

การเปรียบเทียบวิธีการถดถอยที่มีความแข็งแกร่งสำหรับตัวแบบการถดถอยเชิงเส้นแบบพหุ

Comparison of Robust Regression Methods in Multiple Linear Regression Model

อรพรรณ ต้นตระกูล (Orapan Tantrakul)* ประสิทธิ์ พยัคฆพงษ์ (Prasit Payakkapong)**

บุญอ้อม โฉมทิ (Boonorm Chomtee)***

บทคัดย่อ

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้นแบบพหุ โดยใช้ตัวแปรอิสระ 2 ตัวแปร และทำการศึกษาวิธีการประมาณ 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM ทั้งนี้เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (MSE) ซึ่งการเปรียบเทียบดังกล่าว กระทำภายใต้เงื่อนไขของข้อมูลที่มีค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ในตัวแปรอิสระและตัวแปรตาม สำหรับข้อมูลที่ใช้ในการวิจัยได้จากการจำลองด้วยเทคนิคมอนติคาร์โล กระทำซ้ำ 1,000 ครั้ง ในแต่ละกรณี ด้วยภาษาโปรแกรม R ผลการวิจัยสรุปได้ดังนี้ โดยทั่วไปวิธีตัวประมาณ GM ให้ค่า MSE ต่ำที่สุด แต่บางสถานการณ์วิธีกำลังสองน้อยที่สุด และวิธีตัวประมาณ GM ให้ค่า MSE ใกล้เคียงกัน นอกจากนี้ ค่า MSE จะแปรผกผันกับปัจจัยต่อไปนี้ ได้แก่ องศาความเป็นอิสระ ขนาดตัวอย่าง และอัตราส่วนค่าผิดปกติของตัวแปรอิสระ

ABSTRACT

The purpose of this research was to compare the robust estimation methods in the multiple linear regression model with two independent variables. The methods composed of Ordinary Least Squares method, Least Trimmed Squares estimator method and Generalized M-estimator method. The approach of estimator comparison was Mean square error (MSE) which these methods were conducted under mild outlier conditions in independent and dependent variables. For the data, it was generated by Monte Carlo simulation technique, repeated 1,000 times for each case with R programming language. The results of this research summarized that in general case, Generalized M-estimator method indicated the lowest of MSE but in some cases, Ordinary Least Squares method and Generalized M-estimator method illustrated the similar efficiency. In addition, MSE would inverse with degree of freedom, sample size and outlier proportion of independent variables.

คำสำคัญ: การถดถอยที่มีความแข็งแกร่ง ค่าผิดปกติ วิธีตัวประมาณ GM

Key Words: Robust regression, Outliers, Generalized M-estimator method

* นิสิต หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

** รองศาสตราจารย์ สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

*** ผู้ช่วยศาสตราจารย์ สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

บทนำ

การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการทางสถิติอย่างหนึ่งที่ถูกนำมาใช้ในการศึกษาความสัมพันธ์ระหว่าง 2 ตัวแปรหรือมากกว่า 2 ตัวแปร โดยในการวิเคราะห์ห้มุ่งที่จะพยากรณ์ตัวแปรหนึ่งซึ่งเรียกว่าตัวแปรตาม (Dependent Variable) จากตัวแปรอีกตัวหนึ่งหรือมากกว่าที่มีอิทธิพลกับตัวแปรตามซึ่งเรียกว่าตัวแปรอิสระ (Independent Variables) วิธีการดังกล่าวเป็นที่นิยมใช้กันอย่างกว้างขวาง ไม่ว่าจะเป็นทางด้านธุรกิจ เศรษฐศาสตร์ เกษตรศาสตร์ สังคมศาสตร์ และวิทยาศาสตร์ เป็นต้น ถ้ามีตัวแปรอิสระมากกว่า 1 ตัวขึ้นไป การวิเคราะห์การถดถอยจะถูกเรียกว่า การวิเคราะห์การถดถอยแบบพหุ (Multiple Regression Analysis) และในที่นี้ผู้วิจัยสนใจศึกษาวิเคราะห์การถดถอยแบบพหุที่มีลักษณะความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามในรูปแบบเชิงเส้น ซึ่งเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นแบบพหุ โดยลักษณะของความสัมพันธ์มีรูปแบบดังนี้

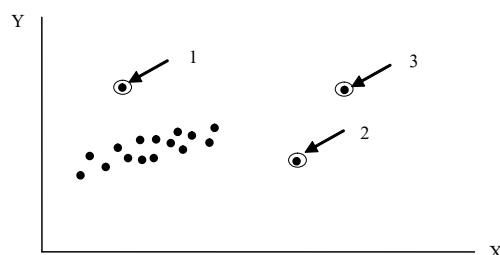
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

$$; i = 1, 2, \dots, n$$

ซึ่งมีข้อตกลงเบื้องต้น ดังนี้ 1. ค่าเฉลี่ยของค่าความคลาดเคลื่อนมีค่าเท่ากับศูนย์ คือ $E(\varepsilon_i) = 0$ 2. ความแปรปรวนของค่าความคลาดเคลื่อนมีค่าคงที่ คือ $V(\varepsilon_i) = \sigma^2$ 3. ค่าความคลาดเคลื่อนไม่มีความสัมพันธ์กัน คือ $E(\varepsilon_i \varepsilon_j) = 0$ เมื่อ $i \neq j$ 4. ตัวแปรอิสระไม่มีความสัมพันธ์กัน

