บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบัน สถิติเป็นความรู้แขนงหนึ่งที่เจริญเติบโตและเป็นที่ยอมรับในแทบทุกสาขา ดังจะเห็นได้จากการนำเอาวิธีทางสถิติไปใช้ประโยชน์ในด้านต่างๆ โดยเฉพาะอย่างยิ่งทางด้านการศึกษาและงานวิจัย เพื่อให้ได้มาซึ่งข้อมูลที่มีคุณภาพ ดังนั้นนอกจากผู้วิจัยจะต้องมีความรู้ ความเข้าใจในเรื่องที่ศึกษาเป็นอย่างดีแล้ว ยังต้องมีความรู้ทางด้านสถิติพอสมควร นั่นคือ จะต้องทราบเกี่ยวกับข้อตกลงเบื้องต้นและข้อจำกัดต่างๆ อีกทั้งจะต้องพิจารณาว่าวิธีการที่นำมาใช้นั้นมีความเหมาะสมที่จะนำไปประยุกต์ใช้กับข้อมูลมากน้อยเพียงไร เช่น ในการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) ซึ่งเป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Dependent Variable : Y) 1 ตัวกับตัวแปรอิสระ (Independent Variable : X) 1 ตัว และนำความสัมพันธ์ที่ได้ไปใช้ประโยชน์ในการพยากรณ์ ดังแสดงในรูปแบบความสัมพันธ์ดังนี้

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , i = 1,2,...,n$$

เมื่อ Y_i คือ ตัวแปรตามของค่าสังเกตที่ i

X_i คือ ตัวแปรต้นหรือตัวแปรอิสระของค่าสังเกตที่ i

β_0 คือ ส่วนตัดแกน Y หรือค่าของ Y เมื่อ X มีค่าเป็นศูนย์ ซึ่งเป็นพารามิเตอร์ที่ไม่ทราบค่า

β_1 คือ ความชัน (Slope) ของเส้นตรง ซึ่งเป็นค่าที่แสดงถึงอัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย และเป็นพารามิเตอร์ที่ไม่ทราบค่า

ε_i คือ ความคลาดเคลื่อน (Error Term) ของค่าสังเกตที่ i นั่นคือ เป็นค่าที่แสดงให้เห็นความแตกต่างระหว่างค่าจริงของ Y และค่า Y บนเส้นถดถอย (Predicted Value : $\hat{Y}$)

n คือ ขนาดตัวอย่างที่ได้รับการสุ่มมาจากประชากร

ซึ่งในที่นี้ X เป็นตัวที่ทราบค่า หรือเป็นค่าคงที่ (known or fixed values) ขณะที่ Y เป็นตัวแปรสุ่ม (Random Variable) อย่างไรก็ตามในการวิเคราะห์ความถดถอยนี้จะต้องมีข้อกำหนดเบื้องต้นหรือข้อสมมติของความคลาดเคลื่อน (The Assumption of Error Term : ε_i) ดังนี้

1. ε_i เป็นตัวแปรสุ่มที่มีการแจกแจงปกติ (Normality)
2. ε_i มีค่าเฉลี่ยเท่ากับศูนย์ นั่นคือ $E(\varepsilon_i) = 0$

3. ε_i และ ε_j เป็นอิสระกันหรือไม่มีความสัมพันธ์กัน (Independence) ทุกค่า i และ j เมื่อ $i \neq j$ นั่นคือ $\mathbf{cov}(\varepsilon_i, \varepsilon_j) = 0$ โดยที่ **cov** คือ covariance
4. ความคลาดเคลื่อนมีความแปรปรวนคงที่ (Homoscedasticity) หรือมีค่าเท่ากันทุกระดับของ X และมีค่าเป็น σ^2 นั่นคือ $V(\varepsilon_i) = \sigma^2$

โดยปกติแล้วการประมาณค่าและการทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์นั้น ผู้วิเคราะห์มักจะเลือกใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method : OLS) เนื่องจากด้วยข้อสมมติเหล่านี้ วิธี OLS จะให้ตัวประมาณค่าที่ดีที่เรียกว่า Best Linear Unbiased Estimator (BLUE) และให้สถิติทดสอบ (Test Statistics) ที่มีอำนาจการทดสอบ (Power of the test) สูง

