การประมาณค่าที่แกร่งโดยใช้การประมาณค่า M และการประมาณค่า S

Robust Estimation Using M Estimation and S Estimation

นิธิภัทร กมลสุข[1*]
Nithipat Kamolsuk[1*]

บทคัดย่อ

ในการวิเคราะห์การถดถอยเชิงเส้น วิธีกำลังสองน้อยที่สุดเป็นวิธีการประมาณค่าพารามิเตอร์ที่ให้ค่าประมาณเป็นค่าไม่เอนเอียงดีที่สุด อย่างไรก็ตามเมื่อเกิดค่านอกเกณฑ์ขึ้นกับข้อมูลที่นำมาวิเคราะห์จะมีผลต่อค่าประมาณจากวิธีกำลังสองน้อยที่สุดนี้ จึงต้องใช้วิธีการอื่นคือ การวิเคราะห์การถดถอยที่แกร่ง ที่จะให้ค่าประมาณของพารามิเตอร์ที่เหมาะสมกับตัวแบบที่วิเคราะห์มากกว่า โดยบทความนี้จะแสดงวิธีการประมาณค่าหลายวิธี แต่จะมุ่งเน้นไปที่วิธีการประมาณค่า M และวิธีการประมาณค่า S

คำสำคัญ: การวิเคราะห์การถดถอยที่แกร่ง, วิธีการประมาณค่า M, วิธีการประมาณค่า S

Abstract

In linear regression analysis, the ordinary least squares (OLS) estimators of parameters have always turned out to be the best linear unbiased estimators. However, if the data contain outliers, this may affect the least-squares estimates. So, an alternative approach; the so-called robust regression methods, is needed to obtain a better fit of the model or more precise estimates of parameters. In this article, various robust regression methods have been reviewed. The focus is on the robust estimation using M estimation and S estimation.

Keywords: robust regression methods, M estimation, S estimation

[1] สำนักการศึกษาทั่วไป สถาบันการจัดการปัญญาภิวัฒน์

[1] Office of General Education, Panyapiwat Institute of Management

[*] Corresponding author. Tel. 08-3834-1999 E-Mail: nithipatkam@pim.ac.th

บทนำ

การวิเคราะห์การถดถอยเชิงเส้น

วิธีการหนึ่งที่ใช้ศึกษาความสัมพันธ์เชิงเส้น (Linear) ระหว่างตัวแปรต้น (Independent Variable) หรือตัวแปรทำนาย (Predictor Variable) กับตัวแปรตาม (Dependent Variable) หรือตัวแปรเกณฑ์ (Criterion Variable) คือ การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression Analysis) ซึ่งแสดงในรูปตัวแบบเชิงเส้นตรง (Linear Model) ที่เป็นความสัมพันธ์ในรูปเวกเตอร์ (Vector) และเมทริกซ์ (Matrix) ได้ตามสมการที่ 1

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{1}$$

หรือเขียนได้ดังนี้

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ X_{31} & X_{32} & \cdots & X_{3p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_p \end{bmatrix}_{p\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

เมื่อ $\underline{Y}$ แทนเวกเตอร์ของตัวแปรตาม

$\underline{X}$ แทนเมทริกซ์ของตัวแปรอิสระ

$\underline{\beta}$ แทนเวกเตอร์ของค่าพารามิเตอร์ หรือสัมประสิทธิ์การถดถอยที่ไม่ทราบค่าในตัวแบบถดถอย

$\underline{\varepsilon}$ แทนเวกเตอร์ของค่าความคลาดเคลื่อน (Error)

n แทนค่าสังเกตทั้งหมด

p แทนจำนวนพารามิเตอร์หรือสัมประสิทธิ์การถดถอยทั้งหมด

นอกจากการศึกษาความสัมพันธ์ระหว่างตัวแปรแล้ว วัตถุประสงค์ข้อหนึ่งของการวิเคราะห์การถดถอย คือ การสร้างสมการถดถอย (Regression Equation) ที่เป็นผลคูณระหว่างตัวแปรอิสระกับค่าสัมประสิทธิ์การถดถอย (Regression Coefficient) ซึ่งในเบื้องต้นจะใช้วิธีกำลังสองน้อยที่สุด (Least square) มาเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธีการนี้มีหลักการประมาณค่าสัมประสิทธิ์การถดถอยจากผลรวมกำลังสองของความคลาดเคลื่อนกำลังสองหรือ ε_i^2 ที่น้อยที่สุดตามสมการที่ 2

$$Min_{\underline{\hat{\beta}}} \sum_{i=1}^{n} \varepsilon_i^2 = Min_{\underline{\hat{\beta}}} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \tag{2}$$

เมื่อ Y_i แทนค่าสังเกตที่ i จากตัวแปรตาม และ $\hat{Y}_i$ แทนค่าสังเกตที่ i จากตัวแปรตามที่ประมาณขึ้นจากสมการถดถอย โดยที่ $i = 1,2,3,\ldots,n$

ถ้าให้ $\hat{\beta}_j$ แทนค่าประมาณของสัมประสิทธิ์การถดถอยที่ j เมื่อ $j = 1,2,3,\ldots,p$ จะได้สมการถดถอยตามสมการที่ 3

$$\hat{Y}_i = X_{i1}\hat{\beta}_1 + X_{i2}\hat{\beta}_2 + X_{i3}\hat{\beta}_3 + \cdots + X_{ip}\hat{\beta}_p \quad (3)$$

สำหรับเงื่อนไขของการวิเคราะห์การถดถอยเชิงเส้นนั้น จะเป็นเงื่อนไขของค่า $\underline{\varepsilon}$ ดังนี้

1. ค่า $\underline{\varepsilon}$ จะต้องมีการแจกแจงปรกติ (Normal Distribution) ที่มีค่าเฉลี่ยเป็นศูนย์และมีความแปรปรวนคงที่เท่ากับ σ^2

2. ค่า ε_i และ ε_j เมื่อ $i, j = 1,2,3, \ldots, n$ โดยที่ $i \neq j$ จะต้องเป็นอิสระกัน

