รายงานการวิจัย

เรื่อง การเปรียบเทียบประสิทธิภาพของการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย ด้วยวิธีกำลังสองน้อยที่สุด วิธีของ Theil และวิธีของ Brown – Mood

A Comparison on efficiency of Ordinary Least Squares Method, Theil's Method and Brown – Mood's Method in Simple Linear Regression Analysis.

โดย รศ.อุมาพร จันทศร

ผศ.วราพร เหลือสินทรัพย์

ภาควิชาสถิติประยุกต์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

สารบัญ

			หน้า
บทที่ 1	บทนำ		
	1.1	ความสำคัญ ที่มาของปัญหา	1
	1.2	วัตถุประสงค์ของโครงการวิจัย	2
	1.3	ขอบเขตของงานวิจัย	2
	1.4	ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย	3
บทที่ 2	ทฤษฎีและงานวิจัยที่เกี่ยวข้องและวิธีดำเนินการวิจัย		
	2.1	ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
	2.2	วิธีดำเนินการวิจัย	7
บทที่ 3	ผลการวิจัย		
	3.1	ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(34, 144) และ ความคลาดเคลื่อนเป็นแบบ N(0, 36)	8
	3.2	ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(34, 144) และ ความคลาดเคลื่อนเป็นแบบ LN(-5.5, 7.3)	9
	3.3	ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(34, 144) และ ความคลาดเคลื่อนเป็นแบบ W(0.5, 1)	9
	3.4	ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(3.3, 3.8) และ ความคลาดเคลื่อนเป็นแบบ N(0, 36)	10
	3.5	ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(3.3, 3.8) และ ความคลาดเคลื่อนเป็นแบบ LN(-5.5, 7.3)	11

หน้า

3.6 ผลการวิจัยเมื่อตัวแปรอิสระเป็นแบบ N(3.3, 3.8) และ
ความคลาดเคลื่อนเป็นแบบ W(0.5, 1) 12

บทที่ 4 สรุปผลและอภิปรายผล

4.1 สรุปผล 14

4.2 การอภิปรายผล 15

เอกสารอ้างอิง 16

สารบัญตาราง

หน้า

ตารางที่ 1 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(34,144)$, $e_i \sim N(0,36)$ 8

ตารางที่ 2 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(34,144)$, $e_i \sim LN(-5.5,7.3)$ 9

ตารางที่ 3 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(34,144)$, $e_i \sim W(0.5,1)$ 10

ตารางที่ 4 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(3.3,3.8)$, $e_i \sim N(0,36)$ 11

ตารางที่ 5 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(3.3,3.8)$, $e_i \sim LN(-5.5,7.3)$ 12

ตารางที่ 6 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนแต่ละขนาดตัวอย่าง จากวิธีการต่าง ๆ และเมื่อ $X_i \sim N(3.3,3.8)$, $e_i \sim W(0.5,1)$ 13

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย โดยใช้วิธีการประมาณค่า 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีทิล และวิธีบราวน์และมูด เกณฑ์การเปรียบเทียบใช้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (MSE) การเปรียบเทียบทำภายใต้เงื่อนไขของขนาดตัวอย่าง การแจกแจงของตัวแปรอิสระ (แบบปกติและไวบูลล์) และการแจกแจงของความคลาดเคลื่อน (แบบปกติ ลอกนอร์มอล และไวบูลล์) ซึ่งจะมีรวมทั้งสิ้น 30 สถานการณ์ ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากเทคนิคมอนติคาร์โล และทำการทดลองซ้ำ ๆ กัน 500 ครั้ง ในแต่ละสถานการณ์ที่กำหนด ผลการวิจัยสรุปได้ว่า

1. ในทุกสถานการณ์ พบว่า ขนาดตัวอย่างไม่มีผลต่อค่า MSE นั่นคือ ได้ค่าที่ใกล้เคียงกันในแต่ละสถานการณ์
2. ในทุกสถานการณ์ พบว่า การแจกแจงของตัวแปรอิสระไม่มีผลต่อค่า MSE
3. ในทุกสถานการณ์ พบว่า วิธีการของทิลและวิธีการของบราวน์และมูด ให้ค่า MSE ใกล้เคียงกับวิธีกำลังสองน้อยที่สุด โดยวิธีกำลังสองน้อยที่สุด ให้ค่า MSE ต่ำสุด อีก 2 วิธีให้ค่ามากกว่าเพียงเล็กน้อย

คำสำคัญ : การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย วิธีกำลังสองน้อยที่สุด วิธีการของทิล วิธีการของบราวน์และมูด

Abstract

This study aimed to compare parameter estimation methods in simple linear regression with ordinary least squares method, Theil's method and Brown – Mood's method by using the mean square error (MSE). The comparisons were done under conditions of sample size, distribution of independent variable (normal and weibull distribution) and distribution of error term (normal, log – normal and weibull distribution) ; altogenter in 30 situations. The data for this study was yielded from simulation, by the method of Monte Corlo, 500 iterations were used in each situation. The results are summarized as follows :

1. In all situations revealed that the sample size has no effect on MSE ; that means values in each situation are similar.
2. In all situations revealed that the distribution of in dependent variable has no effect on MSE.
3. In all situations revealed that Theil's method and Borwn – Mood method gave similar values of MSE to least squares method. The least squares has the minimum MSE ; While the two methods have slightly higher values.