ความคลาดเคลื่อนมีความแปรปรวนคงที่ (Homoscedasticity) เป็นข้อตกลงเบื้องต้นประการหนึ่งที่มีความสำคัญมากในการวิเคราะห์การถดถอย ดังที่ได้กล่าวมาแล้วในข้างต้น อย่างไรก็ตามในทางปฏิบัติเรามักจะพบว่า ข้อมูลที่ถูกนำมาใช้วิเคราะห์นั้นมีจำนวนไม่น้อยที่ไม่เป็นไปตามข้อตกลงเบื้องต้นดังกล่าว โดยในบางครั้งความคลาดเคลื่อนอาจจะมีการผันแปรค่าไปตามค่าของตัวแปรอิสระ ตัวแปรตาม ขนาดตัวอย่าง หรือแม้แต่ผันแปรไปตามกาลเวลา เราเรียกสถานการณ์เช่นนี้ว่า ความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ (heteroscedasticity) ซึ่งมักจะเกิดขึ้นบ่อยๆ เมื่อชุดตัวอย่างมีข้อมูลในช่วงกว้างมาก เช่น รายได้ของผู้บริโภคหรือราคาของสินค้า เป็นต้น และมักเกิดกับข้อมูลภาคตัดขวาง (cross section) เนื่องจากข้อมูลประเภทนี้จะมีความแตกต่างกันตามขนาดหรือลำดับ นอกจากนี้เราสามารถแบ่งประเภทของแหล่งความแปรปรวนของความคลาดเคลื่อนไม่คงที่ ได้ดังนี้

1. Pure heteroscedasticity เป็นส่วนที่เกิดขึ้นเองโดยธรรมชาติด้วยลักษณะการแจกแจงของตัวแปรคลาดเคลื่อนเอง
2. Impure heteroscedasticity เป็นส่วนที่เกิดจากความผิดพลาดในการออกแบบจำลอง (error in specification) เช่น การละเลยตัวแปรสำคัญไปจากแบบจำลองหรือการสร้างรูปแบบสมการผิดพลาด

และบ่อยครั้งที่ผู้วิเคราะห์มักจะละเลย และถ้ายังฝืนใช้ข้อมูลโดยไม่คำนึงถึงข้อกำหนดเบื้องต้นในข้อนี้ ผลที่ตามมาก็คือ การประมาณค่าพารามิเตอร์ β ด้วยวิธี OLS จะให้ค่าประมาณที่มีลักษณะไม่เอนเอียง นั่นคือ $E(\hat{\beta}) = \beta$ แต่ความแปรปรวนของค่าประมาณดังกล่าว ($Var(\hat{\beta})$) อาจจะเล็กหรือใหญ่เกินไปกว่าความเป็นจริง ทำให้ตัวสถิติที่ใช้ในการทดสอบไม่น่าเชื่อถือ และค่าพยากรณ์ของ Y ก็จะไม่มีประสิทธิภาพ อันจะส่งผลให้ผลสรุปที่ได้จากการวิเคราะห์ผิดไปจากความเป็นจริง ดังนั้น การใช้วิธี OLS ในการประมาณค่า β ในกรณีที่ข้อมูลยังมีความแปรปรวนของความคลาดเคลื่อนไม่คงที่จึงไม่เหมาะสม จึงควรหาเทคนิคในการประมาณค่าพารามิเตอร์โดยวิธีอื่นๆ เพื่อให้สอดคล้องกับข้อตกลงเบื้องต้นดังกล่าว เช่น วิธี GLS (Generalized Least Squares) หรือ หากจะใช้วิธีกำลังสองน้อยที่สุดก็จำเป็นต้องแปลงข้อมูล เพื่อให้ความแปรปรวนอยู่ในลักษณะ

ที่ใช้วิธีการของ OLS ได้ เช่น วิธี WLS (Weighted Least Squares) ดังนั้นสิ่งที่สำคัญอย่างยิ่งของผู้วิเคราะห์ที่จะละเลยไม่ได้ คือ การตรวจสอบดูว่าความคลาดเคลื่อนมีความแปรปรวนคงที่หรือไม่

