

การเปรียบเทียบตัวประมาณกำลังสองน้อยสุด กับตัวประมาณค่าสัมบูรณ์น้อยสุดในแบบถดถอยเชิงเส้นอย่างง่าย กรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรตาม

ปัทมาดี นันทนาเนตร์

เป็นที่ทราบกันดีว่า ในงานวิจัยด้านต่างๆ ไม่ว่าจะเป็นด้านวิทยาศาสตร์ ด้านสังคมศาสตร์ ด้านอุตสาหกรรม หรือในด้านธุรกิจนั้น หากต้องการ คาดคะเนหรือพยากรณ์ค่าของตัวแปรตาม (Dependent Variable) ที่เราสนใจ ว่าขึ้นอยู่กับ กลุ่มของตัวแปรอิสระ (Independent Variables) อย่างไร หรือมีความสัมพันธ์กันอย่างไร เช่น ในการ ศึกษาเพื่อพยากรณ์ค่าของต้นทุนการผลิตนั้น กลุ่ม ของตัวแปรอิสระหรือปัจจัยที่มีอิทธิพลต่อต้นทุน การผลิตได้แก่ จำนวนการผลิต (Productivity) ประสิทธิภาพการผลิต (Efficiency) ทักษะคติของ พนักงาน (Attitude) เป็นต้น โดยทั่วไปจะใช้

เทคนิคที่เรียกว่าการวิเคราะห์การถดถอยพหุคูณ (Multiple Regression Analysis) และนิยม ประเมินค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอย ของตัวแบบ ด้วยตัวประมาณกำลังสองน้อยสุด (Least Squares Estimator) เนื่องจากตัวประมาณ ดังกล่าวจะมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียง และในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้นทั้ง หมด ตัวประมาณกำลังสองน้อยสุดจะให้ค่าความ แปรปรวนต่ำสุด ซึ่งเรียกว่ามีคุณสมบัติ BLUE (Best Linear Unbiased Estimator) สำหรับตัว แบบการถดถอยเชิงเส้นพหุคูณเขียนในรูปทั่วไป ได้ดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_j x_{ij} + \varepsilon_i \quad ; i = 1, 2, \dots, n \quad , j = 1, 2, \dots, k$$

