บทที่ 1

บทนำ

ที่มาและความสำคัญของปัญหา

การวิเคราะห์การถดถอย เป็นวิธีการทางสถิติอย่างหนึ่งที่ถูกนำมาใช้ในการคาดคะเนหรือการพยากรณ์ค่าของตัวแปรตัวหนึ่งที่สนใจ โดยอาศัยรูปแบบความสัมพันธ์ในลักษณะเหตุและผล ระหว่างตัวแปรที่ต้องการพยากรณ์หรืออาจเรียกว่าตัวแปรตาม (Dependent Variable) กับตัวแปรอื่น ๆ ที่มีอิทธิพลกับตัวแปรตามซึ่งจะเรียกว่า ตัวแปรอิสระ (Independent Variables) ถ้าตัวแปรอิสระมีจำนวนมากกว่าหนึ่งตัวแล้ว การวิเคราะห์การถดถอยจะถูกเรียกว่า การวิเคราะห์การถดถอยพหุ (Multiple Regression) และในที่นี้สนใจศึกษาการวิเคราะห์การถดถอยพหุที่ลักษณะของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามมีรูปแบบเชิงเส้น ซึ่งเรียกว่า การวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Linear Regression Analysis) โดยลักษณะของความสัมพันธ์อยู่ในรูปของสมการดังนี้

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \quad : i = 1, 2, \ldots, n$$

หรือ

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon$$

เมื่อ $\mathbf{y}$ คือ เวกเตอร์ของตัวแปรตาม ขนาด nx1

$\mathbf{x}$ คือ เมตริกซ์ของตัวแปรอิสระ ขนาด nx(p+1)

β คือ เวกเตอร์ของค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอย ขนาด (p+1)x1

ε คือ เวกเตอร์ของความคลาดเคลื่อน ขนาด nx1

n คือ จำนวนค่าสังเกต

p คือ จำนวนตัวแปรอิสระ

ซึ่งมีข้อตกลงเบื้องต้น ดังนี้

1. ค่าเฉลี่ย (Expected Value) ของความคลาดเคลื่อนมีค่าเท่ากับศูนย์ คือ $E(\varepsilon) = \underline{0}$ หรือ $E(\varepsilon_i) = 0$

2. ความแปรปรวนของความคลาดเคลื่อนมีค่าคงที่ คือ $V(\varepsilon_i) = \sigma^2$

3. ความคลาดเคลื่อนไม่มีความสัมพันธ์กัน คือ $E(\varepsilon_i \varepsilon_j) = 0$ เมื่อ $i \neq j$

4. $\mathbf{X}$ เป็นเมตริกซ์ของค่าคงที่ โดยที่ระหว่างตัวแปรอิสระไม่มีความสัมพันธ์กัน

ถ้าข้อมูลที่ถูกนำมาใช้ในการวิเคราะห์การถดถอยมีลักษณะสอดคล้องกับข้อตกลงเบื้องต้นแล้ว ในการสร้างสมการตัวแบบการพยากรณ์จะต้องทำการประมาณค่าสัมประสิทธิ์การถดถอย โดยปกติจะใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method : OLS Method) ซึ่งวิธีการนี้จะหาตัวประมาณของ β ที่ทำให้ผลบวกกำลังสองของความผิดพลาดมีค่าน้อยที่สุด คือ

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 \Big|_{\beta = \hat{\beta}}$$

เมื่อ $\mathbf{X}_i^T = (1 \quad x_{1i} \quad x_{2i} \quad \ldots \quad x_{pi})$

และ $\hat{\beta}$ คือ เวกเตอร์ของตัวประมาณของสัมประสิทธิ์การถดถอย β ขนาด (p+1)x1 ซึ่งตัวประมาณที่ได้จะมีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียง เชิงเส้นที่ดีที่สุด (Best Linear Unbiased Estimator : BLUE) และถ้า ε_i มีการแจกแจงแบบปกติ (Normal Distribution) แล้ว ตัวประมาณจากวิธี OLS จะเป็นตัวประมาณที่มีรูปแบบเดียวกันกับตัวประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator : MLE) และทำให้ทราบการแจกแจงของ $\hat{\beta}$ เพื่อนำไปใช้หาช่วงความเชื่อมั่น และทดสอบสมมุติฐานต่อไป