จากเงื่อนไขข้างต้นสามารถเขียนเป็นสัญลักษณ์คือ $\underline{\varepsilon} \sim NID(\underline{0}, \sigma^2)$ โดยที่ค่าประมาณของสัมประสิทธิ์การถดถอย หรือ $\hat{\beta}$ จะมีสมบัติเป็นตัวประมาณเชิงเส้นไม่เอนเอียงดีที่สุด (Best Linear Unbiased Estimates) ถ้า $\underline{\varepsilon}$ เป็นไปตามเงื่อนไขนี้ทั้งหมด แต่หากพบว่า $\underline{\varepsilon}$ ไม่เป็นไปตามเงื่อนไข แสดงว่าวิธีกำลังสองน้อยที่สุดไม่ได้เป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ดี ซึ่งสาเหตุหนึ่งที่ทำให้ $\underline{\varepsilon}$ ไม่เป็นไปตามเงื่อนไข คือ เกิดค่านอกเกณฑ์ (Outlier) ขึ้นกับค่าสังเกต โดยอิทธิพลของค่านอกเกณฑ์ที่เกิดขึ้นนี้จะส่งผลต่อการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด (Yarmohammadi & Mahmoudvand, 2010) ซึ่งจะต้องใช้การวิเคราะห์การถดถอยที่แกร่ง (Robust Regression) มาเป็นวิธีประมาณค่าสัมประสิทธิ์การถดถอยแทนวิธีกำลังสองน้อยที่สุด

การวิเคราะห์การถดถอยที่แกร่ง

เมื่อเกิดค่านอกเกณฑ์ขึ้นกับค่าสังเกตที่นำมาวิเคราะห์ การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด จะทำให้ได้ $\hat{\beta}$ ที่ไม่มีสมบัติเป็นตัวประมาณเชิงเส้นไม่เอนเอียงดีที่สุดตามที่ได้กล่าวมาแล้ว จึงต้องใช้การวิเคราะห์การถดถอยที่แกร่งมาเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแทนวิธีกำลังสองน้อยที่สุด โดยในที่นี้ได้นำเสนอวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และวิธีการประมาณค่า S (S-estimation) ดังนี้

วิธีภาวะน่าจะเป็นสูงสุด หรือการประมาณค่า M (M Estimation) เป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อเกิดค่านอกเกณฑ์ขึ้นกับค่าสังเกต นำเสนอครั้งแรกโดย Huber ในปี ค.ศ.1964 ซึ่งเป็นการประมาณค่าสัมประสิทธิ์การถดถอยจากฟังก์ชันของส่วนเหลือ (Residual Function) หรือ $\rho(r_i)$ ที่น้อยที่สุดตามสมการที่ 6

$$Minimize_\beta \sum_{i=1}^{n} \rho(r_i) = Minimize_\beta \sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \quad (6)$$

จากสมการที่ 6 สามารถหาคำตอบได้โดยการแก้สมการอนุพันธ์ย่อยอันดับที่หนึ่ง (First Order Partial Derivative) ของฟังก์ชันส่วนเหลือ $\rho(r_i)$ เทียบกับค่าพารามิเตอร์ β_j เมื่อ $j = 0,1, \ldots, p$ ซึ่งจะได้ระบบสมการจำนวน $k = p + 1$ สมการตามสมการที่ 7

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{Y_i - X_i'\beta}{s}\right) = 0 \quad (7)$$

เมื่อ $j = 0,1, \ldots, k$ ส่วนค่า ψ คืออนุพันธ์ย่อยอันดับที่หนึ่งของ $\rho(r_i)$ ค่า s คือค่าเบี่ยงเบนของส่วนเหลือ และ X_{ij} เป็นค่าสังเกตจากตัวแปรอิสระที่ i จากสมการถดถอยที่ j

จากสมการที่ 7 สามารถจัดรูปแบบที่ใช้สำหรับการหาคำตอบของสมการได้ตามสมการที่ 8 และจะใช้วิธีกำลังสองน้อยที่สุดที่ถูกถ่วงน้ำหนักอย่างซ้ำ (Iteratively Reweighted Least Squares: IRLS) ซึ่งเป็นระเบียบวิธีการหาคำตอบของสมการที่ถูกนำไปใช้อย่างแพร่หลาย พัฒนาขึ้นครั้งแรกโดย Beaton และ Tukey ในปี ค.ศ.1974 มาเป็นวิธีการหาคำตอบ หรือนำมาใช้หาค่าประมาณของสัมประสิทธิ์การถดถอย

$$\sum_{i=1}^{n} \frac{X_{ij}\left\{\psi\left[\frac{(y_i - x_i'\hat{\beta}_0)}{s}\right](y_i - x_i'\hat{\beta}_0)/s\right\}(y_i - x_i'\hat{\beta}_0)}{s} = 0 \qquad (8)$$

ถ้าให้ W เป็นเมทริกซ์ของค่าน้ำหนัก (Weight Matrix) ที่มีมิติ $n \times n$ โดยที่

$W = \psi\left[\frac{(y_i - x_i'\hat{\beta}_0)}{s}\right](y_i - x_i'\hat{\beta}_0)/s$ และถ้าให้ $\sum_{i=1}^{n} X_{ij} W_0 (y_i - x_i'\hat{\beta}_0) = 0$ จะได้ค่าน้ำหนักเริ่มต้น W_0 มีสมาชิกแทนด้วย w_{i0} มีค่าดังนี้

$$w_{i0} = \begin{cases} \dfrac{\psi\left[\dfrac{(y_i - x_i'\hat{\beta}_0)}{s}\right]}{\dfrac{(y_i - x_i'\hat{\beta}_0)}{s}} & ;\ y_i \neq x_i'\hat{\beta}_0 \\ 1 & ;\ y_i = x_i'\hat{\beta}_0 \end{cases}$$