หรือเขียนในรูปเมทริกซ์ ได้ดังนี้

$$\tilde{y} = \tilde{X} \tilde{\beta} + \tilde{\varepsilon}$$

$\tilde{y}$ แทนเวกเตอร์ของตัวแปรตามที่มีขนาด $(n \times 1)$

$\tilde{X}$ แทนเมทริกซ์ของตัวแปรอิสระที่มีขนาด $(n \times q)$

β แทนเวกเตอร์ของสัมประสิทธิ์ถดถอยที่มีขนาด $(q \times 1)$

ε แทนเวกเตอร์ของความคลาดเคลื่อนที่มีขนาด $(n \times 1)$

n แทนขนาดตัวอย่าง

k แทนจำนวนตัวแปรอิสระในตัวแบบ

และ q แทนจำนวนพารามิเตอร์ที่ต้องการประมาณค่าในตัวแบบ $= k+1$

ซึ่งตัวประมาณกำลังสองน้อยสุดที่กล่าวถึงข้างต้นอยู่ในรูปแบบดังนี้

$$\hat{\beta}_{LS} = (X'X)^{-1} \cdot X'y \quad (1.1)$$

แต่ในการประมาณค่าพารามิเตอร์ ด้วยตัวประมาณกำลังสองน้อยสื่อนั้น ลักษณะของข้อมูลจะต้องเป็นไปตามข้อตกลงเบื้องต้น (Assumptions) คือ ความคลาดเคลื่อนสุ่ม (Random Error) $\varepsilon_i; i=1, \dots, n$ มีค่าเฉลี่ยเท่ากับศูนย์ ความแปรปรวนเท่ากับ σ^2 คงที่ทุกค่า i และความแปรปรวนร่วมของ $(\varepsilon_i, \varepsilon_j); i \neq j$ ทุกคู่เท่ากับศูนย์หรือไม่มีสหสัมพันธ์กัน (Uncorrelation) ดังนั้น ในกรณีที่ข้อมูลไม่เป็นไปตามข้อสมมติดังกล่าว การใช้ตัวประมาณกำลังสองน้อยสุดจะทำให้ตัวประมาณค่าพารามิเตอร์ขาดคุณสมบัติ BLUE นอกจากนี้ ตัวประมาณกำลังสองน้อยสุดอาจขาดคุณสมบัติ BLUE อันเนื่องมาจากสาเหตุอื่นได้เช่นกัน เช่น กรณีที่ข้อมูลมีค่าผิดปกติ (Outliers) เป็นต้น ดังนั้น จึงได้มีการ

คิดค้นหาตัวประมาณอื่นแทนตัวประมาณกำลังสองน้อยสุดเมื่อเกิดกรณีที่ ไม่เหมาะสมที่จะใช้ตัวประมาณกำลังสองน้อยสุด ตัวประมาณอื่นๆ เช่น ตัวประมาณการถดถอยที่มีความแกร่ง (Robust Regression Estimator) เป็นต้น

ได้กล่าวแล้วว่า ตัวประมาณกำลังสองน้อยสุดจากสมการ (1.1) มีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียง และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Squares Error) ต่ำสุดในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้นทั้งหมด แต่ในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณด้วยตัวประมาณกำลังสองน้อยสื่อนั้น ในทางปฏิบัติ ลักษณะของข้อมูลที่นำมาใช้ในการวิเคราะห์การถดถอยอาจมีค่าสังเกตบางค่าที่มีค่าสูงหรือต่ำกว่าค่าสังเกตส่วนใหญ่ของข้อมูลโดยปกติ ซึ่งค่าสังเกตดังกล่าว เรียกว่า ค่าผิดปกติ (Outliers)

สาเหตุการเกิดข้อมูลผิดปกติอาจเกิดจาก ความคลาดเคลื่อนจากการวัดค่า เช่น การบันทึกข้อมูลที่ไม่ถูกต้องของผู้บันทึกข้อมูล การใช้เครื่องมือวัดค่าที่มีคุณภาพต่ำ เป็นต้น หรืออาจเกิดจากเหตุการณ์ผิดปกติที่ไม่สามารถคาดการณ์ล่วงหน้าได้ เช่น อุบัติเหตุ หรือ อุทกภัยต่างๆ หรืออาจเกิดจากการที่มีบางหน่วยตัวอย่างของข้อมูลไม่ได้มาจากการแจกแจงเดียวกันกับการแจกแจงของประชากรที่ศึกษา เป็นต้น การที่ข้อมูลมีค่าผิดปกติเกิดขึ้นนั้น อาจส่งผลให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยตัวประมาณกำลังสองน้อยสุดไม่เหมาะสมได้ เนื่องจาก สมการการถดถอยเชิงเส้นที่ได้หรือเส้นสมการถดถอย (Regression Line) จะถูกปรับทิศทางไปตามข้อมูลที่ผิดปกติ และอาจทำให้ค่าประมาณของความแปรปรวนของความคลาดเคลื่อนมีค่าสูงขึ้นกว่าปกติ ดังนั้น ค่าประมาณความแปรปรวนของ

$\hat{\beta}$ จึงมีค่าสูงขึ้นด้วย และเมื่อนำไปทดสอบ

สมมติฐานของพารามิเตอร์ ก็จะทำให้ผลการทดสอบผิดพลาดได้ แต่อย่างไรก็ตาม ค่าผิดปกติบางค่าอาจให้ข้อมูลที่เป็นประโยชน์ต่อการหาสมการถดถอยเชิงเส้นที่เหมาะสมได้