การสร้างสมการถดถอยเริ่มจากการหาค่าประมาณ $\hat{\beta}_j$ ของพารามิเตอร์ β_j โดยถ้าข้อมูลที่นำมาวิเคราะห์สอดคล้องกับข้อตกลงเบื้องต้นของตัวแบบการถดถอยแล้วการประมาณค่าพารามิเตอร์ที่นิยมใช้กันมาก คือ วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares method: OLS) โดยเป็นตัวประมาณที่คุณสมบัติเป็น BLUE (Best Linear Unbiased Estimators) กล่าวคือ มีลักษณะเป็นเชิงเส้นที่ไม่เอนเอียง และมีความแปรปรวนต่ำที่สุด

ในทางปฏิบัติลักษณะของข้อมูลหรือค่าสังเกตที่เก็บรวบรวมเพื่อนำมาวิเคราะห์การถดถอยเชิงเส้นแบบพหุ อาจมีบางค่าสูงกว่าหรือต่ำกว่าค่าสังเกตอื่นๆมาก ซึ่งข้อมูลลักษณะดังกล่าวเรียกว่า ข้อมูลที่มีค่าผิดปกติ (Outliers) โดยค่าผิดปกตินี้อาจจะเป็นค่าผิดปกติเนื่องจากตัวแปรอิสระหรือเนื่องจากค่าของตัวแปรตาม หรือจากค่าของตัวแปรอิสระและตัวแปรตาม ดังตัวอย่างของค่าผิดปกติแสดงในภาพที่ 1



ภาพที่ 1 แผนภาพการกระจายกรณีข้อมูลมีค่าผิดปกติ

จากภาพที่ 1 จะเห็นว่า

จุดที่ 1 แสดงค่าผิดปกติเนื่องจากค่าของ Y โดย X มีค่าอยู่ในช่วงการพิจารณาแต่ Y มีค่าอยู่นอกช่วงการพิจารณา

จุดที่ 2 แสดงค่าผิดปกติเนื่องจากค่าของ X โดย Y มีค่าอยู่ในช่วงการพิจารณาแต่ X มีค่าอยู่นอกช่วงการพิจารณา

จุดที่ 3 แสดงค่าผิดปกติเนื่องจากค่าของ X และ Y เพราะทั้ง X และ Y มีค่าอยู่นอกช่วงการพิจารณาหรือต่างจากกลุ่ม (ทรงศิริ, 2548)

ในกรณีที่ข้อมูลมีค่าผิดปกติ การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด แสดงให้เห็นถึงอันตรายที่เกิดจากการมีค่าผิดปกติ โดยที่ค่าผิดปกติมักจะเกิดขึ้นบ่อยๆ ในข้อมูลจริงและยังเป็นที่สังเกตได้ยากอีกด้วย เนื่องจากปัจจุบันข้อมูลจำนวนมากจะผ่านกระบวนการวิเคราะห์โดยใช้คอมพิวเตอร์ ซึ่งผู้วิเคราะห์มักจะละเลยการตรวจสอบข้อมูล ทำให้เมื่อเกิดค่าผิดปกติขึ้น ค่าความคลาดเคลื่อนที่เกิดจากวิธีกำลังสองน้อยที่สุด จะมีค่าสูง ซึ่งส่งผลให้การประมาณมี

ประสิทธิภาพต่ำ (Rousseeuw and Leroy, 2003) ดังนั้นจึงควรพิจารณาว่าค่าสังเกตที่สงสัยนั้นเป็นค่าผิดปกติหรือไม่ หากพบว่าเป็นค่าผิดปกติจะต้องทำการแก้ปัญหา โดยการกลับไปตรวจสอบหาสาเหตุที่ทำให้เกิดค่าผิดปกติ และถ้าค่าผิดปกติเกิดจากบุคคลหรือเครื่องมือวัดต่างๆ จะไม่นำค่าผิดปกตินั้นมาวิเคราะห์ แต่ถ้าค่าผิดปกติเกิดจากธรรมชาติของข้อมูลต้องนำค่าผิดปกตินั้นมาร่วมในการวิเคราะห์การถดถอยด้วย (ทรงศิริ, 2548)

เมื่อเกิดปัญหาจากการวิเคราะห์ข้อมูลด้วยวิธีกำลังสองน้อยที่สุด การวิเคราะห์การถดถอยที่มีความแกร่ง (Robust Regression) เป็นทางเลือกหนึ่งที่ใช้ในการวิเคราะห์การถดถอยเชิงเส้นแบบพหุ ซึ่งเป็นการวิเคราะห์การถดถอยที่ใช้หลักการลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลง โดยนักสถิติหลายท่านได้ทำการคิดวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งหลายวิธี เช่น Rousseeuw (1984) ได้เสนอวิธีตัวประมาณ LTS (Least Trimmed Squares estimator method) โดยวิธีตัวประมาณ LTS คือวิธีกำลังสองน้อยที่สุดที่ทำการปรับแต่งและตัดค่าความคลาดเคลื่อนบางตัวที่มีค่าตัวเลขเกินจากช่วงที่ต้องการประมาณค่าออกไป จะใช้เฉพาะข้อมูลที่มีค่าความคลาดเคลื่อนใกล้เคียงกันและจะตัดข้อมูลที่มีค่าความคลาดเคลื่อนมากออกไปจากเส้นแนวโน้ม (Knez and Ready, 1997) และ Huber (1964 อ้างถึงใน จิตรวิ, 2539) ได้เสนอวิธีตัวประมาณ M (M-estimator method) ซึ่งเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับกรณีข้อมูลที่มีค่าผิดปกติในตัวแปรตาม หรือกรณีที่ค่าความคลาดเคลื่อนมีค่าผิดปกติ และถ้าข้อมูลที่น่ามาวิเคราะห์เกิดมีค่าผิดปกติในตัวแปรอิสระและตัวแปรตามพร้อมๆ กัน อาจมีผลทำให้วิธีตัวประมาณ M ให้ค่าประมาณสัมประสิทธิ์การถดถอยที่ไม่เหมาะสม ดังนั้น Simpson และ Montgomery (1996) จึงได้พัฒนาจากวิธีตัวประมาณ M มาเป็นวิธีตัวประมาณ GM (Generalized M-estimator Method) เพื่อทำการสร้างตัวถ่วงน้ำหนัก