ในทางปฏิบัติ ลักษณะของข้อมูลหรือค่าสังเกตที่ถูกเก็บรวบรวมเพื่อนำมาวิเคราะห์การถดถอย อาจมีค่าสูงกว่าหรือต่ำกว่าค่าสังเกตส่วนใหญ่มาก หรือเป็นค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกันกับค่าสังเกตส่วนใหญ่ ซึ่งลักษณะของข้อมูลดังกล่าวจะเรียกว่า ข้อมูลผิดปกติ (Outliers) โดยสาเหตุการเกิดข้อมูลผิดปกติอาจเกิดจากความคลาดเคลื่อนของการวัด เช่น การบันทึกข้อมูล การใช้เครื่องมือที่มีคุณภาพต่ำ เป็นต้น หรือเกิดจากเหตุการณ์ผิดปกติที่ไม่สามารถคาดการณ์ได้ล่วงหน้า เช่น ไฟไหม้ น้ำท่วม เป็นต้น หรืออาจเกิดจากการที่หน่วยตัวอย่างที่ให้ข้อมูลมาจากประชากรอื่นที่มีการแจกแจงแตกต่างจากการแจกแจงของประชากรที่สนใจศึกษา

ลักษณะของข้อมูลที่ผิดปกติมีโอกาสเกิดขึ้นได้กับข้อมูลของตัวแปรตามและข้อมูลของตัวแปรอิสระ และอาจส่งผลทำให้การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุดไม่เหมาะสม เนื่องจากสมการถดถอยที่ได้จะถูกปรับทิศทางไปตาม

ข้อมูลที่ผิดปกติ และอาจทำให้ค่าประมาณของความแปรปรวนของ ε_i มีค่าสูงขึ้นกว่าปกติ ดังนั้นค่าประมาณความแปรปรวนของ $\hat{\beta}$ จึงมีค่าสูงตามไปด้วย และเมื่อนำค่าเหล่านี้ไปใช้ในการทดสอบสมมุติฐานของ β ก็จะได้ผลการทดสอบที่ผิดพลาด อย่างไรก็ตาม ค่าผิดปกติบางค่าอาจให้ข้อมูล (Information) ที่เป็นประโยชน์ต่อการหาสมการถดถอยที่เหมาะสม

ดังนั้นถ้านำข้อมูลที่ใช้ในการวิเคราะห์การถดถอยมาตรวจสอบค่าผิดปกติ แล้วพบว่า มีข้อมูลที่มีค่าผิดปกติและมีอิทธิพลต่อการประมาณค่าสัมประสิทธิ์การถดถอยรวมอยู่ด้วย วิธีแก้ไขมีดังนี้ คือ ผู้วิเคราะห์ไม่ควรที่จะทำการตัดข้อมูลที่มีค่าผิดปกติออกทันที แต่ควรที่จะพิจารณาถึงสาเหตุของการเกิดค่าผิดปกติ เช่น

ถ้าทราบว่าเกิดจากการบันทึกข้อมูลผิดพลาด วิธีแก้ไขอาจใช้ การตัดข้อมูลที่มีค่าผิดปกติออกไป หรืออาจทำการปรับแก้ค่าให้เหมาะสม จากนั้นนำข้อมูลที่เหลือมาทำการวิเคราะห์การถดถอยต่อไป โดยใช้วิธีกำลังสองน้อยที่สุดในการประมาณค่าสัมประสิทธิ์การถดถอย

ถ้าข้อมูลที่มีค่าผิดปกติเหล่านี้ถูกบันทึกมาอย่างถูกต้องและผู้วิเคราะห์ทราบถึงสาเหตุของการเกิดค่าผิดปกติ เช่น เกิดจากรูปแบบของสมการถดถอยที่ใช้ไม่เหมาะสม วิธีแก้ไข คือ ผู้วิเคราะห์ควรทำการปรับปรุงรูปแบบของสมการถดถอยให้มีความเหมาะสม แต่ถ้าผู้วิเคราะห์ไม่สามารถอธิบายถึงสาเหตุของการเกิดค่าผิดปกติแล้ว อาจทำการแก้ไขโดยการแปลงข้อมูล (Data Transformations) เพื่อลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลง และนอกจากวิธีการแปลงข้อมูลแล้ว มีนักสถิติหลายท่านเสนอวิธีการประมาณค่าสัมประสิทธิ์การถดถอยหลายวิธี โดยมีหลักการคือ พยายามลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลงโดยไม่มีการตัดข้อมูลทิ้ง วิธีการเหล่านี้จะจัดอยู่ในพวกการวิเคราะห์การถดถอยที่มีความแกร่ง (Robust Regression)

การวิเคราะห์การถดถอยที่มีความแกร่ง เป็นการวิเคราะห์การถดถอยซึ่งใช้หลักการลดอิทธิพลของข้อมูลที่ผิดปกติลง โดยสร้างสมการถดถอยสำหรับข้อมูลส่วนใหญ่แล้วจะทำการตรวจสอบข้อมูลที่ผิดปกติโดยใช้สมการถดถอยที่สร้างเป็นเกณฑ์ และมีนักสถิติหลายท่านได้ทำการคิดวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งหลายวิธี ซึ่งในที่นี้สนใจที่จะทำการศึกษา วิธีตัวประมาณ M (M-Estimator Method) ซึ่งเสนอโดย Huber (1964) วิธีตัวประมาณ M เป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยสำหรับกรณีที่ข้อมูลมีค่าผิดปกติในตัวแปรตาม หรือกรณีที่ความคลาดเคลื่อนมีค่าผิดปกติ