ดังนั้น หากนำข้อมูลที่ใช้ในการวิเคราะห์การถดถอยมาตรวจสอบค่าผิดปกติ แล้วพบว่าข้อมูลมีค่าผิดปกติและมีอิทธิพลต่อการประมาณค่าสัมประสิทธิ์การถดถอยรวมอยู่ ผู้วิเคราะห์ไม่ควรตัดค่าผิดปกตินั้นออกทันที แต่ควรพิจารณาถึงสาเหตุของการเกิดค่าผิดปกติดังกล่าว เช่น หากเกิด

จากการบันทึกข้อมูลผิดพลาด วิธีการแก้ไขอาจทำได้โดยการตัดข้อมูลที่มีค่าผิดปกติออกไป หรืออาจทำการปรับค่าให้เหมาะสม จากนั้น จึงนำข้อมูลที่เหลือมาวิเคราะห์การถดถอยต่อไปโดยใช้ตัวประมาณกำลังสองน้อยสุดในการประมาณค่าพารามิเตอร์ แต่หากค่าผิดปกติถูกบันทึกมาอย่างถูกต้องและทราบถึงสาเหตุของการเกิดค่าผิดปกติ เช่น เกิดจากรูปแบบของสมการถดถอยที่ใช้ไม่เหมาะสม ผู้วิเคราะห์ควรทำการปรับปรุงตัวแบบถดถอยให้มีความเหมาะสมมากขึ้น แต่หากไม่สามารถอธิบายถึงสาเหตุของการเกิดค่าผิดปกติได้ อาจแก้ไขโดยการแปลงข้อมูล (Data Transformation) เพื่อลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลง และนอกจากวิธีการแปลงข้อมูลแล้ว มีนักสถิติหลายท่านได้คิดค้นตัวประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นไว้เป็นจำนวนมาก โดยมีหลักการคือ พยายามลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลง โดยไม่มีการตัดค่าผิดปกติดังกล่าวทิ้ง หรือเป็นการให้น้ำหนักของค่าสังเกตที่เป็นค่าผิดปกติน้อยกว่าค่าสังเกตที่เป็นข้อมูลส่วนใหญ่ ซึ่งตัวประมาณที่สร้างจากหลักการดังกล่าว เรียกว่าตัวประมาณการถดถอยที่มีความแกร่ง (Robust Regression Estimator) ซึ่งการวิจัยในครั้งนี้จะทำการศึกษาโดยใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด (Least Absolute Value Estimator) ซึ่งตัวประมาณการถดถอยดังกล่าว จะหาค่าประมาณของพารามิเตอร์ที่ทำให้ผลรวมของค่าสัมบูรณ์ของความคลาดเคลื่อนมีค่าต่ำสุด กล่าวคือ

$$\min_{\beta} \sum_{i=1}^n |y_i - x_i' \beta| \quad (1.2)$$

ซึ่งการวิจัยในครั้งนี้ จะใช้เทคนิคโปรแกรมเชิงเส้น (Linear Programming Technique) ในการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบถดถอยเชิงเส้นพหุคูณที่ทำให้ผลรวมของค่าสัมบูรณ์ของความคลาดเคลื่อนมีค่าต่ำสุด และใช้วิธีการซิมเพล็กซ์ (Simplex Method) ในการหาค่าประมาณพารามิเตอร์ในตัวแบบ

ดังนั้น ในการวิจัยครั้งนี้ ผู้วิจัยจะศึกษาในกรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรตาม โดยทำการศึกษาตัวประมาณกำลังสองน้อยสุดซึ่งอยู่ในรูปแบบสมการที่ (1.1) และตัวประมาณค่าสัมบูรณ์น้อยสุดซึ่งสร้างโดยใช้หลักการจากสมการที่ (1.2)