ซึ่งจะพิจารณาทั้งค่าผิดปกติที่เกิดขึ้นในตัวแปรอิสระและตัวแปรตามด้วย

จากการศึกษาการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งข้างต้น ผู้วิจัยมีความสนใจที่จะทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยเชิงเส้นแบบพหุ โดยใช้วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM ภายใต้สถานการณ์แบบต่างๆ

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยเชิงเส้นแบบพหุ เมื่อข้อมูลมีค่าผิดปกติระดับไม่รุนแรง (Mild Outliers) ปะปนอยู่ในตัวแปรอิสระและตัวแปรตาม ด้วยวิธีการประมาณ 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM

2. เพื่อศึกษาปัจจัยต่างๆ ได้แก่ ขนาดตัวอย่าง อัตราส่วนค่าผิดปกติของตัวแปรอิสระ และค่าพารามิเตอร์ของการแจกแจงของค่าความคลาดเคลื่อน ได้แก่ อนุภาคความเป็นอิสระ ที่มีผลต่อประสิทธิภาพของการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี

อุปกรณ์และวิธีการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง (Simulation Research) ทำการจำลองข้อมูลที่ใช้ในการวิจัยด้วยเทคนิคมอนติคาร์โล (Monte Carlo Simulation technique) จากภาษาโปรแกรม R และจำลองข้อมูลซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ โดยขั้นตอนการวิจัยจะแบ่งเป็น 5 ขั้นตอน ตามเงื่อนไขดังนี้

1. สร้างข้อมูลของตัวแปรอิสระ X_1 และ X_2 เมื่อกำหนดให้ตัวแปรอิสระแต่ละตัวมีค่าผิดปกติระดับไม่รุนแรง ตามเงื่อนไขของการตรวจสอบค่าผิดปกติโดยใช้กราฟแบบ Box และ Whisker คือ ค่าของตัวแปรอิสระที่อยู่ในช่วง $[Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$ หรือ

$[Q_3 + 1.5(IQR), Q_3 + 3(IQR)]$ โดยกำหนดให้มีอัตราส่วนค่าผิดปกติของตัวแปรอิสระ (PX) เท่ากับ 0.10 และ 0.15 และขนาดตัวอย่าง (n) เท่ากับ 20, 50, 100 และ 200

2. สร้างข้อมูลของค่าความคลาดเคลื่อน โดยมีการแจกแจงแบบที่ ท้องศาความเป็นอิสระ (df) ต่างกัน 3 ระดับ คือ 1, 3 และ 5

3. สร้างข้อมูลของตัวแปรตาม ตามรูปแบบของความสัมพันธ์ดังนี้

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

เมื่อกำหนดให้ $\beta_0 = 1, \beta_1 = 1$ และ $\beta_2 = 1$

4. ประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยทั้ง 3 วิธี ดังนี้

4.1 วิธีกำลังสองน้อยที่สุด ในกรณีการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดจากสมการความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ คือ $Y = X\beta + \varepsilon$

เมื่อ Y เป็นเวกเตอร์ของตัวแปรตาม

X เป็นเมทริกซ์ของตัวแปรอิสระ

β เป็นเวกเตอร์ของพารามิเตอร์หรือสัมประสิทธิ์การถดถอยที่ไม่ทราบค่า

ε เป็น เวกเตอร์ ของ ค่า ความคลาดเคลื่อน

n เป็นจำนวนค่าสังเกต

k เป็นจำนวนตัวแปรอิสระ

โดยที่ ε เป็นตัวแปรสุ่มที่เป็นอิสระกันมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่เท่ากับ σ^2

จากหลักการของวิธีกำลังสองน้อยที่สุด จะหาตัวประมาณของ β หรือ $\hat{\beta}$ ที่ทำให้ผลบวกกำลังสองของความแตกต่างระหว่างค่าจริงกับค่าประมาณหรืออาจเรียกว่าผลบวกกำลังสองของความคลาดเคลื่อน (Sum of Square of Error: SSE) มีค่าน้อยที่สุดซึ่งก็คือ

$$\min_{\beta} \sum_{i=1}^n (Y_i - X_i' \hat{\beta})^2 \text{ หรือ } \min_{\beta} \sum_{i=1}^n (e_i)^2 \quad (1)$$

เมื่อ

$$SSE = e'e$$

$$= (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

$$= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

และการแก้สมการ (1) เพื่อหา $\hat{\beta}$ ทำได้โดยการหาอนุพันธ์ของสมการ (1) เทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับศูนย์ จะได้

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (2)$$

เมื่อ $\hat{\beta}$ คือ เวกเตอร์ของตัวประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอย (พิษณุ, 2549)

4.2 วิธีตัวประมาณ LTS Rousseeuw (1984) ได้เสนอวิธีตัวประมาณ LTS โดยมีหลักการดังนี้

$$\min \sum_{i=1}^h (e_i^2)_{i:n} \quad (3)$$

โดยที่ $(e^2)_{1:n} \leq (e^2)_{2:n} \leq \dots \leq (e^2)_{n:n}$ คือ ลำดับของค่าความคลาดเคลื่อนกำลังสอง และ h คือ ค่าคงที่ในการตัดแต่ง (trimming constant) หรืออีกนัยหนึ่งคือจำนวนของข้อมูลที่ใช้ในการประมาณค่าด้วยวิธีตัวประมาณ LTS และ $h = [n/2] + [(p+1)/2]$