สำหรับการทดสอบว่า ความแปรปรวนของความคลาดเคลื่อนคงที่หรือไม่นั้นในปัจจุบันมีอยู่หลายวิธี เช่น Szroeter's test, Goldfild-Quandt test, Breusch-Pagan test, Glejser test, Bartlett test, White's test, และ Park test เป็นต้น ซึ่งแต่ละวิธีมีกระบวนการและสถิติทดสอบที่แตกต่างกันไป และผู้วิจัยจำเป็นจะต้องเลือกสถิติทดสอบให้เหมาะสมด้วย เนื่องจากการเลือกใช้สถิติทดสอบที่เหมาะสม จะทำให้ผลสรุปของการวิจัยเป็นไปอย่างถูกต้อง ซึ่งในการพิจารณาความเหมาะสมของสถิติทดสอบนั้น สิ่งที่ควรพิจารณา ได้แก่ อำนาจของการทดสอบ (Power of the test) และความแกร่ง (Robustness) ดังคำกล่าวของ Box และ Anderson ที่ว่า เกณฑ์ที่สำคัญในการเลือกใช้สถิติทดสอบ คือ สถิตินั้นจะต้องมีความไว (Sensitivity) ต่อการเปลี่ยนแปลงของปัจจัยที่ต้องการทดสอบ และจะต้องไม่มีความไว (Insensitivity) ต่อการเปลี่ยนแปลงของสิ่งอื่นไม่ใช่ปัจจัยที่ต้องการทดสอบ ซึ่งก็คือ อำนาจการทดสอบ และความแกร่ง ตามลำดับ และจากคำกล่าวของ Neyman ซึ่งกล่าวว่า เมื่อต้องการที่จะเลือกใช้สถิติทดสอบ จะต้องเริ่มพิจารณาถึงความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ก่อน แล้วจึงพิจารณาถึงความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 2 โดยมีขั้นตอนดังนี้ คือ ให้ความน่าจะเป็นที่จะยอมรับให้เกิดความคลาดเคลื่อนประเภทที่ 1 ไม่เกินระดับนัยสำคัญที่กำหนดไว้ (α) และเมื่อเป็นไปตามเงื่อนไขดังกล่าวแล้ว ทำการเลือกสถิติทดสอบที่ให้อำนาจการทดสอบสูงที่สุด นั่นคือ เลือกสถิติทดสอบที่มีโอกาสน้อยที่สุดที่จะยอมรับสมมติฐาน H_0 เมื่อสมมติฐาน H_0 นั้นไม่เป็นจริง อย่างไรก็ตามการทดสอบ heteroscedasticity อาศัยการสังเกตและการทดสอบสมมติฐานร่วมกันไป โดยวิธีที่ง่ายที่สุด คือ การเขียนกราฟการกระจายระหว่างเศษตกค้าง (residual) จากตัวแบบการถดถอยกับตัวแปรอิสระ หรือ กับค่าประมาณที่ได้ (Predicted value) ซึ่งผลที่ได้นั้นขึ้นอยู่กับผู้วิเคราะห์ ส่วนวิธีการทดสอบที่นิยมใช้และพบบ่อยที่สุด คือ White's test , Breusch-Pagan's test และ Szroeter's test ซึ่งเสนอโดย White (1980) , Breusch-Pagan (1979) และ Szroeter (1978) ตามลำดับ ซึ่งในการวิจัยครั้งนี้ได้เสนอวิธีการใหม่เพื่อปรับปรุงการทดสอบดังกล่าว ซึ่งได้แก่ วิธีบูทสแตรป

วิธีบูทสแตรป (Bootstrap Method) เป็นวิธีทางนอนพาราเมตริก (Nonparametric) โดยเป็นวิธีการสุ่มตัวอย่างซ้ำกันหลายๆ ครั้งแบบใส่คืน (Resampling with Replacement) ซึ่งได้รับการเสนอโดยแบรดเลย์ เอฟรอน (Bradley Efron : 1979) วิธีบูทสแตรปนี้เป็นวิธีที่มักจะนำมาแก้ปัญหาการที่ไม่สามารถหาค่าประมาณได้ ในกรณีที่ข้อมูลไม่เป็นไปตามข้อตกลงเบื้องต้น(Assumption) แต่ในที่นี้จะใช้วิธีบูทสแตรปช่วยในการปรับปรุงการตรวจสอบปัญหาความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ แนวคิดที่สำคัญของวิธีบูทสแตรป คือ ตัวอย่างเป็นสิ่งที่เราทราบทั้งหมดเกี่ยวกับประชากร และตัวอย่างแต่ละตัวอย่างจะสามารถอธิบายลักษณะของประชากรด้วยความน่าจะเป็นที่เท่าๆ กัน ซึ่งแนวคิดนี้อาจจะทำให้ได้ข้อสรุปที่ดีเกี่ยวกับลักษณะของประชากร โดยที่อาจจะดี

