ประสิทธิภาพของตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติในการถดถอยเชิงเส้นพหุคูณ

Efficiency of Outlier Detection Statistics in Multiple Linear Regression

วนิดา พงษ์ศักดิ์ชาติ* และ แพรวนภา เหมือนสมัย

Vanida Pongsakchat* and Preawnapa Muansamai

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา

Department of Mathematics, Faculty of Science, Burapha University

Received : 12 June 2017

Accepted : 17 July 2017

Published online : 20 July 2017

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติในการถดถอยเชิงเส้นพหุคูณจำนวน 5 วิธี ได้แก่ leverage value (h_{ii}), studentized deleted residual (t_i), Cook's distance (D_i), $DFFITS_i$ และ covariance ratio (CVR_i) ลักษณะของค่าผิดปกติในชุดข้อมูลที่ศึกษามี 3 ลักษณะ คือ ค่าผิดปกติในตัวแปรอิสระ ในตัวแปรตาม และทั้งในตัวแปรอิสระและตัวแปรตาม ขนาดตัวอย่างคือ 30, 50 และ 100 จำนวนค่าผิดปกติในแต่ละชุดข้อมูลเท่ากับ 1 ค่าสังเกต และร้อยละ 10, 20 และ 30 ของขนาดตัวอย่าง และเกณฑ์ที่ใช้ในการพิจารณาประสิทธิภาพของตัวสถิติทั้งหา 5 วิธี คือสัดส่วนที่ตัวสถิติเหล่านี้ตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดทุกค่าที่มีในชุดข้อมูลจากการทำซ้ำ 10,000 ครั้ง ผลการศึกษาพบว่า ในทุกขนาดตัวอย่างเมื่อมีค่าผิดปกติ 1 ค่าสังเกตในชุดข้อมูล h_{ii} และ CVR_i มีประสิทธิภาพในการตรวจสอบค่าผิดปกติดีที่สุดเมื่อค่าผิดปกติอยู่ในตัวแปรอิสระ ในขณะที่เมื่อค่าผิดปกติอยู่ในตัวแปรตาม t_i, D_i, $DFFITS_i$ และ CVR_i มีประสิทธิภาพในการตรวจสอบค่าผิดปกติมากที่สุด และเมื่อค่าผิดปกติอยู่ทั้งในตัวแปรอิสระและตัวแปรตาม h_{ii}, D_i และ $DFFITS_i$ มีประสิทธิภาพในการตรวจสอบค่าผิดปกติมากที่สุด อย่างไรก็ตาม ตัวสถิติเหล่านี้จะมีประสิทธิภาพในการตรวจสอบค่าผิดปกติลดลงเมื่อจำนวนค่าผิดปกติเพิ่มขึ้น

คำสำคัญ : ค่าผิดปกติ leverage, studentized deleted residual, Cook's distance, covariance ratio

***Corresponding author. E-mail** : vanida@buu.ac.th

Abstract

The objective of this research was to compare the performance of 5 outlier detecting statistics in the multiple linear regression which are leverage value (h_{ii}), studentized deleted residual (t_i), Cook's distance (D_i), $DFFITS_i$ and covariance ratio (CVR_i). There were three types of outliers: outliers in independent variables, in the dependent variable and in both independent and dependent variables. Sample sizes were 30, 50 and 100 and number of outliers in each dataset were 1 observation and 10, 20 and 30 percent of the sample size. The criterion used for considering the performance of these statistics was proportion of correctly detect all of outliers in the dataset from 10,000 replications. The results are shown as follows: for all sample sizes, in single-outlier case, h_{ii} and CVR_i had the highest performance when an outlier was in the independent variables, whereas t_i, D_i, $DFFITS_i$ and CVR_i had the highest performance when an outlier was in the dependent variable, and when an outlier was in both the independent variables and the dependent variable, h_{ii}, D_i and $DFFITS_i$ had the best performance. However, the performance of these statistics decreased as the number of outliers in the dataset increased.