เมื่อ n เป็นจำนวนค่าสังเกต

p เป็นจำนวนพารามิเตอร์ที่ใช้ในการประมาณค่า

วิธีตัวประมาณ LTS คือวิธีกำลังสองน้อยที่สุดที่ทำการปรับแต่งและตัดค่าความคลาดเคลื่อนบางตัวที่มีค่าตัวเลขเกินจากช่วงที่ต้องการประมาณค่าออกไป ซึ่งการตัดแต่งค่าความคลาดเคลื่อนจะมีตัวกำหนดจำนวนของค่าความคลาดเคลื่อนที่ใช้ในการประมาณค่าของวิธีตัวประมาณ LTS คือ ค่า h โดยจะเลือกประมาณค่าเฉพาะในช่วงที่มีการกระจายตัวของค่าความคลาดเคลื่อนมากที่สุด และจะไม่ประมาณค่าในช่วงที่มีการกระจายตัวของค่าความคลาดเคลื่อนน้อย

4.3 วิธีตัวประมาณ GM Simpson และ Montgomery (1996) ได้พัฒนาวิธีตัวประมาณ GM เพื่อทำการสร้างตัวถ่วงน้ำหนัก ซึ่งจะพิจารณาทั้งค่าผิดปกติที่เกิดขึ้นในตัวแปรอิสระและตัวแปรตาม โดยตัว

ประมาณ GM มาจากหลักการพื้นฐานของการประมาณแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator) นั่นคือหาตัวประมาณ $\hat{\beta}$ ที่ทำให้ฟังก์ชันภาวะน่าจะเป็นของค่าความคลาดเคลื่อนมีค่าสูงสุด ดังนี้

$$\max \sum_{i=1}^n \ln f(e_i) = \max \sum_{i=1}^n \ln f(Y_i - X_i' \hat{\beta}) \quad (4)$$

เมื่อ $\sum_{i=1}^n \ln f(Y_i - X_i' \hat{\beta})$ เป็นฟังก์ชันภาวะน่าจะเป็น

โดยที่ X_i' เป็นเวกเตอร์ของตัวแปรอิสระ k ตัวของค่าสังเกตที่ i หรือ $X_i' = [1 \ X_{1i} \ X_{2i} \ \dots \ X_{ki}]$

จากสมการ (4) หา $\hat{\beta}$ โดยการหาอนุพันธ์ของ $\sum_{i=1}^n \ln f(Y_i - X_i' \hat{\beta})$ เทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับศูนย์

$$\sum_{i=1}^n X_i' \frac{f(Y_i - X_i' \hat{\beta})}{f(Y_i - X_i' \hat{\beta})} = 0 \quad \text{เมื่อ } f(e_i) = \frac{df(e_i)}{de_i}$$

จากหลักการของวิธีภาวะน่าจะเป็นสูงสุด จำเป็นต้องทราบรูปแบบของฟังก์ชันความหนาแน่นของค่าความคลาดเคลื่อนคือ $f(e_i)$ เมื่อ e_i คือ $Y_i - X_i' \hat{\beta}$ ถ้า e_i มีการแจกแจงแบบปกติแล้ว $\hat{\beta}$ จะมีรูปแบบเดียวกันกับตัวประมาณจากวิธีกำลังสองน้อยที่สุด แต่ในกรณีที่ไม่มีทราบรูปแบบฟังก์ชันการแจกแจงของค่าความคลาดเคลื่อน จะทำให้ไม่สามารถหา $f(e_i)/f(e_i)$ และ $\hat{\beta}$ ได้ ดังนั้น Huber ได้กำหนดฟังก์ชัน ψ ซึ่งเป็นตัวแทนที่ดีในการอธิบายลักษณะการแจกแจงของค่าความคลาดเคลื่อน และการกำหนดฟังก์ชัน ψ จะมีผลต่อความแกร่งและประสิทธิภาพของตัวประมาณ GM โดยหลักการของตัวประมาณ GM คือหาตัวประมาณของ β ที่ทำให้ผลรวมของฟังก์ชันความผิดพลาด (ρ) มีค่าน้อยที่สุด ซึ่งรูปแบบ คือ

$$\min \sum_{i=1}^n \rho \left(\frac{Y_i - X_i' \hat{\beta}}{\pi_i s} \right) \pi_i \quad (5)$$

เมื่อ ρ เป็นฟังก์ชันของความผิดพลาดที่ถูกเลือกให้เหมาะสม

s เป็นตัวประมาณที่แกร่งของสเกล (robust estimator of scale)

π_i เป็นค่าถ่วงน้ำหนักซึ่งสามารถ

คำนวณจาก $\pi_i = \text{median} |z_i| / z_i$; $z_i = \left(\frac{e_i}{s} \right)$

จากสมการ (5) หา $\hat{\beta}$ ที่เหมาะสมโดยการหา

อนุพันธ์บางส่วนของ $\sum_{i=1}^n \rho \left(\frac{Y_i - X_i' \hat{\beta}}{\pi_i s} \right) \pi_i$ เทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับศูนย์

$$\sum_{i=1}^n \psi \left(\frac{e_i}{\pi_i s} \right) \pi_i X_i = 0 \quad (6)$$

เมื่อ $\psi(u) = \frac{\partial}{\partial u} \rho(u) = \rho'$ และถ้า $\psi(u) = u$ แล้ววิธี