ดังนั้นจากสมการ 8 สามารถเขียนใหม่ได้เป็น $\sum_{i=1}^{n} x_{ij} W_0 (y_i - x_i'\hat{\beta}) = 0$ เมื่อจัดรูปแบบใหม่ที่เป็นรูปแบบทั่วไปให้อยู่ในรูปแบบเมทริกซ์ ได้ดังสมการที่ 9

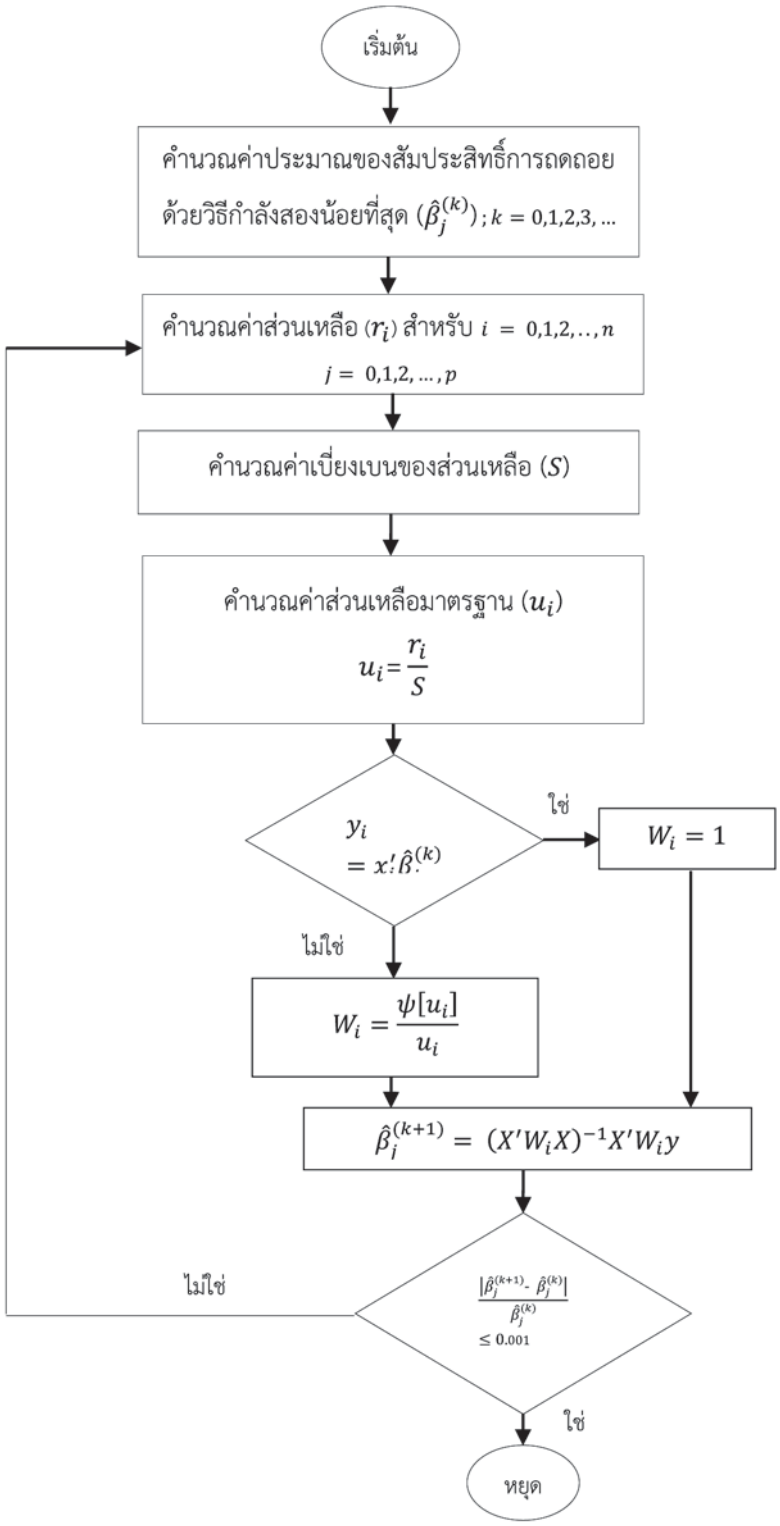
$$X'W_0X\hat{\beta} = X'W_0y \qquad (9)$$

เมื่อ W_0 มีสมาชิกทั้งหมดเป็น $w_{10}, w_{20}, w_{30}, \ldots, w_{n0}$ เมื่อนำไปถ่วงน้ำหนักกับค่าสังเกตจะได้ค่าประมาณของสัมประสิทธิ์การถดถอยในรอบแรกคือ $\hat{\beta}_1$ ตามสมการที่ 10

$$\hat{\beta}_1 = (X'W_0X)^{-1}X'W_0y \qquad (10)$$

ค่า $\hat{\beta}_1$ ที่ได้จะนำไปเป็นค่าเริ่มต้นของการคำนวณรอบถัดไป และจะกระทำซ้ำไปจนกระทั่งลู่เข้า (Converge) สู่คำตอบ ทั้งนี้ Panik (2009: 293) แนะนำให้ใช้อัตราการเปลี่ยนแปลงของค่าสัมประสิทธิ์การถดถอย $\hat{\beta}_j^{(k+1)}$ กับ $\hat{\beta}_j^{(k)}$ หรือค่าสัมประสิทธิ์การถดถอยที่คำนวณได้ในรอบปัจจุบัน (รอบที่ k+1) กับค่าสัมประสิทธิ์การถดถอยที่คำนวณได้ในรอบก่อนหน้า (รอบที่ k) หรือคำนวณหาค่า $\delta = \frac{\left|\hat{\beta}_j^{(k+1)} - \hat{\beta}_j^{(k)}\right|}{\hat{\beta}_j^{(k)}}$ ถ้าพบว่าค่า δ เกินกว่า 0.001 ให้หยุดคำนวณ แต่ถ้าไม่เป็นไปตามนี้ให้ทำซ้ำ โดยเริ่มจากคำนวณค่าส่วนเหลือจากค่าประมาณของสัมประสิทธิ์การถดถอยที่ได้ในรอบใหม่ จากนั้นให้นำค่าเบี่ยงเบนของส่วนเหลือที่ได้ไปเป็นค่าตั้งต้นสำหรับคำนวณค่าถ่วงน้ำหนักที่จะนำไปใช้สำหรับหาค่าประมาณของสัมประสิทธิ์การถดถอยในรอบถัดไป

