

อุปกรณ์และวิธีการ

อุปกรณ์

เครื่องไมโครคอมพิวเตอร์มีหน่วยความจำขนาด 40 GB มีความเร็วในการประมวลผล 256 MHz ที่ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

วิธีการ

การวิจัยครั้งนี้เป็นการศึกษาเปรียบเทียบการประมาณเส้นการถดถอยที่ดีที่สุดสำหรับการถดถอยเชิงเส้นอย่างง่ายด้วยวิธีกำลังสองน้อยที่สุด วิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด วิธีการถดถอยแบบเอ็่ม และวิธีการถดถอยนอนพาราเมตริกซ์โดยการจัดตำแหน่งของตัวเลขแทนที่ข้อมูล เมื่อการแจกแจงของความคลาดเคลื่อนไม่เป็นแบบปกติ โดยกำหนดให้ความคลาดเคลื่อนมีการแจกแจงแบบต่าง ๆ ด้วยค่าพารามิเตอร์ต่างๆ กัน ตามที่ระบุในขอบเขตการวิจัย งานวิจัยนี้แบ่งการดำเนินการออกเป็น 2 ส่วน คือ (1) ส่วนของการจำลองแบบด้วยเทคนิคมอนติคาร์โล จำนวน 1000 ครั้ง โดยใช้โปรแกรม Statistical Analysis System (SAS) และ (2) ส่วนของการศึกษาโดยข้อมูลตัวอย่าง

ในการจำลองแบบมีขั้นตอนดังนี้

1. สร้าง (Generate) ข้อมูลแต่ละชุดตามลักษณะการแจกแจง และขนาดตัวอย่าง (n) ที่กำหนด โดยมีขั้นตอนดังต่อไปนี้
 - 1.1 สร้างค่าความคลาดเคลื่อน (e_i) จำนวน n ค่า ให้มีการแจกแจงตามที่กำหนด
 - 1.2 สร้างค่าตัวแปรอิสระ (X) จำนวน n ค่า มีการแจกแจงแบบยูนิฟอร์มในช่วง $(0,1)$
 - 1.3 สร้างค่าตัวแปรตาม (Y) จำนวน n ค่า จากความสัมพันธ์เชิงเส้นกับตัวแปรอิสระตามตัวแบบ

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad ; i = 1, 2, \dots, n$$

2. การสร้างสมการถดถอยด้วยการประมาณค่าพารามิเตอร์ a และ b

- 2.1 การประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยที่สุด (Least square method : LS)

จากขั้นตอนที่ 1.2 และ 1.3 นำตัวแปรอิสระและตัวแปรตาม มาคำนวณตัวประมาณค่าพารามิเตอร์ a และ b ด้วยวิธีกำลังสองน้อยที่สุด ดังขั้นตอนต่อไปนี้

$$2.1.1 \text{ คำนวณค่าความชัน } b \text{ จาก } b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$2.1.2 \text{ คำนวณค่าคงที่ } a \text{ จาก } a = \bar{y} - b\bar{x}$$

$$2.1.3 \text{ เส้นการถดถอยโดยวิธีกำลังสองน้อยที่สุด คือ } \hat{Y}_i = a + bX_i$$

2.2 การประมาณค่าพารามิเตอร์ด้วยวิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด (Least absolute deviations regression : LAD)

จากขั้นตอนที่ 1.2 และ 1.3 นำตัวแปรอิสระและตัวแปรตาม มาคำนวณตัวประมาณค่าพารามิเตอร์ a และ b ด้วยวิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด ดังขั้นตอนต่อไปนี้

2.2.1 นำข้อมูลเริ่มต้นในแต่ละรอบที่สร้างขึ้น ของตัวแปรอิสระและตัวแปรตาม โดยยังไม่เรียงลำดับมาคำนวณค่าความชัน $b_{i1} = \frac{y_i - y_1}{x_i - x_1}$ เมื่อ $i = 1, 2, \dots, n$ เรียงลำดับความชันจากค่าน้อยที่สุดไปยังค่ามากที่สุด

2.2.2 หาค่าผลรวมสะสมของค่าสัมบูรณ์ของข้อมูลแต่ละค่าลบด้วยข้อมูลเริ่มต้นในแต่ละรอบ และนำค่าที่ได้หารด้วย 2 $\left(\frac{\sum |x_i - x_1|}{2} \right)$

2.2.3 นำค่า $\frac{\sum |x_i - x_1|}{2}$ มาเปรียบเทียบกับค่า $\sum |x_i - x_i|$ ถ้าค่าของ $\frac{\sum |x_i - x_1|}{2}$ มากกว่าค่าใดของ $\sum |x_i - x_i|$ ให้เลือกค่านั้นเป็นจุดเริ่มต้นถัดไป

2.2.4 ดำเนินการเช่นเดียวกับข้อ 2.2.1 ถึงข้อ 2.2.4 จนกระทั่งได้ข้อมูลที่ซ้ำกับจุดข้อมูลก่อนหน้าจึงหยุดขั้นตอน ดังนั้นจะได้ความชันของเส้นถดถอย คือ $b = (\text{จุดข้อมูล } Y \text{ ก่อนหน้า} - \text{จุดข้อมูล } Y \text{ ปัจจุบัน}) / (\text{จุดข้อมูล } X \text{ ก่อนหน้า} - \text{จุดข้อมูล } X \text{ ปัจจุบัน})$ และค่าคงที่ (Intercept) คือ $a = \text{จุดข้อมูล } Y \text{ ก่อนหน้า} - b(\text{จุดข้อมูล } X \text{ ก่อนหน้า})$ ดังนั้น เส้นการถดถอยโดยวิธีการถดถอยแบบค่าสัมบูรณ์ของการเบี่ยงเบนต่ำสุด คือ $\hat{Y}_i = a + bX_i$