และเปรียบเทียบประสิทธิภาพของตัวประมาณในการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้นอย่างง่าย โดยใช้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Squares Error : RMSE) ซึ่งตัวประมาณใดที่ให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ต่ำสุด จะถือว่าเป็นตัวประมาณที่มีประสิทธิภาพสูงสุดในแต่ละสถานการณ์ และข้อมูลที่ใช้ในการศึกษาครั้งนี้ได้จากการจำลองด้วยเทคนิคการจำลองมอนติคาร์โล (Monte Carlo Simulation Technique) โดยทำการจำลองซ้ำจำนวน 1,000 ครั้ง ในแต่ละสถานการณ์

ขอบเขตของการวิจัย

1. ในการวิจัยครั้งนี้ ตัวแบบการถดถอยเชิงเส้นอย่างง่ายอยู่ในรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i \quad ; i = 1, 2, \dots, n$$

หรือเขียนในรูปเมทริกซ์ ได้ดังนี้

$$\underset{\sim}{y} = \underset{\sim}{X} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

- เมื่อ $\underset{\sim}{y}$ แทนเวกเตอร์ของตัวแปรตามที่มีขนาด $(n \times 1)$
 $\underset{\sim}{X}$ แทนเมทริกซ์ของตัวแปรอิสระที่มีขนาด $(n \times q)$
 $\underset{\sim}{\beta}$ แทนเวกเตอร์ของสัมประสิทธิ์ถดถอยที่มีขนาด $(q \times 1)$
 $\underset{\sim}{\varepsilon}$ แทนเวกเตอร์ของความคลาดเคลื่อนที่มีขนาด $(n \times 1)$
 n แทนขนาดตัวอย่าง

- และ q แทนจำนวนพารามิเตอร์ที่ต้องการประมาณค่าในตัวแบบซึ่งเท่ากับ 2

โดยมีข้อกำหนด (Assumptions) ว่า ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่เป็นอิสระกันและมีการแจกแจงเดียวกันที่มีค่าเฉลี่ย $E(\varepsilon) = 0$ และ เมทริกซ์ความแปรปรวน - ความแปรปรวนร่วม

$$E(\varepsilon \varepsilon') = \sigma^2 I \text{ ซึ่งศึกษาในกรณีที่ } \underset{\sim}{\beta} = 1$$

2. การแจกแจงของความคลาดเคลื่อนสุ่ม แบ่งออกเป็น 2 การแจกแจง ดังนี้

2.1 การแจกแจงปกติ (Normal Distribution)

$$f(\varepsilon) = \frac{1}{\sigma^2 \sqrt{2\pi}} \cdot \exp\left\{\frac{-1}{2\sigma^2}(\varepsilon - \mu)^2\right\} \quad ; -\infty < \mu < \infty, \sigma^2 > 0$$

เมื่อ μ แทน ค่าเฉลี่ยของความคลาดเคลื่อนสุ่ม
 σ^2 แทน ความแปรปรวนของความคลาดเคลื่อนสุ่ม
 งานวิจัยในครั้งนี้ จะศึกษาในกรณีที่ $\mu = 0$ และ $\sigma^2 = 3^a$

2.2 การแจกแจงปกติปลอมปน (Contaminated Normal Distribution)

$$f(\varepsilon) = (1 - P)N(\mu, \sigma^2) + PN(\mu, C^2\sigma^2)$$

เมื่อ P แทน สัดส่วนการปลอมปน (Percent of Contamination)

และ C แทน สเกลแฟกเตอร์ (Scale Factor)

งานวิจัยในครั้งนี้ จะศึกษาในกรณีที่ $\mu = 0$, $\sigma^2 = 3^a$, $P = 5\%$, 10% และ 15% และมีระดับค่าผิดปกติ 3 ระดับ คือ ระดับเล็กน้อย ปานกลางและรุนแรง โดยใช้เกณฑ์การกำหนดขนาดค่าผิดปกติด้วย Box Plot จะได้ว่า