ตัวประมาณ GM ก็คือวิธีกำลังสองน้อยที่สุด โดยที่

$$u = \left(\frac{e_i}{\pi_i s} \right)$$

และการคำนวณตัวประมาณที่แกร่งของสเกล (s) ซึ่งเป็นตัวประมาณที่ไม่เอนเอียงของส่วนเบี่ยงเบนมาตรฐาน ในที่นี้ใช้ตัวประมาณแบบมัธยฐานของค่าสัมบูรณ์ของความเบี่ยงเบน (Median Absolute Deviation: MAD) และถูกปรับด้วยค่าคงที่ 1.4826 ซึ่งทำให้ s เป็นตัวประมาณที่ไม่เอนเอียงเมื่อ n มีขนาดใหญ่และการแจกแจงของค่าความคลาดเคลื่อนเป็นแบบปกติ ดังนั้นจะได้ s มีรูปแบบดังนี้

$$s = 1.4826 \text{ median}_i (|e_i - \text{median}_i(e_i)|)$$

จากสมการ (6) พบว่าฟังก์ชัน ψ ไม่เป็นเชิงเส้น ดังนั้นวิธีการแก้สมการนี้จะอาศัยเทคนิคการทำซ้ำเพื่อหาค่าที่เหมาะสม ซึ่งวิธีที่ใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยส่วนใหญ่นิยมใช้วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักชนิดทำซ้ำ (Iteratively Reweighted Least Square : IRLS) ซึ่งการคำนวณวิธี IRLS จะต้องมีการกำหนดค่าเริ่มต้นของ $\hat{\beta}$ และคำนวณ s แล้วหาค่าถ่วงน้ำหนักที่เหมาะสมของแต่ละค่าสังเกตเพื่อหาตัวประมาณค่า $\hat{\beta}$ จากวิธี Weighted Least Square (WLS)

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

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

$$\sum_{i=1}^n W_i \left(\frac{e_i}{\pi_i s} \right) \pi_i X_i = 0$$

$$\text{เมื่อ } W_i = \psi \left(\frac{e_i}{\pi_i s} \right) / \left(\frac{e_i}{\pi_i s} \right)$$

ดังนั้น ตัวประมาณของ β จากวิธี WLS อยู่ใน

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

เมื่อ W เป็นเมตริกซ์ทแยงมุมขนาด $n \times n$ ที่มี w_i เป็นสมาชิกแนวทแยงมุม

สรุปขั้นตอนการคำนวณเมื่อใช้วิธี IRLS ดังนี้

1) คำนวณค่าเริ่มต้นของ $\hat{\beta}$ จากวิธีกำลังสองน้อยที่สุด

$$\hat{\beta}^{(0)} = (X'X)^{-1} X'Y$$

2) เข้าสู่อัลกอริทึมซ้ำรอบที่ 1 เพื่อหาค่า $e_i^{(1)}$, $s^{(1)}$ และ $\pi_i^{(1)}$

$$\text{จาก } e_i^{(1)} = Y_i - X_i' \hat{\beta}^{(0)}$$

$$s^{(1)} = 1.4826 \text{ median } (|e_i^{(1)} - \text{median}(e_i^{(1)})|)$$

$$\pi_i^{(1)} = \text{median} |z_i^{(1)}| / z_i^{(1)} \quad ; \quad z_i^{(1)} = \left(\frac{e_i^{(1)}}{s^{(1)}} \right)$$

3) ทำซ้ำรอบที่ 1 เพื่อหาน้ำหนักของแต่ละค่าสังเกตตามเงื่อนไขของเกณฑ์ความแรงหรือรูปแบบฟังก์ชัน ψ ที่ถูกเลือกจากสมการ (6) ถ้าเลือกเกณฑ์ความแรงที่เหมาะสมจะทำให้ค่าสัมประสิทธิ์การถดถอยที่ได้มีความแรงและมีประสิทธิภาพ

โดยงานวิจัยครั้งนี้ได้ใช้เกณฑ์ความแรงของ Huber ซึ่งมีรูปแบบฟังก์ชัน ρ ฟังก์ชัน ψ และฟังก์ชัน W ดังต่อไปนี้

$$\rho(u) = \begin{cases} \frac{u^2}{2} & ; |u| \leq c_H \\ c_H |u| - \frac{c_H^2}{2} & ; |u| > c_H \end{cases}$$

$$\text{เมื่อ } \psi(u) = \frac{\partial}{\partial u} \rho(u) \text{ แล้ว}$$

$$\psi(u) = \begin{cases} -c_H & ; u < -c_H \\ u & ; -c_H \leq u \leq c_H \\ c_H & ; u > c_H \end{cases}$$

$$\text{และ } W(u) = \frac{\psi(u)}{u} \text{ แล้ว}$$

$$W(u) = \begin{cases} -\frac{c_H}{u} & ; u < -c_H \\ 1 & ; -c_H \leq u \leq c_H \\ \frac{c_H}{u} & ; u > c_H \end{cases}$$

เมื่อ u เป็นค่าความคลาดเคลื่อนมาตรฐาน $\left(\frac{e_i}{\pi_i s} \right)$ และ

c_H เป็นค่าคงที่ (Tuning Constant) มีค่าเท่ากับ 1.345

ซึ่งกำหนดมาเพื่อให้ตัวประมาณที่ได้มีประสิทธิภาพ

ตามที่ต้องการเมื่อเทียบกับตัวประมาณจากวิธีกำลังสองน้อยที่สุด

จากลักษณะของฟังก์ชัน ρ ฟังก์ชัน ψ และฟังก์ชัน W พบว่ามีการกำหนดขอบเขตของค่าความคลาดเคลื่อนมาตรฐาน โดยถือว่าค่าความคลาดเคลื่อนมาตรฐานที่อยู่ภายใต้ขอบเขตจะมีการแจกแจงปกติมาตรฐานจึงมีการให้น้ำหนักเท่ากับหนึ่งกับค่าสังเกตเหล่านี้ ส่วนค่าความคลาดเคลื่อนมาตรฐานที่อยู่ภายนอกขอบเขตจะถูกลดอิทธิพลลงโดยให้น้ำหนักที่น้อยกว่าหนึ่ง