2.3 การประมาณค่าพารามิเตอร์ด้วยวิธีการถดถอยแบบเอ็ม (M regression : M)

จากขั้นตอนที่ 1.2 และ 1.3 นำตัวแปรอิสระและตัวแปรตาม มาคำนวณตัวประมาณค่าพารามิเตอร์ a และ b ด้วยวิธีการถดถอยแบบเอ็ม ดังขั้นตอนต่อไปนี้

2.3.1 คำนวณตัวประมาณค่าพารามิเตอร์ a และ b ด้วยวิธีกำลังสองน้อยที่สุด นำค่าที่ได้แทนลงในสมการ $\hat{Y}_i = a + bX_i$ จะได้ค่า $\hat{y}_i$ แต่ละค่า

2.3.2 คำนวณค่าความเบี่ยงเบน e_i จาก $y_i - \hat{y}_i$

2.3.3 คำนวณค่ามัธยฐานของ $|e_i|$

2.3.4 กำหนดค่า $\hat{e}_i$ โดยนำค่ามัธยฐานของ $|e_i|$ มาพิจารณาตามเกณฑ์ ดังนี้ ถ้า $-1.5\hat{\sigma} < e_i < 1.5\hat{\sigma}$ จะได้ $\hat{e}_i = e_i$, ถ้า $e_i \geq 1.5\hat{\sigma}$ จะได้ $\hat{e}_i = 1.5\hat{\sigma}$, ถ้า $e_i \leq -1.5\hat{\sigma}$ จะได้ $\hat{e}_i = -1.5\hat{\sigma}$

2.3.5 ปรับค่า y_i โดยนำ $\hat{y}_i + \hat{e}_i$ จะได้ค่าใหม่ คือ $\hat{y}_i$

2.3.6 นำค่า $\hat{y}_i$ ไปคำนวณซ้ำในขั้นตอนที่ 2.3.1 ถึงขั้นตอนที่ 2.3.6 ไปเรื่อยๆ จนได้ตัวประมาณค่าพารามิเตอร์ a และ b คงที่ ดังนั้น เส้นการถดถอยโดยวิธีการถดถอยแบบเอ็ม คือ $\hat{Y}_i = a + bX_i$

2.4 การประมาณค่าพารามิเตอร์ด้วยวิธีการถดถอยนอนพารามेटริกซ์โดยการจัด

ตำแหน่งของตัวเลข (Nonparametric regression using the ranks of numbers : NP)

จากขั้นตอนที่ 1.2 และ 1.3 นำตัวแปรอิสระและตัวแปรตาม มาคำนวณตัวประมาณค่าพารามิเตอร์ a และ b ด้วยวิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข ดังขั้นตอนต่อไปนี้

2.4.1 หาจำนวนคู่ที่แตกต่างกันของจุดข้อมูล จาก $\frac{n(n-1)}{2}$ จะได้จำนวนคู่ทั้งหมดที่จะต้องคำนวณหาความชัน

2.4.2 คำนวณค่าความชันจาก $b_{ij} = \frac{y_i - y_j}{x_i - x_j}$ โดยที่ $i < j$ จากนั้นเรียงลำดับค่า

ความชันจากน้อยไปมาก

2.4.3 คำนวณค่าน้ำหนักถ่วงของแต่ละคู่ จาก $w_{ij} = \frac{|x_i - x_j|}{\sum |x_i - x_j|}$ โดยที่ $i < j$

เรียงค่า w_{ij} ตามค่า b_{ij}

2.4.4 คำนวณผลรวมสะสมของน้ำหนักถ่วง

2.4.5 หาผลรวมสะสมของน้ำหนักถ่วงค่าแรกที่มีมากกว่า 0.5 จะได้ ตัวประมาณค่าพารามิเตอร์ b คือค่าของ b_{ij}

2.4.6 คำนวณค่า $y_i - bx_i$ ที่ทุกค่าของ i

2.4.7 นำค่าที่ได้จาก 2.4.6 มาหาค่ามัธยฐาน ซึ่งค่าที่ได้ คือ ตัวประมาณค่าพารามิเตอร์ a ดังนั้น เส้นการถดถอยโดยวิธีการถดถอยนอนพารามेटริกซ์โดยการจัดตำแหน่งของตัวเลข คือ $\hat{Y}_i = a + bX_i$

3. การสรุปผลในแต่ละลักษณะจากการทดลองซ้ำ 1000 ครั้ง ประกอบด้วย

3.1 คำนวณ MSE และ MAD ของเส้นการถดถอยจากทั้ง 4 วิธีที่ได้จากขั้นตอนที่ 2

3.2 ทำซ้ำตามขั้นตอนที่ 1.1 ถึง 3.1 จำนวน 1000 ครั้ง

3.3 สรุปผลโดยการเปรียบเทียบค่า MSE และค่า MAD เส้นการถดถอยที่มีค่า MSE และค่า MAD ต่ำที่สุด จะเป็นเส้นที่มีประสิทธิภาพดีที่สุด

4. ทำการทดลองตามขั้นตอนที่ 1–4 สำหรับขนาดตัวอย่างและการแจกแจงของความคลาดเคลื่อนแบบอื่น รวบรวมผลการวิเคราะห์และสรุปผล

สถานที่และระยะเวลาทำการวิจัย

สถานที่ทำการวิจัย ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน มีระยะเวลาทำการวิจัยตั้งแต่เดือนพฤศจิกายน 2548 ถึง มีนาคม 2549