Key Words : Simple Linear Regression, Ordinary least Squares Method, Theil's Method, Brown – Mood's Method.

บทที่ 1

บทนำ

1.1 ความสำคัญ ที่มาของปัญหาที่วิจัย

การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) มีรูปแบบคือ $y_i = \beta_0 + \beta_1 X_i + e_i$ (i = 1, 2, ..., n) เมื่อ y เป็นตัวแปรตาม และ x เป็นตัวแปรอิสระ วิธีการที่นิยมใช้ในการประมาณค่าพารามิเตอร์ของการถดถอยเชิงเส้น คือวิธีกำลังสองน้อยที่สุด (Least Squares Method : LS) ที่มีข้อกำหนดเบื้องต้นว่าเทอมความคลาดเคลื่อน e_i ต้องมาจากการแจกแจงปกติด้วยค่าเฉลี่ย = 0 และความแปรปรวนคงที่ ดังนั้นการจะนำสมการพยากรณ์ไปประยุกต์ใช้จึงต้องระมัดระวังถึงข้อกำหนดดังกล่าวนี้ แม้ว่าจะมีการแปลงข้อมูลเพื่อให้เป็นไปตามข้อกำหนดเบื้องต้นดังกล่าว เช่น แปลงเป็นค่าล็อค (Log) หรือส่วนกลับ $\left(\frac{1}{x_i}\right)$ หรืออื่น ๆ แต่การแปลผลในขั้นสุดท้ายจะเข้าใจได้ยาก รวมทั้งวิธีการกำลังสองน้อยที่สุดยังไว (Sensitive) ต่อค่าผิดปกติจากกลุ่ม (Outlier) มาก นั่นคือจะได้เส้นถดถอยพยากรณ์ที่ให้ความคลาดเคลื่อนสูง

จากปัญหาดังกล่าวข้างต้น ถ้าสามารถหาวิธีการวิเคราะห์การถดถอยที่มีคุณสมบัติแกร่ง (Robustness) และไม่จำเป็นต้องมีข้อกำหนดเกี่ยวกับการแจกแจงของเทอมความคลาดเคลื่อน ก็จะทำให้ผู้ใช้มั่นใจในผลการพยากรณ์ได้สูงขึ้น ในที่นี้จะเสนอวิธีการของสถิติที่ไม่ใช้พารามิเตอร์ คือวิธีการของ Theil และของ Brown - Mood โดยจะเปรียบเทียบกับวิธีกำลังสองน้อยที่สุด เมื่อใช้เกณฑ์การเปรียบเทียบคือค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (MSE) ที่ต่ำที่สุด โดยจะใช้วิธีการจำลองแบบโดยใช้เทคนิคมอนติคาร์โล พิจารณาในกรณีที่เทอมความคลาดเคลื่อนมีการแจกแจงแบบอื่น ๆ เช่น Log Normal และ Weibull และตัวแปรอิสระมีการแจกแจงแบบ Normal และ Weibull เป็นต้น โดยคาดหวังว่าจะมีความคลาดเคลื่อนใกล้เคียงหรือต่ำกว่า วิธีกำลังสองน้อยที่สุด ซึ่งจะเป็นประโยชน์ในการนำไปประยุกต์ใช้อย่างแท้จริง

1.2 วัตถุประสงค์ของโครงการวิจัย

1.2.1 เพื่อศึกษาเปรียบเทียบวิธีการวิเคราะห์การถดถอย ด้วยวิธีกำลังสองน้อยที่สุดกับวิธีการของ Theil และวิธีการของ Brown - Mood โดยศึกษาจากความคลาดเคลื่อน จากตัวแบบการถดถอยแบบเชิงเส้นอย่างง่าย (Simple Linear Model)

1.2.2 เพื่อหาผลสรุปว่า วิธีการวิเคราะห์การถดถอยวิธีใดที่จะให้ความคลาดเคลื่อนต่ำที่สุด จากตัวแบบการถดถอยเชิงเส้นอย่างง่าย ที่เทอมความคลาดเคลื่อนไม่เป็น Normal ระหว่างวิธีการของ Theil หรือวิธีการของ Brown - Mood

1.2.3 เพื่อหากรณีด้อยของการวิเคราะห์การถดถอย ด้วยวิธีการทางพาราเมตริก และนอนพารามิเตอร์ในการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้จะศึกษาภายใต้สถานการณ์ต่าง ๆ ดังนี้

1.3.1 ตัวแบบที่ใช้เป็นเชิงเส้นอย่างง่าย คือ $y_i = \beta_0 + \beta_1 X_i + e_i$

1.3.2 ขนาดตัวอย่างที่จะทดลองใช้มีขนาดเล็ก ปานกลาง และใหญ่ ดังนี้ 20, 30, 40, 50 และ 60

1.3.3 ตัวแปรอิสระ (X_i) มีการแจกแจงแบบปกติด้วยค่าพารามิเตอร์ 34 และ 144 $X_i \sim N(34,144)$ และมีการแจกแจงแบบไวบูลล์ด้วยพารามิเตอร์ 3.3 และ 3.8 $(X_i \sim W(3.3,3.8))$