Keywords : outliers, leverage, studentized deleted residual, Cook's distance, covariance ratio

บทนำ

การถดถอยเชิงเส้นพหุคูณ (multiple linear regression) เป็นวิธีเชิงสถิติที่ใช้ในการศึกษารูปแบบความสัมพันธ์ระหว่างตัวแปรตาม (dependent variable), y และตัวแปรอิสระ (independent variables), $x_1, x_2, \ldots, x_k$ เมื่อ k คือจำนวนตัวแปรอิสระที่ศึกษา โดยมีตัวแบบคือ $y_i = \beta_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$ เมื่อ $i = 1,2,\ldots,n$ และ n คือ จำนวนค่าสังเกตในชุดข้อมูล การประมาณค่าพารามิเตอร์ ($\beta_0, \beta_1, \ldots, \beta_k$) ด้วยวิธีกำลังสองน้อยที่สุด (ordinary least-square, OLS) ซึ่งเป็นวิธีหนึ่งที่นิยมใช้มากที่สุด เนื่องจากเป็นวิธีที่ทำให้ตัวประมาณที่ได้มีคุณสมบัติที่ดีคือเป็นตัวประมาณเชิงเส้นไม่เอนเอียงดีที่สุด (best linear unbiased estimator, BLUE) และสามารถเขียนตัวแบบการถดถอยนี้ในรูปเวกเตอร์และเมทริกซ์ได้ดังนี้

$$\underline{y} = \mathrm{x}\underline{\beta} + \underline{\varepsilon} \qquad (1)$$

เมื่อ $\underline{y}$ คือเวกเตอร์ขนาด $n \times 1$ ของค่าสังเกตของตัวแปรตาม x คือเมทริกซ์ขนาด $n \times (k+1)$ ของตัวแปรอิสระจำนวน k ตัวแปร $\underline{\beta}$ คือเวกเตอร์ขนาด $(k+1) \times 1$ ของสัมประสิทธิ์การถดถอย และ $\underline{\varepsilon}$ คือเวกเตอร์ขนาด $n \times 1$ ของความคลาดเคลื่อนที่มีข้อสมมุติ $N(\underline{0}, \sigma^2 \mathrm{I})$ เมื่อ σ^2 คือความแปรปรวนของค่าคลาดเคลื่อน และ I คือเมทริกซ์เอกลักษณ์ขนาด $n \times n$

อย่างไรก็ตาม หากมีค่าผิดปกติ (outlier) อยู่ในชุดข้อมูลตัวอย่าง ไม่ว่าจะเป็นค่าผิดปกติในตัวแปรอิสระ ค่าผิดปกติในตัวแปรตาม หรือค่าผิดปกติทั้งในตัวแปรอิสระและตัวแปรตาม อาจทำให้ $\hat{\underline{\beta}}$ ซึ่งเป็นตัวประมาณของ $\underline{\beta}$ จากวิธี OLS ไม่ใช่ตัวประมาณค่าที่ดีที่สุดอีกต่อไป เนื่องจากค่าผิดปกติอาจทำให้ข้อสมมุติเกี่ยวกับการแจกแจงปรกติของการถดถอยเชิงเส้นพหุคูณผิดไป ดังนั้น การตรวจสอบค่าผิดปกติในชุดข้อมูลจึงมีความสำคัญมาก ซึ่งวิธีการตรวจสอบค่าผิดปกติมีอยู่ด้วยกันหลายวิธี