ซึ่งอาจเป็นผลมาจากการที่ความคลาดเคลื่อนมีการแจกแจงแบบไม่ปกติ (Nonnormal Distributions) เช่น ความคลาดเคลื่อนมีการแจกแจงแบบหางหนา (Heavy-Tailed Distributions) เป็นต้น และในการประมาณค่า β ด้วยวิธีตัวประมาณ M จะใช้หลักการลดอิทธิพลของข้อมูลที่มีค่าความคลาดเคลื่อนผิดปกติลงตามเงื่อนไขของเกณฑ์ความแกร่ง

และถ้าข้อมูลที่นำมาวิเคราะห์เกิดมีค่าผิดปกติในตัวแปรตามและตัวแปรอิสระพร้อม ๆ กัน อาจมีผลให้วิธีตัวประมาณ M ให้ค่าประมาณสัมประสิทธิ์การถดถอยที่ไม่เหมาะสม ดังนั้นจึงมีนักสถิติหลายท่านได้เสนอวิธีการประมาณสัมประสิทธิ์การถดถอยที่เหมาะสมกว่า โดยวิธีการเหล่านี้มีหลักการพื้นฐานคือ พยายามลดอิทธิพลของข้อมูลที่มีค่าผิดปกติที่เกิดในความคลาดเคลื่อนและตัวแปรอิสระไปพร้อม ๆ กัน วิธีการเหล่านี้มีชื่อเรียกว่า วิธีตัวประมาณ Bounded-Influence (Bounded-Influence Estimator Method : BI Estimator Method)

จากการศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีความแกร่งข้างต้น ผู้วิจัยจึงมีความสนใจที่จะทำการศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยในสมการถดถอยเชิงเส้นพหุ โดยใช้วิธีกำลังสองน้อยที่สุด วิธีตัวประมาณ M และวิธีตัวประมาณ BI เมื่อใช้เกณฑ์ความแกร่งของ Huber และของ Tukey ภายใต้สถานการณ์ที่ข้อมูลมีค่าผิดปกติแบบต่าง ๆ แล้วทำการเปรียบเทียบประสิทธิภาพของทุกวิธีด้วยค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (Mean Square Error : MSE) ของการประมาณค่าสัมประสิทธิ์การถดถอย เนื่องจากค่า MSE เป็นค่าใช้วัดคุณสมบัติของตัวประมาณซึ่งได้แก่ความเอนเอียงและความแม่นยำ ดังนั้นถ้าวิธีการประมาณค่าสัมประสิทธิ์การถดถอยวิธีใดมีค่า MSE ต่ำที่สุดจะถือว่า วิธีนั้นจะมีประสิทธิภาพมากที่สุด

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยในสมการถดถอยเชิงเส้นพหุ เมื่อข้อมูลมีค่าผิดปกติด้วยวิธีการประมาณ 5 วิธี คือ

   1.1 วิธีกำลังสองน้อยที่สุด

   1.2 วิธีตัวประมาณ M เมื่อใช้เกณฑ์ความแกร่งของ Huber

   1.3 วิธีตัวประมาณ M เมื่อใช้เกณฑ์ความแกร่งของ Tukey

   1.4 วิธีตัวประมาณ BI เมื่อใช้เกณฑ์ความแกร่งของ Huber

   1.5 วิธีตัวประมาณ BI เมื่อใช้เกณฑ์ความแกร่งของ Tukey

และตัวประมาณสเกลของความคลาดเคลื่อนจะใช้มัธยฐานของค่าสัมบูรณ์ของความเบี่ยงเบน (Median Absolute Deviation : MAD)

2. เพื่อเปรียบเทียบวิธีการประมาณทั้ง 5 วิธีด้วยการเปรียบเทียบค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของการประมาณค่าสัมประสิทธิ์การถดถอยที่ได้จากแต่ละวิธี

ข้อตกลงเบื้องต้น

1. สมการถดถอยที่ใช้ในการศึกษาครั้งนี้จะใช้สมการถดถอยเชิงเส้นพหุ (Multiple Linear Regression Equation) โดยมีรูปแบบสมการดังนี้

