การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอย สำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีค่านอกเกณฑ์ในตัวแปรตาม

Efficiency Comparison of Regression Coefficient Estimation Methods for Multiple Linear Regression Model when Data Contain Outliers in Dependent Variable

กฤตพร ธิตะจารี*, จุฑาภรณ์ สินสมบูรณ์ทอง และ ธิดาพร ศุภภากร

Kritaporn Thitacharee*, Juthaphorn Sinsomboonthong and Thidaporn Supapakorn

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

Department of Statistics, Faculty of Science, Kasetsart University

Received : 3 May 2018

Accepted : 13 June 2018

Published online : 25 June 2018

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีค่านอกเกณฑ์ระดับไม่รุนแรงในตัวแปรตาม โดยศึกษาวิธีการประมาณ 5 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธี LTS วิธี M เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews และ Welsch และ วิธี GM โดยใช้ฟังก์ชันถ่วงน้ำหนักของ Huber ทั้งนี้เกณฑ์ในการเปรียบเทียบประสิทธิภาพ คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ข้อมูลที่ใช้ในการวิจัยได้จากการจำลองด้วยเทคนิคมอนติคาร์โล จำนวน 78 สถานการณ์ ทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ ผลการวิจัยสรุปได้ดังนี้ กรณีที่ไม่เกิดค่านอกเกณฑ์ในตัวแปรตาม พบว่าวิธีกำลังสองน้อยที่สุดมีประสิทธิภาพสูงที่สุด กรณีเกิดค่านอกเกณฑ์ในตัวแปรตามเมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติ สำหรับขนาดตัวอย่างเท่ากับ 10, 20 และ 30 ส่วนใหญ่วิธี M โดยใช้ฟังก์ชันถ่วงน้ำหนัก Welsch มีประสิทธิภาพสูงที่สุด และสำหรับขนาดตัวอย่างเท่ากับ 50, 100 และ 150 วิธี M โดยใช้ฟังก์ชันถ่วงน้ำหนัก Andrews มีประสิทธิภาพสูงที่สุด อย่างไรก็ตามเมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่องศาเสรีเท่ากับ 1 วิธี GM ให้ประสิทธิภาพสูงที่สุดในทุกขนาดตัวอย่าง และเมื่อองศาเสรีเพิ่มสูงขึ้น พบว่า วิธี M โดยใช้ฟังก์ชันถ่วงน้ำหนักของ Welsch มีแนวโน้มให้ประสิทธิภาพสูงที่สุด เมื่อขนาดตัวอย่างไม่เกิน 30 แต่เมื่อขนาดตัวอย่างมากกว่า 30 ส่วนใหญ่พบว่าวิธี M โดยใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews มีแนวโน้มให้ประสิทธิภาพสูงที่สุด

คำสำคัญ : การถดถอยเชิงเส้นพหุคูณ, ค่านอกเกณฑ์, สัมประสิทธิ์การถดถอย, วิธีกำลังสองน้อยที่สุด, ความคลาดเคลื่อนกำลังสองเฉลี่ย

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

Abstract

The purpose of this research was to compare the efficiency of five regression coefficient estimation methods for multiple linear regression model when data containing mild outliers in dependent variable. The five methods composed of ordinary least squares method, least trimmed squares method, M method using Andrews and Welsch weight functions and GM method using Huber weight function. The criterion for efficiency comparison was estimated mean square error (EMSE). The data was generated by Monte Carlo simulation technique for 78 situations and repeated 1,000 times for each situation. The results of this research were as follow: in case of no outliers in dependent variable, ordinary least squares method was the most efficient method. In case of outliers in dependent variable and random error was normally distributed, when sample size was 10, 20 and 30, M method using Welsch weight function provided the most efficient estimator. In addition, when sample size was 50, 100 and 150, M method using Andrews weight function provided the most efficient estimator. However, when random error was t-distributed with 1 degree of freedom, GM tended to be the most efficient estimator for all situations. Moreover, when degree of freedom increased and sample size was not greater than 30, M method using Welsch weight function was likely to be the most efficient estimator. However, when sample size was greater than 30, M method using Andrews weight function tended to be the most efficient estimator.

Keywords : multiple linear regression model, outliers, regression coefficient, ordinary least squares method, mean square error

บทนำ

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Analysis) เป็นการศึกษาความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ (Independent Variable) ที่มากกว่า 1 ตัว กับตัวแปรตาม (Dependent Variable) ที่สนใจศึกษา 1 ตัวโดยทั่วไปในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณจะกำหนดข้อสมมุติพื้นฐานเกี่ยวกับความคลาดเคลื่อนสุ่มในตัวแบบการถดถอยเชิงเส้นพหุคูณ คือ ความคลาดเคลื่อนสุ่มเป็นอิสระต่อกันมีการแจกแจงปรกติ ค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนคงที่ เมื่อเป็นไปตามข้อสมมุติพื้นฐานของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ จะสามารถประมาณค่าพารามิเตอร์ หรือค่าสัมประสิทธิ์การถดถอยจากวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method: OLS Method) ซึ่งเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่นิยมกันอย่างกว้างขวาง เนื่องจากตัวประมาณที่ได้มีคุณสมบัติเป็นตัวประมาณเชิงเส้นไม่เอนเอียงดีที่สุด (Best Linear Unbiased Estimator: BLUE) แต่ในทางปฏิบัติความคลาดเคลื่อนสุ่มอาจมีสาเหตุที่ทำให้ไม่เป็นไปตามข้อสมมุติพื้นฐานของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ เช่น การเกิดค่านอกเกณฑ์ (Outliers) ซึ่งค่านอกเกณฑ์ คือ ค่าของตัวแปรตามหรือตัวแปรอิสระบางตัวที่มีค่าสูงผิดปกติหรือต่ำผิดปกติจากค่าสังเกตส่วนใหญ่ในชุดข้อมูล โดยสามารถแบ่งค่านอกเกณฑ์ออกเป็น 2 ระดับ (Barnett & Lewis, 1995) ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์โดยใช้แผนภาพกล่อง (Box and Whisker Plot) ได้ดังนี้ ค่านอกเกณฑ์ระดับไม่รุนแรง (Mild Outliers) คือ ค่าที่อยู่ในช่วง $[Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$ หรือ $[Q_3 + 1.5(IQR), Q_3 + 3(IQR)]$ และค่านอกเกณฑ์ระดับรุนแรง (Extreme Outliers) คือ

ค่าที่อยู่ในช่วง $[-\infty, Q_1 - 3(IQR)]$ หรือ $[Q_3 + 3(IQR), \infty]$ โดยที่ Q_1 เป็นค่าควอร์ไทล์ที่ 1 ของตัวแปรอิสระหรือตัวแปรตาม Q_3 เป็นค่าควอร์ไทล์ที่ 3 ของตัวแปรอิสระหรือตัวแปรตาม และ IQR เป็นพิสัยระหว่างควอร์ไทล์ นั่นคือ IQR เป็นผลต่างระหว่างควอร์ไทล์ที่ 3 กับควอร์ไทล์ที่ 1