สำหรับขั้นตอนการหาค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด จากการหาคำตอบด้วยวิธี IRLS แสดงได้ตามภาพที่ 1



ภาพที่ 1 ขั้นตอนวิธีภาวะน่าจะเป็นสูงสุด

ทั้งนี้ Rousseeuw and Leroy (2003: 136) ได้กล่าวถึงลักษณะของฟังก์ชันส่วนเหลือ หรือ $\rho(r)$ ว่าต้องมีความสมมาตร (Symmetry) นั่นคือ $\rho(r) = \rho(-r)$ และต้องเป็นฟังก์ชันต่อเนื่อง (Continuous) ที่มี $\rho(0) = 0$ โดยที่สามารถหาค่าคงที่ c ที่มากกว่าศูนย์ ที่ทำให้ $\rho(r)$ มีค่าเพิ่มขึ้นบนช่วง $[0, c]$ และคงที่บนช่วง $[c, \infty)$ นอกจากนี้ Koller and Stahel (2011) ยังได้กล่าวถึงสมบัติของฟังก์ชันส่วนเหลือเพิ่มเติมดังนี้

1. ค่า $lim_{r\to\infty}\rho(r) = lim_{r\to-\infty}\rho(r) = 1$
2. ถ้าให้ K เป็นค่าคงที่ และ $K > 0$ จะได้ $\lim\limits_{r\to\infty}\frac{\rho(Kr)}{\rho(r)} = 1$
3. ค่า $\lim\limits_{|r|\to\infty}\frac{d\rho(r)}{dr} = 0$

นอกจากนี้ Montgomery (2006: 374) ยังได้สรุปตัวอย่างฟังก์ชันส่วนเหลือที่ปรับเป็นค่ามาตรฐานแล้วแทนด้วย $\rho(u)$ เมื่อ $u_i = \frac{r_i}{s}$ โดยที่ s คือค่าเบี่ยงเบนของส่วนเหลือ ซึ่งฟังก์ชันเหล่านี้พัฒนาโดยฮูเบอร์ และมีผู้พัฒนาต่อเนื่องดังแสดงตามตารางที่ 1

ตารางที่ 1 ฟังก์ชันของส่วนเหลือ

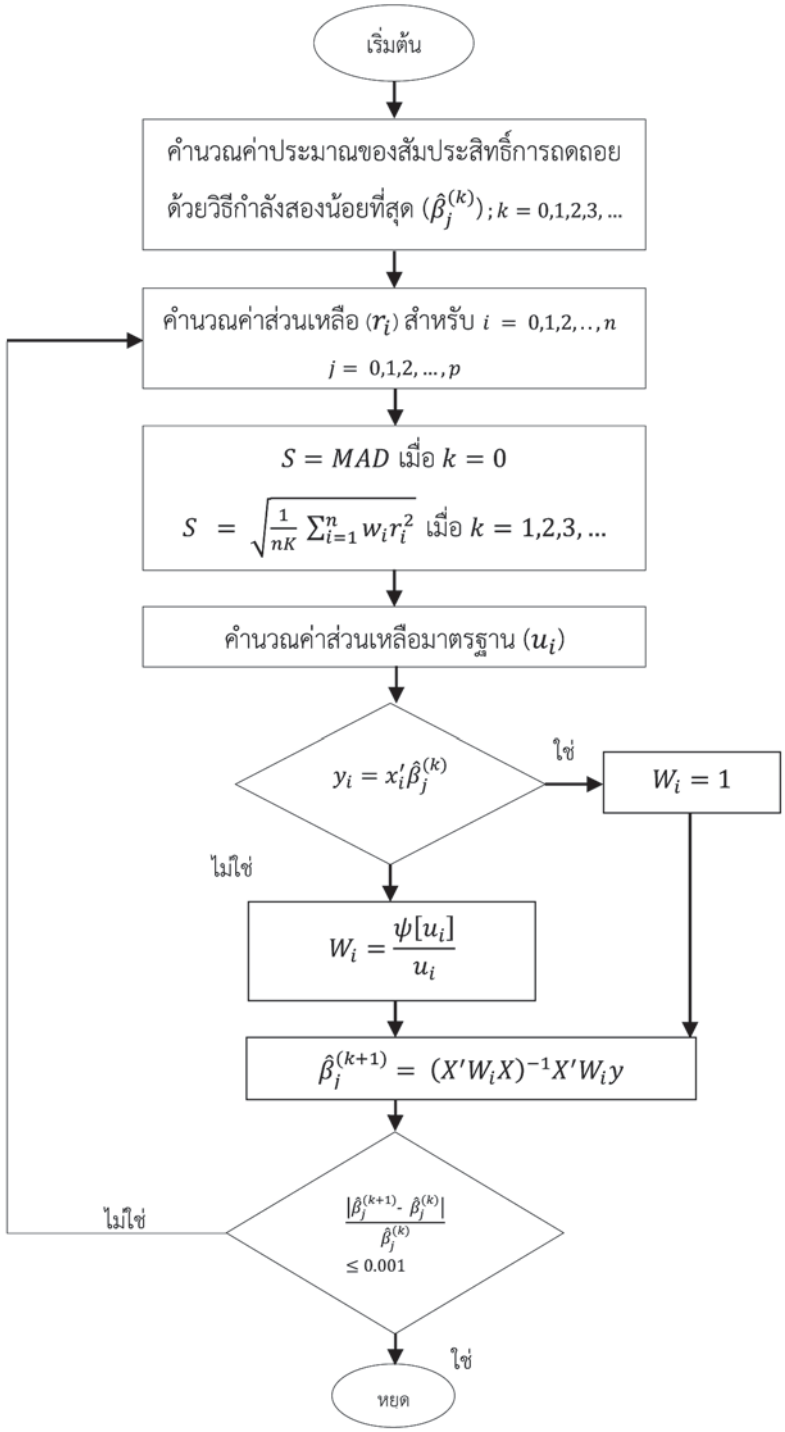
ชื่อฟังก์ชัน	ฟังก์ชันของส่วนเหลือ $\rho(u)$	อนุพันธ์ย่อยอันดับที่หนึ่งของส่วนเหลือ $(\psi(u))$	ฟังก์ชันน้ำหนัก $(W(u))$	ขอบเขต
Least squares	$\frac{1}{2}u^2$	u	1	$\|u\| < \infty$
ฟังก์ชัน t ของ Huber	$\frac{1}{2}u^2$	u	1	$\|u\| \le t$
ฟังก์ชัน t ของ Huber เมื่อ $t = 2$	$\|u\|t - \frac{1}{2}t^2$	$t\ sign\ (u)$	$\frac{t}{\|u\|}$	$\|u\| > t$
ฟังก์ชัน E_a ของ Ramsay เมื่อ $a = 0.3$	$a^{-2}[1 - exp(-a\|u\|).(1 + a\|u\|)]$	$u\ exp(-a\|u\|)$	$exp(-a\|u\|)$	$\|u\| < \infty$
ฟังก์ชันคลื่นของ Andrews	$a[1 - cos(u/a)]$	$sin(u/a)$	$\frac{sin(u/a)}{u/a}$	$\|u\| \le a\pi$
	$2a$	0	0	$\|u\| > a\pi$