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad , i = 1, 2, \ldots, n$$

เมื่อ y_i เป็นตัวแปรตาม

x_{1i}, x_{2i} เป็นตัวแปรอิสระตัวที่ 1 และตัวแปรอิสระตัวที่ 2

β_b เป็นพารามิเตอร์ที่ไม่ทราบค่าหรือสัมประสิทธิ์การถดถอย , b = 0,1,2

ε_i เป็นความคลาดเคลื่อน

n เป็นจำนวนตัวอย่าง

2. ความคลาดเคลื่อน (ε) เป็นตัวแปรสุ่มที่มีลักษณะดังนี้

2.1 $E(\varepsilon_i) = 0$

2.2 $V(\varepsilon_i) = \sigma^2$

2.3 $E(\varepsilon_i \varepsilon_j) = 0 \quad , i \neq j$

2.4 $\varepsilon_i \sim N(0, \sigma^2)$

3. ข้อมูลที่มีค่าผิดปกติจะกำหนดให้เกิดในกรณีต่าง ๆ ดังนี้

3.1 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรตาม (ความคลาดเคลื่อน) ตัวอย่างของข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรตามแสดงไว้ในรูปที่ 1.1

3.2 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระและตัวแปรตาม (ความคลาดเคลื่อน)

3.2.1 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_1 และตัวแปรตาม

3.2.2 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_2 และตัวแปรตาม

3.2.3 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_1 ตัวแปรอิสระ x_2

และตัวแปรตาม

ตัวอย่างของข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระ x_1 ตัวแปรอิสระ x_2 และตัวแปรตามแสดงไว้ในรูปที่ 1.2

3.3 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระและตัวแปรตาม (ความคลาดเคลื่อน) ณ ตำแหน่งเดียวกัน

3.3.1 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_1 และตัวแปรตาม ณ ตำแหน่งเดียวกัน

3.3.2 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_2 และตัวแปรตาม ณ ตำแหน่งเดียวกัน

3.3.3 กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ x_1 หรือตัวแปรอิสระ x_2 และตัวแปรตาม ณ ตำแหน่งเดียวกัน

ตัวอย่างของข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระ x_1 หรือ ตัวแปรอิสระ x_2 และตัวแปรตาม ณ ตำแหน่งเดียวกัน แสดงไว้ในรูปที่ 1.3

4. กรณีที่ข้อมูลเกิดค่าผิดปกติในตัวแปรอิสระ ในที่นี้ผู้วิจัยจะกำหนดให้ตัวแปรอิสระแต่ละตัวประกอบด้วยค่าผิดปกติ 2 ระดับ ตามเงื่อนไขของการตรวจสอบค่าผิดปกติโดยใช้กราฟแบบ Box และ Whisker[1] คือ

1. ค่าผิดปกติระดับปานกลาง (Mild Outliers) คือ ค่าของตัวแปรอิสระที่อยู่ในช่วง $(Q_1 - 3\,(IQR),\ Q_1 - 1.5\,(IQR))$ หรือ $(Q_3 + 1.5\,(IQR),\ Q_3 + 3\,(IQR))$