ซึ่งเมื่อเกิดปัญหาข้อมูลมีค่านอกเกณฑ์ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ พบว่าวิธีกำลังสองน้อยที่สุดจะทำให้ค่าประมาณสัมประสิทธิ์การถดถอยที่ได้ไม่ถูกต้อง โดยสาเหตุการเกิดค่านอกเกณฑ์อาจเนื่องมาจากในปัจจุบันมีการบันทึกข้อมูลจำนวนมากลงในคอมพิวเตอร์อาจเกิดการบันทึกข้อมูลผิดพลาดได้ หรือ เกิดจากความผิดพลาดในการวัด (Measurement Error) ความผันแปรที่มีอยู่แล้วในหน่วยทดลอง (Inherent Variability) ข้อผิดพลาดในการดำเนินการ (Execution Error) เป็นต้น (Anscombe, 1960) และถ้าผู้วิจัยละเลยต่อการตรวจสอบหาค่านอกเกณฑ์ เมื่อมีค่านอกเกณฑ์ปลอมปนอยู่ในชุดข้อมูลจะมีผลต่อความแม่นยำที่ลดลงของการประมาณพารามิเตอร์โดยวิธีกำลังสองน้อยที่สุด ดังนั้นการวิเคราะห์การถดถอยที่มีความแกร่ง (Robust) จึงเป็นอีกทางเลือกหนึ่งในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ซึ่งมีผู้คิดค้นวิธีการวิเคราะห์การถดถอยที่มีความแกร่งไว้อย่างมากมาย ดังต่อไปนี้ ในปี ค.ศ. 1964 Huber ได้เสนอวิธี M (M Method) ซึ่งวิธีนี้จะนำฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกอย่างเหมาะสมมาแทนที่ค่าคลาดเคลื่อนกำลังสองที่เรานิยมใช้กันในวิธีกำลังสองน้อยที่สุดและวิธี M สามารถหาตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณได้หลายแบบขึ้นอยู่กับฟังก์ชันถ่วงน้ำหนักที่แตกต่างกัน เช่น ฟังก์ชันถ่วงน้ำหนักของ Andrews, Cauchy, Tukey, Huber, Fair, Ramsay และ Welsch (Holland & Welsch, 1977) เป็นต้น ในที่นี้ผู้วิจัยสนใจศึกษาฟังก์ชันถ่วงน้ำหนักของ Andrews และ Welsch เนื่องจากเมื่อเกิดค่านอกเกณฑ์ของแต่ละขนาดตัวอย่างฟังก์ชันถ่วงน้ำหนักของ Andrews เป็นวิธีที่มีประสิทธิภาพมากกว่าฟังก์ชันถ่วงน้ำหนักอื่น ๆ (Hadara, 2006) และฟังก์ชันถ่วงน้ำหนักของ Welsch ยังไม่มีผู้วิจัยท่านใดเคยศึกษาใช้กับวิธี M มาก่อน ต่อมาในปี ค.ศ. 1984 Rousseeuw ได้เสนอวิธี LTS (Least Trimmed Squares Method) ซึ่งเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีหลักการคำนวณเช่นเดียวกับวิธีกำลังสองน้อยที่สุดแต่วิธีนี้จะเลือกชุดข้อมูลที่มีค่าคลาดเคลื่อนที่ใกล้เคียงกันมาวิเคราะห์จำนวนหนึ่งเท่านั้น และในปี ค.ศ. 1996 Simpona and Montgomery ได้พัฒนาวิธี M มาเป็นวิธี GM (Generalized M Method) โดยทำการปรับฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกอย่างเหมาะสม และในงานวิจัยนี้ได้เลือกใช้ฟังก์ชันถ่วงน้ำหนักของ Huber (Holland & Welsch, 1977) สำหรับการคำนวณโดยวิธี GM เนื่องจากเมื่อมีอัตราการเกิดค่านอกเกณฑ์ในแต่ละขนาดตัวอย่าง พบว่าการใช้ฟังก์ชันถ่วงน้ำหนักของ Huber สำหรับวิธี GM เป็นวิธีที่มีประสิทธิภาพดีที่สุด (Tantrakul, 2012)

จากการศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งดังกล่าวข้างต้นนี้ ผู้วิจัยมีความสนใจที่จะศึกษาเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณที่มีตัวแปรอิสระ 2 ตัว เมื่อข้อมูลไม่มีค่านอกเกณฑ์ในตัวแปรอิสระและตัวแปรตาม และเมื่อข้อมูลมีค่านอกเกณฑ์ระดับไม่รุนแรงในตัวแปรตาม โดยใช้วิธี OLS วิธี LTS วิธี M เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews (M-Andrews) และฟังก์ชันถ่วงน้ำหนักของ Welsch (M-Welsch)และ วิธี GM เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Huber ภายใต้สถานการณ์แตกต่างกันรวมทั้งหมด 78 สถานการณ์ โดยใช้เกณฑ์ตาม

งานวิจัยของ Ozkale and Arican (2015) ในการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการประมาณกล่าวคือ พิจารณาจากค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ซึ่งวิธีที่ให้ค่า EMSE ต่ำที่สุดถือว่าเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยมีประสิทธิภาพมากที่สุด

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง (Simulation Research) ซึ่งจำลองข้อมูลทั้งหมด 78 สถานการณ์ ด้วยเทคนิคมอนติคาร์โล โดยใช้โปรแกรม R และทำซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ วิธีการดำเนินการวิจัยมี ดังนี้

1. สร้างข้อมูลตัวแปรอิสระ X_1 และ X_2 มีการแจกแจงเอกรูปต่อเนื่อง (Continuous Uniform Distribution) ตามรูปแบบ (Faris & Al-Amleh, 2016) ดังนี้

$$X_1 \sim U(-5,5)$$

$$X_2 \sim U(-3,3)$$

กำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีค่านอกเกณฑ์ ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์โดย ค่าของตัวแปรอิสระอยู่ในช่วง $[Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)]$

2. สร้างข้อมูลความคลาดเคลื่อนสุ่มตามกรณีการเกิดค่านอกเกณฑ์ในตัวแปรตาม ดังนี้

2.1 กรณีไม่มีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม ให้ความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติ (Normal Distribution) โดยมีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1

2.2 กรณีมีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม ให้ความคลาดเคลื่อนสุ่มมีการแจกแจง 2 แบบ คือ แบบที่ 1 การแจกแจงปรกติมาตรฐาน และแบบที่ 2 ความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที (t-Distribution) ที่องศาเสรี (df) ต่างกัน 5 ระดับ คือ 1, 3, 6, 10 และ 30

3. สร้างข้อมูลตัวแปรตาม (Y_i) โดยมีรูปแบบดังนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \text{เมื่อ } i = 1, 2, \ldots, n$$

หรือสามารถเขียนในรูปเมทริกซ์ได้ ดังนี้

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}_{n\times 3} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

หรือ $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

เมื่อ $\boldsymbol{\beta}$ คือเมทริกซ์ของค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ขนาด 3×1

$\mathbf{Y}$ คือเมทริกซ์ของตัวแปรตาม ขนาด $n \times 1$

$\mathbf{X}$ คือเมทริกซ์ของตัวแปรอิสระ ขนาด $n \times 3$

$\boldsymbol{\varepsilon}$ คือเมทริกซ์ของความคลาดเคลื่อนสุ่ม ขนาด $n \times 1$

โดย n คือขนาดตัวอย่างซึ่งเท่ากับ 10, 20, 30, 50, 100 และ 150 และ ε_i คือความคลาดเคลื่อนสุ่ม โดย $i = 1, 2, ..., n$ ในที่นี้ ε_i มีการแจกแจง 2 แบบตามกรณีการเกิดค่านอกเกณฑ์ในตัวแปรตาม กล่าวคือ กรณีไม่มีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตามจะกำหนดให้ความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติมาตรฐาน และกรณีมีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตามจะกำหนดให้ความคลาดเคลื่อนสุ่มมีการแจกแจงเช่นเดียวกับข้อ 2.2