1.3.4 ความคลาดเคลื่อน (e_i) มีการแจกแจงแบบปกติด้วยพารามิเตอร์ 0 และ 36 $(e_i \sim N(0,36))$ การแจกแจงลอกนอร์มอลด้วยพารามิเตอร์ -5.5 และ 7.3 $(e_i \sim LN(-5.5,7.3))$ และการแจกแจงแบบไวบูลล์ด้วยพารามิเตอร์ 0.5 และ 1 $(e_i \sim W(0.5,1))$

1.3.5 ใช้เทคนิคมอนติคาร์โล โดยกระทำซ้ำ 500 ครั้ง ในแต่ละสถานการณ์ รวมทั้งสิ้น 30 สถานการณ์

1.4 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1.4.1 จะได้ผลสรุปว่าการวิเคราะห์ด้วยวิธีการทั้งสามในแต่ละสถานการณ์นั้นให้ผลสรุปการวิเคราะห์ด้วยค่าความคลาดเคลื่อนต่างกันหรือไม่อย่างไร

1.4.2 เป็นแนวทางสำหรับนักวิจัยที่จะเลือกใช้วิธีการวิเคราะห์การถดถอยเชิงเส้นอย่างง่ายในกรณีที่เทอมความคลาดเคลื่อนมีการแจกแจงไม่เป็น Normal

1.4.3 เพื่อเป็นแนวทางในการศึกษาเกี่ยวกับการวิเคราะห์การถดถอยเชิงซ้อน (Multiple Regression Analyses) ต่อไป

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้องและวิธีดำเนินการวิจัย

2.1 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

วิธีการกำลังสองน้อยที่สุด จะใช้หลักการหาค่าประมาณของเส้นถดถอยจากการพยายามทำให้เทอมกำลังสองของความคลาดเคลื่อน $\left(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right)$ มีค่าต่ำที่สุด และต้องมีข้อกำหนดเบื้องต้นว่าเทอมความคลาดเคลื่อน $(e_i=y_i-\hat{y}_i)$ มีการแจกแจงแบบ Normal เพื่อทำการทดสอบนัยสำคัญของค่าประมาณของเส้นถดถอยในลำดับถัดไป แต่วิธีการของ Theil จะใช้ค่ามัธยฐานของค่า slope จากทุกคู่ที่เป็นไปได้ เป็นค่าประมาณของ slope ซึ่งค่ามัธยฐานจะไม่ไว (Insensitive) ต่อค่าผิดปกติในกลุ่ม ดังนั้นจึงเป็นสถิติที่มีคุณสมบัติแกร่ง วิธีการทดสอบนัยสำคัญของค่า slope ของเส้นถดถอย จะใช้ค่าสัมประสิทธิ์สหสัมพันธ์แบบเคนดอล (The Kendall Rank correlation coefficient) ซึ่งไม่จำเป็นต้องใช้ข้อกำหนดเบื้องต้นเกี่ยวกับการแจกแจงแบบ Normal ของเทอมความคลาดเคลื่อน ส่วนวิธีการของ Brown - Mood ก็ยังคงใช้ค่ามัธยฐานเป็นสถิติที่ใช้คำนวณหาค่าประมาณของเส้นถดถอย กล่าวคือ จะแบ่งข้อมูลตัวอย่างออกเป็น 2 กลุ่ม กลุ่มที่หนึ่งจะคือข้อมูลคู่ที่มีค่า X_i ที่มีค่าน้อยกว่ามัธยฐานของ X_i กลุ่มที่สองคือข้อมูลคู่ที่มีค่า X_i มากกว่าค่ามัธยฐานของ X_i ในแต่ละส่วนหาค่ามัธยฐานของตัวแปร X_i และ y_i และลากเส้นตรงผ่านจุดทั้งสองดังกล่าว และเส้นถดถอยที่จะได้ในขั้นสุดท้าย คือเส้นถดถอยที่ได้ค่ามัธยฐานของความคลาดเคลื่อน $(y_i-\hat{y}_i)=0$ และการทดสอบนัยสำคัญของค่า slope จะอาศัยการแจกแจงแบบไคสแควร์

ในลำดับแรกจะกล่าวถึงวิธีการของ Theil และ Brown & Mood โดยละเอียดก่อน ดังนี้

วิธีการของ Theil

สมการถดถอยที่ต้องการสร้าง มีโมเดลดังนี้ $y_i = \alpha+\beta X_i+e_i$, i = 1, ..., n เมื่อ X_i เป็นค่าคงที่ที่ทราบค่า, α และ β เป็นค่าพารามิเตอร์ที่ไม่ทราบค่า และ y_i เป็นค่าสังเกตที่มีค่าต่อเนื่องของตัวแปร y ณ ค่าตัวแปร X_i และอาจมีค่าหลายค่าที่ค่า X_i หนึ่ง ๆ ค่า e_i เป็นอิสระกัน ค่า X_i มีค่าไม่ซ้ำกันให้ $X_1<X_2<X_3...X_n$