เมื่อ Q_1 คือ ค่าควอไทล์ที่ 1 (The First Quartile) ของตัวแปรอิสระ

Q_3 คือ ค่าควอไทล์ที่ 3 (The Third Quartile) ของตัวแปรอิสระ

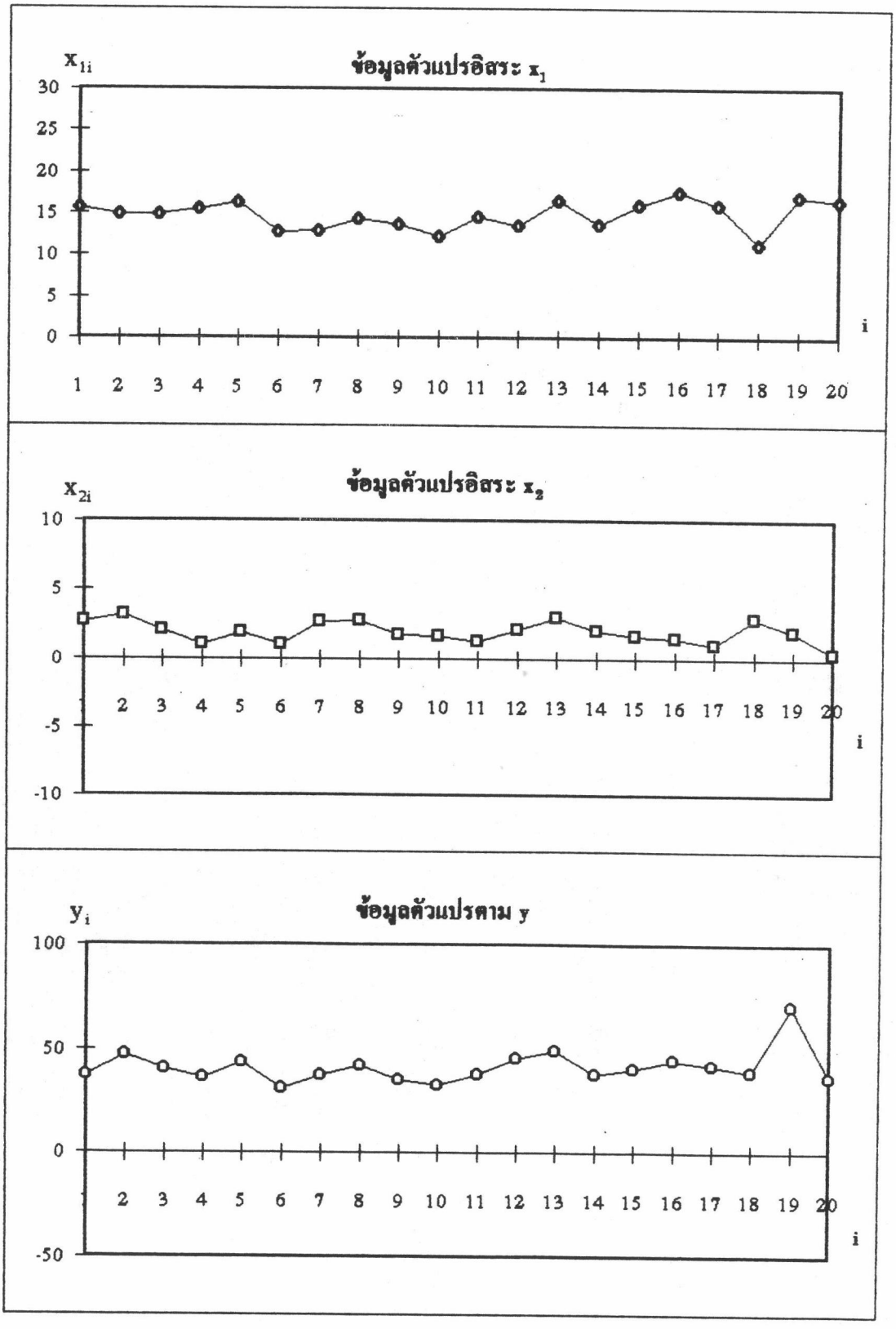
และ IQR คือ ระยะห่างระหว่างควอไทล์ (The Interquartile Range) ซึ่งเท่ากับ $Q_3 - Q_1$

2. ค่าผิดปกติระดับรุนแรง (Extreme Outliers) คือ ค่าของตัวแปรอิสระที่อยู่ในช่วง $(-\infty,\ Q_1 - 3\,(IQR))$ หรือ $(Q_3 + 3\,(IQR),\ \infty)$