เมื่อกำหนดให้ $\beta_0 = 1$, $\beta_1 = 1$ และ $\beta_2 = 1$ และกำหนดให้ตัวแปรตามมีค่านอกเกณฑ์ระดับไม่รุนแรง ตามเงื่อนไขของการตรวจสอบค่านอกเกณฑ์ของแผนภาพกล่อง คือ ค่าของตัวแปรตามอยู่ในช่วง $[Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$ หรือ $[Q_3 + 1.5(IQR), Q_3 + 3(IQR)]$ และมีร้อยละการเกิดค่านอกเกณฑ์ในตัวแปรตาม เท่ากับ 0, 10 และ 20 ของแต่ละขนาดตัวอย่าง (n) ตัวอย่างเช่น เมื่อขนาดตัวอย่างเท่ากับ 10 กรณีไม่มีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม ความคลาดเคลื่อนสุ่มจะมีการแจกแจงปรกติมาตรฐาน และร้อยละการเกิดค่านอกเกณฑ์ในตัวแปรตามเท่ากับ 0 ส่วนกรณีมีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม ความคลาดเคลื่อนสุ่มจะมีการแจกแจงเช่นเดียวกับข้อ 2.2 และเมื่อกำหนดให้ร้อยละการเกิดค่านอกเกณฑ์ในตัวแปรตามเท่ากับ 10 นั่นคือจะมีจำนวนค่านอกเกณฑ์ของตัวแปรตามเท่ากับ 1 ตัว เป็นต้น

4. ประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณทั้ง 5 วิธี ดังนี้

4.1 วิธีกำลังสองน้อยที่สุด มีหลักการ คือทำให้ผลบวกกำลังสองของความคลาดเคลื่อน (Sum Square Error : SSE) มีค่าน้อยที่สุด (Jitthavech, 2015) กล่าวคือ

$$\min_{\beta}(SSE) \text{ หรือ } \min_{\beta}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) \qquad (1)$$

เมื่อ $SSE = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$

และทำการหาอนุพันธ์ย่อยลำดับที่ 1 ของ SSE เทียบกับ $\hat{\boldsymbol{\beta}}$ แล้วกำหนดเท่ากับศูนย์จะได้

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

เมื่อ $\hat{\boldsymbol{\beta}}$ คือ เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอย ขนาด $(k+1)\times 1$ และ k คือจำนวนตัวแปรอิสระ

$\mathbf{Y}$ คือเมทริกซ์ของตัวแปรตาม ขนาด $n\times 1$

$\mathbf{X}$ คือเมทริกซ์ของตัวแปรอิสระ ขนาด $n\times(k+1)$

4.2 วิธี LTS ในปี ค.ศ. 1984 Rousseeuw ได้เสนอวิธี LTS ซึ่งมีหลักการคล้ายกับวิธีกำลังสองน้อยที่สุด แต่วิธี LTS จะตัดค่าคลาดเคลื่อนกำลังสองบางตัวที่ให้ค่าสูงเกินไปออกจากช่วงที่ต้องการ ซึ่งตัวที่กำหนดจำนวนค่าคลาดเคลื่อนกำลังสองที่ต้องการ คือ พิจารณาจากค่า h โดยที่ $h=(n/2)+((p+1)/2)$ และ p คือ จำนวนพารามิเตอร์ในสมการถดถอยเชิงเส้นพหุคูณ นั่นคือ การประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณโดยวิธี LTS สามารถหาได้โดยทำให้ผลบวกกำลังสองของความคลาดเคลื่อนของข้อมูลจำนวน h ค่าที่ถูกเลือกมีค่าน้อยที่สุดดังสมการที่ (2)

$$\min_{\beta}\sum_{i=1}^{h} e_{(i)}^{2} \tag{2}$$

โดยที่ $e_{(1)}^{2}\le e_{(2)}^{2}\le\ldots\le e_{(n)}^{2}$ คือ ค่าคลาดเคลื่อนกำลังสองที่ทำการเรียงลำดับจากค่าน้อยไปหาค่ามาก

4.3 วิธี M ในปี ค.ศ.1964 Huber ได้ศึกษาวิธีการหาตัวประมาณแบบแกร่งที่เรียกว่า วิธี M โดยมีพื้นฐานมาจากตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator) ซึ่งหลักการของวิธี M คือหาค่าประมาณสัมประสิทธิ์การถดถอยที่ทำให้ผลรวมของฟังก์ชัน ρ มีค่าน้อยที่สุด เมื่อ ρ เป็นฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกอย่างเหมาะสมเพื่อมาแทนที่ค่าคลาดเคลื่อนกำลังสอง และในปี ค.ศ. 2012 Montgomery, Peck, and Vining ได้ศึกษาวิธี M เพิ่มเติมจาก Huber กล่าวคือ

$$\min_{\beta}\sum_{i=1}^{n}\rho(e_i) = \min_{\beta}\sum_{i=1}^{n}\rho\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}\right) \tag{3}$$

เมื่อ $\hat{\boldsymbol{\beta}}$ เป็นเวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอย ขนาด $(k+1)\times 1$ และ k คือจำนวนตัวแปรอิสระ

$\mathbf{x}_i'$ เป็นเมทริกซ์ของตัวแปรอิสระในแถวที่ i ขนาด $1\times(k+1)$

y_i เป็นตัวแปรตามตัวที่ i

e_i เป็นค่าคลาดเคลื่อนของค่าสังเกตที่ i โดย $i=1, 2, \ldots, n$

เนื่องจาก ρ มีความสัมพันธ์กับฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) จึงทำการปรับเปลี่ยนค่าคลาดเคลื่อนในสมการ (3) ให้เป็นค่ามาตรฐาน โดยการหารด้วยตัวประมาณสเกลที่มีความแกร่ง (Robust Estimator of Scale) s ดังนั้นสมการ (3) จึงจัดรูปได้เป็นสมการ (4) ดังนี้

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = \min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{s}\right) \quad (4)$$

เมื่อ $s = \left(median_i \left| e_i - median_i (e_i) \right| \right) \Big/ 0.6745$

จากสมการที่ (4) หา $\hat{\boldsymbol{\beta}}$ ที่เหมาะสมโดยการหาอนุพันธ์บางส่วนของ $\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{s}\right)$ เทียบกับ $\hat{\boldsymbol{\beta}}$ แล้วกำหนดให้เท่ากับศูนย์จะได้

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{s}\right) x_{ij} \qquad \text{โดย} \quad j = 0, 1, \ldots, k \quad (5)$$

กำหนดให้ $\psi(z) = \frac{\partial}{\partial z}\rho(z)$ โดยที่ $z = \frac{e_i}{s}$ และ x_{ij} เป็นตัวแปรอิสระตัวที่ j ของค่าสังเกตตัวที่ i เมื่อมีตัวแปรอิสระทั้งหมด k ตัว

