บทความวิจัย

การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นอย่างง่าย ด้วยวิธีทีล วิธีควอนไทล์ และวิธีกำลังสองน้อยที่สุด เมื่อข้อมูลในตัวแปรอิสระและตัวแปรตามมีค่าผิดปกติ

ชัยวัฒน์ สุวรรณาภินันท์* ธิดาพร ศุภภากร และ ลี่ลี อิงศรีสว่าง

บทคัดย่อ

การวิจัยนี้ศึกษาเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นอย่างง่าย เมื่อข้อมูลในตัวแปรอิสระ และตัวแปรตามมีค่าผิดปกติ โดยเปรียบเทียบค่าสัมประสิทธิ์การถดถอยเชิงเส้น 3 วิธีคือ วิธีทีล วิธีควอนไทล์ และวิธีกำลังสองน้อยที่สุด โดยกำหนดขนาดตัวอย่างแบ่งเป็น 3 กลุ่ม คือ ขนาดเล็ก (10, 20) ขนาดกลาง (30, 40) และขนาดใหญ่ (70, 90) กำหนดระดับความผิดปกติในตัวแปรอิสระ 2 ระดับ คือ ระดับปานกลาง และระดับรุนแรง สัดส่วนค่าผิดปกติในตัวแปรอิสระ คือ 0.1, 0.2 และ 0.3 ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3, 5 และ 7 เกณฑ์ที่ใช้ในการศึกษาคือค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ของตัวประมาณพารามิเตอร์ โดยทำการจำลองทั้งหมด 1,000 ครั้งในแต่ละสถานการณ์

ผลการศึกษาพบว่า เมื่อข้อมูลไม่มีค่าผิดปกติวิธีกำลังสองน้อยที่สุดมีประสิทธิภาพดีที่สุด เมื่อข้อมูลมีค่าผิดปกติ ในขนาดตัวอย่างเล็ก พบว่าวิธีทีลมีประสิทธิภาพดีที่สุด ยกเว้นกรณีขนาดตัวอย่างเท่ากับ 10 ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 ค่าผิดปกติระดับรุนแรง สัดส่วนค่าผิดปกติเท่ากับ 0.1 วิธีกำลังสองน้อยที่สุดมีประสิทธิภาพดีที่สุด และในขนาดตัวอย่างกลางและใหญ่ พบว่าวิธีทีลมีประสิทธิภาพดีที่สุดในทุกๆ กรณี

คำสำคัญ: วิธีทีล วิธีควอนไทล์ วิธีกำลังสองน้อยที่สุด การถดถอยเชิงเส้นอย่างง่าย

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

*ผู้นิพนธ์ประสานงาน, e-mail: chaiyawat_big@hotmail.com

A Comparison on Efficiency of Theil, Quantile and Ordinary Least Square Estimation Methods of Simple Linear Regression Coefficients with Outliers on Independent and Dependent Variables

Chaiyawat Suwannapinant*, Thidaporn Supapakorn and Lily Ingsrisawang

ABSTRACT

The objective of this paper is to compare the efficiency of Theil, Quantile and Ordinary Least Square (OLS) estimation methods of simple linear regression coefficients with outliers on independent and dependent variables. The sample sizes are set into 3 levels; small (10, 20), medium (30, 40) and large (70, 90). There are two levels of outliers; mild and extreme. The percentages of outliers on independent variable are defined as 0.1, 0.2 and 0.3. At the position of outlier, the error term is normal distribution with means equal to 3, 5 and 7 and variance equal to 1. The criterion of this study is the Mean Square Error (MSE) of estimated parameters. The data used in research is obtained by simulating 1,000 times for each situation.

The result shows that; in case of no outlier, OLS provides the most efficient method. In case of situation with outliers at the same position, Theil is the most efficient, except when the sample size is 10, the level of outlier is extreme, the outlier proportion of independent variable and error terms is 0.1, and at the position of outliers, the error term has normal distribution with mean of 3, the most efficient method is the OLS. In medium and large sample sizes. Theil is the most efficient method.

Keywords: Theil Method, Quantile Method, Ordinary Least Square Method, Simple Linear Regression

Department of Statistics, Faculty of Science, Kasetsart University

*Corresponding author, e-mail: chaiyawat_big@hotmail.com

บทนำ

การวิเคราะห์การถดถอย เป็นวิธีการทางสถิติที่แสดงถึงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามโดยสามารถนำไปใช้พยากรณ์หรือวิเคราะห์ความสัมพันธ์ โดยพื้นฐานการศึกษาการวิเคราะห์การถดถอยจะเริ่มศึกษาจากการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย เพื่อให้ผู้ศึกษาเข้าใจถึงรูปแบบความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระหนึ่งตัวแปร ในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นอย่างง่ายมีวิธีการประมาณค่าหลายวิธีที่สามารถเลือกใช้ ซึ่งในทางปฏิบัติการได้มาของข้อมูลอาจจะเกิดข้อผิดพลาดในการได้มา [1] ทำให้ข้อมูลมีค่าผิดปกติ การแก้ไขข้อมูลที่มีค่าผิดปกติเบื้องต้นเราต้องสืบหาแหล่งที่มาและแก้ไขในส่วนที่ผิด แต่เมื่อเราสืบค้นแหล่งที่มาแล้วข้อมูลเกิดผิดปกติจริงๆ การนำมาคำนวณเพื่อประมาณค่าสัมประสิทธิ์ควรที่จะเลือกใช้วิธีให้เหมาะสม การประมาณค่าสัมประสิทธิ์เราสามารถแบ่งออกเป็น 3 กลุ่มใหญ่ คือ การประมาณค่าด้วยวิธีพาราเมตริก การประมาณค่าด้วยวิธีกึ่งพาราเมตริก การประมาณค่าด้วยวิธีนอนพาราเมตริก โดยในงานวิจัยนี้จะได้กล่าวถึงเฉพาะ 2 กลุ่ม คือ การประมาณค่าด้วยวิธีพาราเมตริกและการประมาณค่าด้วยวิธีนอนพาราเมตริก