เมื่อค่า $C = 4 - 6$ ตัวแปรตามจะมีค่าผิดปกติระดับเล็กน้อย - ปานกลาง
 และ เมื่อค่า $C = 12 - 13$ ตัวแปรตามจะมีค่าผิดปกติระดับรุนแรง
 ดังนั้น งานวิจัยในครั้งนี้ เลือกกำหนดค่า $C = 4$ เมื่อตัวแปรตามมีค่าผิดปกติระดับเล็กน้อย $C = 6$ เมื่อตัวแปรตามมีค่าผิดปกติระดับปานกลาง และ $C = 13$ เมื่อตัวแปรตามมีค่าผิดปกติระดับรุนแรง

3. จำนวนตัวแปรอิสระที่ใช้ในการวิจัยเท่ากับ 1^b ตัวแปร

4. ขนาดตัวอย่าง(n) มีจำนวน 5 ระดับ คือ 20, 30, 40, 50 และ 60

5. ในการวิจัยครั้งนี้ทำการจำลองข้อมูลโดยใช้เทคนิคการจำลองมอนติคาร์โล(Monte Carlo Simulation Technique) โดยทำการจำลองจำนวน 1,000 ครั้ง ในแต่ละสถานการณ์

6. เกณฑ์ที่ใช้ในการพิจารณา

ในการวิจัยครั้งนี้ จะทำการเปรียบเทียบประสิทธิภาพของตัวประมาณพารามิเตอร์แต่ละตัว โดยการเปรียบเทียบค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณการถดถอยเชิงเส้น (Root Mean Squares Error : RMSE) โดยมีสูตรในการคำนวณดังต่อไปนี้

$$RMSE = \sqrt{\sum_{j=1}^{1,000} \frac{(\beta_{ij} - \hat{\beta}_{ij})^2}{1,000}}$$

เมื่อ β_{ij} แทน ค่าจริงของพารามิเตอร์ในสมการถดถอยตัวที่ i ในการจำลองรอบที่ j
เมื่อ $i = 1, 2$ และ $j = 1, 2, \dots, 1,000$

$\hat{\beta}_{ij}$ แทน ค่าประมาณของพารามิเตอร์ในสมการถดถอยตัวที่ i ในการจำลองรอบที่ j

a หมายถึง เมื่อทดลองใช้ ค่าอื่นๆ จะให้ผลการเปรียบเทียบตัวประมาณไม่แตกต่างกัน

b หมายถึง เมื่อทดลองใช้จำนวนตัวแปรอิสระเท่ากับ 2 และ 3 ตัวแปร จะให้ผลการเปรียบเทียบ ตัวประมาณไม่แตกต่างกัน

ผลการทดลอง

ตารางที่ 1 แสดงค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) จากการใช้ตัวประมาณกำลังสองน้อยสุด (LS) และตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) ในการประมาณค่าพารามิเตอร์ เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ โดยจำแนกตามระดับขนาดตัวอย่าง (n)

ระดับค่า ผิดปกติ	สัดส่วนการ ปลอมปน	ตัวประ มาณ	ขนาดตัวอย่าง(n)				
			20	30	40	50	60
ไม่มีค่า ผิดปกติ	0%	LS	0.007265*	0.005668*	0.004718*	0.004336*	0.003989*
		LAV	0.014677	0.010714	0.008571	0.007371	0.006634

หมายเหตุ *แทนค่า RMSE ที่มีค่าต่ำสุดในแต่ละสถานการณ์

ตารางที่ 2 แสดงค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง(RMSE) จากการใช้ตัวประมาณกำลังสองน้อยสุด(LS)และตัวประมาณค่าสัมบูรณ์น้อยสุด(LAV)ในการประมาณค่าพารามิเตอร์ เมื่อความคลาดเคลื่อนมีการแจกแจงปกติปลอมปน โดยจำแนกตามระดับค่าผิดปกติ สัดส่วนการปลอมปน และระดับขนาดตัวอย่าง (n)