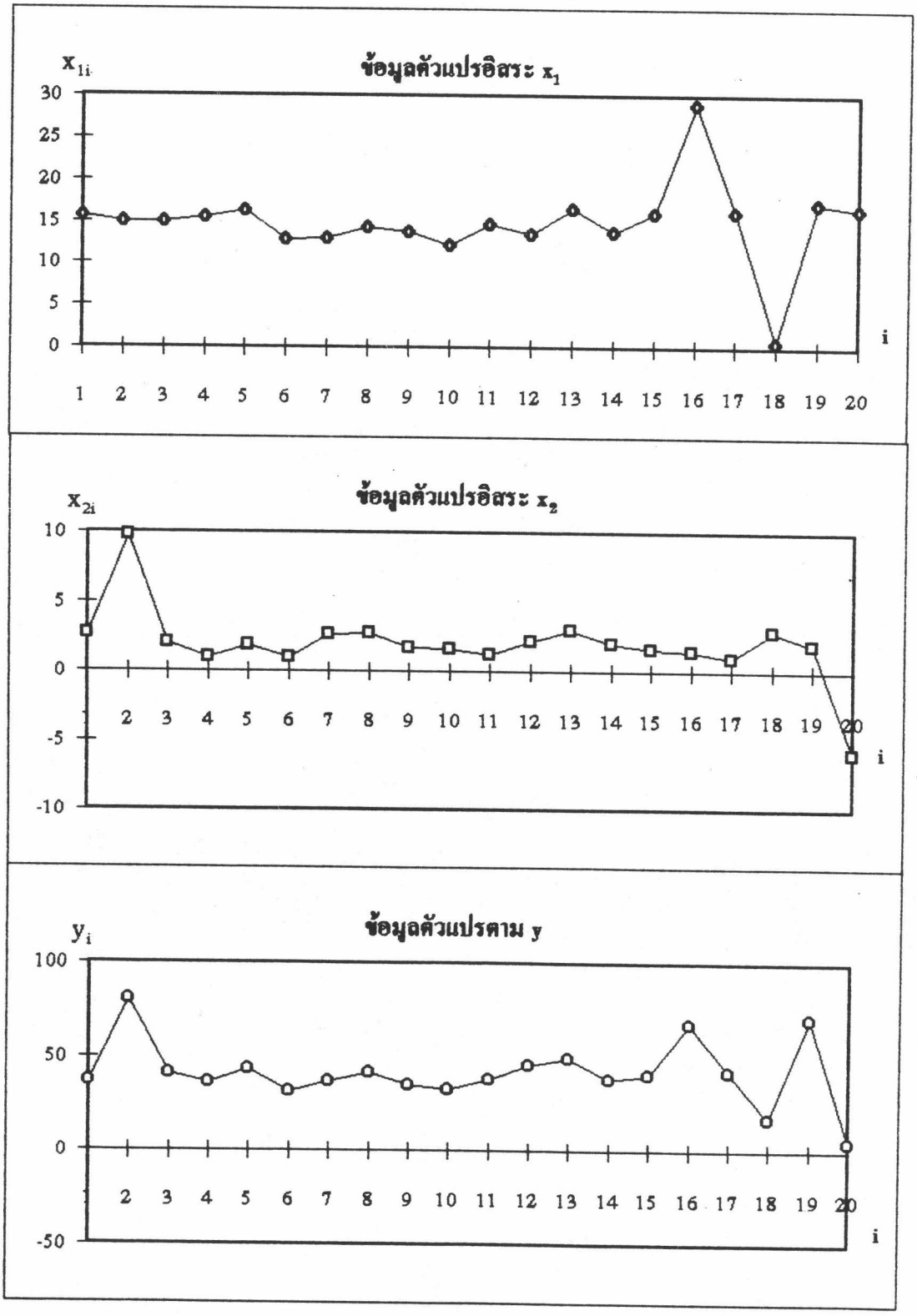
---

[1] กราฟแบบ Box และ Whisker คือ กราฟที่แสดงค่ามัธยฐาน (Median) , ค่าควอไทล์ที่ 1 (The First Quartile) , ค่าควอไทล์ที่ 3 (The Third Quartile) , ค่าที่เล็กที่สุด และค่าที่ใหญ่ที่สุดของข้อมูลซึ่งจะช่วยในการอธิบายลักษณะของข้อมูล เช่น การกระจาย (Spread) ความเบ้ (Skewness) รวมถึงการตรวจสอบว่ามีข้อมูลผิดปกติ (Outliers) หรือไม่ เป็นต้น

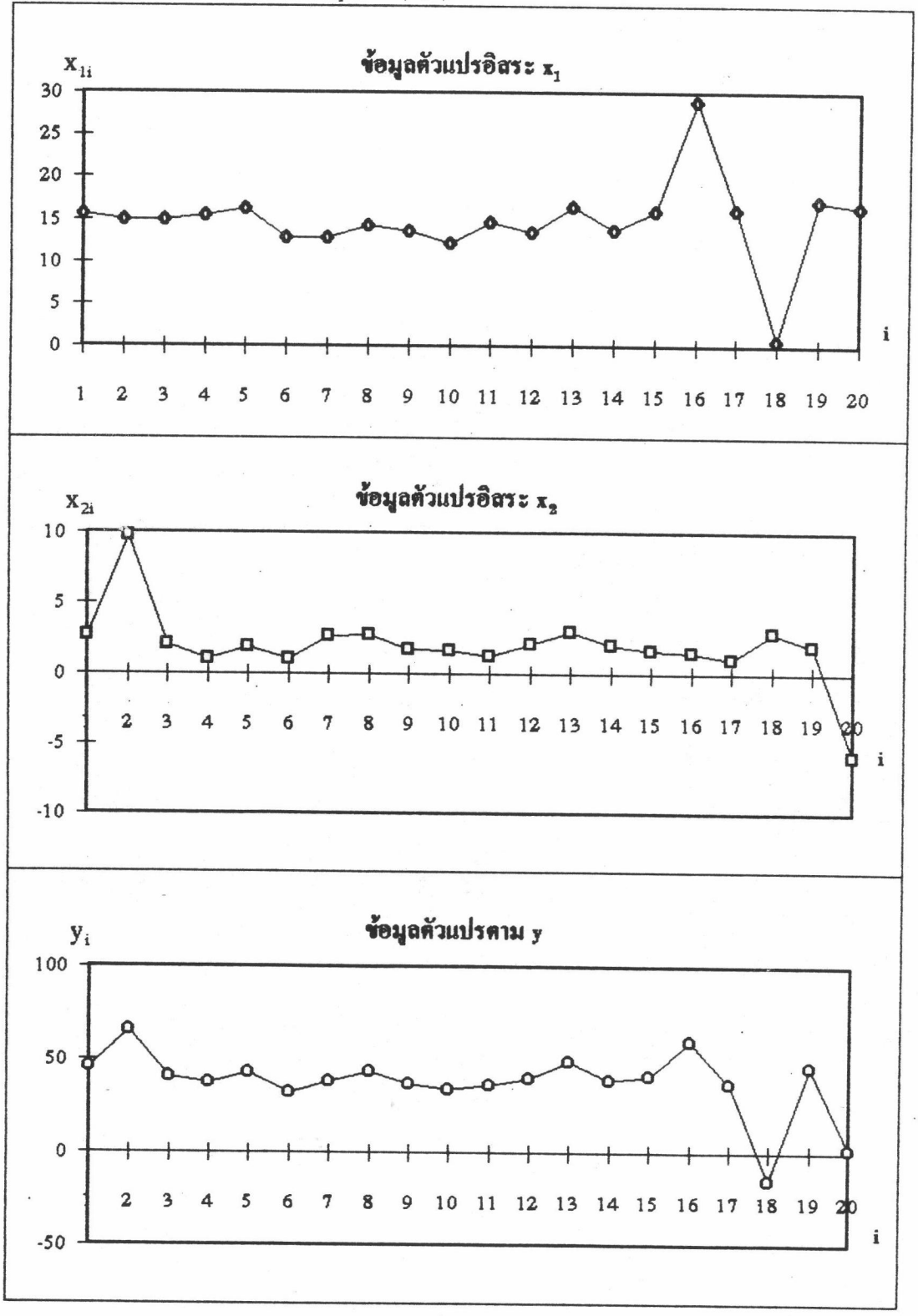
รูปที่ 1.1 แสดงตัวอย่างของข้อมูลกรณีเกิดค่าผิดปกติในตัวแปรตาม (ความคลาดเคลื่อน)
เมื่อ $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, i =1, 2, ... , n โดยที่ $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 5$, n = 20 และ $\varepsilon_i \sim CN(0.10,100)$