Ampanthong and Suwattee (2009) ได้ศึกษาเปรียบเทียบตัวสถิติที่ใช้สำหรับตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอย 8 วิธี ได้แก่ standardized residuals, studentized residuals, PRESS residuals, leverage, Cook's distance, R-Student, DFFITS และ Mahalanobis distance และพบว่า Mahalanobis distance สามารถตรวจสอบค่าผิดปกติในตัวแปรอิสระ และค่าผิดปกติทั้งในตัวแปรอิสระและตัวแปรตามได้ดีกว่าวิธีอื่น สำหรับค่าผิดปกติในตัวแปรตาม Cook's distance และ DFFITS เป็นตัวสถิติที่มีประสิทธิภาพมากกว่าวิธีอื่น ๆ ต่อมา Zakaria, et al. (2014) ได้เสนอวิธีการตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอยคือ Coefficient of Determination Ratio (CDR) และเปรียบเทียบประสิทธิภาพของ CDR กับตัวสถิติที่ใช้สำหรับตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอย 5 วิธี คือ leverage, studentized deleted residuals, Cook's distance, DFFITS และ covariance ratio โดยพบว่า CDR, Cook's distance และ DFFITS มีประสิทธิภาพในการตรวจสอบค่าผิดปกติได้ใกล้เคียงกัน และ Marubini and Orenti (2014) ได้นำเสนอวิธีการตรวจสอบค่าผิดปกติในการวิเคราะห์การถดถอยโดยใช้วิธี robust two-stage procedure ซึ่งมีการตรวจสอบค่าผิดปกติในขั้นตอนแรกด้วยค่า robust distance และใช้ตัวสถิติ leverage, Cook's distance และ studentized deleted residuals ที่ใช้เกณฑ์ cut-off points จากวิธี jackknife-after-bootstrap ในการตรวจสอบค่าผิดปกติในขั้นตอนที่ 2 โดยพบว่าวิธีที่นำเสนอนี้สามารถตรวจสอบค่าผิดปกติได้ดีทั้งในตัวแปรตามและตัวแปรอิสระ

จากงานวิจัยที่ศึกษามาในข้างต้น พบว่ามีตัวสถิติที่สามารถใช้ในการตรวจสอบค่าผิดปกติในการถดถอยเชิงเส้นอยู่ด้วยกันหลายวิธี ทั้งที่เป็นวิธีที่นิยมใช้กันอย่างแพร่หลาย เช่น leverage, Cook's distance, studentized deleted residual และ DFFITS และวิธีที่พัฒนาขึ้นมาใหม่ อย่างไรก็ตาม วิธีที่มีการพัฒนาขึ้นใหม่นั้นมักจะมีความยุ่งยากในการนำมาใช้ และไม่มีในโปรแกรมสำเร็จรูปทางสถิติทั่ว ๆ ไป จึงไม่เหมาะกับผู้ใช้ทั่ว ๆ ไป ดังนั้น ในการศึกษานี้จึงต้องการเปรียบเทียบประสิทธิภาพของตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติสำหรับการถดถอยเชิงเส้น 5 วิธี ซึ่งเป็นวิธีที่นิยมใช้กันอย่างแพร่หลายและมีในโปรแกรมสำเร็จรูปทางสถิติ เพื่อเป็นแนวทางในการเลือกใช้ตัวสถิติเหล่านี้ให้กับผู้ใช้ทั่ว ๆ ไป ซึ่งตัวสถิติที่ศึกษา ได้แก่

1. Leverage value (h_{ii}) ซึ่งเป็นสมาชิกในแนวทแยงมุมหลักที่ i ของเมทริกซ์ H โดยที่

$$H = X(X'X)^{-1}X' \tag{2}$$

และค่าสังเกตที่ i จะถูกบ่งชี้ว่าเป็นค่าผิดปกติหาก $h_{ii} > 2(k+1)/n$ (Rousseeuw and Leroy, 2003, p.220)

2. Studentized deleted residual (t_i)

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \tag{3}$$

$i = 1,2,\ldots,n$ เมื่อ t_i มีการแจกแจงที่ มีองศาเสรีเท่ากับ $n-k-2$, $e_i = y_i - \hat{y}_i$ คือตัวประมาณของ ε_i , $s_{(i)}$ คือ ตัวประมาณของ σ ที่ได้จากการประมาณสมการถดถอยโดยใช้ชุดข้อมูลที่ไม่มีค่าสังเกตที่ i และค่าสังเกตที่ i จะถูกบ่งชี้ว่าเป็นค่าผิดปกติหาก $\left|t_i\right| > t_{(\alpha/2n),(n-k-2)}$