4) การทำซ้ำรอบที่ 1 เพื่อหา $\hat{\beta}^{(1)}$ จาก

$$\hat{\beta}^{(1)} = (X'W^{(1)}X)^{-1} X'W^{(1)}Y$$

โดยที่ $W^{(1)}$ คือ เมตริกซ์ทแยงมุมของน้ำหนักในรอบที่ 1 ขนาด $n \times n$

$$W^{(1)} = \begin{bmatrix} w_1^{(1)} & 0 & \dots & 0 \\ 0 & w_2^{(1)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n^{(1)} \end{bmatrix}$$

5) หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณของสัมประสิทธิ์การถดถอยเริ่มต้นกับค่าประมาณของสัมประสิทธิ์การถดถอยรอบที่ 1 ของสัมประสิทธิ์การถดถอยทุกค่า

6) ถ้าค่าสัมบูรณ์จากข้อ (5) มีอย่างน้อยหนึ่งค่ามากกว่า 0.001 (ค่าความคลาดเคลื่อนจากค่าคงที่ ที่ยอมให้เกิดขึ้นได้ในการคำนวณวิธี IRLS เพื่อให้ได้ค่า $\hat{\beta}$ ก่อนข้างคงที่) ให้กลับไปทำซ้ำรอบที่ r ในข้อ (7)

ต่อไป แต่ถ้านำค่าสัมบูรณ์จากข้อ (5) ของสัมประสิทธิ์การถดถอยทุกค่าไม่มากกว่า 0.001 นั่นคือจะได้ค่าสัมประสิทธิ์การถดถอยจากรอบที่ 1

7) เข้าสู่การทำซ้ำรอบที่ r เมื่อ $r = 2, 3, \dots$ เพื่อหาค่า $e_i^{(r)}$, $s^{(r)}$ และ $\pi_i^{(r)}$

$$e_i^{(r)} = Y_i - X_i' \hat{\beta}^{(r-1)}$$

$$s^{(r)} = 1.4826 \text{ median } (|e_i^{(r)} - \text{median}(e_i^{(r)})|)$$

$$\pi_i^{(r)} = \text{median } |z_i^{(r)}| / z_i^{(r)} \quad ; \quad z_i^{(r)} = \left(\frac{e_i^{(r)}}{s^{(r)}} \right)$$

และคำนวณค่าประมาณสัมประสิทธิ์การถดถอยในการทำซ้ำรอบที่ r

$$\hat{\beta}^{(r)} = (X'W^{(r)}X)^{-1}X'W^{(r)}Y$$

8) หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณของสัมประสิทธิ์การถดถอยรอบที่ $r-1$ กับค่าประมาณของสัมประสิทธิ์การถดถอยรอบที่ r ของสัมประสิทธิ์การถดถอยทุกค่า ถ้านำค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยรอบที่ $r-1$ กับค่าประมาณของสัมประสิทธิ์การถดถอยรอบที่ r ของสัมประสิทธิ์การถดถอยทุกค่า มีอย่างน้อยหนึ่งค่ามากกว่า 0.001 ให้กลับไปทำซ้ำในข้อ (7) ต่อไปเรื่อยๆ จนกระทั่งค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณสัมประสิทธิ์การถดถอยรอบที่ $r-1$ กับค่าประมาณของสัมประสิทธิ์การถดถอยรอบที่ r ของสัมประสิทธิ์การถดถอยทุกค่าไม่มากกว่า 0.001 นั่นคือจะได้ค่าสัมประสิทธิ์การถดถอยจากรอบที่ r

5. คำนวณค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (MSE) ของการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี ภายใต้สถานการณ์ต่างๆ

$$MSE = \frac{\sum_{j=0}^k \left[\sum_{i=1}^{1,000} \frac{(\beta_{ij} - \hat{\beta}_{ij})^2}{1,000} \right]}{k+1}$$

เมื่อ β_{ij} เป็นค่าสัมประสิทธิ์การถดถอยที่กำหนด ตัวที่ j ของการทำซ้ำรอบที่ i

$\hat{\beta}_{ij}$ เป็นค่าประมาณสัมประสิทธิ์การถดถอย ตัวที่ j ของการทำซ้ำรอบที่ i