ระดับค่าผิดปกติ	สัดส่วนการปลอมปน	ตัวประมาณ	ขนาดตัวอย่าง(n)				
			20	30	40	50	60
เล็กน้อย	5%	LS	0.023703	0.015948	0.012974	0.010542	0.008677
		LAV	0.014320*	0.010816*	0.008444*	0.006971*	0.005945*
	10%	LS	0.027236	0.018336	0.014938	0.012188	0.010033
		LAV	0.015084*	0.011393*	0.008896*	0.007396*	0.006309*
	15%	LS	0.028158	0.018970	0.015474	0.012678	0.010437
		LAV	0.015653*	0.011866*	0.009267*	0.007732*	0.006596*
ปานกลาง	5%	LS	0.036386	0.023651	0.015474	0.014191	0.011637
		LAV	0.017525*	0.013144*	0.009267*	0.008298*	0.007054*
	10%	LS	0.042556	0.027662	0.020747	0.016598	0.013610
		LAV	0.018722*	0.014040*	0.010812*	0.008865*	0.007536*
	15%	LS	0.044796	0.029121	0.021838	0.017470	0.014325
		LAV	0.019707*	0.014781*	0.011831*	0.009332*	0.007933*
รุนแรง	5%	LS	0.080777	0.050614	0.034411	0.026963	0.021993
		LAV	0.028741*	0.020298*	0.015990*	0.012945*	0.010933*
	10%	LS	0.096177	0.060303	0.041079	0.032033	0.026131
		LAV	0.031453*	0.023307*	0.017515*	0.014008*	0.011832*
	15%	LS	0.103030	0.064648	0.044113	0.034242	0.027934
		LAV	0.033896*	0.024979*	0.018779*	0.014931*	0.012614*

หมายเหตุ * แทนค่า RMSE ที่มีค่าต่ำสุดในแต่ละสถานการณ์

ตารางที่ 3 แสดงตัวประมาณค่าพารามิเตอร์ที่ควรเลือกใช้ในแต่ละสถานการณ์ของการแจกแจงของความคลาดเคลื่อน ระดับค่าผิดปกติ สัดส่วนการปลอมปน และระดับขนาดตัวอย่าง (n)

การแจกแจงของ ความคลาดเคลื่อน	ระดับค่าผิดปกติ	สัดส่วนการ ปลอมปน	ขนาดตัวอย่าง(n)				
			20	30	40	50	60
การแจกแจง ปกติ	ไม่มีค่าผิดปกติ	0%	LS	LS	LS	LS	LS
การแจกแจง ปกติปลอมปน	เล็กน้อย	5%	LAV	LAV	LAV	LAV	LAV
		10%	LAV	LAV	LAV	LAV	LAV
		15%	LAV	LAV	LAV	LAV	LAV
	ปานกลาง	5%	LAV	LAV	LAV	LAV	LAV
		10%	LAV	LAV	LAV	LAV	LAV
		15%	LAV	LAV	LAV	LAV	LAV
	รุนแรง	5%	LAV	LAV	LAV	LAV	LAV
		10%	LAV	LAV	LAV	LAV	LAV
		15%	LAV	LAV	LAV	LAV	LAV

ผลการทดลองดังแสดงได้จากตารางที่ 1 ถึง 3 ซึ่งแสดงค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) จากการใช้ตัวประมาณกำลังสองน้อยสุด (LS) และตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) ในการประมาณค่าพารามิเตอร์ เมื่อความคลาดเคลื่อนมีการแจกแจงปกติและการแจกแจงปกติปลอมปน โดยจำแนกตามระดับค่าผิดปกติ สัดส่วนการปลอมปน และระดับขนาดตัวอย่าง(n) และแสดงตัวประมาณค่าพารามิเตอร์ที่ควรเลือกใช้ในแต่ละสถานการณ์ นำมาสรุปผลการทดลองได้ดังนี้

1. เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ (ไม่มีค่าผิดปกติในตัวแปรตาม)

ตัวประมาณกำลังสองน้อยสุด (LS) จะให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลัง

สอง (RMSE) ต่ำกว่าตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) ในทุกระดับขนาดตัวอย่าง