วิธีกำลังสองน้อยที่สุดเป็นวิธีการประมาณค่าแบบพาราเมตริกที่เป็นพื้นฐานในการหาค่าสัมประสิทธิ์โดยเป็นวิธีที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อนมีค่าต่ำที่สุด โดยในการหาค่าสัมประสิทธิ์ด้วยวิธีนี้จะต้องอยู่บนข้อสมมุติที่ว่าด้วยความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเป็นแบบเส้นตรง ค่าคาดหวังของความคลาดเคลื่อนเป็นศูนย์ และค่าความแปรปรวนของความคลาดเคลื่อนคงที่ [2]

วิธีทีลและวิธีควอนไทล์เป็นวิธีการประมาณค่าแบบนอนพาราเมตริก วิธีทีลและวิธีควอนไทล์สามารถละทิ้งข้อกำหนดเบื้องต้นบางข้อของวิธีกำลังสองน้อยที่สุดได้ เช่น การแจกแจงของค่าคลาดเคลื่อนมีการแจกแจงแบบปกติรวมถึงวิธีทีลเป็นวิธีนอนพาราเมตริกที่มีความไวต่อข้อมูลที่ผิดปกติน้อยกว่าวิธีกำลังสองน้อยที่สุดโดยมีจุดความทนต่อค่าผิดปกติที่ 29.3% [3] และวิธีควอนไทล์สามารถเลือกตำแหน่งมัธยฐานในการวิเคราะห์เพื่อลดความอ่อนไหวต่อค่าผิดปกติในตัวแปรตาม [4]

ดังนั้นผู้วิจัยจึงสนใจศึกษาเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นอย่างง่าย ด้วยวิธีทีลและวิธีควอนไทล์ซึ่งเป็นวิธีการประมาณค่าแบบนอนพาราเมตริก และวิธีกำลังสองน้อยที่สุดซึ่งเป็นวิธีการประมาณค่าแบบพาราเมตริก เมื่อข้อมูลในตัวแปรอิสระ และตัวแปรตามมีค่าผิดปกติ โดยจากการตรวจเอกสารพบว่า อุมาพร และ วราพร [5] ได้ทำการเปรียบเทียบประสิทธิภาพของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายด้วยวิธีกำลังสองน้อยที่สุด วิธีของ Theil และวิธีของ Brown-Mood ซึ่งกำหนดสถานการณ์ให้ตัวแปรอิสระมีการแจกแจงแบบปกติและไวบูลล์ ความคลาดเคลื่อนมีการแจกแจงแบบปกติ ลอกนอร์มอล และไวบูลล์ ทำการสร้างสถานการณ์ทดลองซ้ำ 500 ครั้ง ผลที่ได้พบว่าวิธีกำลังสองน้อยที่สุดให้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนที่ต่ำที่สุดในทุกกรณี และวิธีทีลให้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนที่ใกล้เคียงวิธีกำลังสองน้อยที่สุด ทำให้ผู้วิจัยมีความสนใจในกรณีที่แตกต่างกัน โดยผู้วิจัยจะพิจารณาจากค่าคลาดเคลื่อนกำลังสองเฉลี่ย ในวิธีกำลังสองน้อยที่สุด และวิธีทีล โดยทั้ง 3 วิธีเป็นที่นิยมในการนำเอาไปใช้วิเคราะห์ข้อมูลในด้านต่างๆ อาทิ การประเมินความยั่งยืนของระบบชลประทานด้วยการประมาณค่าวิธีทีลในข้อมูลอนุกรมเวลา [6] งานวิจัยนี้ใช้วิธีทีลในการประมาณค่าสัมประสิทธิ์ปริมาณน้ำ พื้นที่เพาะปลูกพืชชนิดต่างๆ การประเมินค่าเชิงเส้นโดยไม่ใช้วิธีกำลังสองน้อยที่สุด การประยุกต์การประมาณค่าด้วยวิธีทีล [7] งานวิจัยนี้มีการนำวิธีทีลมาประมาณค่าสัมประสิทธิ์ รายได้

หนี้สิน สินทรัพย์ การประมาณสมการภาคที่ใช้ข้อมูลภาคตัดขวาง ดังสมการ $EPS_{t+1} = aP_t + bEPS_t + u_{t+1}$ ระหว่างประเทศจีนและสหรัฐอเมริกาว่าเพื่อค้นหาค่าสัมประสิทธิ์ที่ประมาณได้ให้เป็นไปตามข้อกำหนดทางเศรษฐศาสตร์ [8] งานวิจัยนี้ใช้วิธีทีลในการประมาณค่าสัมประสิทธิ์ของอัตราส่วนกำไรสุทธิต่อหุ้น และราคา การศึกษาปัญหาที่ทำให้เกษตรกรได้ผลผลิตต่ำจากการปลูกข้าวโพดโดยใช้ควอนไทล์รีเกรสชั่นในฟิลิปปินส์ [9] โดยงานวิจัยที่นำเอาวิธีทีลและควอนไทล์มาใช้ งานวิจัยที่กล่าวมาข้างต้นส่วนมากต้องการตัวประมาณค่าที่มีความแกร่ง เนื่องจากข้อมูลที่ได้มานั้นมีการแกว่งตัวสูง และการประมาณค่าด้วยวิธีควอนไทล์สามารถวิเคราะห์ถึงผลกระทบของข้อมูลได้หลายกลุ่ม

วัตถุประสงค์

เพื่อศึกษาเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าพารามิเตอร์ในรูปแบบถดถอยเชิงเส้นอย่างง่าย เมื่อข้อมูลไม่มีค่าผิดปกติ เมื่อข้อมูลมีค่าผิดปกติในตัวแปรอิสระ และความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3, 5 และ 7 ความแปรปรวนเท่ากับ 1

