

การศึกษาเปรียบเทียบวิธีการประมาณเส้นการถดถอยเชิงเส้นอย่างง่าย เมื่อการแจกแจงของความคลาดเคลื่อนไม่เป็นแบบปกติ

A Comparative Study of Simple Linear Regression Estimation Methods with Non - Normality Error

คำนำ

การวิเคราะห์การถดถอย (regression analysis) เป็นวิธีทางสถิติวิธีหนึ่งที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Y) และตัวแปรอิสระ (X) รูปแบบความสัมพันธ์เชิงเส้นระหว่างตัวแปรตามและตัวแปรอิสระเพียงตัวเดียวเรียกว่าการถดถอยเชิงเส้นอย่างง่าย (Simple linear regression) มีรูปแบบคือ

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad \text{สำหรับ } i = 1, 2, \dots, n$$

โดยที่ Y_i คือ ตัวแปรตาม
 X_i คือ ตัวแปรอิสระ
 β_0 คือ พารามิเตอร์ที่แสดงจุดตัดแกน Y (Y-intercept)
 β_1 คือ พารามิเตอร์ที่แสดงความชัน (slope) ของเส้นถดถอย
 e_i คือ ความคลาดเคลื่อน
 n คือ ขนาดตัวอย่าง

วิธีการที่นิยมใช้ในการประมาณค่าพารามิเตอร์ของการถดถอยเชิงเส้น คือ วิธีกำลังสองน้อยที่สุด (Least squares method) เป็นวิธีที่ให้ตัวประมาณที่มีคุณสมบัติที่ดี คือ เป็นตัวประมาณที่ไม่เอนเอียง (unbiasedness) มีความสอดคล้อง (consistency) มีความพอเพียง (sufficiency) และมีประสิทธิภาพ (efficiency) สูงสุด ในกลุ่มของตัวประมาณที่ไม่เอนเอียงทั้งหมดภายใต้ข้อตกลงเบื้องต้น (assumption) ของการวิเคราะห์การถดถอย โดยข้อตกลงเบื้องต้น ดังกล่าว คือ ความคลาดเคลื่อนแต่ละค่าเป็นอิสระกัน มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนเท่ากัน (ทรงศิริ, 2548)

ในบางครั้งอาจพบว่า ข้อมูลที่นำมาวิเคราะห์มีลักษณะไม่ตรงตามข้อตกลงเบื้องต้นดังกล่าว สาเหตุที่พบบ่อยมากในการวิเคราะห์การถดถอย ได้แก่ ความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติ ซึ่งตัวประมาณค่าพารามิเตอร์ของสมการถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุดไม่มีคุณสมบัติความแกร่ง (robust) กับข้อตกลงเบื้องต้นนี้ (ปรียาร์ตน์, 2534) ซึ่งกรณีความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติอาจจะเป็นกรณีใดกรณีหนึ่ง คือ การแจกแจงของความคลาดเคลื่อนที่มีหางหนา (heavy tail) หรือหางยาว (long tail) กว่าหางของการแจกแจงแบบปกติ ทำให้มีโอกาสที่จะได้ค่าความคลาดเคลื่อนบางค่าสูงหรือต่ำผิดปกติ หรือความคลาดเคลื่อนมีการแจกแจงแบบเบ้ ดังนั้น เมื่อความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติ จะมีผลทำให้ตัวสถิติต่าง ๆ เช่น ค่าเฉลี่ย ความแปรปรวน และตัววัดความสัมพันธ์ระหว่างตัวแปร เป็นต้น ขาดคุณสมบัติของตัวประมาณพารามิเตอร์ที่ดี ซึ่งมีผลทำให้สมการถดถอยที่ประมาณได้มีความคลาดเคลื่อนจากความเป็นจริงค่อนข้างสูง และนำไปสู่การวิเคราะห์และการสรุปผลที่ผิดพลาด

ดังนั้น เมื่อพบว่าความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติจึงควรใช้วิธีการประมาณค่าพารามิเตอร์วิธีอื่น ๆ แทนวิธีกำลังสองน้อยที่สุด ซึ่งมีผู้เสนอวิธีการประมาณค่าพารามิเตอร์ไว้หลายวิธี ดังนี้

1. วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด (Least absolute deviations regression) Gauss and Laplace (1787) ได้พัฒนาทฤษฎีและการคำนวณวิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด จากทฤษฎีของ Boscovich (1757) และนำมาใช้ในการถดถอยเชิงเส้น โดยมีหลักการ คือ ต้องการค่าประมาณของพารามิเตอร์ที่ทำให้ ผลรวมของค่าสัมบูรณ์ของเศษตกค้าง (sum of the absolute values of the residual; $\sum |e_i|$) มีค่าน้อยที่สุด ซึ่งวิธีนี้มีข้อดีกว่าวิธีกำลังสองน้อยที่สุด คือ สามารถใช้แก้ปัญหาความไม่เป็นหนึ่งเดียว (nonuniqueness)

2. วิธีการถดถอยแบบเอ็ม (M regression) Huber (1964) ได้พัฒนาตัวประมาณแบบเอ็ม ซึ่งเป็นตัวประมาณพารามิเตอร์ที่มีคุณสมบัติความแกร่ง (robust) กว่าวิธีกำลังสองน้อยที่สุด และถูกนำมาประยุกต์ใช้ในการถดถอยเชิงเส้น โดยมีหลักการ คือ ต้องการค่าประมาณของพารามิเตอร์ที่ทำให้ผลรวมของฟังก์ชันของเศษตกค้าง ($\sum \rho(e_i)$) มีค่าน้อยที่สุด

3. วิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข (Nonparametric regression using the ranks of numbers) Härdle (1990) ได้แนะนำวิธีการนอนพารามेटริกซ์โดยมีหลักการ คือ ต้องการค่าประมาณของพารามิเตอร์ที่ทำให้สัมประสิทธิ์ของเศษตกค้างที่ได้จาก

ผลรวมของอันดับของค่าสัมบูรณ์ของเศษตกค้างคูณกับค่าสัมบูรณ์ของเศษตกค้าง $(\sum \text{rank}(|\hat{e}_i|)|\hat{e}_i|)$ หรือ $\sum \left[\text{rank}(\hat{e}_i) - \frac{n+1}{2} \right] \cdot \hat{e}_i$ มีค่าต่ำสุด วิธีนี้มีข้อดีกว่าวิธีกำลังสองน้อยที่สุด คือ สามารถนำมาใช้ประมาณเส้นการถดถอยได้ดีในกรณีที่ความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติ ดังนั้นในการวิจัยครั้งนี้ ผู้วิจัยมีวัตถุประสงค์ที่จะเปรียบเทียบวิธีการประมาณเส้นการถดถอยสำหรับการถดถอยเชิงเส้นอย่างง่าย 4 วิธี ในกรณีที่การแจกแจงของความคลาดเคลื่อนไม่เป็นแบบปกติ คือ

1. วิธีกำลังสองน้อยที่สุด (Least squares method : LS)
2. วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด (Least absolute deviations regression : LAD)
3. วิธีการถดถอยแบบเอ็ม (M regression : M)
4. วิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข (Nonparametric regression using the ranks of numbers : NP)

โดยศึกษาจากเกณฑ์ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (mean square error : MSE) และเกณฑ์ค่าเฉลี่ยสัมบูรณ์ของความแตกต่างของค่าประมาณเส้นการถดถอยกับเส้นการถดถอยที่แท้จริง (mean absolute deviation : MAD) เส้นการถดถอยที่ดีที่สุดจะมีค่า MSE และค่า MAD ต่ำสุด

วัตถุประสงค์

1. เพื่อศึกษาถึงอิทธิพลของค่าของพารามิเตอร์ และขนาดตัวอย่างที่มีผลต่อการประมาณเส้นการถดถอยที่ดีที่สุด โดยใช้วิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด วิธีการถดถอยแบบเอ็ม และวิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข
2. เพื่อศึกษาเปรียบเทียบวิธีการประมาณเส้นการถดถอยที่ดีที่สุดโดยวิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด วิธีการถดถอยแบบเอ็ม และวิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข เมื่อความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติ

ขอบเขตการวิจัย

การศึกษาเปรียบเทียบวิธีการประมาณเส้นการถดถอยที่ดีที่สุด โดยวิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด วิธีการถดถอยแบบเอ็่ม และวิธีการถดถอยนอนพารามตริกซ์โดยการจัดตำแหน่งของตัวเลข มีขอบเขตการศึกษา ดังนี้

1. กำหนดขนาดตัวอย่าง 4 ระดับ คือ 10, 20, 30 และ 50

2. กำหนดให้ความคลาดเคลื่อนแต่ละค่าเป็นอิสระกันและมีการแจกแจงดังต่อไปนี้

2.1 การแจกแจงแบบลอการมอรัล (Lognormal distribution) ที่มีส่วนเบี่ยงเบนมาตรฐาน (σ) เท่ากับ 0.1, 0.5, 1 และ 1.5 และกำหนดค่าเฉลี่ย (μ) ให้เป็นค่าคงที่เท่ากับ 0 ฟังก์ชันความหนาแน่นของ x คือ

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} & ; x > 0, \sigma > 0, -\infty < \mu < \infty \\ 0 & ; \text{อื่นๆ} \end{cases}$$

2.2 การแจกแจงแบบเบต้า (Beta distribution) กำหนดค่าพารามิเตอร์ α เท่ากับ 4, 6 และ 8 และค่าพารามิเตอร์ β เท่ากับ 1, 3 และ 5 ฟังก์ชันความหนาแน่นของ x คือ

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & ; 0 < x < 1, \alpha > 0, \beta > 0 \\ 0 & ; \text{อื่นๆ} \end{cases}$$

2.3 การแจกแจงแบบไวบูลล์ (Weibull distribution) กำหนดค่าพารามิเตอร์ α เท่ากับ 2, 3 และ 5 และค่าพารามิเตอร์ β เท่ากับ 0.5, 1 และ 3 ฟังก์ชันความหนาแน่นของ x คือ

$$f(x) = \begin{cases} \alpha \beta x^{\beta-1} e^{-\alpha x^\beta} & ; 0 < x < 1, \alpha > 0, \beta > 0 \\ 0 & ; \text{อื่นๆ} \end{cases}$$

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อให้ทราบถึงปัจจัยที่มีอิทธิพลต่อการประมาณเส้นการถดถอยที่ดีที่สุด โดยวิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด วิธีการถดถอยแบบเอ็่ม และวิธีการถดถอยนอนพาราเมตริกซ์โดยการจัดตำแหน่งของตัวเลข
2. เพื่อเป็นแนวทางให้กับผู้วิจัยที่พบปัญหาเกี่ยวกับความคลาดเคลื่อนที่มีการแจกแจงไม่เป็นแบบปกติ สามารถเลือกวิธีการประมาณเส้นการถดถอยที่เหมาะสมกับลักษณะของข้อมูล อันจะทำให้เกิดความน่าเชื่อถือและความแม่นยำในการทำนายค่ามากขึ้น
3. เพื่อเป็นแนวทางสำหรับผู้สนใจศึกษาเกี่ยวกับวิธีการประมาณเส้นการถดถอยแบบอื่นๆ ในกรณีที่ความคลาดเคลื่อนมีการแจกแจงไม่เป็นแบบปกติต่อไป