ชื่อฟังก์ชัน	ฟังก์ชันของส่วนเหลือ $\rho(u)$	อนุพันธ์ย่อยอันดับที่หนึ่งของส่วนเหลือ $(\psi(u))$	ฟังก์ชันน้ำหนัก $(W(u))$	ขอบเขต
ฟังก์ชันคลื่นของ Andrews เมื่อ $a = 1.339$				
ฟังก์ชัน 17A ของ Hampel $a = 1.7$ $b = 3.4$ $c = 8.5$	$\frac{1}{2}u^2$	u	1	$\lvert u\rvert \le a$
	$a\lvert u\rvert - \frac{1}{2}a^2$	$asin(u)$	$a/\lvert u\rvert$	$a < \lvert u\rvert \le b$
	$\frac{a\left(c\lvert u\rvert - \frac{1}{2}u^2\right)}{c-b} - (7/6)a^2$	$\frac{asign(u)(c-}{c-b}$	$\frac{a(c-\lvert u\rvert)}{\lvert u\rvert(c-b)}$	$b < \lvert u\rvert \le c$
	$a(b+c-a)$	0	0	$\lvert u\rvert > c$

นอกจากนี้ Rousseeuw และ Yohai ยังได้นำเสนอวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ต่อยอดมาจากวิธีภาวะน่าจะเป็นสูงสุด คือวิธีการประมาณค่า S หรือการประมาณค่า S (S Estimation) ในปี ค.ศ.1984 และถูกพัฒนาอย่างต่อเนื่องโดย Rousseeuw และ Leroy ในปี ค.ศ.1987 ซึ่งมีหลักการหาค่าประมาณของสัมประสิทธิ์การถดถอยจากค่าเบี่ยงเบนของส่วนเหลือที่น้อยที่สุด หรือ $min_\beta \hat{\sigma}_s(r_1, r_2, \dots, r_n)$ เมื่อ $\hat{\sigma}_s = s(r_1, r_2, \dots, r_n) = s$ โดยวิธีการนี้เริ่มจากนำค่าส่วนเหลือที่ได้จากค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด ประมาณค่าเบี่ยงเบนของส่วนเหลือจากค่ามัธยฐานส่วนเบี่ยงเบนสัมบูรณ์ (Median Absolute Deviation: MAD) แล้วนำค่าเบี่ยงเบนนี้มาใช้เป็นค่าเริ่มต้นของการคำนวณค่าน้ำหนักเพื่อนำมาใช้หาค่าประมาณของสัมประสิทธิ์การถดถอยในรอบแรก โดยค่าประมาณของสัมประสิทธิ์การถดถอยในแต่ละรอบจะถูกนำมาเปรียบเทียบกัน หรือนำมาคำนวณค่า จนกระทั่งลู่เข้าสู่คำตอบ ($\delta \le 0.001$) แต่ถ้าค่า δ นี้ไม่ได้เป็นไปตามเงื่อนไข ให้นำค่าส่วนเหลือและค่าน้ำหนักที่ได้มาคำนวณค่าเบี่ยงเบนของส่วนเหลือรอบต่อไปตามสมการที่ 11 ซึ่งค่าเบี่ยงเบนของส่วนเหลือในแต่ละรอบจะนำมาเป็นค่าเริ่มต้นของการคำนวณค่าน้ำหนัก เพื่อนำมาใช้หาค่าประมาณของสัมประสิทธิ์การถดถอยในรอบต่อไป ซึ่งค่าประมาณของสัมประสิทธิ์การถดถอยในแต่ละรอบจะนำมาคำนวณค่า δ และจะกระทำซ้ำต่อไปจนกระทั่งลู่เข้าสู่คำตอบ

$$S = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i r_i^2} \quad (11)$$

เมื่อ K เป็นค่าคาดหมาย (Expected Value) ของฟังก์ชัน $\rho(u)$ หรือ $E_\phi[\rho(u)]$ และ u มีการแจกแจงปรกติมาตรฐาน (Standard Normal Distribution) (Rousseww & Leroy, 2003: 139) สำหรับขั้นตอนการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณค่า S ภายใต้การหาคำตอบโดยวิธี IRLS แสดงได้ตามภาพที่ 2



ภาพที่ 2 ขั้นตอนวิธีตัวประมาณค่า S

การหาค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณค่า S นั้น Rousseeuw and Yohai (1984) ได้แนะนำให้ใช้ฟังก์ชันส่วนเหลือของ Turky ที่มีฟังก์ชันดังนี้

$$\rho(u_i) = \begin{cases} \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^4} & ; |u_i| \le c \\ \frac{c^6}{6} & ; |u_i| > c \end{cases}$$