3. Cook's distance (D_i)

$$D_i = \frac{\left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)'(\mathrm{x}'\mathrm{x})\left(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(i)}\right)}{(k+1)s^2} \tag{4}$$

$i = 1,2,\ldots,n$ เมื่อ $\underline{\hat{\beta}}$ คือ ตัวประมาณของ $\underline{\beta}$, $\underline{\hat{\beta}}_{(i)}$ คือตัวประมาณของ $\underline{\beta}$ ที่ประมาณจากชุดข้อมูลที่ไม่มีค่าสังเกตที่ i และ s^2 คือ ตัวประมาณของ σ^2 และค่าสังเกตที่ i จะถูกบ่งชี้ว่าเป็นค่าผิดปกติหาก $D_i > 4/(n-k-1)$ (Zakaria et al., 2014)

4. $DFFITS_i$

$$DFFITS_i = \frac{e_i\sqrt{h_{ii}}}{s_{(i)}(1-h_{ii})} \tag{5}$$

$i = 1,2,\ldots,n$ และค่าสังเกตที่ i จะถูกบ่งชี้ว่าเป็นค่าผิดปกติหาก $\left|DFFITS_i\right| > 2\sqrt{(k+1)/n}$ (Rousseeuw and Leroy, 2003, p.228)

5. Covariance ratio (CVR_i)

$$CVR_i = \left(\frac{s_{(i)}^2}{s^2}\right)^k \left(\frac{1}{1-h_{ii}}\right) \tag{6}$$

$i = 1,2,\ldots,n$ และค่าสังเกตที่ i จะถูกบ่งชี้ว่าเป็นค่าผิดปกติหาก $\left|CVR_i - 1\right| > 3k/n$ (Zakaria *et al.*, 2014)

วิธีดำเนินการวิจัย

ตัวแบบที่ใช้ในการศึกษานี้ คือ ตัวแบบการถดถอยพหุคูณที่มีตัวแปรอิสระ 2 ตัวแปร $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ โดยกำหนดให้ $\beta_0 = 2$ และ $\beta_1 = \beta_2 = 1$ และ ε เป็นตัวแปรสุ่มที่มีการแจกแจงปรกติและเป็นอิสระกัน กำหนดขนาดตัวอย่างที่ใช้ในการศึกษา (n) เท่ากับ 30, 50 และ 100 และจำนวนค่าผิดปกติในแต่ละชุดข้อมูล (m) เท่ากับ 1 ค่าสังเกต และเป็นร้อยละ 10, 20 และ 30 ของขนาดตัวอย่าง โดยค่าผิดปกติที่ศึกษาจำแนกเป็น 3 ลักษณะ คือ ค่าผิดปกติในตัวแปรอิสระ ในตัวแปรตาม และทั้งในตัวแปรอิสระและตัวแปรตาม จำลองข้อมูลด้วยเทคนิคมอนติคาร์โลและหาค่าสถิติที่ใช้สำหรับตรวจสอบค่าผิดปกติ 5 วิธี ได้แก่ leverage value (h_{ii}), studentized deleted residual (t_i), Cook's distance (D_i), $DFFITS_i$ และ covariance ratio (CVR_i) ด้วยโปรแกรม R และทำการทดลองซ้ำ 10,000 ครั้ง ในแต่ละสถานการณ์ โดยการกำหนดสถานการณ์ต่าง ๆ ในการศึกษานี้ เป็นการกำหนดให้สอดคล้องกับงานวิจัยของ Ampanthong, P. and Suwattee, P. (2009)