เกณฑ์ที่ใช้ในการศึกษา

ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ของตัวประมาณพารามิเตอร์ [10] ทั้ง 2 วิธี จำนวน 1,000 ครั้งในแต่ละสถานการณ์ โดยวิธีที่ให้ค่า MSE ต่ำกว่าจะเป็นวิธีที่มีประสิทธิภาพมากกว่า

$$MSE = \frac{\sum_{i=0}^{p}\sum_{t=1}^{1000}\left[\frac{\left(\beta_{it}-\hat{\beta}_{it}\right)^2}{p+1}\right]}{1000}$$

โดยที่ β_{it} คือ ค่าพารามิเตอร์ตัวที่ i ในการทำซ้ำรอบที่ t

$\hat{\beta}_{it}$ คือ ค่าประมาณพารามิเตอร์ตัวที่ i ในการทำซ้ำรอบ ที่ t

p คือ จำนวนตัวแปรอิสระในรูปแบบการถดถอย โดยในงานวิจัยนี้ $p = 1$

t คือ รอบที่ทำซ้ำ โดยในงานวิจัยนี้ $t = 1, 2, 3, \ldots, 1000$

i คือ จำนวนตัวพารามิเตอร์ โดยในงานวิจัยนี้ $i = 0, 1$

วิธีดำเนินการศึกษา

1. ในการวิจัยครั้งนี้ใช้เทคนิคมอนติคาร์โลในการจำลองข้อมูล กำหนดขนาดตัวอย่าง มีค่าเท่ากับ 10, 20, 30, 40, 70 และ 90 กำหนดสัดส่วนการปลอมปนของค่าผิดปกติในข้อมูลตัวแปรอิสระเท่ากับ 0, 0.1, 0.2 และ 0.3 โดยระดับความผิดปกติในตัวแปรอิสระแบ่งเป็น 3 ระดับ คือ ไม่ผิดปกติ ระดับปานกลาง และระดับรุนแรง โดยการสร้างค่าผิดปกติ มีวิธีการดังนี้

ขั้นตอนที่ 1 จำลองข้อมูลของตัวแปรอิสระให้มีการแจกแจงแบบปกติมาตรฐาน คือ $X \sim N(0,1)$ จำนวน n ค่า แล้วนำค่าจากการจำลองมาตรวจสอบค่าผิดปกติด้วยวิธีกราฟ Box and Whisker

โดยข้อมูลไม่มีค่าผิดปกติจะต้องอยู่ในช่วง $(Q_1 - 1.5(IQR),\ Q_3 + 1.5(IQR))$ แล้วเราจะได้ค่าข้อมูลของตัวแปรอิสระที่ไม่มีค่าผิดปกติ

ขั้นตอนที่ 2 นำข้อมูลจากขั้นตอนที่ 1 มาเรียงจากน้อยไปหามากดังนี้ $X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)}$ เมื่อ $X_{(i)}$ เป็นค่าตัวแปรอิสระในลำดับที่ i โดยที่ $i = 1, 2, \ldots, n$

ขั้นตอนที่ 3 แทนค่าของตัวแปรอิสระ ที่มีค่ามากที่สุดด้วยค่าคงที่ค่าหนึ่งโดยให้มีค่ามากขึ้นกว่าเดิม หรือแทนค่าของตัวแปรอิสระ ที่มีค่าน้อยที่สุดด้วยค่าคงที่ค่าหนึ่งโดยให้มีค่าน้อยลงกว่าเดิม โดยกำหนดให้ค่าผิดปกติของตัวแปรอิสระมีค่าดังนี้

$$X_{(i)} = Q_1 - L(QR) \quad \text{เมื่อตัวแปรอิสระมีค่าน้อย}$$
$$X_{(i)} = Q_3 + L(IQR) \quad \text{เมื่อตัวแปรอิสระมีค่ามาก}$$

โดย Q_1 คือ ค่าควอไทล์ที่ 1 ของตัวแปรอิสระ

Q_3 คือ ค่าควอไทล์ที่ 3 ของตัวแปรอิสระ

$IQR = Q_3 - Q_1$

$L = 2$ เมื่อกำหนดให้ค่าผิดปกติระดับปานกลาง

$L = 5$ เมื่อกำหนดให้ค่าผิดปกติระดับรุนแรง

ขั้นตอนที่ 4 จำลองข้อมูลค่าคลาดเคลื่อน โดยทำการสร้างข้อมูลของตัวแปรอิสระให้มีค่าผิดปกติ ตามขั้นตอนที่ 2 แล้วกำหนดให้ความคลาดเคลื่อนที่เกิดขึ้น ณ ตำแหน่งที่มีค่าของตัวแปรอิสระผิดปกติมีการแจกแจงแบบปกติที่มีค่าคาดหวังเท่ากับ M และค่าความแปรปรวนเท่ากับ 1 จะได้ว่า $\varepsilon_i \sim N(M, 1)$; $M = 3,5,7$ และกำหนดให้ความคลาดเคลื่อนที่เกิดขึ้น ณ ตำแหน่งที่มีค่าของตัวแปรอิสระปกติมีการแจกแจงแบบปกติมาตรฐาน $\varepsilon_i \sim N(0, 1)$

ขั้นตอนที่ 5 จำลองข้อมูลตัวแปรตาม ตามรูปแบบความสัมพันธ์สมการถดถอยเชิงเส้นตรงอย่างง่าย ดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\ ;\ i = 1,2\ ,\ldots,\ n$$

เมื่อ β_0, β_1 คือ ค่าสัมประสิทธิ์การถดถอย ซึ่งในการวิจัยครั้งนี้กำหนดให้ $\beta_0 = \beta_1 = 1$

2. ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีการประมาณ 2 วิธี คือ วิธีกำลังสองน้อยที่สุด และวิธีทีล

2.1 วิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method : OLS) [2]

รูปแบบการถดถอยเชิงเส้น ที่แสดงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม คือ