k เป็นจำนวนตัวแปรอิสระในสมการถดถอย และในที่นี้ $k = 2$

i เป็นจำนวนรอบในการทำซ้ำ
; $i = 1, 2, \dots, 1000$ และ $j = 0, 1, 2$

ผลการวิจัย

ผลการวิจัยสรุปได้ดังนี้

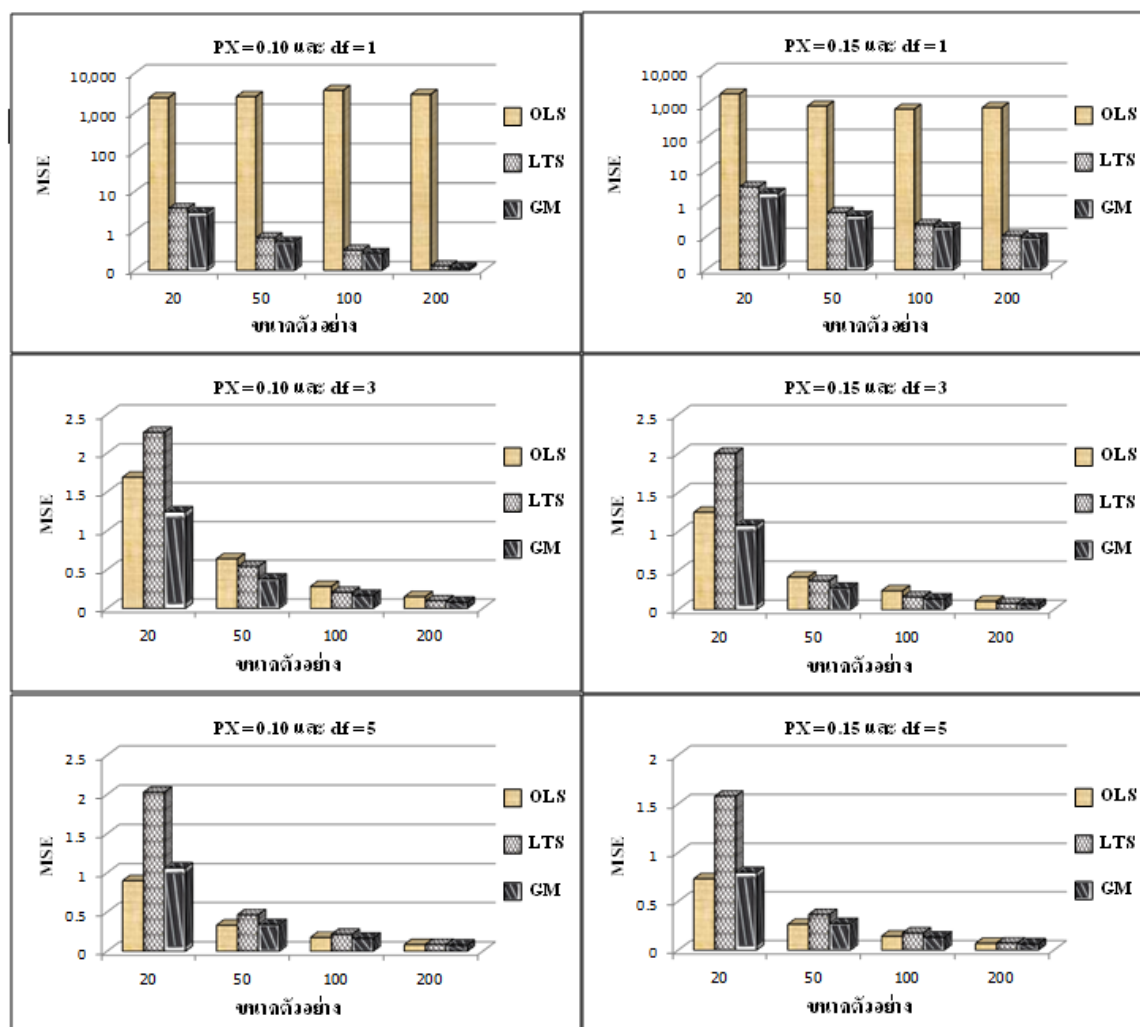
1. องศาความเป็นอิสระที่แตกต่างกันของการแจกแจงแบบที มีผลต่อค่า MSE โดยค่า MSE มีแนวโน้มลดลงเมื่อระดับองศาความเป็นอิสระเพิ่มขึ้น เมื่อองศาความเป็นอิสระเท่ากับ 1 สำหรับทุกขนาดตัวอย่างและทุกอัตราส่วนค่าผิดปกติของตัวแปรอิสระ วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM ให้ค่า MSE ใกล้เคียงกัน โดยวิธีตัวประมาณ GM ให้ค่า MSE ต่ำที่สุด ส่วนที่องศาความเป็นอิสระเท่ากับ 3 สำหรับทุกขนาดตัวอย่างและทุกอัตราส่วนค่าผิดปกติของตัวแปรอิสระ วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM ให้ค่า MSE ใกล้เคียงกัน โดยวิธีตัวประมาณ GM ให้ค่า MSE ต่ำที่สุด และที่องศาความเป็นอิสระเท่ากับ 5 ในทุกสถานการณ์วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM ให้ค่า MSE ใกล้เคียงกัน โดยขนาดตัวอย่างเท่ากับ 20 และ 50 วิธีกำลังสองน้อยที่สุด ให้ค่า MSE ต่ำที่สุด และเมื่อขนาดตัวอย่างเพิ่มขึ้นเท่ากับ 100 และ 200 วิธีตัวประมาณ GM ให้ค่า MSE ต่ำที่สุด

2. ขนาดตัวอย่างมีผลต่อค่า MSE กล่าวคือ ค่า MSE มีแนวโน้มลดลงเมื่อขนาดตัวอย่างเพิ่มมากขึ้น ยกเว้นเมื่อ $df = 1$ สำหรับวิธีกำลังสองน้อยที่สุด พบว่าขนาดตัวอย่างไม่มีผลต่อค่า MSE

3. อัตราส่วนค่าผิดปกติของตัวแปรอิสระมีผลต่อค่า MSE โดยค่า MSE มีแนวโน้มลดลงเมื่ออัตราส่วนค่าผิดปกติของตัวแปรอิสระเพิ่มมากขึ้น รายละเอียดดังตารางที่ 1 และภาพที่ 2