เนื่องจากสมการที่ (5) เป็นสมการแบบไม่เชิงเส้น (Nonlinear) ดังนั้นการแก้สมการจะอาศัยเทคนิค คือ วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักซ้ำหลายรอบ (Iteratively Reweight Least Square : IRLS) เพื่อหาค่าฟังก์ชัน ψ ที่เหมาะสมที่สุด ซึ่งวิธี IRLS นี้ เริ่มต้นด้วยการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด เพื่อหาค่าคลาดเคลื่อนของค่าสังเกตที่ i (e_i) แล้วคำนวณหาค่า s และหาค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดที่ถูกถ่วงน้ำหนัก (Weight Least Square Method : WLS) ดังนั้นจากสมการที่ (5) สามารถแก้สมการได้ดังนี้

$$\sum_{i=1}^{n} x_{ij} w_i \left(y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}} \right) = \mathbf{0} \qquad \text{โดย} \quad j = 0, 1, \ldots, k \quad (6)$$

กำหนดให้ $w_i(z) = \frac{\psi(z)}{z}$

ดังนั้นตัวประมาณสัมประสิทธิ์การถดถอยของ $\boldsymbol{\beta}$ โดยวิธี WLS สามารถคำนวณได้ดังนี้

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

เมื่อ $\mathbf{W}$ เป็นเมทริกซ์ทแยงมุม (Diagonal Matrix) ขนาด $n \times n$ ที่มี $w_i(z)$ เป็นสมาชิกในแนวทแยงมุม เมื่อ $i = 1, 2, \ldots, n$ โดยที่ $w_i(z)$ เป็นฟังก์ชันถ่วงน้ำหนักตัวที่ i โดยในงานวิจัยนี้ วิธี M ได้ใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews และ Welsch ซึ่งปรากฏในสมการที่ (7) และ (8) ตามลำดับ

ในการคำนวณด้วยวิธี IRLS นี้จะหยุดทำการคำนวณเมื่อค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ $m-1$ กับค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ m มีค่าแตกต่างกันไม่เกิน 0.001 นั่นคือจะได้ตัวประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี M

สามารถเขียนสรุปขั้นตอนการประมาณค่าสัมประสิทธิ์การถดถอย $\boldsymbol{\beta}$ ด้วยวิธี IRLS ได้ดังนี้

ขั้นตอนที่ 1 คำนวณค่าประมาณสัมประสิทธิ์การถดถอยเริ่มต้นของ $\boldsymbol{\beta}$ หรือ $\hat{\boldsymbol{\beta}}^{(0)}$ จากวิธีกำลังสองน้อยที่สุด นั่นคือ

$$\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

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

$$e_i^{(1)} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)} \qquad \text{โดยที่} \quad i = 1, 2, \ldots, n$$

$$s^{(1)} = \left(\underset{i}{median} \left| e_i^{(1)} - \underset{i}{median}\left(e_i^{(1)}\right) \right| \right) / 0.6745$$

ขั้นตอนที่ 3 นำค่า $e_i^{(1)}$ และ $s^{(1)}$ จากขั้นตอนที่ 2 มาคำนวณ z ในรอบที่ 1 กล่าวคือ $z^{(1)} = \dfrac{e_i^{(1)}}{s^{(1)}}$ เพื่อหาน้ำหนักของแต่ละค่าสังเกตตามเงื่อนไขของเกณฑ์ความแกร่งหรือแบบฟังก์ชัน ρ และฟังก์ชัน ψ ที่เลือก โดยงานวิจัยครั้งนี้ได้ใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews มีการกำหนดฟังก์ชัน $\rho(z)$ ฟังก์ชัน $\psi(z)$ และฟังก์ชันถ่วงน้ำหนัก $w(z)$ ดังนี้

เกณฑ์ความแกร่งของ Andrews (Montgomery, peck, & Vining, 2012) กำหนดดังนี้

$$\rho(z) = \begin{cases} c\left[1 - \cos\left(\dfrac{z}{c}\right)\right] & \text{เมื่อ} \quad |z| \le c\pi \\ 2c & \text{เมื่อ} \quad |z| > c\pi \end{cases}$$

ให้ $\psi(z) = \dfrac{\partial}{\partial z}\rho(z)$ จะได้ฟังก์ชัน $\psi(z)$ ดังนี้

$$\psi(z) = \begin{cases} \sin\left(\dfrac{z}{c}\right) & \text{เมื่อ} \quad |z| \le c\pi \\ 0 & \text{เมื่อ} \quad |z| > c\pi \end{cases}$$

ให้ $w(z) = \dfrac{\psi(z)}{\left(\dfrac{z}{c}\right)}$ จะได้ฟังก์ชัน $w(z)$ ดังนี้

$$w(z) = \begin{cases} \dfrac{\sin\left(\dfrac{z}{c}\right)}{\left(\dfrac{z}{c}\right)} & \text{เมื่อ} \quad |z| \le c\pi \\ 0 & \text{เมื่อ} \quad |z| > c\pi \end{cases} \qquad (7)$$

เมื่อ z คือค่าคลาดเคลื่อนมาตรฐานที่คำนวณได้จาก $z = \frac{e_i}{s}$ และ c มีค่าเท่ากับ 1.339 กล่าวคือ เมื่อค่าคลาดเคลื่อนมาตรฐานมีการแจกแจงปรกติแล้ว ตัวประมาณวิธี M ที่ใช้เกณฑ์ความแกร่งของ Andrews โดยกำหนดค่า c ดังกล่าวนี้ จะมีประสิทธิภาพ 95% เมื่อเทียบกับตัวประมาณวิธี OLS (Holland & Welsch, 1977)

ขั้นตอนที่ 4 จากน้ำหนักของแต่ละค่าสังเกตที่ได้จากขั้นตอนที่ 3 นำมาคำนวณค่าประมาณสัมประสิทธิ์การถดถอยของ $\boldsymbol{\beta}$ ในรอบที่ 1 ดังนี้ $\hat{\boldsymbol{\beta}}^{(1)} = \left(\mathbf{X}'\mathbf{W}^{(1)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(1)}\mathbf{Y}$

โดยที่ $\mathbf{W}^{(1)}$ เป็นเมทริกซ์ทแยงมุมของน้ำหนักในรอบที่ 1 ขนาด $n \times n$

$$\mathbf{W}^{(1)} = \begin{bmatrix} w_1^{(1)}(z) & 0 & \cdots & 0 \\ 0 & w_2^{(1)}(z) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n^{(1)}(z) \end{bmatrix}_{n\times n}$$

เมื่อ $w_i(z)$ เป็นสมาชิกในแนวทแยงมุม เมื่อ $i = 1, 2, ..., n$ โดยที่ $w_i(z)$ เป็นฟังก์ชันถ่วงน้ำหนัก ในงานวิจัยนี้วิธี M ได้ใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews และ Welsch ซึ่งปรากฏในสมการที่ (7) และ (8) ตามลำดับ

ขั้นตอนที่ 5 หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยเริ่มต้นกับค่าประมาณสัมประสิทธิ์การถดถอยรอบที่ 1 นั่นคือ $\left|\hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(0)}\right|$

ขั้นตอนที่ 6 ถ้า $\left|\hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(0)}\right|$ จากขั้นตอนที่ 5 มีค่ามากกว่า 0.001 ให้ไปทำซ้ำในรอบที่ m ในขั้นตอนที่ 7 ต่อไป แต่ถ้า $\left|\hat{\boldsymbol{\beta}}^{(1)} - \hat{\boldsymbol{\beta}}^{(0)}\right|$ จากขั้นตอนที่ 5 พบว่าค่าประมาณสัมประสิทธิ์การถดถอยทุกค่าไม่มากกว่า 0.001 นั่นคือจะได้ค่าประมาณสัมประสิทธิ์การถดถอยจากรอบที่ 1 และสิ้นสุดการคำนวณ

