเท่ากับ 5 ส่วนใหญ่วิธี M-Andrews ให้ค่า MSE ต่ำที่สุด ยกเว้นกรณีขนาดตัวอย่างเท่ากับ 30 ร้อยละ 10 ของค่าผิดปกติ วิธี OLS ให้ค่า MSE ต่ำที่สุด และกรณีขนาดตัวอย่างเท่ากับ 50 ร้อยละ 20 ของค่าผิดปกติ กรณีขนาดตัวอย่างเท่ากับ 100 ทุกร้อยละค่าผิดปกติ วิธี GM-Huber ให้ค่า MSE ต่ำที่สุด อย่างไรก็ตาม ผู้วิจัยเห็นว่าควรศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีค่าผิดปกติด้วยวิธีการอื่น ๆ ได้แก่ วิธีกำลังสองของมัธยฐานน้อยที่สุด (least median of square estimator) วิธีประมาณ modified GM6 วิธี Bounded-Influence เป็นต้น รวมถึงควรศึกษากรณีความคลาดเคลื่อนมีการแจกแจงแบบอื่น ๆ ได้แก่ การแจกแจงแบบบีต้า การแจกแจงแบบเอกซ์โพเนนเชียล

การแจกแจงแบบแกมมา เป็นต้น

5. References

- [1] Huber, P.J., 1964, Robust estimation of a location parameter, Ann. Math. Stat. 35: 73-101.
- [2] Simpsona, J.R. and Montgomery, D.C., 1996, A biased-robust regression technique for the combined outlier multicollinearity problem, J. Stat. Comput. Simul. 56: 1-22.
- [3] Tantrakul, O. and Payakkapong, P., 2012, Comparison of robust regression methods in multiple linear regression model, Nat. Grad. Res. Conf. 13(1): 291-300. (in Thai)
- [4] Thitacharee, K., Sinsomboonthong, J. and Supapakorn, T., 2018, Robust comparison of regression coefficient estimation methods for multiple linear regression model when dataset contains outliers, Univ. Thai Chamber Commerce 2(1): 1686-1699. (in Thai)
- [5] Montgomery, D. C., Peck, E. A. and Vining, G.G., 2012, Introduction to Linear Analysis, 5th Ed., John Wiley & Son, Inc., New York.
- [6] Holland, P.W. and Welsch, R.E., 1977, Robust regression using iteratively reweight least-squares, Commun. Stat. Theory Meth. 6: 813-827.