ในการจำลองข้อมูลได้แบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนที่ 1 คือส่วนที่ไม่ใช่ค่าผิดปกติจำนวน $n - m$ ค่า และส่วนที่ 2 คือส่วนที่เป็นค่าผิดปกติจำนวน m ค่า ข้อมูลในส่วนที่ 1 ตัวแปรอิสระ x_1, x_2 และความคลาดเคลื่อน ε มีการแจกแจงปรกติมาตรฐาน ส่วนข้อมูลในส่วนที่ 2 ที่เป็นค่าผิดปกติ ตัวแปรอิสระ x_1, x_2 และความคลาดเคลื่อน ε มีการแจกแจงปรกติที่มีค่าเฉลี่ยเท่ากับ 10 และค่าเบี่ยงเบนมาตรฐานเท่ากับ 1 เพื่อให้ค่าสังเกตในส่วนนี้มีค่าที่แตกต่างจากค่าสังเกตในส่วนที่ 1 อย่างชัดเจน และแน่ใจได้ว่าค่าที่ได้เป็นค่าผิดปกติจริง

การเปรียบเทียบประสิทธิภาพของตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติทั้ง 5 วิธี พิจารณาจากสัดส่วนที่ตัวสถิติเหล่านี้ตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดทุกค่าที่มีอยู่ในชุดข้อมูล ซึ่งหาได้จาก

สัดส่วนของการตรวจสอบค่าผิดปกติถูกต้อง = จำนวนครั้งที่ตัวสถิติตรวจสอบค่าผิดปกติได้ถูกต้องทุกค่า/10,000

หากตัวสถิติใดมีสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงแสดงว่าเป็นตัวสถิติที่มีประสิทธิภาพในการตรวจสอบค่าผิดปกติในชุดข้อมูล

ผลการวิจัยและวิจารณ์ผล

กรณีชุดข้อมูลมีค่าผิดปกติในตัวแปรอิสระ จากตารางที่ 1 จะเห็นได้ว่าในทุกขนาดตัวอย่าง เมื่อมีค่าผิดปกติ 1 ค่า ตัวสถิติ h_{ii} และ CVR_i ให้ค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุด (1.000) และเมื่อจำนวนค่าผิดปกติเท่ากับร้อยละ 10 ของขนาดตัวอย่าง ตัวสถิติ h_{ii} มีประสิทธิภาพดีที่สุด รองลงมาคือ CVR_i และประสิทธิภาพของตัวสถิติจะลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น อย่างไรก็ตาม เมื่อร้อยละของจำนวนค่าผิดปกติเป็นร้อยละ 20 และ 30 ของขนาดตัวอย่าง ตัวสถิติทุกชนิดไม่สามารถตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมด (0.000)

ตารางที่ 1 สัดส่วนของการตรวจสอบค่าผิดปกติในตัวแปรอิสระได้ถูกต้องทั้งหมดของตัวสถิติ 5 วิธี

ขนาดตัวอย่าง	จำนวนค่าผิดปกติ	h_{ii}	t_i	D_i	$DFFITS_i$	CVR_i
30	1	1.0000	0.0015	0.8109	0.8176	1.0000
	3	0.9998	0.0000	0.0405	0.0515	0.6778
	6	0.0005	0.0000	0.0000	0.0000	0.0155
	9	0.0000	0.0000	0.0000	0.0000	0.0000
50	1	1.0000	0.0008	0.8125	0.8154	1.0000
	5	0.9986	0.0000	0.0024	0.0039	0.4342
	10	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000
100	1	1.0000	0.0002	0.8045	0.8065	1.0000
	10	0.9966	0.0000	0.0000	0.0000	0.1110
	20	0.0000	0.0000	0.0000	0.0000	0.0000
	30	0.0000	0.0000	0.0000	0.0000	0.0000

กรณีชุดข้อมูลมีค่าผิดปกติในตัวแปรตาม จากตารางที่ 2 จะเห็นได้ว่าในทุกขนาดตัวอย่าง เมื่อมีค่าผิดปกติ 1 ค่า ตัวสถิติ t_i , D_i, $DFFITS_i$ และ CVR_i ให้ค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุด (1.000) และเมื่อจำนวนค่าผิดปกติเท่ากับร้อยละ 10 ของขนาดตัวอย่าง ตัวสถิติ CVR_i ให้ค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุด รองลงมาคือ $DFFITS_i$ และประสิทธิภาพของตัวสถิติเหล่านี้จะลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น อย่างไรก็ตาม เมื่อจำนวนค่าผิดปกติเป็นร้อยละ 20 และ 30 ตัวสถิติทุกชนิดไม่สามารถตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมด (0.000)