ขั้นตอนที่ 7 การทำซ้ำรอบที่ m เมื่อ $m = 2,3,...$ เพื่อหาค่า $e_i^{(m)}$ และ $s^{(m)}$ จาก

$$e_i^{(m)} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(m-1)} \qquad \text{โดยที่} \quad i = 1, 2, ..., n$$

$$s^{(m)} = \left(\underset{i}{median}\left|e_i^{(m)} - \underset{i}{median}\left(e_i^{(m)}\right)\right|\right) / 0.6745$$

และคำนวณหาค่าประมาณสัมประสิทธิ์การถดถอยในการทำซ้ำรอบที่ m จาก

$$\hat{\boldsymbol{\beta}}^{(m)} = \left(\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(m)}\mathbf{Y}$$

โดยที่ $\mathbf{W}^{(m)}$ เป็นเมทริกซ์ทแยงมุมของน้ำหนักในรอบที่ m ขนาด $n \times n$

ขั้นตอนที่ 8 หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ $m-1$ กับรอบที่ m ของค่าประมาณสัมประสิทธิ์การถดถอยทุกตัว ถ้ามีค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ $m-1$ กับรอบที่ m อย่างน้อยหนึ่งค่าที่มากกว่า 0.001 ให้กลับไปทำซ้ำในขั้นตอนที่ 7 ทำเช่นนี้จนกระทั่งค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยรอบที่ $m-1$ กับรอบที่ m ของค่าประมาณสัมประสิทธิ์การถดถอยทุกตัวมีค่าไม่มากกว่า 0.001 นั่นคือจะได้ค่าประมาณสัมประสิทธิ์การถดถอยจากรอบที่ m

4.4 วิธี M โดยใช้ฟังก์ชันถ่วงน้ำหนักของ Welsch ซึ่งมีขั้นตอนการดำเนินการเช่นเดียวกับวิธี M ในข้อ 4.3 ซึ่งเกณฑ์ความแกร่งของ Welsch (Holland & Welsch, 1977) กำหนดดังนี้

$$\rho(z)=\left(\frac{c^2}{2}\right)\left(1-\exp\left(-\left(\frac{z}{c}\right)^2\right)\right)$$

ให้ $\psi(z)=\frac{\partial}{\partial z}\rho(z)$ จะได้ฟังก์ชัน $\psi(z)$ ดังนี้

$$\psi(z)=z\exp\left(-\left(\frac{z}{c}\right)^2\right)$$

ให้ $w(z)=\frac{\psi(z)}{z}$ จะได้ฟังก์ชัน $w(z)$ ดังนี้

$$w(z)=\exp\left(-\left(\frac{z}{c}\right)^2\right) \tag{8}$$

เมื่อ z คือค่าคลาดเคลื่อนมาตรฐานที่คำนวณได้จาก $z=\frac{e_i}{s}$ และ c มีค่าเท่ากับ 2.985 กล่าวคือ เมื่อค่าคลาดเคลื่อนมาตรฐานมีการแจกแจงปรกติแล้วตัวประมาณวิธี M ที่ใช้เกณฑ์ความแกร่งของ Welsch โดยกำหนดค่า c ดังกล่าวนี้ จะมีประสิทธิภาพ 95% เมื่อเทียบกับตัวประมาณวิธี OLS (Holland & Welsch, 1977)

4.5 วิธี GM ในปี ค.ศ. 1996 Simpsona and Montgomery ได้พัฒนาวิธี GM มาจากวิธี M โดยทำการถ่วงน้ำหนักเพิ่มเข้าไปในฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกให้เหมาะสม (ρ) โดยมีหลักการ คือ หาตัวประมาณค่าสัมประสิทธิ์การถดถอยที่ทำให้ผลรวมของฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกให้เหมาะสม (ρ) มีค่าน้อยที่สุด คือ

$$\min_{\beta}\sum_{i=1}^{n}\rho\left(\frac{y_i-\mathbf{X}_i'\hat{\boldsymbol{\beta}}}{\pi_i s}\right)\pi_i \tag{9}$$

เมื่อ ρ เป็นฟังก์ชันของค่าคลาดเคลื่อนที่ถูกเลือกให้เหมาะสม

π_i เป็นค่าถ่วงน้ำหนักซึ่งสามารถคำนวณจาก $\pi_i = \frac{median_i |z_i|}{z_i}$ เมื่อ $z_i = \left(\frac{e_i}{s}\right)$

e_i เป็นค่าคลาดเคลื่อนของค่าสังเกตที่ i โดย $i = 1, 2, ..., n$

และ s เป็นตัวประมาณสเกลที่มีความแกร่ง ในที่นี้ใช้ s แทนมัธยฐานส่วนเบี่ยงเบนสัมบูรณ์ ซึ่งสามารถหาค่า s ได้ดังนี้

$$s = \left(median_i \left| e_i - median_i (e_i) \right|\right) / 0.6745$$

จากสมการที่ (9) หา $\hat{\boldsymbol{\beta}}$ ที่เหมาะสมโดยการหาอนุพันธ์บางส่วนของ $\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{\pi_i s}\right)\pi_i$ เทียบกับ $\hat{\boldsymbol{\beta}}$ แล้วกำหนดให้เท่ากับศูนย์ จะได้

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}}{\pi_i s}\right)\pi_i \mathbf{X}_i = 0 \qquad (10)$$

เมื่อ $\psi(u) = \frac{\partial}{\partial u}\rho(u)$ และกำหนดให้ $u = \left(\frac{e_i}{\pi_i s}\right)$

จากสมการที่ (10) ทำการแก้สมการได้โดยวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักซ้ำหลายรอบ (IRLS) ซึ่งมีหลักการคล้ายกับวิธี M เพื่อหาค่าฟังก์ชัน ψ ที่เหมาะสมที่สุด ซึ่งวิธี IRLS นี้ เริ่มต้นด้วยการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด เพื่อหาค่าคลาดเคลื่อนของค่าสังเกตที่ i (e_i) แล้วคำนวณหาค่า s และหาค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดที่ถูกถ่วงน้ำหนัก (WLS) ดังนั้นจากสมการที่ (10) จะได้ตัวประมาณของสัมประสิทธิ์การถดถอย $\boldsymbol{\beta}$ ด้วยวิธี WLS ดังนี้

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

เมื่อ $\mathbf{W}$ เป็นเมทริกซ์ทแยงมุม (Diagonal Matrix) ขนาด $n \times n$ ที่มี $w_i(u)$ เป็นสมาชิกในแนวทแยงมุม เมื่อ $i = 1, 2, ..., n$ โดยที่ $w_i(u)$ เป็นฟังก์ชันถ่วงน้ำหนักตัวที่ i โดยในงานวิจัยนี้วิธี GM ได้ใช้ฟังก์ชันถ่วงน้ำหนักของ Huber ซึ่งปรากฏในสมการที่ (11)

ซึ่งในการคำนวณด้วยวิธี IRLS นี้จะหยุดทำการคำนวณเมื่อค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ $m-1$ กับค่าประมาณสัมประสิทธิ์การถดถอยในรอบที่ m มีค่าแตกต่างกันไม่เกิน 0.001 นั่นคือจะได้ตัวประมาณสัมประสิทธิ์ของวิธี GM