ทั้งนี้ Rousseeuw and Leroy (2003: 142) ได้แสดงค่าคงที่ K และ c ที่คำนวณจากค่าร้อยละของจุดแบ่งข้อมูล (ε^*) และค่าประสิทธิภาพของตัวประมาณค่า (Efficiency of Parameter) หรือ e ได้ตามตารางที่ 2

ตารางที่ 2 ค่าคงที่ K และ c จากค่าร้อยละของจุดแบ่งข้อมูลและค่าประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอย

c	K	ε^*	e
1.547	1.995	50%	28.7%
1.756	0.2312	45%	37.0%
1.988	0.2634	40%	46.2%
2.251	0.2957	35%	56.0%
2.560	0.3278	30%	66.1%
2.937	0.3593	25%	75.9%
3.420	0.3899	20%	84.7%
4.096	0.4194	15%	91.7%
5.182	0.4475	10%	96.6%

สำหรับคำนวณหาค่า K จากฟังก์ชันส่วนเหลือของทูกี เมื่อมีจุดเปลี่ยนข้อมูล 50% ที่ค่าคง c = 1.547 จากค่า u ที่มีการแจกแจงปรกติมาตรฐานที่มีฟังก์ชันความหนาแน่นความน่าจะเป็น (Probability Density Function) หรือ $f(u)$ ดังนี้

$$f(u) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{u^2}{2}\right)$$

$$K = \int_{-\infty}^{\infty} \rho(u)\, f(u)dx = 2\int_{0}^{+\infty} \rho(u)\, f(u)dx$$

$$= \quad 2\int_0^{1.547}\left(\frac{u^2}{2}-\frac{u^4}{2\times1.547^2}+\frac{u^6}{6\times1.547^4}\right)\frac{1}{\sqrt{2\pi}}\,exp\left(-\frac{u^2}{2}\right)dx$$

$$+\,2\int_{1.547}^{+\infty}\frac{1.547^2}{6}\frac{1}{\sqrt{2\pi}}\,exp\left(-\frac{u^2}{2}\right)dx$$

$$= \sqrt{\frac{2}{\pi}} \times 0.189 + 0.798 \times (1 - 0.939) = 0.199$$

ตัวอย่างการประมาณค่าสัมประสิทธิ์การถดถอยจากข้อมูลปริมาณผลผลิตข้าวโพดในแต่ละเมืองจากปัจจัยต่างๆ ที่มีต่อผลผลิตตามงานวิจัยของ Susanti and Pratiwi (2014) โดยตัวแปรตาม Y แทนปริมาณผลผลิตข้าวโพด (ตัน) เมื่อ X_1 แทนพื้นที่เพาะปลูก (เฮกตาร์) X_2 แทนปริมาณฝนเฉลี่ยต่อเดือน (มิลลิเมตร) X_3 แทนความชื้นโดยเฉลี่ยต่อเดือน (เปอร์เซ็นต์) X_4 แทนอุณหภูมิเฉลี่ยต่อเดือน (องศาเซลเซียส) X_5 แทนปริมาณแสงอาทิตย์ที่ส่องโดยเฉลี่ยต่อเดือน (เปอร์เซ็นต์) และ X_6 แทนอัตรากำลังของเกษตรกรในการดูแล (คน) ซึ่งผลการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด แสดงในรูปสมการถดถอยตามสมการที่ 12 ดังนี้

$$\hat{y} = 623{,}773 + 4.71x_1 + 25x_2 - 8{,}521x_3 + 14{,}097x_4 - 5{,}668x_5 + 0.155x_6 \quad (12)$$

โดยมี $R^2 = 98.8\%$ $R^2_{\text{adjusted}} = 98.5\%$ และค่าเบี่ยงเบนมาตรฐาน (s) = 131,501 ที่มีค่า $F = 354.81$ เมื่อค่า p (p-value) < 5% แสดงว่าเป็นตัวแบบถดถอยที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทุกตัว แต่เมื่อทดสอบการแจกแจงแบบปรกติของความคลาดเคลื่อนพบว่า ความคลาดเคลื่อนไม่ได้แจกแจงปรกติอย่างมีนัยสำคัญที่ระดับ 0.05 โดยพบว่ามีสาเหตุมาจากเกิดค่านอกเกณฑ์ขึ้นกับค่าสังเกตที่นำมาวิเคราะห์ จึงได้ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดและวิธีการประมาณค่า S จึงได้สมการถดถอยที่ 13 และ 14 ตามลำดับดังนี้

$$\hat{y} = 1{,}483{,}237 + 4.43x_1 + 37x_2 - 5{,}645x_3 + 37{,}101x_4 - 732x_5 + 0.05447x_6 \quad (13)$$

$$\hat{y} = 1{,}885{,}905 + 4.39x_1 + 8.2x_2 - 9{,}306x_3 - 36{,}997x_4 - 236x_5 + 0.0549x_6 \quad (14)$$

หลักการพิจารณาเลือกตัวแบบถดถอยที่ดีที่สุดจะพิจารณาจาก R^2 หรือ R^2_{adjusted} และค่า s โดยตัวแบบถดถอยที่ดีที่สุดจะให้ R^2 หรือ R^2_{adjusted} มีค่ามากที่สุด เมื่อมี s น้อยที่สุด ซึ่งจากข้อมูลตัวอย่างสามารถเปรียบเทียบค่า R^2_{adjusted} และ s จากวิธีประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี ดังตารางที่ 3 ดังนี้

ตารางที่ 3 การเปรียบเทียบค่า R^2_{adjusted} และ s

	กำลังสองน้อยที่สุด	ภาวะน่าจะเป็นสูงสุด	ตัวประมาณค่า S
R^2_{adjusted}	98.8%	99.9%	100%
S	131,501	36,646.5	28,929.5
ตัวแปรอิสระที่มีอยู่ในตัวแบบอย่างมีนัยสำคัญ	$X_1 - X_6$	X_1, X_6	X_1, X_3, X_4, X_5, X_6