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

เมื่อ $\underline{Y}$ เป็นเวกเตอร์ขนาด $n\times 1$ ของค่าสังเกตของตัวแปรตาม

X เป็นเมตริกซ์ขนาด $n\times(p+1)$ ของค่าสังเกตของตัวแปรอิสระ

$\underline{\beta}$ เป็นเวกเตอร์ขนาด $(p+1)\times 1$ ของค่าพารามิเตอร์ β

$\underline{\varepsilon}$ เป็นเวกเตอร์ขนาด $n\times 1$ ของความคลาดเคลื่อน

p เป็นจำนวนตัวแปรอิสระ

n เป็นขนาดตัวอย่าง

จากวิธีกำลังสองน้อยที่สุด สามารถหาค่าประมาณของสัมประสิทธ์การถดถอย ($\underline{\hat{\beta}}$) ได้โดยหาค่าประมาณสัมประสิทธิ์การถดถอยที่ทำให้ผลบวกกำลังสองของความแตกต่าง ระหว่างค่าจริงกับค่าประมาณ หรือผลบวกกำลังสองของความคลาดเคลื่อน (Sum Square of Error : SSE) มีค่าน้อยที่สุด

$$\begin{aligned} SSE &= (\underline{Y}-X\underline{\hat{\beta}})^T(\underline{Y}-X\underline{\hat{\beta}}) \\ &= (\underline{Y}^T\underline{Y}-2\underline{\hat{\beta}}^TX^T\underline{Y}+\underline{\hat{\beta}}^TX^TX\underline{\hat{\beta}} \end{aligned}$$

หาค่า $\underline{\hat{\beta}}$ ได้โดยการหาอนุพันธ์ของ SSE เทียบกับ $\underline{\hat{\beta}}$ แล้วกำหนดให้เท่ากับ 0 จะได้

$$\underline{\hat{\beta}}=(X^TX)^{-1}X^T\underline{Y}$$

2.2 วิธีทีล (Theil)

Theil (1950) [11] เสนอวิธีประมาณค่าความชัน (β_1) ของเส้นถดถอย $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$ เมื่อ $i=1, 2,...,n$ ซึ่งมีขั้นตอนดังนี้

ขั้นตอนที่ 1 คำนวนค่าความชัน

$S_{ij}=\dfrac{y_j-y_i}{x_j-x_i}$; $1\le i<j\le n$ จะได้ค่าความชันทั้งหมด $N=\dfrac{n(n-1)}{2}$ ค่า

ขั้นตอนที่ 2 ประมาณค่า β_1

$$\hat{\beta}_1 = \text{median }(S_{ij}, 1\le i<j\le n)$$

ถ้า N เป็นเลขคี่, $N = 2k + 1$ จะได้ $\hat{\beta}_1 = S^{(k+1)}$

ถ้า N เป็นเลขคู่, $N = 2k$ จะได้ $\hat{\beta}_1 = \dfrac{S^{(k)}+S^{(k+1)}}{2}$

ขั้นตอนที่ 3 ประมาณค่า β_0

$$\hat{\beta}_0 = M_y-\hat{\beta}_1M_x$$

เมื่อ M_y และ M_x เป็นค่ามัธยฐานของข้อมูล X และ Y ตามลำดับ

2.3 วิธีควอนไทล์ (Quantile)

ในวิธีควอนไทล์ได้ประยุกต์ Conditional Mean Function โดยเรียกว่า Generic Conditional Quantile Function [4]

$$\hat{q}_Y(\theta, X) = \underset{Q_Y(\theta,X)}{\operatorname{argmin}} E[Y-Q_Y(\theta, X)] \text{ และ } Q_Y(\theta,X)=Q_\theta\left[Y\middle|X=\underline{x}\right]$$

โดย $\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}}\, E[\rho_\theta(\mathrm{Y} - \mathrm{X}\beta)]$

และ
$$\begin{aligned}\rho_\theta(y) &= \left|[\theta - I(y<0)]y\right| \\ &= [(1-\theta)I(y \le 0) + \theta I(y>0)]\left|y\right|\end{aligned}$$

เมื่อ θ คือตำแหน่งของควอนไทล์และ $0<\theta<1$

วิธีการหาค่า $\hat{\beta}(0)$ ใช้วิธีการโปรแกรมเชิงเส้น (Linear Programming) โดยในงานวิจัยนี้กำหนด θ = 0.5 นั่นคือ

ขั้นตอนที่ 1 กำหนดสมการวัตถุประสงค์ $\min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{n} \left|\beta_0 + \beta_1 x_i - y_i\right|$

ขั้นตอนที่ 2 ปรับสมการวัตถุประสงค์ให้อยู่ในรูปแบบเชิงเส้นตรงจะได้ $\min\limits_{\beta_0,\beta_1} \sum\limits_{i=1}^{n} e_i$ และมีสมการข้อจำกัด ดังนี้

$$e_i \ge \beta_0 + \beta_1 x_i - y_i \qquad ;i=1,2,...,n$$
$$e_i \ge -(\beta_0 + \beta_1 x_i - y_i) \qquad ;i=1,2,...,n$$

ซึ่ง $e_i \ge \max\left\{\beta_0 + \beta_1 x_i - y_i, -\left(\beta_0 + \beta_1 x_i - y_i\right)\right\} = \left|\beta_0 + \beta_1 x_i - y_i\right|$

ขั้นตอนที่ 3 ทำการคำนวณหาค่าด้วยวิธีซิมเพล็กซ์ (Simplex)

3. คำนวณค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error: MSE) ของตัวประมาณพารามิเตอร์ทั้ง 3 วิธี จำนวน 1,000 ครั้งในแต่ละสถานการณ์ ดังตารางที่ 1 โดยวิธีที่ให้ค่า MSE ต่ำกว่าจะเป็นวิธีที่มีประสิทธิภาพดีกว่า