รูปที่ 1.2 แสดงตัวอย่างของข้อมูลกรณีเกิดค่าผิดปกติในตัวแปรอิสระ x_1 ตัวแปรอิสระ x_2 และตัวแปรตาม (ความคลาดเคลื่อน) เมื่อ $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, $i = 1, 2, \ldots, n$ โดยที่ $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 5$, $n = 20$ ตัวแปรอิสระ x_1 และ x_2 เกิดค่าผิดปกติ และ $\varepsilon_i \sim CN(0.10,100)$



รูปที่ 1.3 แสดงตัวอย่างของข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระ x_1 หรือตัวแปรอิสระ x_2 และตัวแปรตาม (ความคลาดเคลื่อน) ณ ตำแหน่งเดียวกัน เมื่อ $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, i = 1, 2, ... , n โดยที่ $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 5$, n = 20 ณ ตำแหน่งที่มีค่าตัวแปรอิสระ x_1 หรือ x_2 ผิดปกติจะกำหนดให้ $\varepsilon_i \sim N(0,100)$ ส่วนตำแหน่งอื่นจะกำหนดให้ $\varepsilon_i \sim N(0,1)$



สมมุติฐานของการวิจัย

ภายใต้สถานการณ์ที่ค่าสังเกตมีค่าผิดปกติในตัวแปรตามและตัวแปรอิสระ วิธีตัวประมาณ BI จะให้ตัวประมาณที่มีประสิทธิภาพมากกว่าตัวประมาณจากวิธีตัวประมาณ M และวิธีกำลังสองน้อยที่สุด

ขอบเขตของการวิจัย

1. กำหนดลักษณะข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรตาม (ความคลาดเคลื่อน)

1.1 ลักษณะการแจงแจกของตัวแปรอิสระ x_1 และตัวแปรอิสระ x_2 มีรูปแบบดังนี้

การแจกแจงแบบปกติ (Normal Distribution)

$$x_{1i} \sim N(15,5)$$

$$x_{2i} \sim N(2,0.5)$$

1.2 ลักษณะการแจกแจงของความคลาดเคลื่อน (ε) มีรูปแบบดังนี้

การแจกแจงแบบปกติ (Normal Distribution)

$$\varepsilon_i \sim N(0,1)$$

การแจกแจงแบบปกติปลอมปน (Contaminated Normal Distribution)

$$CN(PE, C_\varepsilon^2) = (1-PE)\cdot N(0,1) + PE\cdot N(0, C_\varepsilon^2)$$

$$\varepsilon_i \sim CN(PE, C_\varepsilon^2)$$

เมื่อ PE คือ อัตราส่วนปลอมปนของ ε ในที่นี้ PE = 0.01, 0.05 และ 0.10
และ C_ε คือ สเกลแฟคเตอร์ของ ε ในที่นี้ C_ε = 3 และ 10

2. กำหนดลักษณะข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระและตัวแปรตาม (ความคลาดเคลื่อน)

2.1 ลักษณะการแจงแจกของตัวแปรอิสระ x_1 และตัวแปรอิสระ x_2 มีรูปแบบดังนี้

การแจกแจงแบบปกติ (Normal Distribution)

$$x_{1i} \sim N(15,5)$$

$$x_{2i} \sim N(2,0.5)$$

และศึกษาระดับค่าผิดปกติของตัวแปรอิสระ x_1 (VX1) และตัวแปรอิสระ x_2(VX2) 2 ระดับ คือ