ข้อมูลประกอบด้วย n คู่ของค่าสังเกต

$(x_1,y_1),(x_2,y_2),...(x_i,y_i),...(x_n,y_n)$

เพื่อหาค่าประมาณของ β ให้คือ $\hat{\beta}$ = มัธยฐานของ S_{ij}

เมื่อ $S_{ij} = \frac{(y_j - y_i)}{(x_j - x_i)}$ เมื่อ $i < j$ ซึ่งจะมีทั้งหมด $\binom{n}{2}$ ค่า

และค่า $\hat{\alpha}$ = มัธยฐานของทุกค่าของ $\alpha_i = y_i - \hat{\beta}x_i$

วิธีการของ Brown - Mood

ในลำดับแรกหาค่ามัธยฐานของตัวแปร X จากค่า X_i, i = 1, ..., n ก่อน แล้วจึงแบ่งข้อมูลออกเป็น 2 กลุ่ม โดยกลุ่มที่ 1 จะประกอบด้วยคู่ค่าสังเกต (x_i, y_i) ที่มีค่า X_i น้อยกว่าค่ามัธยฐานของ X และกลุ่มที่ 2 จะประกอบด้วยคู่ค่าสังเกต (x_i, y_i) ที่มีค่า X_i มากกว่าค่ามัธยฐานของ X ในแต่ละกลุ่มของ 2 กลุ่มดังกล่าว หาค่ามัธยฐานของค่า X และ Y ดังนั้นในขั้นตอนนี้จะมีค่ามัธยฐานทั้งหมด 4 ค่า สร้างกราฟโดยแกนนอนเป็นค่า x แกนตั้งเป็นค่า y จากค่ามัธยฐานทั้ง 4 ในตอนต้นจะได้จุด 2 จุดในกราฟ ลากเส้นตรงผ่าน 2 จุดนี้ และหาค่าสโลป, $\hat{\beta}'$ ของเส้นตรงนี้ ซึ่งจะเป็นค่าประมาณ β เบื้องต้น ส่วนค่าประมาณ $\hat{\beta}$ ที่จะได้ในขั้นสุดท้าย คือสโลปของเส้นตรงที่ปรับจากเส้นตรงที่ได้สโลป $\hat{\beta}'$ ไปจนกระทั่งได้ค่ามัธยฐานของความแตกต่างระหว่าง y ที่แท้จริง กับ $\hat{y}$ จากเส้นตรงนั้นมีค่าเป็น 0 ในข้อมูลทั้ง 2 กลุ่ม เพื่อความเข้าใจจะขอยกตัวอย่างแสดงดังนี้

จงสร้างเส้นถดถอยเชิงเส้นระหว่างค่า Lipids (X) และค่า Cholesterol (Y) ของคนไข้ที่ผ่านการผ่าตัด ได้ข้อมูลจากตัวอย่างคนไข้ 10 ราย ดังนี้

คนไข้ที่	1	2	3	4	5	6	7	8	9	10
ค่า Lipids (X)	3.81	2.10	0.79	1.99	1.03	2.07	0.74	3.88	1.43	0.41
ค่า Cholesterol (Y)	1.90	1.03	0.44	1.18	0.62	1.29	0.39	2.30	0.93	0.29

ขั้นแรกหาค่ามัธยฐานของค่า X ได้ค่า = 1.71

แบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มที่มีค่า X ต่ำกว่า และสูงกว่ามัธยฐาน จะได้ค่า (x, y) ใน 2 กลุ่ม ดังนี้

กลุ่มที่ 1 ; (0.79, 0.44), (1.03, 0.62), (0.74, 0.39), (1.43, 0.93), (0.41, 0.29)

กลุ่มที่ 2 ; (3.81, 1.90), (2.10, 1.03), (1.99, 1.18), (2.07, 1.29), (3.88, 2.30)

ในกลุ่มที่ 1 หาค่ามัธยฐานของ x, y จะได้ค่า 0.79, 0.44ตามลำดับ

ในกลุ่มที่ 2 หาค่ามัธยฐานของ x, y จะได้ค่า 2.10, 1.29ตามลำดับ

ดังนั้นในกราฟ จะมีจุด 2 จุด นี้ ลากเส้นตรงผ่าน 2 จุดนี้ ได้ค่าสโลปในเบื้องต้น,

$$\hat{\beta}' = \frac{1.29 - 0.44}{2.10 - 0.79} = 0.6487$$

หาค่าเบี่ยงเบนในแกนตั้งคือค่า y ที่แท้จริง และ $\hat{y}$ ที่ได้จากเส้นถดถอยนี้ พบว่าค่ามัธยฐานของค่าเบี่ยงเบนยังไม่เท่ากับ 0 ใน 2 กลุ่ม จึงจำเป็นต้องปรับเส้นตรงนี้ จนกระทั่งอยู่ในตำแหน่งที่ได้ค่ามัธยฐานของค่าเบี่ยงเบนมีค่าเป็น 0