จากตารางที่ 3 พบว่า วิธีภาวะน่าจะเป็นสูงสุด และวิธีการประมาณค่า S ให้ค่า R^2_{adjusted} มากกว่าวิธีกำลังสองน้อยที่สุด และมีค่า s น้อยกว่าวิธีกำลังสองน้อยที่สุด จึงเป็นตัวแบบถดถอยที่เหมาะสมมากกว่าตัวแบบถดถอยที่มาจากวิธีกำลังสองน้อยที่สุด ทั้งนี้จะพบว่าวิธีการประมาณค่า S ให้ค่า R^2_{adjusted} มากที่สุดที่มี s น้อยที่สุด ดังนั้นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยวิธีการประมาณค่า S จึงเหมาะสมที่สุดกับข้อมูลชุดนี้ โดยพบว่าประกอบด้วยตัวแปรอิสระ X_1, X_3, X_4, X_5 และ X_6 ในตัวแบบถดถอย จึงได้สมการถดถอยใหม่ดังนี้

$$\hat{y} = -1{,}876{,}912 + 4.39x_1 - 9{,}205x_3 - 36{,}855x_4 - 2{,}388x_5 + 0.0553x_6 \quad (15)$$

เมื่อ $R^2_{\text{adjusted}} = 100\%$ *ที่มีค่า* $F = 60499.55$ เมื่อค่า *p* (p-value) < 5% แสดงว่าเป็นตัวแบบถดถอยที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระเหล่านี้ทุกตัว

ทั้งนี้นอกจากวิธีภาวะน่าจะเป็นสูงสุดและวิธีการประมาณค่า S แล้วยังมีวิธีการอื่นอีก เช่น วิธีมัธยฐานของกำลังสองที่น้อยที่สุด (Least Median of Square) หรือวิธี LMS ที่ Hampel เป็นผู้พัฒนาขึ้นในปี ค.ศ.1974 และ Rousseeuw ได้พัฒนาต่อเนื่องในปี ค.ศ.1984 โดยหลักการของวิธี LMS คือ การประมาณค่าพารามิเตอร์จากค่ามัธยฐานของส่วนตกค้างที่มีค่าน้อยที่สุด หรือ $min_{\hat{\beta}}\, med_i r_i^2$ (Massart, Kaufman, Rousseeuw, & Leroy, 1986)

วิธีกำลังสองของส่วนเหลือที่ถูกตัดค่าน้อยที่สุด (Least Trimmed Square) หรือ LTS พัฒนาโดย Rousseeuw ในปี ค.ศ.1984 ที่จะเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยจากค่าส่วนเหลือที่ถูกตัดค่ายกกำลังสองที่น้อยที่สุด หรือ $min \sum_{i=1}^{h} (r^2)_{i:n}$ เมื่อ $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ เป็นลำดับของค่าส่วนเหลือกำลังสอง และ h คือค่าคงที่ของการตัดค่า (Trimming Constant) ซึ่งเป็นจำนวนส่วนเหลือที่ใช้ประมาณค่าด้วยวิธี LTS เมื่อ $h = \frac{n+p+1}{2}$ โดยที่ n เป็นค่าสังเกต และ p เป็นจำนวนพารามิเตอร์หรือค่าสัมประสิทธิ์การถดถอยทั้งหมด (Nguyen & Welsch, 2010)

วิธีตัวประมาณค่า MM (MM-estimator) เป็นวิธีการที่ต่อยอดมาจากวิธีภาวะน่าจะเป็นสูงสุด และวิธีการประมาณค่า S พัฒนาโดย Yohai ในปี ค.ศ.1987 ซึ่งวิธีนี้เป็นการหาค่าประมาณของสัมประสิทธิ์การถดถอยจากสมการที่ 16

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - \sum_{j=0}^{k} X_{ij}\hat{\beta}_j}{s_{MM}}\right) X_{ij} = 0 \quad (16)$$

เมื่อ s_{MM} แทนค่าเบี่ยงเบนมาตรฐานของส่วนเหลือ จากส่วนเหลือที่ได้มาด้วยวิธีการประมาณค่า S และ ρ แทนฟังก์ชันส่วนเหลือของ Tukey (Susanti & Pratiwi, 2014)

วิธี Generalized M-estimator หรือวิธี GM พัฒนาโดย Simpson, Ruppert และ Carroll ในปี ค.ศ.1992 ที่พัฒนาจากวิธีภาวะน่าจะเป็นสูงสุด ที่มีหลักการหาค่าประมาณของสัมประสิทธิ์การถดถอยที่ทำให้ผลรวมของฟังก์ชันส่วนตกค้างในสมการที่ 17 มีค่าน้อยที่สุด

$$min \sum_{i=1}^{n} \rho\left(\frac{r_i}{\pi_i s}\right) \pi_i \quad (17)$$

ค่า π_i เป็นค่าถ่วงน้ำหนักที่คำนวณได้จาก $\pi_i = median_i \frac{|u_i|}{u_i}$ โดยที่ $u_i = \frac{r_i}{s}$ เมื่อหาอนุพันธ์ย่อย (Partial Differential) ของ $min \sum_{i=1}^{n} \rho\left(\frac{r_i}{\pi_i s}\right) \pi_i$ เทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับศูนย์ จะได้ฟังก์ชันที่ใช้สำหรับหาค่าประมาณของสัมประสิทธิ์การถดถอยใหม่ ตามสมการที่ 18

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\pi_i s}\right) \pi_i X_i = 0 \quad (18)$$

เมื่อ $\psi(u) = \frac{\partial}{\partial u} p(u) = p'$ ถ้า $\psi(u) = u$ แล้ว วิธีตัวประมาณ GM คือวิธีกำลังสองน้อยที่สุด โดยที่ $u = \frac{r_i}{\pi_i s}$ และ S เป็นค่าเบี่ยงเบนของส่วนเหลือคำนวณได้จากค่า MAD ที่ถูกปรับด้วยค่าคงที่ 1.4826 ซึ่งเป็นการทำให้ S เป็นตัวประมาณที่ไม่เอนเอียง (Biased) เมื่อ n มีขนาดใหญ่และการแจกแจงของความคลาดเคลื่อนเป็นแบบปกติที่คำนวณได้ตามสมการที่ 19

$$s = 1.4826\ median_i\left(\left|r_i - \underset{i}{median}(r_i)\right|\right) \quad (19)$$