มากกว่าการกำหนดข้อตกลงเบื้องต้นเกี่ยวกับประชากร ซึ่งวิธีบูทสแตรปมีหลักเกณฑ์ดังนี้ คือ เราจะให้ตัวอย่างที่ถูกเก็บรวบรวมมาจากประชากรเปรียบเสมือนประชากร แล้วทำการสุ่มตัวอย่างจากตัวอย่างที่มีอยู่แบบใส่คืน (Resampling with Replacement) ด้วยจำนวนครั้ง (Bootstrap Replications) ที่มากเพียงพอ เพื่อสร้างการแจกแจงของตัวสถิติตัวอย่าง (Sampling Distribution) นำไปใช้ประมาณค่าพารามิเตอร์ที่สนใจและแก้ไขปัญหาที่ต้องการศึกษา โดยใช้ประโยชน์จากคอมพิวเตอร์ในการคำนวณ และในการวิจัยครั้งนี้ได้ทำการศึกษาโดยอาศัยวิธีการจำลองแบบที่เรียกว่า วิธีมอนติคาร์โล (Monte Carlo Methods)

1.2 วัตถุประสงค์ของการวิจัย

สำหรับงานวิจัยครั้งนี้ เป็นการศึกษาเปรียบเทียบอำนาจการทดสอบของตัวสถิติที่ใช้ในการตรวจสอบว่าความแปรปรวนของความคลาดเคลื่อนเป็นค่าคงที่หรือไม่ ซึ่งตัวสถิติที่จะนำมาเปรียบเทียบ ได้แก่

1.2.1 การทดสอบบูทสแตรปของไวท์ (Bootstrapped White's Test)

1.2.2 การทดสอบบูทสแตรปของบรูชและพาแกน (Bootstrapped Breusch-Pagan's Test)

1.2.3 การทดสอบบูทสแตรปของสโรเตอร์ (Bootstrapped Szroeter's Test)

1.3 สมมติฐานของการวิจัย

เมื่อใช้วิธีบูทสแตรปในการปรับปรุงการทดสอบปัญหาความคลาดเคลื่อนมีความแปรปรวนไม่คงที่แล้ว

(1) ตัวสถิติบูทสแตรปของสโรเตอร์จะมีอำนาจการทดสอบสูงที่สุด รองลงมา คือ ตัวสถิติบูทสแตรปของบรูชและพาแกน และตัวสถิติบูทสแตรปของไวท์ ตามลำดับ สำหรับความรุนแรงของปัญหาความคลาดเคลื่อนอยู่ในระดับต่ำ

(2) ในกรณีที่ความรุนแรงของปัญหาความคลาดเคลื่อนอยู่ในระดับที่สูงขึ้น ตัวสถิติบูทสแตรปของบรูชและพาแกนจะมีอำนาจการทดสอบใกล้เคียงกับตัวสถิติบูทสแตรปของสโรเตอร์

1.4 ข้อตกลงเบื้องต้น

ในงานวิจัยชิ้นนี้ทำการศึกษาตัวแบบการถดถอยเชิงเส้น (Linear Regression Model) ซึ่งมีจำนวนตัวแปรอิสระเท่ากับ 1 และ 2 นั่นคือ

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i \quad , i = 1,2,\ldots,n$$

เมื่อ Y_i คือ ตัวแปรตามของค่าสังเกตที่ i

p คือ จำนวนตัวแปรอิสระ

X_{ji} คือ ตัวแปรต้นหรือตัวแปรอิสระตัวที่ j ของค่าสังเกตที่ i

β_0 คือ ส่วนตัดแกน Y หรือค่าของ Y เมื่อ X มีค่าเป็นศูนย์ ซึ่งเป็นพารามิเตอร์ที่ไม่ทราบค่า