ตารางที่ 1 แสดงค่า MSE ของการประมาณค่าสัมประสิทธิ์การถดถอย วิธี OLS วิธีตัวประมาณ LTS และวิธีตัวประมาณ GM เมื่อเกิดค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ตัวแปรอิสระและตัวแปรตาม โดยจำแนกตามระดับองศาความเป็นอิสระ (df) อัตราส่วนค่าผิดปกติ (PX) และขนาดตัวอย่าง (n)

ขนาดตัวอย่าง	อัตราส่วนค่าผิดปกติ	วิธี	องศาความเป็นอิสระ		
			1	3	5
20	0.10	OLS	2396.628	1.6887	0.8983
		LTS	3.7264	2.2654	2.0280
		GM	2.9209	1.2425	1.0607
	0.15	OLS	2176.887	1.2455	0.7299
		LTS	3.3209	2.0027	1.5762
		GM	2.1214	1.0793	0.7944
50	0.10	OLS	2565.468	0.6387	0.3306
		LTS	0.6741	0.5379	0.4644
		GM	0.5232	0.3769	0.3387
	0.15	OLS	929.1444	0.4164	0.2622
		LTS	0.5509	0.3688	0.3639
		GM	0.4270	0.2725	0.2709
100	0.10	OLS	3710.49	0.2813	0.1753
		LTS	0.3262	0.2054	0.2147
		GM	0.2680	0.1689	0.1651
	0.15	OLS	750.6822	0.2391	0.1421
		LTS	0.2377	0.1627	0.1747
		GM	0.1943	0.1366	0.1324
200	0.10	OLS	2932.374	0.1495	0.0845
		LTS	0.1294	0.0954	0.0857
		GM	0.1146	0.0797	0.0778
	0.15	OLS	835.0604	0.1068	0.0680
		LTS	0.1093	0.0764	0.0709
		GM	0.0921	0.0664	0.0631



ภาพที่ 2 การเปรียบเทียบค่า MSE ของการประมาณค่าสัมประสิทธิ์การถดถอยทั้ง 3 วิธี เมื่อเกิดค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ตัวแปรอิสระและตัวแปรตาม โดยจำแนกตามระดับองศาความเป็นอิสระ (df) อัตราส่วนค่าผิดปกติ (PX) และขนาดตัวอย่าง (n)

สรุปผลการวิจัย

ในการประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยในแบบการถดถอยเชิงเส้นแบบพหุ เมื่อข้อมูลมีค่าผิดปกติระดับไม่รุนแรงปะปนอยู่ในตัวแปรอิสระและตัวแปรตาม พบว่า วิธีตัวประมาณ GM มีประสิทธิภาพมากที่สุด ในทุกระดับองศาความเป็นอิสระ ทุกขนาดตัวอย่างและทุกอัตราส่วนค่าผิดปกติของตัวแปรอิสระ ยกเว้น เมื่อระดับองศาความเป็นอิสระเท่ากับ 5 ขนาดตัวอย่างเท่ากับ 20 และ 50 ในทุกอัตราส่วนค่าผิดปกติของตัวแปรอิสระ

วิธีกำลังน้อยที่สุดมีประสิทธิภาพมากที่สุด และปัจจัยที่มีผลต่อประสิทธิภาพของการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี ได้แก่ องศาความเป็นอิสระ ขนาดตัวอย่าง และอัตราส่วนค่าผิดปกติของตัวแปรอิสระ โดยจะแปรผกผันกับค่า MSE

ข้อเสนอแนะ

ควรมีการศึกษาเพิ่มเกี่ยวกับการประมาณค่าพารามิเตอร์ สำหรับตัวแบบการถดถอยเชิงเส้นแบบพหุ เมื่อข้อมูลมีค่าผิดปกติ ด้วยวิธีประมาณ modified GM6 ที่ศึกษาโดย Bagheri, Midi, Ganjali and Eftekhari (2010) หรือศึกษากรณีค่าความคลาดเคลื่อนมีการแจกแจงแบบอื่นๆ เช่น การแจกแจงแบบโคชี การแจกแจงแบบเอกซ์โพเนนเชียล เป็นต้น

เอกสารอ้างอิง

- จิตรวี วีระประดิษฐ์. 2539. การประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นพหุ เมื่อข้อมูลมีค่าผิดปกติ. วิทยานิพนธ์ปริญญาศิลปศาสตรมหาบัณฑิต สาขาวิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย.
- ทรงศิริ แต่สมบัติ. 2548. การวิเคราะห์การถดถอย. พิมพ์ครั้งที่ 3. กรุงเทพฯ: สำนักพิมพ์มหาวิทยาลัยเกษตรศาสตร์.
- พิษณุ เจียวคุณ. 2549. การวิเคราะห์การถดถอย. เชียงใหม่: สถาบันการวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเชียงใหม่.

- Arezoo Bagheri, Habshah Midi, Mojtaba Ganjali and Samaneh Eftekhari. 2010. A Comparison of Various Influential Points Diagnostic Methods and Robust Regression Approaches: Reanalysis of Interstitial Lung Disease Data. Applied Mathematical Sciences. 4(28): 1367-1386.
- Knez, P. and Ready, M. 1997. On the Robustness of Size and Book-to-Market in Cross-Sectional Regressions. Journal of Finance. 52(4): 1355-82.
- Rousseeuw, P. J. 1984. Least median of squares regression. Journal of the American Statistical Association. 79(388): 871-880.
- Rousseeuw, P. J. and Leroy, A. M. 2003. Robust Regression and Outlier Detection. New York: John Wiley and Sons.
- Simpson, J. R. and Montgomery, D. C. 1996. A biased-robust regression technique for the combined outlier-multicollinearity problem. Journal of Statistical Computation and Simulation. 52(1): 1-22.