2. เมื่อความคลาดเคลื่อนมีการแจกแจงปกติปลอมปน (มีค่าผิดปกติในตัวแปรตาม)

ตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) จะให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ต่ำกว่าตัวประมาณกำลังสองน้อยสุด (LS) ในทุกระดับขนาดตัวอย่าง

3. ผลกระทบจากปัจจัยต่าง ๆ ที่มีต่อค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) พบว่า ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) จะแปรผันตามกับระดับค่าผิดปกติและสัดส่วนการปลอมปน และแปรผกผันกับระดับขนาดตัวอย่าง เมื่อกำหนดให้ปัจจัยอื่น ๆ คงที่

การประยุกต์ใช้กับข้อมูลจริง

การที่จะนำผลการศึกษาที่ได้ในครั้งนี้ไปประยุกต์ใช้กับข้อมูลจริงในการประมาณค่าหรือพยากรณ์ค่าของตัวแปรตามที่ต้องการ มีหลักในการปฏิบัติดังนี้

1. ผลการวิจัยครั้งนี้จะนำไปใช้กับข้อมูลจริงได้ในกรณีที่ความคลาดเคลื่อนสุ่มมีการแจกแจงปกติหรือการแจกแจงปกติปลอมปนเท่านั้น ดังนั้นจึงควรทดสอบการแจกแจงของความคลาดเคลื่อนสุ่มก่อนว่ามีการแจกแจงปกติหรือไม่ โดยสามารถทดสอบการแจกแจงของตัวแปรตามแทนได้ เนื่องจากหากความคลาดเคลื่อนสุ่มมีการแจกแจงปกติตัวแปรตามก็จะมีแจกแจงปกติเช่นกัน ในการทดสอบว่าตัวแปรตามมีการแจกแจงปกติหรือไม่นั้นใช้สถิติทดสอบ Kolmogorov – Smirnov Test (K-S Test) หรือ Shapiro – Wilk Test

2. ในการที่จะทดสอบว่ามีค่าผิดปกติในตัวแปรตามหรือไม่ โดยมีค่าผิดปกติอยู่ในระดับใดและมีสัดส่วนของค่าผิดปกติเท่าไรในข้อมูลที่จะนำมาวิเคราะห์นั้น ทดสอบโดยใช้ Box Plot โดยที่ถ้าทดสอบแล้วพบว่า ไม่มีค่าผิดปกติในตัวแปรตามควรใช้ตัวประมาณกำลังสองน้อยสุด (LS) ในการประมาณค่าพารามิเตอร์ในตัวแบบ ถ้าทดสอบแล้วพบว่า มีค่าผิดปกติในตัวแปรตาม ในทุกระดับค่าผิดปกติ (เล็กน้อย ปานกลาง และรุนแรง) และทุกระดับของสัดส่วนการเกิดค่าผิดปกติ (5%, 10% และ 15%) ควรใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) ในการประมาณค่าพารามิเตอร์ในตัวแบบ ซึ่งในการเลือกใช้ตัวประมาณค่าพารามิเตอร์ที่เหมาะสม จะสามารถทำให้ได้ผลการพยากรณ์ที่มีความถูกต้องมากยิ่งขึ้น

บรรณานุกรม

- Askin, R. G., and Montgomery, D. C. (1980). Augmented robust estimators. *Technometrics*, 22, pp. 333–341.
- Gillett, B. E. (1976). *Introduction to operations research : A computer - oriented algorithmic approach*. New York : McGraw – Hill, Inc.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Annual Mathematics and Statistics*, 35, pp.73–101.
- Lawrence, Kenneth D. and Arthur, Jeffrey L. (Eds.). (1990). *Robust regression : analysis and applications*. New York : M. Dekker.
- Neter, John and Wasserman, William. (1990). *Applied Linear Statistical Models*. Homewood, Illinois : Richard D. Irwin.
- Rousseeuw, Peter J. and Leroy, Annick M. (2003). *Robust regression and outlier detection*. Hoboken [NY] : John Wiley & Sons.