ตารางที่ 1 สถานการณ์ในการจำลองค่าคลาดเคลื่อนกำลังสองเฉลี่ย

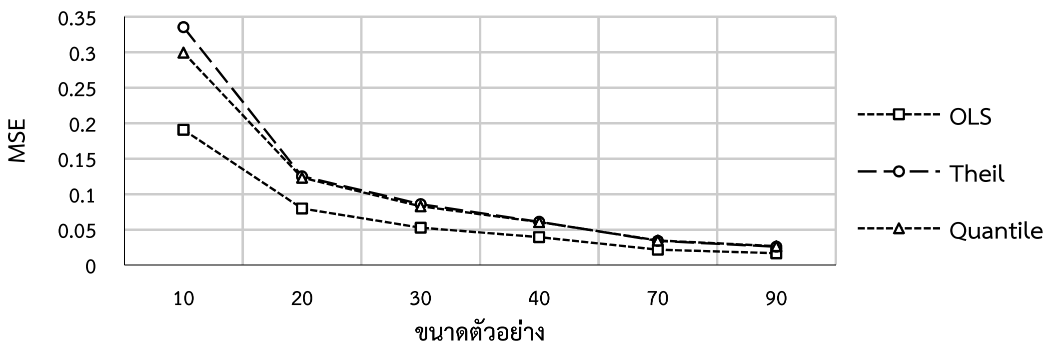
จำนวนค่าสังเกต (n)	**สัดส่วนค่าผิดปกติในตัวแปรอิสระ และความคลาดเคลื่อน**	**ระดับความผิดปกติในตัวแปรอิสระ**	**การแจกแจงของค่าผิดปกติของความคลาดเคลื่อน**
1. กรณีค่าสังเกตไม่มีค่าผิดปกติ 6 สถานการณ์			
ขนาดเล็ก (10, 20)	0	ไม่ผิดปกติ	N(0, 1)
ขนาดกลาง (30, 40)	0	ไม่ผิดปกติ	N(0, 1)
ขนาดใหญ่ (70, 90)	0	ไม่ผิดปกติ	N(0, 1)
2. กรณีค่าสังเกตของตัวแปรอิสระและความคลาดเคลื่อนผิดปกติ ณ ตำแหน่งเดียวกัน 108 สถานการณ์			
ขนาดเล็ก (10, 20)	0.1, 0.2 และ 0.3	ปานกลาง รุนแรง	N(3, 1), N(5, 1), N(7, 1)
ขนาดกลาง (30, 40)	0.1, 0.2 และ 0.3	ปานกลาง รุนแรง	N(3, 1), N(5, 1), N(7, 1)
ขนาดใหญ่ (70, 90)	0.1, 0.2 และ 0.3	ปานกลาง รุนแรง	N(3, 1), N(5, 1), N(7, 1)

ผลการศึกษา

การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นอย่างง่าย ด้วยวิธีทีล และวิธีกำลังสองน้อยที่สุด เมื่อข้อมูลในตัวแปรอิสระ และตัวแปรตามมีค่าผิดปกติ โดยใช้โปรแกรม R เวอร์ชัน 3.2.0 ในการจำลองข้อมูลและคำนวณค่าได้ผลการวิจัยดังนี้

1. เมื่อข้อมูลไม่มีค่าผิดปกติ

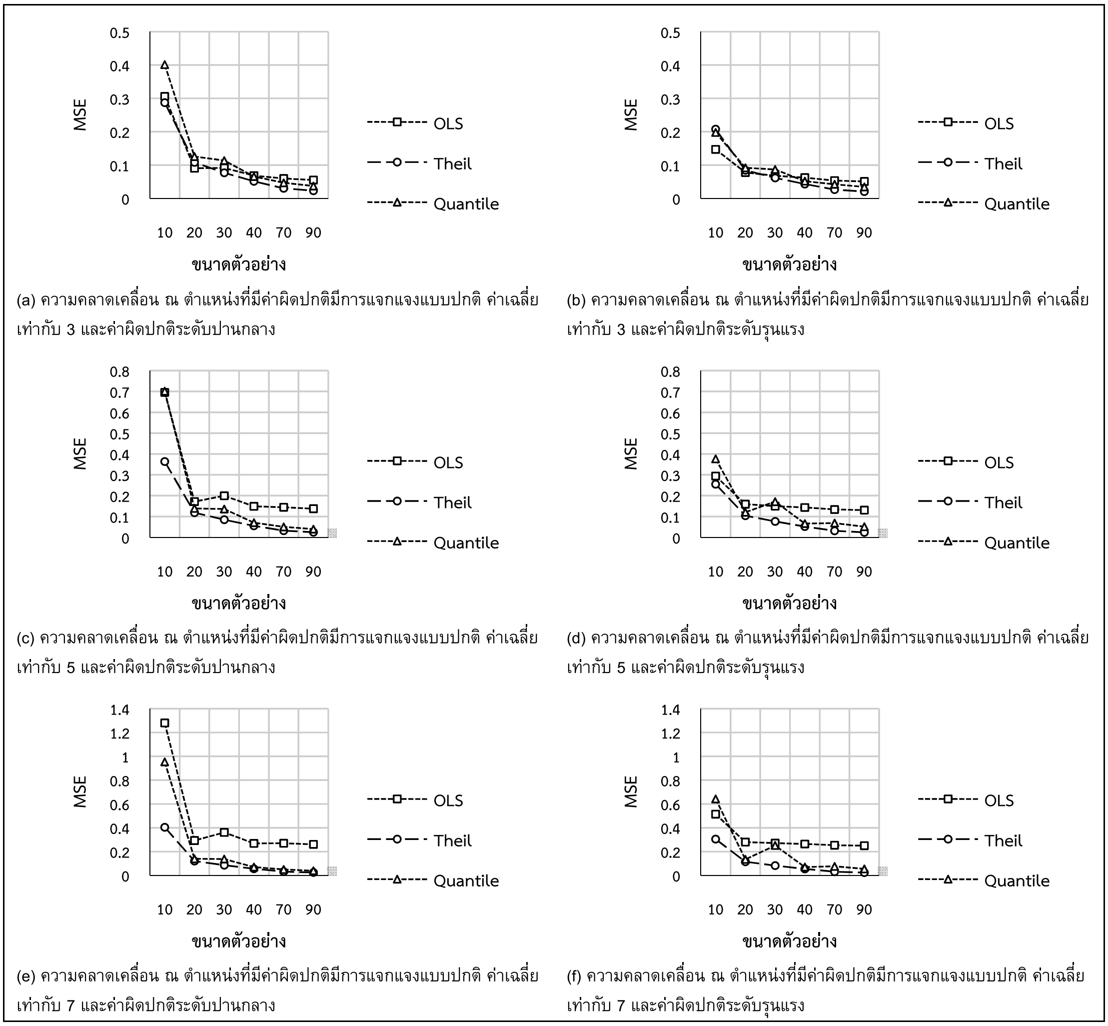
กรณีข้อมูลมีไม่มีค่าผิดปกติค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของวิธีกำลังสองน้อยที่สุดมีค่าต่ำกว่าในทุกๆ ขนาดตัวอย่าง แต่เมื่อขนาดตัวอย่างเพิ่มขึ้นจากขนาดกลาง (30, 40) จนถึงขนาดใหญ่ (70, 90) ค่า MSE ของวิธีทีลและวิธีควอนไทล์มีค่าเข้าใกล้วิธีกำลังสองน้อยที่สุดจนแทบจะไม่แตกต่างกันเลย ดังภาพที่ 1