โดยงานวิจัยครั้งนี้ได้ใช้ฟังก์ชันถ่วงน้ำหนักของ Huber (Holland & Welsch, 1977) ซึ่งมีการกำหนดฟังก์ชัน $\rho(u)$ ฟังก์ชัน $\psi(u)$ และฟังก์ชันถ่วงน้ำหนัก $w(u)$ ดังนี้

$$\rho(u) = \begin{cases} \frac{u^2}{2} & \text{เมื่อ} \quad |u| \le c \\ c|u| - \frac{c^2}{2} & \text{เมื่อ} \quad |u| > c \end{cases}$$

กำหนดให้ $\psi(u) = \frac{\partial}{\partial u}\rho(u)$ จะได้ฟังก์ชัน y (u) ดังนี้

$$\psi(u) = \begin{cases} -c & \text{เมื่อ} \quad u < -c \\ u & \text{เมื่อ} \quad -c > u > c \\ c & \text{เมื่อ} \quad u > c \end{cases}$$

กำหนดให้ $w(u) = \frac{\psi(u)}{u}$ จะได้ฟังก์ชัน $w(u)$ ดังนี้

$$w(u) = \begin{cases} \frac{-c}{u} & \text{เมื่อ} \quad u < -c \\ 1 & \text{เมื่อ} \quad -c > u > c \\ \frac{c}{u} & \text{เมื่อ} \quad u > c \end{cases} \qquad (11)$$

เมื่อ u คือค่าคลาดเคลื่อนมาตรฐานที่คำนวณได้จาก $u = \frac{e_i}{\pi_i s}$ และ c มีค่าเท่ากับ 1.345 กล่าวคือ เมื่อค่าคลาดเคลื่อนมาตรฐานมีการแจกแจงปรกติแล้ว ตัวประมาณวิธี GM ที่ใช้เกณฑ์ความแกร่งของ Huber โดยกำหนดค่า c ดังกล่าวนี้ จะมีประสิทธิภาพ 95% เมื่อเทียบกับตัวประมาณวิธี OLS (Holland & Welsch, 1977)

5. คำนวณค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยของ $\hat{\boldsymbol{\beta}}$ โดยใช้เกณฑ์ตามงานวิจัยของ Ozkale and Arican (2015) สำหรับการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 5 วิธี ดังนี้

$$\text{EMSE} = \frac{1}{1,000}\sum_{t=1}^{1,000}\sum_{j=0}^{p-1}\left(\hat{\beta}_{j(t)} - \beta_j\right)^2$$

เมื่อ β_j คือค่าสัมประสิทธิ์การถดถอยที่กำหนด ตัวที่ j โดยที่ $j = 0, 1, \ldots, p-1$

$\hat{\beta}_j$ คือค่าของตัวประมาณสัมประสิทธิ์การถดถอย ตัวที่ j

p คือจำนวนพารามิเตอร์ในสมการถดถอยเชิงเส้นพหุคูณ ในที่นี้ $p = 3$

ผลการวิจัยและวิจารณ์ผล

ตารางที่ 1 ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยทั้ง 5 วิธี เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติมาตรฐาน

ขนาดตัวอย่าง (n)	ร้อยละการเกิดค่านอกเกณฑ์	วิธี				
		OLS	LTS	M-Andrews	M-Welsch	GM
10	0	0.18930**	0.25940	0.22479	0.21842	0.27502
	10	5.80748	0.58298	0.36664	0.35326**	0.37514
	20	7.39896	2.12849	2.38502	2.41691	1.66495**
20	0	0.07805**	0.10808	0.08565	0.08443	0.12084
	10	2.50150	0.37260	0.10801	0.10645**	0.13796
	20	6.15123	0.46455	0.31784	0.30625	0.17841**
30	0	0.05307**	0.06938	0.05654	0.05587	0.08013
	10	1.66124	0.34191	0.05872	0.05869**	0.07944
	20	3.97920	0.29111	0.07791	0.07754**	0.08450
50	0	0.02982**	0.03975	0.03343	0.03312	0.04674
	10	1.30758	0.33046	0.03579**	0.03579**	0.04571
	20	2.68233	0.25439	0.04109**	0.04172	0.04906
100	0	0.01456**	0.01802	0.01510	0.01505	0.01967
	10	0.91249	0.29418	0.01700**	0.01710	0.02197
	20	1.59080	0.22219	0.01987**	0.02002	0.02383
150	0	0.00964**	0.01233	0.01020	0.01017	0.01349
	10	0.79223	0.27624	0.01159**	0.01163	0.01501
	20	1.05783	0.20250	0.01234**	0.01237	0.01496

หมายเหตุ ** หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ตารางที่ 2 ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยทั้ง 5 วิธี เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่ระดับองศาเสรี (df) เท่ากับ 1, 3, 6, 10 และ 30

องศาเสรี (df)	ขนาดตัวอย่าง (n)	ร้อยละการเกิดค่านอกเกณฑ์	วิธี				
			OLS	LTS	M-Andrews	M-Welsch	GM
1	10	10	13.72602	6.94997	4.37706	4.37384	4.29153**
		20	14.49435	6.73331	9.53247	8.90270	5.62580**
	20	10	4.56518	1.36057	0.45580	0.44413	0.31104**
		20	8.23472	1.64072	1.49184	1.40655	0.48114**
	30	10	3.29677	0.86382	0.23478	0.23817	0.17676**
		20	5.56974	1.08093	0.42964	0.46045	0.20862**
	50	10	2.34864	0.63513	0.12056	0.12049	0.08933**
		20	3.94535	0.68094	0.16968	0.17955	0.10370**
	100	10	1.65969	0.43695	0.04894	0.04875	0.03821**
		20	2.46151	0.41967	0.06871	0.07080	0.04997**
	150	10	1.58732	0.42132	0.03360	0.03396	0.02705**
		20	1.62599	0.29894	0.04266	0.04536	0.02941**
3	10	10	6.18936	1.20100	0.76998	0.73134	0.64208**
		20	8.55811	2.88635	3.30387	3.10518	2.25293**
	20	10	2.88026	0.66619	0.16778	0.16626**	0.18984
		20	6.92107	0.94834	0.47047	0.48650	0.23340**
	30	10	1.97905	0.53034	0.09777	0.09658**	0.10622
		20	4.43030	0.50747	0.12400	0.12241	0.11232**
	50	10	1.50608	0.46169	0.05806	0.05789**	0.05807
		20	2.93502	0.38922	0.06637	0.06724	0.06441**
	100	10	1.03945	0.34767	0.02666	0.02639	0.02630**
		20	1.66573	0.31673	0.03197	0.03209	0.03113**
	150	10	0.89803	0.31137	0.01757	0.01746	0.01728**
		20	1.14767	0.25834	0.02011	0.02012	0.01941**

หมายเหตุ ** หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ตารางที่ 2 (ต่อ) ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยทั้ง 5 วิธี เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่ระดับองศาเสรี (df) เท่ากับ 1, 3, 6, 10 และ 30