ตารางที่ 2 สัดส่วนของการตรวจสอบค่าผิดปกติในตัวแปรตามได้ถูกต้องทั้งหมดของตัวสถิติ 5 วิธี

ขนาดตัวอย่าง	จำนวนค่าผิดปกติ	h_{ii}	t_i	D_i	$DFFITS_i$	CVR_i
30	1	0.0780	1.0000	1.0000	1.0000	1.0000
	3	0.0002	0.0003	0.4053	0.7802	0.8240
	6	0.0000	0.0000	0.0001	0.0020	0.0000
	9	0.0000	0.0000	0.0000	0.0000	0.0000
50	1	0.0786	1.0000	1.0000	1.0000	1.0000
	5	0.0000	0.0000	0.2325	0.4841	0.7393
	10	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000
100	1	0.0785	1.0000	1.0000	1.0000	1.0000
	10	0.0000	0.0000	0.0582	0.1248	0.5830
	20	0.0000	0.0000	0.0000	0.0000	0.0000
	30	0.0000	0.0000	0.0000	0.0000	0.0000

กรณีชุดข้อมูลมีค่าผิดปกติทั้งในตัวแปรอิสระและตัวแปรตาม จากตารางที่ 3 จะเห็นได้ว่าเมื่อมีค่าผิดปกติ 1 ค่าและขนาดตัวอย่างเท่ากับ 30 (ในค่าสังเกตที่ 30 ตัวแปรตามและตัวแปรอิสระจะมีค่าเป็นค่าผิดปกติทั้งคู่) ตัวสถิติ h_{ii} จะให้ค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุด (1.000) รองลงมาคือ D_i และ $DFFITS_i$ ในขณะที่เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 ตัวสถิติ h_{ii} , D_i และ $DFFITS_i$ ให้ค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุด (1.000) เมื่อจำนวนค่าผิดปกติเท่ากับร้อยละ 10 ของขนาดตัวอย่าง ตัวสถิติ h_{ii} มีค่าสัดส่วนของการตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดสูงที่สุดที่ทุกขนาดตัวอย่าง และประสิทธิภาพจะลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น อย่างไรก็ตาม เมื่อจำนวนค่าผิดปกติเป็นร้อยละ 20 และ 30 ตัวสถิติทุกชนิดไม่สามารถตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมด (0.000)

ตารางที่ 3 *สัดส่วนของการตรวจสอบค่าผิดปกติทั้งในตัวแปรอิสระและตัวแปรตามได้ถูกต้องทั้งหมดของตัวสถิติ 5 วิธี*

ขนาดตัวอย่าง	จำนวนค่าผิดปกติ	h_{ii}	t_i	D_i	$DFFITS_i$	CVR_i
30	1	1.0000	0.4796	0.9992	0.9993	0.8944
	3	0.9998	0.0000	0.0861	0.1032	0.5632
	6	0.0004	0.0000	0.0000	0.0000	0.0161
	9	0.0000	0.0000	0.0000	0.0000	0.0000
50	1	1.0000	0.7736	1.0000	1.0000	0.8878
	5	0.9991	0.0000	0.0055	0.0069	0.2971
	10	0.0000	0.0000	0.0000	0.0000	0.0000
	15	0.0000	0.0000	0.0000	0.0000	0.0000
100	1	1.0000	0.9734	1.0000	1.0000	0.8580
	10	0.9961	0.0000	0.0000	0.0000	0.0527
	20	0.0000	0.0000	0.0000	0.0000	0.0000
	30	0.0000	0.0000	0.0000	0.0000	0.0000