ในขั้นสุดท้าย จะได้ค่า $\hat{\beta} = 0.5939$ และ $\hat{\alpha} = 0$

ดังนั้นสมการถดถอยที่ได้คือ $\hat{y}_i = 0.00 + 0.5939\, X_i$

มีผลงานวิจัยเกี่ยวกับวิธีการของ Theil ดังนี้ Dritz (1989) ได้ศึกษาถึงตัวประมาณค่าสโลปแบบสถิติที่ไม่ใช้พารามิเตอร์หลายวิธี รวมทั้งวิธีของ Theil พบว่าค่าประมาณสโลปตามวิธีของ Theil มีคุณสมบัติแกร่ง (Robust) คำนวณได้ง่าย และได้ค่า Mean Square error ใกล้เคียงกับวิธีประมาณแบบอื่น ๆ ส่วน Sen (1968) ได้ศึกษากรณีค่าตัวแปรอิสระ $\left(X_i\right)$ มีค่าไม่แตกต่างกัน (Nondistinct x – values) และคำนวณค่าประมาณสโลปตามวิธีของ Theil พบว่าค่าประมาณนั้นจะเหมือนกับค่าประมาณของ Hodges – Lehman กรณี 2 กลุ่มตัวอย่างอิสระกัน ที่ใช้ค่าจาก 2 กลุ่มตัวอย่างจาก $y_1, \ldots y_m$ และ $y_{m+1} \ldots y_{m+q}$ รวมทั้งศึกษาค่า ARE (Asymptotic Relative efficiency) เมื่อเทียบกับค่าประมาณจากวิธีกำลังสองน้อยที่สุด และ Hussain, SS and Sprent, P. (1983) ได้จำลองค่าความคลาดเคลื่อนให้มีการแจกแจงแบบหางยาว (Long tailed) พบว่าวิธีการของ Theil มีประสิทธิภาพสูงกว่าวิธีกำลังสองน้อยที่สุด โดยเฉพาะเมื่อขนาดตัวอย่างน้อยกว่า 30 รวมทั้งพบว่าค่าประมาณของ y – intercept มีประสิทธิภาพที่ดีขึ้น (Marked improvement) แม้ว่าโดยทั่วไปจะสนใจค่าประมาณนี้น้อยกว่าค่าสโลปก็ตาม และมีประสิทธิภาพใกล้เคียงวิธีกำลังสองน้อยที่สุด เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ นอกจากนี้ Jorge Adrover and Ruben H. Zamar ได้ศึกษาถึงคุณสมบัติที่เรียกว่า Biase performance ของค่าประมาณการถดถอยอย่างง่าย จากวิธีการประมาณ 3 วิธี คือ Brown & Mood, Theil และวิธีของ Sen พบว่าค่าประมาณที่ได้จากทั้ง 3 วิธี มีค่าที่น่าสนใจ (outstanding) เนื่องจากมี

การเอนเอียงเพียงเล็กน้อย (Small biases) เหมาะสมกับคุณสมบัติของ Robust Inference และยังศึกษารายละเอียดเกี่ยวกับค่า y – intercept ในคุณสมบัติ Maximum Asymptotic biases ที่มีการศึกษาในวงแคบอยู่

2.2 วิธีดำเนินการวิจัย

จะอาศัยการจำลอง (Simulation) ด้วยเทคนิคมอนติคาร์โล ตามลำดับดังนี้

2.2.1 สร้างข้อมูลตัวอย่างจากตัวแบบ $y_i = \beta_0 + \beta_1 X_i + e_i$ เมื่อกำหนดให้ X_i มีการแจกแจงแบบปกติ และ Weibull และ e_i มีการแจกแจงแบบ Normal, Log Normal, Weibull ด้วยขนาดตัวอย่างที่แยกเป็นค่าน้อย ปานกลาง มาก คือ 20, 30, 40, 50 และ 60 โดยสร้างข้อมูลมา 1,000 ชุด จากแต่ละการแจกแจง

2.2.2 สุ่มตัวอย่างข้อมูลจากขั้นที่ 1 มาด้วยขนาดเล็ก ปานกลาง มาก เพื่อทำการวิเคราะห์หาเส้นถดถอย ด้วยวิธีกำลังสองน้อยที่สุด วิธีการของ Theil และวิธีการของ Brown - Mood

2.2.3 จากเส้นถดถอยที่ได้ในขั้นที่ 2 คำนวณหาค่าพยากรณ์จากแต่ละวิธีการวิเคราะห์ คำนวณหาค่า MSE (Mean Square Error) และเปรียบเทียบจากทั้ง 3 วิธี โดยใช้เกณฑ์ของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้มีค่าน้อยที่สุด

2.2.4 ในแต่ละขั้นตอนจะใช้โปรแกรมซึ่งจำเป็นต้องเขียนขึ้นมาในการวิเคราะห์

บทที่ 3

ผลการวิจัย

การเปรียบเทียบค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีกำลังสองน้อยที่สุด วิธีการของ Theil และวิธีการของ Brown – Mood เมื่อกำหนดให้ตัวแปรอิสระ (X_i) และความคลาดเคลื่อน (e_i) มีการแจกแจงแบบต่าง ๆ และขนาดตัวอย่างต่าง ๆ แยกเป็นกรณีต่าง ๆ กัน