จากการที่ค่า ψ ไม่เป็นเส้นตรง (Non-linear) ดังนั้นวิธีการแก้สมการเพื่อหาค่าประมาณของสัมประสิทธิ์การถดถอยจะใช้วิธี IRLS โดยเริ่มจากค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด และค่า S จากนั้นจึงคำนวณค่าถ่วงน้ำหนัก เพื่อนำมาใช้หาค่าประมาณของสัมประสิทธิ์การถดถอยจากวิธีกำลังสองที่น้อยที่สุดที่ถูกถ่วงน้ำหนัก (Weighted Least Square: WLS) ตามสมการที่ 20

$$\sum_{i=1}^{n} W\left(\frac{r_i}{\pi_i s}\right) \pi_i X_i \quad = \quad 0 \quad (20)$$

เมื่อ $W_i = \psi\left(\frac{r_i}{\pi_i S}\right) / \left(\frac{r_i}{\pi_i S}\right)$ ดังนั้นค่าประมาณของสัมประสิทธิ์การถดถอยจากวิธี WLS แสดงได้ตามสมการที่ 21

$$\hat{\beta} = (X'WX)^{-1}X'WY \quad (21)$$

โดยที่ W เป็นเมทริกซ์ทแยงมุมขนาด $n \times n$ ที่มี W_i เป็นสมาชิกแนวทแยงมุม (Thomas & Mili, 2007)

นอกจากวิธีการประมาณค่าสัมประสิทธิ์การถดถอยต่างๆ ตามที่ได้กล่าวมาแล้ว ในปัจจุบันยังได้มีการพัฒนาอย่างต่อเนื่อง อาทิ Milhano, Sequera and Sotto (2013) ได้ปรับแก้ขั้นตอนการคำนวณหาค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณค่า S โดยใช้หลักการ การคัดเลือกส่วนเหลือใหม่ที่มีค่าเบี่ยงเบนน้อยที่สุด เพื่อนำมาใช้คำนวณหาค่าเบี่ยงเบน และค่าถ่วงน้ำหนักในแต่ละรอบ โดยผลการวิจัยพบว่า วิธีการใหม่ที่พัฒนาขึ้นนี้ ใช้เวลาของการคำนวณในแต่ละรอบที่น้อยที่สุดมีค่าเท่ากับ $(1 - \varepsilon_N^*)N\Delta t$ เมื่อ ε_N^* คือจุดเปลี่ยนข้อมูล N คือค่าสังเกตทั้งหมดและ Δt เวลาที่เกิดขึ้นจากการปรับวิธีการใหม่นี้ใน*แต่ละรอบ* Smirnov and Shevlyakov (2014) ได้ปรับแก้วิธีตัวประมาณค่า M โดยใช้ค่าสถิติ Q_n ของ Rousseeuw and Croux (1993) มาหาค่าเบี่ยงเบนของส่วนเหลือ ผลจากการจำลองสถานการณ์พบว่า วิธีตัวประมาณค่า M ปรับแก้ ใช้เวลาในการคำนวณหาค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธี IRLS น้อยกว่าวิธีการเดิม และพบว่าตัวประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับแก้นั้นมีสมบัติที่ดีกว่าวิธีเดิม และ Ollerer, Alfons, and Croux (2016) ที่ได้ปรับแก้วิธีการประมาณค่า S โดยปรับฟังก์ชันส่วนเหลือของ Turky และ Huber ผลการเปรียบเทียบประสิทธิภาพของค่าประมาณที่ได้ พบว่าวิธีการใหม่มีประสิทธิภาพมากกว่า เมื่อใช้สถานการณ์จำลองที่มีค่านอกเกณฑ์ร้อยละ 1 ร้อยละ 2 ร้อยละ 5 และร้อยละ 10

เอกสารอ้างอิง

Koller, M., & Stahel, W. A. (2011). Sharpening wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis,* 55(8): 2504–2515.

Massart, D. L., & Kaufman, L., M., Rousseeuw, P. J., & Leroy, A. (1986). Least median of squares: A robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta,* 187: 171–179.

Milhano, T., Sequera, J., & Sotto, E. D. (2013). Using S-estimators in Parameter Identification. In *Proceedings of the Information Fusion International Conference 2013,* (pp. 1058–1065). Istanbul: Turkey.

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

Nguyen, T. D., and Welsch, R. (2010). Outlier detection and least trimmed squares approximation using semidefinite programming. Computational Statistics & Data Analysis, 54: 3212–3226.

Ollerer, V., Alfons, A., & Croux, C. (2016). The shooting S-estimator for robust regression. *Computational Statistical,* 31(3): 829–844.

Panik, M. (2009). *Regression Modeling Methods, Theory, and Computation with SAS*. New York: Taylor & Francis Group.

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection* (Vol. 589): John wiley & sons.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. *Robust and nonlinear time series analysis* (pp. 256-272): Springer.

Smirnov, P. O., & Shevlyakov, G. L. (2014). Fast highly efficient and robust one-step M-estimators of scale based on Qn. *Computational Statistics & Data Analysis,* 78: 153-158.

Susanti, Y., & Pratiwi, H. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics,* 91(3): 349-360.

Thomas, L., & Mili, L. (2007). A Robust GM-Estimator for the Automated Detection of External Defects on Barked Hardwood Logs and Stems. *IEEE TRANSACTIONS ON SIGNAL PROCESSING,* 15(7): *3568-3576.*

Yarmohammadi, M., & Mahmoudvand, R. (2010). The Effect Of Outliers On Robust And Resistant Coefficient Of Determination In The Linear Regression Models. *International Journal of Academic Research,* 2(3).