สรุปผลการวิจัย

การตรวจสอบค่าผิดปกติในชุดข้อมูลสำหรับการวิเคราะห์การถดถอยเป็นสิ่งที่ผู้ศึกษาไม่ควรละเลย เนื่องจากการมีค่าผิดปกติในชุดข้อมูลอาจส่งผลกระทบให้ผลการวิเคราะห์การถดถอยที่ได้ผิดไปจากที่ควรจะเป็น ในการศึกษานี้จึงได้มีการเปรียบเทียบประสิทธิภาพของตัวสถิติที่ใช้สำหรับการตรวจสอบค่าผิดปกติในชุดข้อมูล 5 วิธี ซึ่งเป็นวิธีที่นิยมใช้กันอย่างแพร่หลาย โดยจากการศึกษาพบว่า ในทุกขนาดตัวอย่างเมื่อมีค่าผิดปกติ 1 ค่าสังเกตในชุดข้อมูล ตัวสถิติ h_{ii} และ CVR_i มีประสิทธิภาพในการตรวจสอบค่าผิดปกติดีที่สุดเมื่อค่าผิดปกติอยู่ในตัวแปรอิสระ ในขณะที่เมื่อค่าผิดปกตินั้นอยู่ในตัวแปรตามตัวสถิติ t_i , D_i , $DFFITS_i$ และ CVR_i มีประสิทธิภาพในการตรวจสอบค่าผิดปกติมากที่สุด และเมื่อค่าผิดปกติอยู่ทั้งในตัวแปรอิสระและตัวแปรตามตัวสถิติ h_{ii} , D_i และ $DFFITS_i$ มีประสิทธิภาพในการตรวจสอบค่าผิดปกติมากที่สุด อย่างไรก็ตาม ตัวสถิติเหล่านี้จะมีประสิทธิภาพในการตรวจสอบค่าผิดปกติลดลงเมื่อจำนวนค่าผิดปกติเพิ่มขึ้นเป็นร้อยละ 10 ของชุดข้อมูล และไม่สามารถตรวจสอบค่าผิดปกติได้ถูกต้องทั้งหมดทุกค่าเมื่อมีค่าผิดปกติจำนวนมากในชุดข้อมูล ซึ่งสอดคล้องกับ Rousseeuw and Leroy (2003, p. 227-234) ที่กล่าวว่าตัวสถิติเหล่านี้สามารถใช้ตรวจสอบค่า

ผิดปกติในการวิเคราะห์การถดถอยได้ดีเมื่อมีค่าผิดปกติเพียง 1 ค่าในชุดข้อมูล สำหรับกรณีที่มีค่าผิดปกติในชุดข้อมูลหลายค่า Rousseeuw and Leroy (2003) ได้แนะนำตัวสถิติที่พัฒนาจากตัวสถิติที่ใช้ในการตรวจสอบค่าผิดปกติ 1 ค่า ได้แก่ ตัวสถิติ D_i ของ Cook and Weisberg ที่นำเสนอในปี 1982 และ Andrews and Pregibon ได้นำเสนอตัวสถิติ Andrew-Pregibon (AP_i) ในปี 1978 และ Marubini and Orenti (2014) ได้เสนอวิธีการตรวจสอบค่าผิดปกติหลายค่าในการวิเคราะห์การถดถอยโดยใช้วิธี robust two-stage procedure ซึ่งผู้วิจัยจะได้ทำการศึกษาและพัฒนาหาวิธีที่มีประสิทธิภาพสำหรับการตรวจสอบค่าผิดปกติหลายค่าในการวิเคราะห์การถดถอยเชิงเส้นต่อไป

เอกสารอ้างอิง

Ampanthong, P. & Prachoom, S. (2009). A comparative study of outlier detection procedures in multiple linear regression. In *Proceeding of the International MultiConference of Engineers and Computer Scientists 2009*. Hong Kong.

Marubini, E. and Orenti, A. (2014). Detecting outliers and/or leverage points: a robust two-stage procedure with bootstrap cut-off points. *Epidemiology Biostatistics and Public Health*, *11*(3).

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust regression and outlier detection*. New Jersey: John Wiley & Sons.

Zakaria, A., Howard, N. K. and Nkansah, B. K. (2014). On the detection of influential outliers in linear regression analysis. *American Journal of Theoretical and Applied Statistics*, *3*(4), 100-106.