β_1 คือ ความชัน (Slope) ของเส้นตรง ซึ่งเป็นค่าที่แสดงถึงอัตราการเปลี่ยนแปลงของ Y เมื่อ X เปลี่ยนไป 1 หน่วย และเป็นพารามิเตอร์ที่ไม่ทราบค่า

n คือ ขนาดตัวอย่างที่ได้รับการสุ่มมาจากประชากร

ε_i คือ ความคลาดเคลื่อนอิสระ (Independent Error Term) ของค่าสังเกตที่ i ที่มีค่าเฉลี่ยเป็น 0 และกำหนดความแปรปรวนของ ε_i ซึ่งมีค่าขึ้นอยู่กับตัวแปรอิสระ 1 ตัว ดังนี้

1. รูปแบบการคูณ (Multiplicative model)

$$\sigma_i^2 = V(\varepsilon_i) = E(\varepsilon_i^2) = KX_i^r$$

2. รูปแบบการบวก (Additive model)

$$\sigma_i^2 = V(\varepsilon_i) = E(\varepsilon_i^2) = K^2(1 + \lambda X_i)^2$$

เมื่อ K เป็นค่าคงที่ใดๆ r และ λ เป็นพารามิเตอร์ที่ไม่ทราบค่า

1.5 ขอบเขตของการวิจัย

สำหรับงานวิจัยชิ้นนี้ จะทำการศึกษาเปรียบเทียบสถิติทดสอบ 3 ชนิด นั่นคือ ระหว่างตัวสถิติทดสอบบูทสแตรปของไวท์ ตัวสถิติทดสอบบูทสแตรปของบรูชและพาแกน และตัวสถิติทดสอบบูทสแตรปของสโรเตอร์ ภายใต้สถานการณ์ที่ว่า ความคลาดเคลื่อนมีความแปรปรวน 2 รูปแบบข้างต้น โดยจะศึกษาอำนาจในการทดสอบและความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ของสถิติทดสอบดังกล่าว ณ ความรุนแรงของปัญหาที่ระดับต่างๆ เมื่อจำนวนตัวแปรอิสระเท่ากับ 1 และ 2 และมีขนาดตัวอย่างเท่ากับ 20 , 50 และ 100 ที่ระดับนัยสำคัญ 0.01 และ 0.05 โดยใช้เทคนิคมอนติคาร์โล (Monte Carlo Method) ช่วยในการคำนวณ

1.6 คำจำกัดความ

1.6.1 ความคลาดเคลื่อนประเภทที่ 1 (Type I error)

เป็นความผิดพลาดที่เกิดจากการปฏิเสธสมมติฐาน H_0 เมื่อสมมติฐาน H_0 เป็นจริง

1.6.2 ความคลาดเคลื่อนประเภทที่ 2 (Type II error)

เป็นความผิดพลาดที่เกิดจากการยอมรับสมมติฐาน H_0 เมื่อสมมติฐาน H_0 ไม่เป็นจริง

1.6.3 อำนาจการทดสอบ (Power of the test)

หมายถึง ความน่าจะเป็นที่จะปฏิเสธสมมติฐาน H_0 เมื่อสมมติฐาน H_0 ไม่เป็นจริง

1.6.4 ปัญหาความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ (Heteroscedasticity)

หมายถึง ปัญหาที่เกิดจากความแปรปรวนของความคลาดเคลื่อนมีค่าแตกต่างกัน ซึ่งระดับความรุนแรงของปัญหาขึ้นอยู่กับระดับความแตกต่างของความแปรปรวน นั่นคือ ถ้าความแปรปรวนมีค่าแตกต่างกันมาก ความรุนแรงของปัญหาความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ก็จะอยู่ในระดับสูง

1.7 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อปรับปรุงตัวสถิติทดสอบของไวท์ ตัวสถิติของบรูชและพาแกน และตัวสถิติของสโรเตอร์ในการตรวจสอบปัญหาความคลาดเคลื่อนมีความแปรปรวนไม่คงที่ ให้มีอำนาจการทดสอบที่สูงขึ้น โดยวิธีการที่เรียกว่า บูทสแตรป อีกทั้งเป็นประโยชน์ในการเลือกใช้ตัวสถิติที่เหมาะสม