องศาเสรี (df)	ขนาดตัวอย่าง (n)	ร้อยละการเกิดค่านอกเกณฑ์	วิธี				
			OLS	LTS	M-Andrews	M-Welsch	GM
6	10	10	5.69069	0.82636	0.49352	0.46135**	0.48875
		20	7.16199	1.96808	2.87019	2.65510	1.78535**
	20	10	2.71310	0.50423	0.12277	0.12130**	0.15255
		20	6.21710	0.76607	0.46586	0.46314	0.19095**
	30	10	1.79232	0.42118	0.07964	0.07893**	0.09416
		20	4.37417	0.40178	0.09379**	0.09663	0.09848
	50	10	1.38964	0.39461	0.04323**	0.04324	0.04999
		20	2.91680	0.32846	0.05254**	0.05289	0.05362
	100	10	0.94707	0.31827	0.02064	0.02049**	0.02186
		20	1.56692	0.27923	0.02483**	0.02530	0.02712
	150	10	0.82285	0.29658	0.01412**	0.01413	0.01590
		20	1.14576	0.23761	0.01670**	0.01674	0.01796
10	10	10	5.36813	0.66744	0.42460	0.40108**	0.42027
		20	8.15881	2.37220	3.96939	3.74980	1.61733**
	20	10	2.59439	0.46770	0.10836	0.10703**	0.13788
		20	5.91810	0.56076	0.34218	0.33279	0.22982**
	30	10	1.86902	0.42997	0.06572	0.06518**	0.08236
		20	3.95488	0.33587	0.08303**	0.08426	0.09236
	50	10	1.24225	0.35608	0.04216	0.04202**	0.05253
		20	2.96959	0.32178	0.04653**	0.04690	0.05395
	100	10	0.98309	0.31963	0.01859**	0.01862	0.02217
		20	1.55426	0.25299	0.02365**	0.02385	0.02556
	150	10	0.85873	0.30177	0.01349	0.01343**	0.01567
		20	1.08415	0.22858	0.01535**	0.01548	0.01693

หมายเหตุ ** หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ตารางที่ 2 (ต่อ) ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยทั้ง 5 วิธี เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่ระดับองศาเสรี (df) เท่ากับ 1, 3, 6, 10 และ 30

องศาเสรี (df)	ขนาดตัวอย่าง (n)	ร้อยละการเกิดค่านอกเกณฑ์	วิธี				
			OLS	LTS	M-Andrews	M-Welsch	GM
30	10	10	5.14800	0.60731	0.37335	0.36104**	0.38538
		20	7.52527	1.65292	2.07091	1.82984	1.49092**
	20	10	2.61156	0.42145	0.10343	0.10203**	0.13354
		20	6.46003	0.47378	0.27057	0.27267	0.14592**
	30	10	1.73819	0.36534	0.06755	0.06702**	0.08752
		20	4.13175	0.32807	0.07415**	0.07874	0.08715
	50	10	1.27780	0.33927	0.03659	0.03649**	0.04625
		20	2.88672	0.29061	0.04418**	0.04441	0.05210
	100	10	0.86603	0.28384	0.01844**	0.01847	0.02313
		20	1.55114	0.24129	0.01970**	0.01982	0.02333
	150	10	0.83284	0.28415	0.01180**	0.01183	0.01481
		20	1.01302	0.21233	0.01287**	0.01303	0.01525

หมายเหตุ ** หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ผลการวิจัยสรุปตารางที่ 1 และ 2 ได้ดังนี้

1. กรณีความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติมาตรฐาน จากตารางที่ 1 พบว่าสำหรับทุกขนาดตัวอย่างเมื่อไม่มีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม วิธีกำลังสองน้อยที่สุด ให้ค่า EMSE ต่ำที่สุด แต่เมื่อมีค่านอกเกณฑ์เกิดขึ้นในตัวแปรตาม พบว่า วิธี M-Welsch และ M-Andrews ให้ค่า EMSE ที่ใกล้เคียงกัน สำหรับขนาดตัวอย่าง 10 และ 20 เมื่อเกิดร้อยละค่านอกเกณฑ์เท่ากับ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด และเมื่อเกิดร้อยละค่านอกเกณฑ์เท่ากับ 20 วิธี GM ให้ค่า EMSE ต่ำที่สุด สำหรับขนาดตัวอย่าง 30 เมื่อเกิดร้อยละค่านอกเกณฑ์ 10 และ 20 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด ที่ขนาดตัวอย่าง 50 ร้อยละการเกิดค่านอกเกณฑ์เท่ากับ 10 พบว่าวิธี M-Welsch และ M-Andrews ให้ค่า EMSE ต่ำที่สุดทั้งสองวิธี และที่ขนาดตัวอย่าง 100 และ 150 พบว่าวิธี M-Andrews ให้ค่าEMSE ต่ำที่สุดในทุกร้อยละการเกิดค่านอกเกณฑ์

2. กรณีความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที จากตารางที่ 2 พบว่า องศาเสรีที่แตกต่างกันของการแจกแจงแบบที มีผลต่อค่า EMSE กล่าวคือเมื่อองศาเสรีเพิ่มขึ้น ค่า EMSE มีแนวโน้มลดลง เมื่อองศาเสรีเท่ากับ 1 พบว่าทุกขนาดตัวอย่างและทุกร้อยละการเกิดค่านอกเกณฑ์ วิธี GM ให้ค่า EMSE ต่ำที่สุด สำหรับองศาเสรีเท่ากับ 3 พบว่าเกือบทุกขนาดตัวอย่างและ

ทุกร้อยละการเกิดค่านอกเกณฑ์ วิธี GM ให้ค่า EMSE ต่ำที่สุด ยกเว้นที่ขนาดตัวอย่างเท่ากับ 20 30 และ 50 ที่ร้อยละการเกิดค่านอกเกณฑ์ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด สำหรับองศาเสรีเท่ากับ 6 เมื่อขนาดตัวอย่างเท่ากับ 10 และ 20 ร้อยละการเกิดค่านอกเกณฑ์ที่ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด ส่วนร้อยละการเกิดค่านอกเกณฑ์ที่ 20 วิธี GM ให้ค่า EMSE ต่ำที่สุด และที่ขนาดตัวอย่างเท่ากับ 30 50 100 และ150 เกือบทุกร้อยละการเกิดค่านอกเกณฑ์ วิธี M-Andrews ให้ค่า EMSE ต่ำที่สุด ยกเว้นที่ขนาดตัวอย่าง 30 และ 100 ร้อยละการเกิดค่านอกเกณฑ์ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด และที่องศาเสรีเท่ากับ 10 และ 30 ค่า EMSE มีแนวโน้มไปในทิศทางเดียวกัน กล่าวคือที่ขนาดตัวอย่างเท่ากับ 10, 20,30 และ 50 ร้อยละการเกิดค่านอกเกณฑ์ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด ส่วนที่ร้อยละการเกิดค่านอกเกณฑ์ 20 ของขนาดตัวอย่างเท่ากับ 10 และ 20 วิธี GM ให้ค่า EMSE ต่ำที่สุด แต่ที่ขนาดตัวอย่าง 30 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 100 และ 150 วิธี M-Andrews ให้ค่า EMSE ต่ำที่สุด ในเกือบทุกระดับของร้อยละการเกิดค่านอกเกณฑ์ ยกเว้นที่ขนาดตัวอย่างเท่ากับ 150 ร้อยละการเกิดค่านอกเกณฑ์ 10 ขององศาเสรีเท่ากับ 10 วิธี M-Welsch ให้ค่า EMSE ต่ำที่สุด

3. ขนาดตัวอย่างมีผลต่อค่า EMSE กล่าวคือ เมื่อขนาดตัวอย่างเพิ่มมากขึ้น ค่า EMSE มีแนวโน้มลดลง ในเกือบทุกสถานการณ์ ยกเว้นความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที องศาเสรีเท่ากับ 30 ขนาดตัวอย่าง 150 ของวิธี LTS

4. จากการศึกษาการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ และการจำลองข้อมูลในสถานการณ์ต่าง ๆ พบว่า เมื่อร้อยละการเกิดค่านอกเกณฑ์เพิ่มขึ้นจาก 10 เป็น 20 สำหรับความคลาดเคลื่อนสุ่มที่มีการแจกแจงปรกติมาตรฐาน พบว่า วิธี M เมื่อใช้ฟังก์ชันถ่วงน้ำหนักของ Andrews มีแนวโน้มให้ค่า EMSE ต่ำที่สุด ซึ่งมีความสอดคล้องกับงานวิจัยของ Hadara (2006) นอกจากนี้เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่องศาเสรีเท่ากับ 1 พบว่าวิธี GM ให้ค่า EMSE ต่ำที่สุด มีความสอดคล้องกับงานวิจัยของ Tantrakul (2012)

สรุปผลการวิจัย

ตารางที่ 3 วิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ที่ให้ค่า EMSE ต่ำที่สุดในสถานการณ์ต่าง ๆ เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติมาตรฐาน

ขนาดตัวอย่าง (n)	ร้อยละการเกิดค่านอกเกณฑ์		
	0	10	20
10	OLS	M-W	GM
20	OLS	M-W	GM
30	OLS	M-W	M-W
50	OLS	M-W, M-A	M-A
100	OLS	M-A	M-A
150	OLS	M-A	M-A

หมายเหตุ M-A แทนวิธี M-Andrews และ M-W แทนวิธี M-Welsch

ตารางที่ 4 วิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณ ที่ให้ค่า EMSE ต่ำที่สุดในสถานการณ์ต่าง ๆ เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่องศาเสรี (df) ต่างกัน 5 ระดับ

	ร้อยละการเกิดค่านอกเกณฑ์							
	10					20		
df / n	1	3	6	10	30	1	3	6, 10, 30
10	GM	GM	M-W	M-W	M-W	GM	GM	GM
20	GM	M-W	M-W	M-W	M-W	GM	GM	GM
30	GM	M-W	M-W	M-W	M-W	GM	GM	M-A
50	GM	M-W	M-A	M-W	M-A	GM	GM	M-A
100	GM	GM	M-W	M-A	M-A	GM	GM	M-A
150	GM	GM	M-A	M-W	M-A	GM	GM	M-A

หมายเหตุ M-A แทนวิธี M-Andrews และ M-W แทนวิธี M-Welsch

ในการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับการถดถอยเชิงเส้นพหุคูณ เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงปรกติ จากตารางที่ 3 พบว่า เมื่อข้อมูลไม่มีค่านอกเกณฑ์ในตัวแปรอิสระและตัวแปรตาม วิธี OLS มีประสิทธิภาพสูงที่สุด อย่างไรก็ตามเมื่อข้อมูลมีค่านอกเกณฑ์ระดับไม่รุนแรง ที่ขนาดตัวอย่าง 10, 20 และ 30 ร้อยละการเกิดค่านอกเกณฑ์ 10 พบว่า วิธี M-Welsch มีประสิทธิภาพสูงที่สุด ส่วนที่ร้อยละการเกิดค่านอกเกณฑ์ 20 ขนาดตัวอย่างเท่ากับ 10 และ 20 วิธี GM มีประสิทธิภาพสูงที่สุด นอกจากนี้เมื่อข้อมูลมีค่านอกเกณฑ์ในตัวแปรตามในทุกระดับร้อยละการเกิดค่านอกเกณฑ์และขนาดตัวอย่างใหญ่ ($n \geq 50$) วิธี M-Andrews มีแนวโน้มให้ประสิทธิภาพสูงที่สุด

จากตารางที่ 4 เมื่อความคลาดเคลื่อนสุ่มมีการแจกแจงแบบที ที่องศาเสรีเท่ากับ 1 พบว่า วิธี GM ให้ประสิทธิภาพสูงที่สุดในทุกระดับของร้อยละการเกิดค่านอกเกณฑ์ และยังให้ประสิทธิภาพสูงที่สุดเมื่อองศาเสรีเท่ากับ 3 ที่ร้อยละการเกิดค่านอกเกณฑ์ 20 ในทุกขนาดตัวอย่าง นอกจากนี้เมื่อองศาเสรีมีแนวโน้มสูงขึ้น ($df = 6,10,30$) ที่ร้อยละการเกิดค่านอกเกณฑ์ 10 ขนาดตัวอย่างเท่ากับ 10, 20 และ 30 ส่วนใหญ่ วิธี M-Welsch มีแนวโน้มให้ประสิทธิภาพสูงที่สุด และเมื่อขนาดตัวอย่างเท่ากับ 50, 100 และ 150 พบว่าส่วนใหญ่วิธี M-Andrews มีแนวโน้มให้ประสิทธิภาพสูงที่สุด และสำหรับร้อยละการเกิดค่านอกเกณฑ์ 20 เมื่อองศาเสรีมีแนวโน้มสูงขึ้น ($df = 6,10,30$) พบว่าที่ขนาดตัวอย่างเท่ากับ 10 และ 20 วิธี GM มีแนวโน้มให้ประสิทธิภาพสูงที่สุด แต่เมื่อขนาดตัวอย่าง 30, 50, 100 และ 150 วิธี M-Andrews มีแนวโน้มให้ประสิทธิภาพสูงที่สุด

เอกสารอ้างอิง

Anscombe, F. J. (1960). Rejection of Outlier. *Journal of American Statistical Association and American Society for Quality*, *2*(2), 123-147.

Barnett, V., & Lewis, T. (1995). *Outlier in Statistical Data*, 3th Edition., New York: John Wiley and Sons.

Faris, M. A., & Al-Amleh, M. A. (2016). *A Comparison between Least Trimmed of Squares and MM-Estimation in Linear Regression Using Simulation Technique*. Retrieved September 8, 2017, from http://iacmc.zu.edu.jo/ar/images/stories/IACMC2016/39.

Hadara, P. (2006). *A Comparison of Methods for Estimation of Parameters in Multiple Linear Regression with Outliers. Master of Science Thesis*, Naresuan University. (in Thai).

Holland, P. W., & Welsch, R. E. (1977). Robust Regression Using Iteratively Reweight Least-Squares. *Communications in Statistics-Theory and Methods*, *6*(9), 813-827.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. Annals of Mathematical Statistics. *35*(1), 73-101

Jitthavech, J. (2015). *Regression Analysis*. Bangkok: WVO Officer of Printing Mull. (in Thai).

Montgomery D.C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, 5th Edition. New York: John Wiley and Sons.

Ozkale, M. R., & Arican, E. (2015). First-order r-d Class Estimator in Binary Logistic Regression Model. *Statistics and Probability Letters.* Retrieved March 12, 2015, from http://booksc.org/book/44176223/fa1f97.

Rousseeuw, P. J. (1984) Least Median of Squares Regression. *Journal of the American Statistics Associatio*n, *79*(388), 871-880.

Simpsona, J. R., & Montgomery, D. C. (1996). A Biased-Robust Regression Technique for the Combined Outlier-Multicolinearity Problem. *Journal of Statistical Computation and Simulation*, *56*(1), 1-22.

Tantrakul, O. (2012). *A Comparison of Robust Regression Coefficient Estimation Methods for Multiple Linear Regression with Outliers. Master of Science Thesis*, Kasetsart University. (in Thai).