3.1 เมื่อตัวแปรอิสระ (X_i) และความคลาดเคลื่อน (e_i) มีการแจกแจงปกติ คือ $X_i \sim N(34,144)$ และ $e_i \sim N(0,36)$ จะได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนในแต่ละขนาดตัวอย่างดังนี้

ตารางที่ 1 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim N(34,144)$ และ $e_i \sim N(0,36)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	5.7056	5.819	6.0871
30	5.9548	6.0251	6.266
40	5.6127	5.6622	5.7795
50	6.1055	6.1461	6.2781
60	5.8409	5.8772	5.9797

ซึ่งสามารถสรุปได้ว่า วิธีกำลังสองน้อยที่สุด ให้ค่า MSE ต่ำที่สุดในทุกขนาดตัวอย่าง แม้ว่าวีการของ Theil และ Brown & Mood จะให้ค่า MSE สูงในลำดับต่อไป แต่ความแตกต่างก็ไม่มากนัก

3.2 เมื่อตัวแปรอิสระ (X_i) มีการแจกแจงปกติ $X_i \sim N(34,144)$ แต่ความคลาดเคลื่อนมีการแจกแจงแบบลอกนอร์มอล คือ $e_i \sim LN(-5.5,7.3)$ จะได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนในแต่ละขนาดตัวอย่างดังนี้

ตารางที่ 2 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim N(34,144)$ และ $e_i \sim LN(-5.5,7.3)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	1,570,538.00	1,645,995.74	1,645,995.68
30	462,066.15	470,900.39	470,900.38
40	725,818.44	753,404.08	753,404.08
50	139,697.48	141,643.37	141,643.36
60	400,792.78	414,224.89	414,224.88

จากตารางที่ 2 จะพบว่า วิธีกำลังสองน้อยที่สุดให้ค่า MSE ต่ำที่สุด ในทุกกรณีของขนาดตัวอย่าง วิธีของ Theil และ Brown – Mood ให้ผลใกล้เคียงกันมาก แต่ให้ค่า MSE ที่มากกว่าวิธีกำลังสองน้อยที่สุด

3.3 เมื่อตัวแปรอิสระ (X_i) มีการแจกแจงปกติ คือ $X_i \sim N(34,144)$ แต่ความคลาดเคลื่อนมีการแจกแจงแบบไวบูลล์ คือ $e_i \sim W(0.5,1)$ ได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนในแต่ละขนาดตัวอย่างดังนี้

ตารางที่ 3 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim N(34,144)$ และ $e_i \sim W(0.5,1)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	0.029838	0.030521	0.037542
30	0.032256	0.032979	0.043612
40	0.034217	0.034815	0.040352
50	0.029899	0.030394	0.039209
60	0.032973	0.033515	0.039236

จากตารางที่ 3 จะพบว่า ได้ค่า MSE ที่น้อยมาก และวิธีกำลังสองน้อยที่สุดให้ค่า MSE น้อยที่สุด โดยขนาดตัวอย่างไม่มีผลต่อค่า MSE ในขณะที่วิธีของ Theil และ Brown – Mood ให้ค่า MSE มากในลำดับถัดไป แต่ความแตกต่างมีเพียงเล็กน้อย

3.4 เมื่อตัวแปรอิสระ (X_i) มีการแจกแจงแบบไวล์บูล คือ $X_i \sim W(3.3,3.8)$ และความคลาดเคลื่อนมีการแจกแจงแบบปกติ คือ $e_i \sim N(0,36)$ ได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนในแต่ละขนาดตัวอย่าง ดังนี้

ตารางที่ 4 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim W(3.3, 3.8)$ และ $e_i \sim N(0, 36)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	5.6949	5.825	6.1131
30	5.9596	6.0305	6.3125
40	5.6207	5.6708	5.8415
50	6.0883	6.1351	6.2414
60	5.8601	5.8969	6.0035

จากตารางที่ 4 จะพบว่า ค่า MSE ที่ได้จะใกล้เคียงกับตารางที่ 1 เมื่อความคลาดเคลื่อน (e_i) มีการแจกแจงปกติเช่นเดียวกัน แม้ว่าตัวแปรอิสระ (X_i) จะมีการแจกแจงที่ต่างจากกรณีนี้ คือ มีการแจกแจงปกติ และยังได้ผลลัพธ์ในทำนองเดียวกันกับตารางที่ 1 คือวิธีกำลังสองน้อยที่สุด ได้ค่า MSE ต่ำที่สุด และไม่มีผลเมื่อขนาดตัวอย่างเปลี่ยนไป ส่วนวิธีของ Theil และ Brown – Mood ให้ค่า MSE มากในลำดับถัดไป แต่ความแตกต่างมีไม่มากนัก

3.5 เมื่อตัวแปรอิสระ (X_i) มีการแจกแจงแบบไวบูลล์ คือ $X_i \sim W(3.3, 3.8)$ และความคลาดเคลื่อนมีการแจกแจงแบบลอกนอร์มอลคือ $e_i \sim LN(-5.5, 7.3)$ จะได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน ดังนี้

ตารางที่ 5 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim W(3.3, 3.8)$ และ $e_i \sim LN(-5.5, 7.3)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	1,596,559.98	1,645,995.74	1,645,995.72
30	462,522.32	470,900.39	470,900.38
40	729,538.90	753,404.08	753,404.08
50	139,199.73	141,643.37	141,643.36
60	409,115.73	414,224.89	414,224.88