ภาพที่ 1 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อข้อมูลไม่มีค่าผิดปกติ

2. เมื่อข้อมูลมีการแจกแจงผิดปกติ โดยสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.1

กรณีขนาดตัวอย่างเล็ก (10, 20) พบว่าทั้ง 3 วิธี ให้ค่า MSE ใกล้เคียงกันทุกๆ กรณี โดยในระดับค่าผิดปกติปานกลางและรุนแรงพบว่ากรณีความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 วิธีกำลังสองน้อยที่สุดให้ค่า MSE ที่ต่ำกว่าวิธีทีลและวิธีควอนไทล์ ขนาดตัวอย่างปานกลาง (30, 40) พบว่าวิธีทีลให้ค่า MSE ที่ต่ำที่สุด และรองลงมาคือวิธีควอนไทล์ โดยค่าผิดปกติระดับปานกลางและสูงและวิธีควอนไทล์ให้ค่า MSE สูงขึ้นกว่าปกติ โดยมีความต่างกันอย่างเห็นได้ชัด ขนาดตัวอย่างใหญ่ (70, 90) พบว่าวิธีทีลให้ค่า MSE ที่ต่ำกว่าอย่างเห็นได้ชัดและวิธีควอนไทล์ให้ค่า MSE ที่ใกล้เคียงวิธีทีล และเมื่อค่าผิดปกติระดับปานกลางและสูง หรือ ความคลาดเคลื่อน ณ ตำแหน่งข้อมูลมีค่าผิดปกติเปลี่ยนแปลงผลที่ได้มีความต่างกันอย่างเห็นได้ชัดระหว่างค่า MSE ของวิธีกำลังสองน้อยที่สุด กับวิธีทีลและควอนไทล์ ดังภาพที่ 2



(a) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับปานกลาง

(b) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับรุนแรง

(c) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับปานกลาง

(d) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับรุนแรง

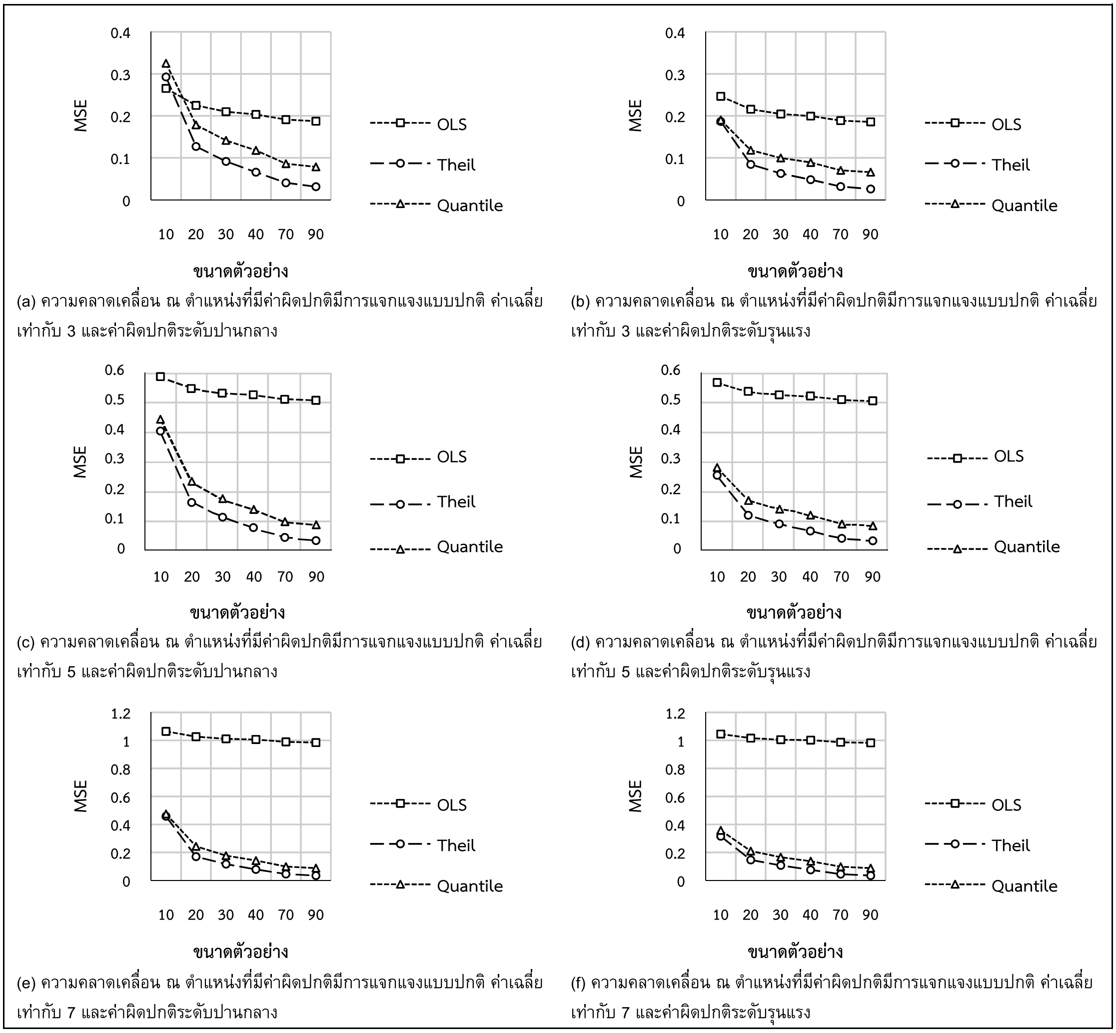
(e) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับปานกลาง

(f) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับรุนแรง

ภาพที่ 2 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อข้อมูลมีสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.1 ในกรณีต่างๆ

3. เมื่อข้อมูลมีการแจกแจงผิดปกติ โดยสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.2

กรณีขนาดตัวอย่างเล็ก (10, 20) พบว่าวิธีทีลให้ค่า MSE ที่ต่ำกว่าในทุกๆ กรณี โดยในระดับค่าผิดปกติรุนแรงพบว่ากรณีความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 วิธีกำลังสองน้อยที่สุดให้ค่า MSE ที่ต่ำกว่าวิธีทีลและวิธีควอนไทล์ ขนาดตัวอย่างปานกลาง (30, 40) พบว่าวิธีทีลให้ค่า MSE ที่ต่ำกว่าอย่างเห็นได้ชัดและวิธีควอนไทล์ให้ค่า MSE ที่ใกล้เคียงวิธีทีลเช่นเดียวกับขนาดตัวอย่างใหญ่ (70, 90) โดยเมื่อความคลาดเคลื่อน ณ ตำแหน่งข้อมูลมีค่าผิดปกติเปลี่ยนแปลง ผลที่ได้มีความต่างกันอย่างเห็นได้ชัดระหว่างค่า MSE ของวิธีกำลังสองน้อยที่สุด กับวิธีทีลและควอนไทล์ ซึ่งระดับค่าผิดปกติปานกลางและสูง มีความใกล้เคียงกันในทุกๆ ขนาดตัวอย่าง ดังภาพที่ 3



(a) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับปานกลาง

(b) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับรุนแรง

(c) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับปานกลาง

(d) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับรุนแรง

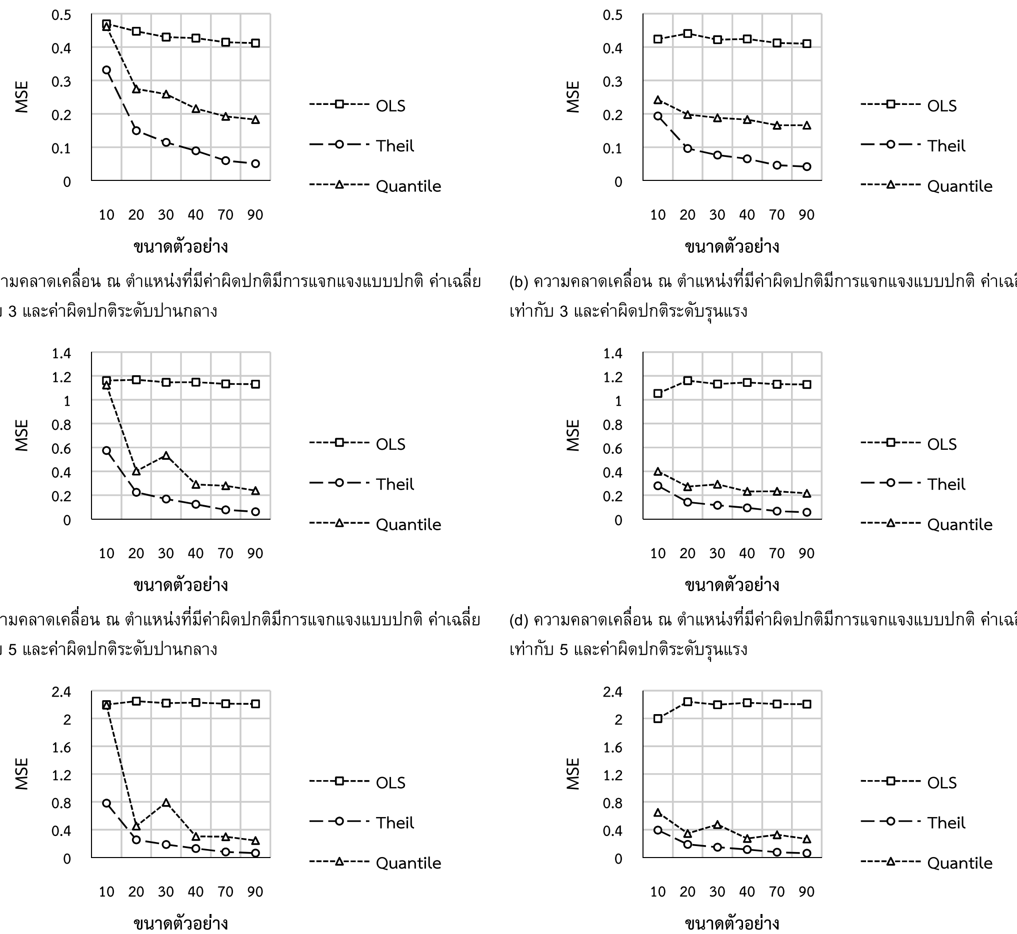
(e) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับปานกลาง

(f) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับรุนแรง

ภาพที่ 3 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อข้อมูลมีสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.2 ในกรณีต่างๆ

4. เมื่อข้อมูลมีการแจกแจงผิดปกติ โดยสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.3

กรณีขนาดตัวอย่างเล็ก (10, 20) พบว่าวิธีทีลและวิธีควอนไทล์ให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ต่ำกว่าในทุกๆ กรณี โดยวิธีทีลให้ค่า MSE ที่ต่ำกว่าอย่างเห็นได้ชัด เช่นเดียวกับขนาดตัวอย่างปานกลาง (30, 40) และขนาดตัวอย่างใหญ่ (70, 90) โดยเมื่อความคลาดเคลื่อน ณ ตำแหน่งข้อมูลมีค่าผิดปกติเปลี่ยนแปลง ผลที่ได้มีความใกล้เคียง เช่นเดียวกับระดับค่าผิดปกติปานกลางและสูง ก็มีความใกล้เคียงกันในทุกๆ ขนาดตัวอย่าง ดังภาพที่ 4



(a) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับปานกลาง

(b) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 3 และค่าผิดปกติระดับรุนแรง

(c) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับปานกลาง

(d) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 5 และค่าผิดปกติระดับรุนแรง

(e) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับปานกลาง

(f) ความคลาดเคลื่อน ณ ตำแหน่งที่มีค่าผิดปกติมีการแจกแจงแบบปกติ ค่าเฉลี่ยเท่ากับ 7 และค่าผิดปกติระดับรุนแรง

ภาพที่ 4 ค่าคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อข้อมูลมีสัดส่วนการปลอมปนของค่าผิดปกติเท่ากับ 0.3 ในกรณีต่างๆ

สรุป

เมื่อข้อมูลไม่มีค่าผิดปกติวิธีกำลังสองน้อยที่สุดมีประสิทธิภาพมากที่สุด เมื่อข้อมูลมีค่าผิดปกติขนาดตัวอย่างเล็ก วิธีทีลมีประสิทธิภาพที่ดีกว่า ในทุกๆ สัดส่วนการปลอมปนของค่าผิดปกติ (0.1, 0.2, 0.3) โดยวิธีควอนไทล์มีประสิทธิภาพรองลงมาในบางกรณี ขนาดตัวอย่างปานกลาง (30, 40) วิธีทีลมีประสิทธิภาพอย่างเห็นได้ชัด วิธีควอนไทล์มีประสิทธิภาพรองลงมาโดยในบางกรณีให้ค่า MSE ที่ใกล้เคียงกับวิธีกำลังสองน้อยที่สุด และวิธีกำลังสองน้อยที่สุดมีประสิทธิภาพแย่ที่สุด ในทุกๆ สัดส่วนการปลอมปนของค่าผิดปกติ (0.1, 0.2, 0.3) เช่นเดียวกับขนาดตัวอย่างใหญ่ (70, 90) จึงสามารถสรุปได้ว่าวิธีทีลมีประสิทธิภาพดีที่สุดในเกือบทุกๆ กรณี และวิธีควอนไทล์มีประสิทธิภาพรองลงมาในบางกรณี เนื่องจาก 2 วิธีมีคุณสมบัติตัวประมาณพารามิเตอร์ที่มีความแกร่ง และวิธีกำลังสองน้อยที่สุดให้ประสิทธิภาพที่แย่ที่สุดในเกือบทุกๆกรณี เนื่องจากข้อมูลมีค่าผิดปกติขัดต่อข้อกำหนดของวิธีกำลังสองน้อยที่สุดในการประมาณค่าพารามิเตอร์

เอกสารอ้างอิง

1. อรพรรณ ตันตระกูล. 2555. การเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่แกร่งสำหรับการถดถอยเชิงเส้นพหุ เมื่อข้อมูลมีค่าผิดปกติ. วิทยานิพนธ์ปริญญาโท. มหาวิทยาลัยเกษตรศาสตร์.
2. Kutner, M. H., Nachtsheim, J.C. Nachtsheim and J. Neter. 2004. Applied Linear Regression Models. 4^{th} edition. McGraw-Hill/Irwin: New York. p. 17-18, 199-201.
3. Wang, X. 2003. The Properties of the Theil-Sen Estimator. Dissertation, Binghamton University.
4. Davino, C., et al. 2013. Quantile Regression: Theory and Applications, Wiley. p. 8, 67.
5. อุมาพร จันทศร, วราพร เหลือสินทรัพย์. 2550. การเปรียบเทียบประสิทธิภาพของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายด้วยวิธีกำลังสองน้อยที่สุด วิธีของ Theil และวิธีของ Brown-Mood. กรุงเทพมหานคร : สำนักงานคณะกรรมการวิจัยแห่งชาติ (วช.).
6. Ohlson, J. A. and Kim, S. 2014. Linear Valuation without OLS: The Theil-Sen Estimation Approach. Working Paper, New York University.
7. Ahmed, S. 2014. Assessment of Irrigation System Sustainability using the Theil-Sen Estimator of Slope of Time Series. *Sustainability Science,* 9(3): 293-302.
8. Ohlson, J. A. 2014. Does the Cross-Sectional Equation $EPS_{t+1} = aP_t + bEPS_t + u_{t+1}$ Differ between China and the US?. *China Accounting and Finance Review,* 16(2), 1-9.
9. Sanglestsawai, S. Rejesus, R.M. and Yorobe, J.M. 2014. Do Lower Yielding Farmers Benefit from Bt Corn? Evidence from Instrumental Variable Quantile Regressions. *Food Policy.* 44: 285-296.
10. Xu, P. and Rummel, R. 1993. A Simulation Study of Smoothness Methods in Recovery of Regional Fields. *Geophys. J. Int.* 117: 472-486.
11. Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. J*ournal of the American Statistical Association,* 63(324): 1379-1389.

ได้รับบทความวันที่ 12 ตุลาคม 2558
ยอมรับตีพิมพ์วันที่ 23 ธันวาคม 2558