1. ระดับปานกลาง (Mild Outliers)
2. ระดับรุนแรง (Extreme Outliers)

และแต่ละระดับจะกำหนดให้มีอัตราส่วนค่าผิดปกติเท่ากับ 0.05 และ 0.10

2.2 ลักษณะการแจกแจงของความคลาดเคลื่อน (ε) มีรูปแบบดังนี้

การแจกแจงแบบปกติ (Normal Distribution)

$$\varepsilon_i \sim N(0,1)$$

การแจกแจงแบบปกติปลอมปน (Contaminated Normal Distribution)

$$CN(PE, C_\varepsilon^2) = (1-PE)\cdot N(0,1) + PE\cdot N(0, C_\varepsilon^2)$$

$$\varepsilon_i \sim CN(PE, C_\varepsilon^2)$$

เมื่อ PE คือ อัตราส่วนปลอมปนของ ε ในที่นี้ PE = 0.01, 0.05 และ 0.10

และ C_ε คือ สเกลแฟคเตอร์ของ ε ในที่นี้ C_ε = 3 และ 10

3. กำหนดลักษณะของข้อมูลกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระและตัวแปรตาม (ความคลาดเคลื่อน) ณ ตำแหน่งเดียวกัน

ผู้วิจัยจะกำหนดให้ ความคลาดเคลื่อน (ε) ที่เกิดขึ้น ณ ตำแหน่งที่มีค่าของตัวแปรอิสระผิดปกติ จะมีการแจกแจงแบบปกติ คือ

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

เมื่อ σ_ε^2 คือ ความแปรปรวนของ ε ในที่นี้ σ_ε^2 = 9 และ 100

ส่วนความคลาดเคลื่อน (ε) ที่เกิด ณ ตำแหน่งที่มีค่าของตัวแปรอิสระ ปกติ จะมีการแจกแจงแบบปกติ คือ

$$\varepsilon_i \sim N(0,1)$$

โดยข้อมูลที่มีค่าผิดปกติในตัวแปรอิสระจะมีลักษณะเช่นเดียวกับข้อ 2.2

4. กำหนดค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยโดย $\beta_0 = 1$, $\beta_1 = 2$ และ $\beta_2 = 5$

5. ขนาดตัวอย่างที่ใช้ในการศึกษาเท่ากับ 20, 30 และ 50

6. การวิจัยครั้งนี้ได้จำลองข้อมูลให้มีสถานการณ์ตามที่กำหนดข้างต้น โดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) จากเครื่องคอมพิวเตอร์ AMDAHL 5860 เขียนด้วยโปรแกรมภาษาฟอร์แทรน (Fortran) และทำการจำลองข้อมูลซ้ำ 500 รอบ ในแต่ละสถานการณ์

เกณฑ์การตัดสินใจ

ในที่นี้วัตถุประสงค์ของการวิเคราะห์การถดถอยคือ เพื่อการอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม ดังนั้นเกณฑ์การตัดสินใจว่าวิธีการใดจะมีประสิทธิภาพมากที่สุดในการประมาณค่าสัมประสิทธิ์การถดถอยภายใต้สถานการณ์ต่าง ๆ ที่กำหนดขึ้น ผู้วิจัยจะพิจารณาโดยการเปรียบเทียบค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (Mean Square Error : MSE) ของการประมาณค่าสัมประสิทธิ์การถดถอยจากทุกวิธี และวิธีใดให้ค่า MSE ที่ต่ำกว่าจะเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีประสิทธิภาพกว่า และค่า MSE คำนวณจากสมการดังนี้

$$\text{MSE} = \frac{\sum_{b=0}^{p}\left[\sum_{t=1}^{500}\frac{(\beta_{bt}-\hat{\beta}_{bt})^2}{500}\right]}{p+1}$$

โดยที่ β_{bt} คือ ค่าสัมประสิทธิ์การถดถอยที่กำหนด ตัวที่ b ในการทำซ้ำรอบที่ t

$\hat{\beta}_{bt}$ คือ ค่าประมาณของสัมประสิทธิ์การถดถอย ตัวที่ b ในการทำซ้ำรอบที่ t

p คือ จำนวนของตัวแปรอิสระในสมการถดถอย และในที่นี้ p = 2

t คือ จำนวนรอบของการทำซ้ำ

และ b = 0, 1 และ 2

ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการตัดสินใจเลือกใช้วิธีการประมาณค่าสัมประสิทธิ์การถดถอยได้อย่างเหมาะสม เมื่อข้อมูลมีค่าผิดปกติ

2. เพื่อเป็นแนวทางในการศึกษาและเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อข้อมูลมีค่าผิดปกติ ในสถานการณ์อื่น ๆ อีกต่อไป