จากตารางที่ 5 จะได้ผลลัพธ์เช่นเดียวกับตารางที่ 2 คือได้ค่า MSE สูงมากจากทุกวิธีการ และเมื่อเปรียบเทียบระหว่างวิธีการก็ได้ผลเช่นเดิมกับตารางอื่น ๆ ที่ผ่านมา คือวิธีกำลังสองน้อยที่สุด จะให้ค่า MSE ต่ำที่สุด ในขณะที่วิธี Theil และ Brown – Mood ให้ผลเท่ากัน ซึ่งมากกว่าวิธีกำลังสองน้อยที่สุด

3.6 เมื่อตัวแปรอิสระ (X_i) มีการแจกแจงแบบไวล์บูล คือ $X_i \sim W(3.3, 3.8)$ และความคลาดเคลื่อนมีการแจกแจงแบบไวล์บูลด้วย เมื่อ $e_i \sim W(0.5, 1)$ ได้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนในแต่ละขนาดตัวอย่างดังนี้

ตารางที่ 6 ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนจากวิธีการต่าง ๆ และขนาดตัวอย่างต่าง ๆ เมื่อ $X_i \sim W(3.3, 3.8)$ และ $e_i \sim W(0.5, 1)$

ขนาดตัวอย่าง, n	วิธีการแบบ		
	กำลังสองฯ	Theil	Brown & Mood
20	0.029893	0.030723	0.034202
30	0.032314	0.033058	0.035641
40	0.034093	0.034815	0.037217
50	0.029872	0.030351	0.03273
60	0.032969	0.033503	0.035529

จะได้ค่า MSE ใกล้เคียงกับตารางที่ 3 ผลลัพธ์ ก็ได้ในทำนองเดียวกันคือวิธีกำลังสองน้อยที่สุด ได้ค่าต่ำที่สุดไม่ขึ้นกับขนาดตัวอย่าง

บทที่ 4

บทสรุปและอภิปรายผล

4.1 สรุปผล

งานวิจัยนี้มีวัตถุประสงค์คือ เพื่อเปรียบเทียบค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (Mean Square Error) จากวิธีกำลังสองน้อยที่สุด (Least Squares Method) กับวิธีการของ Theil และวิธีการของ Brown – Mood เมื่อกำหนดให้ตัวแปรอิสระ (X_i) และความคลาดเคลื่อน (e_i) มีการแจกแจงแบบต่าง ๆ ซึ่งมีลักษณะเบ้ และหางยาว (Long tail) ด้วยเทคนิคมอนติคาร์โล โดยจำลองเหตุการณ์ต่าง ๆ 6 สถานการณ์ และกระทำซ้ำ 500 รอบ ได้ผลสรุปดังนี้

- เมื่อ e_i มีการแจกแจงแบบไวบูลล์ ส่วน X_i มีการแจกแจงแบบปกติหรือไวบูลล์ จะพบว่าวิธีการทั้ง 3 ให้ค่า MSE ต่ำที่สุด โดยขนาดตัวอย่างไม่มีผลต่อค่า MSE แม้จะพบว่าวิธีกำลังสองน้อยที่สุดให้ค่า MSE ที่ต่ำที่สุด แต่พบว่าวิธีการของ Theil และ Brown – Mood ก็ยังคงให้ค่าที่ต่ำมากในทำนองเดียวกัน โดยมีค่ามากกว่าวิธีกำลังสองน้อยที่สุดเพียงเล็กน้อย และวิธีการของ Theil ให้ผลดีเป็นอันดับสอง

- เมื่อ e_i มีการแจกแจงแบบปกติ ส่วน X_i มีการแจกแจงแบบปกติหรือไวบูลล์ วิธีการทั้งสามให้ผลลัพธ์ในทำนองเดียวกัน คือให้ค่า MSE มีค่าในช่วง 5.5 – 6.5 โดยวิธีกำลังสองน้อยที่สุดให้ค่า MSE ต่ำที่สุด ในทุกขนาดตัวอย่าง แม้ว่าวิธีการของ Theil และ Brown – Mood จะให้ค่า MSE สูงในลำดับถัดไป แต่ความแตกต่างก็ไม่มากนัก

- กรณีสุดท้าย คือเมื่อ e_i มีการแจกแจงแบบลอกนอร์มอล และ X_i มีการแจกแจงแบบปกติหรือไวบูลล์ พบว่าวิธีการทั้งสามให้ค่า MSE ที่ใหญ่มาก วิธีการของ Theil และ Brown – Mood ให้ค่า MSE เกือบจะเท่ากัน และมากกว่าวิธีกำลังสองน้อยที่สุด

จากผลที่ได้สามารถสรุปได้ว่า วิธีการของ Theil และ Brown – Mood ให้ค่า MSE ในทำนองเดียวกับวิธีกำลังสองน้อยที่สุดในทุกสถานการณ์ที่ทดลองคือ 6 สถานการณ์ และพบว่าการแจกแจงของตัวแปรอิสระ (X_i) ไม่มีผลต่อค่า MSE นั่นคือ ไม่ว่า X_i จะมีการแจกแจงแบบใด ๆ ค่า MSE ที่ได้จากวิธีการทั้ง 3 จะขึ้นกับการแจกแจงของ e_i เท่านั้น

สำหรับการนำไปประยุกต์ใช้ สามารถให้ข้อแนะนำดังนี้ วิธีการของ Theil สามารถนำไปแทนที่วิธีการกำลังสองน้อยที่สุด (เมื่อคำนึงถึงค่า MSE เป็นประเด็นสำคัญ) ได้ทุกสถานการณ์จาก 6 สถานการณ์ หรืออาจใช้วิธีกำลังสองน้อยที่สุด ซึ่งจะมีคุณสมบัติแกร่งในสถานการณ์เหล่านี้ เพราะยังคงให้ค่า MSE น้อยที่สุด ส่วนวิธีของ Brown – Mood แม้บางสถานการณ์จะให้ค่า MSE แทบจะไม่แตกต่างจากวิธีของ Theil แต่การคำนวณค่อนข้างยากกว่า โดยเฉพาะการปรับเส้นถดถอยในขั้นตอนสุดท้าย นอกจากวิธีการของ Theil จะให้ค่า MSE ที่ใกล้เคียงกับวิธีกำลังสองน้อยที่สุดแล้ว ยังมีคุณสมบัติอื่น ๆ เช่น คุณสมบัติแกร่ง คุณสมบัติ Biase Performance และคุณสมบัติ Maximum Asymptotic Biase ดังกล่าวในบทนำในตอนต้น

4.2 การอภิปรายผล

แม้ว่าวิธีการของ Theil และ Brown – Mood จะถูกคิดค้นตั้งแต่ปี คศ. 1950 และ ปี 1980 แล้วก็ตาม และปัจจุบันนักสถิติได้พัฒนาวิธีการประมาณค่าพารามิเตอร์ของการถดถอยเชิงเส้นอย่างง่ายด้วยวิธีการแบบสถิติที่ไม่ใช้พารามิเตอร์เพิ่มขึ้นมากมาย อาทิเช่น วิธีการปรับโค้งให้เรียบ (Curve Smoothing) ที่มีวิธีการแบบเคอร์เนล (Kernel Method) หรือวิธีการแบบลอสส์ (Loess Method) หรือแม้กระทั่งวิธีแบบ M (M – Estimation) แต่วิธีการของ Theil และ Brown – Mood มีข้อเด่น คือคำนวณได้ง่าย ไม่จำเป็นต้องใช้คณิตศาสตร์ชั้นสูง เหมาะสำหรับนักวิจัยสาขาต่าง ๆ ที่จะนำไปประยุกต์ใช้

ผลสรุปจากงานวิจัยนี้ ให้ผลในทำนองเดียวกับงานวิจัยที่ผ่านมาภายใต้สถานการณ์จำลอง 6 แบบ ที่แตกต่างกันคือการแจกแจงปกติแบบลอกนอร์มอล และแบบไวบูลล์ ตัวแปรอิสระและความคลาดเคลื่อนจะกำหนดพารามิเตอร์ของการแจกแจงไว้ให้คงที่ เช่น ให้ $X_i \sim N(34,144)$ หรือ $e_i \sim N(0,36)$ เป็นต้น จึงขาดผลสรุปในกรณีให้ค่าพารามิเตอร์เปลี่ยนไปจากค่าคงที่เหล่านี้ ซึ่งจะทำให้ผลงานวิจัยที่สมบูรณ์มากยิ่งขึ้น จึงเสนอแนะให้ทำการศึกษาในสถานการณ์เหล่านี้ต่อไป

เอกสารอ้างอิง

1. ศิริรัตน์ เชียงจง และคณะ, 2550 การศึกษาเปรียบเทียบวิธีการประมาณเส้นการถดถอยเชิงเส้นอย่างง่าย เมื่อการแจกแจงของความคลาดเคลื่อนไม่เป็นแบบปกติ, 118 – 121, การประชุมวิชาการสถิติและสถิติประยุกต์ ประจำปี 2550. สถาบันบัณฑิตพัฒนบริหารศาสตร์. กรุงเทพฯ
2. Brown, B.M., 1980. Median estimation in simple linear regression. Austral. J. Statist., 22, 154 – 66.
3. Brown, B.M. and Maritz, J.S., 1982. Distribution – free methods in regression. Austral, J. Statist., 24, 318 – 31.
4. Hussain, S. S. and Sprent, P., 1983. Nonparametric regression. J. Roy. Statist. Soc. A, 146, 182 – 91.
5. Jorge Adrover and Ruben H. Zamar, 2004. Bias robustness of three median – based regression estimates. J. of Statistical Planning and Inference, 122, 203 – 227.
6. Myles Hollander, Douglas A. Wolf, 1999. Nonparametric Statistical Methods, John Wiley & Sons, New York.
7. Theil. H., 1950. A rank invariant method of linear and Polynomial regression analysis, I, II, III. Proc. Kon. Nederl. Akad. Wetensch. A, 53, 386 – 92, 521 – 5, 1397 – 1412.