บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ข้อมูลข่าวสารในปัจจุบันมีความสำคัญและมีการนำไปใช้ประโยชน์ในการดำเนินงานหลายด้าน โดยเฉพาะด้านธุรกิจ อุตสาหกรรม การแพทย์ เศรษฐกิจ การศึกษา การเกษตรกรรม และการเมือง เป็นต้น ข้อมูลที่ได้มักนำมาใช้เป็นเครื่องมือช่วยในการตัดสินใจ ประกอบการวางแผน หรือแก้ปัญหาเกี่ยวกับกิจกรรมต่างๆ ที่อาจเกิดขึ้นในอนาคต ยกตัวอย่างเช่น กรุงเทพมหานครใช้ข้อมูลอุบัติเหตุหรือน้ำท่วมในอดีตมาพยากรณ์หรือวางแผนการแก้ปัญหา นักธุรกิจใช้ข้อมูลยอดขายสินค้ามาช่วยพยากรณ์หรือทำนายเหตุการณ์ล่วงหน้าว่าการดำเนินธุรกิจนั้นจะมีแนวโน้มเป็นอย่างไร ซึ่งจะช่วยนักธุรกิจตัดสินใจได้ว่าจะลงทุนในธุรกิจที่กระทำอยู่นั้นต่อหรือไม่

การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติที่ใช้ในการพยากรณ์ค่าของตัวแปรตาม (Dependent variable) ซึ่งเป็นตัวแปรที่สนใจศึกษา โดยใช้ตัวแปรอื่นที่เกี่ยวข้องที่มีอิทธิพลต่อตัวแปรตามซึ่งเรียกว่า ตัวแปรอิสระ(Independent variable) ตัวแปรอิสระอาจมีเพียงตัวแปรเดียว การวิเคราะห์ลักษณะนี้ เรียกว่า การวิเคราะห์การถดถอยเชิงเดียว (Simple regression) แต่ในบางกรณีตัวแปรอิสระอาจมีตั้งแต่สองตัวขึ้นไป การวิเคราะห์การถดถอยลักษณะเช่นนี้ เรียกว่า การวิเคราะห์การถดถอยเชิงพหุ (Multiple regression) ลักษณะความสัมพันธ์ของตัวแปรตามและตัวแปรอิสระอาจมีความสัมพันธ์ได้ทั้งในรูปแบบเชิงเส้นและไม่เชิงเส้น รูปแบบที่ไม่เชิงเส้นมีหลายรูปแบบ เช่น รูปแบบเลขชี้กำลัง(Exponential) รูปแบบไฮเพอร์โบลา(Hyperbola) เป็นต้น

การวิเคราะห์การถดถอยใช้มากทั้งในสถานการณ์ที่ไม่สามารถควบคุมค่าของตัวแปรอิสระและสถานการณ์ที่สามารถควบคุมค่าของตัวแปรอิสระได้ เช่น การศึกษาความสัมพันธ์ระหว่างผลผลิตข้าว กับปริมาณแสงอาทิตย์ ปริมาณฝน และปริมาณปุ๋ย โดยผลผลิตข้าวเป็นตัวแปรตาม ปริมาณแสงอาทิตย์ ปริมาณฝน เป็นตัวแปรอิสระที่ควบคุมค่าของตัวแปรไม่ได้ ปริมาณปุ๋ยเป็นตัวแปรอิสระที่ควบคุมค่าของตัวแปรได้

การสร้างตัวแบบถดถอยนั้นจะต้องทำภายใต้ข้อตกลงเบื้องต้นเกี่ยวกับความคลาดเคลื่อนสุ่ม 2 ข้อ ดังนี้

(1) ความคลาดเคลื่อนสุ่ม (ε_i) เป็นตัวแปรสุ่มที่มีการแจกแจงปกติ ค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ σ^2 นั่นคือ

$$E(\varepsilon_i) = 0 \quad , \quad V(\varepsilon_i) = \sigma^2$$

(2) ความคลาดเคลื่อนสุ่มไม่มีความสัมพันธ์กัน (Uncorrelated) นั่นคือ

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ เมื่อ } i \neq j \text{ ; } i, j = 1, 2, ..., n$$

ดังนั้นควรตรวจสอบว่าความคลาดเคลื่อนสุ่ม (ε_i) มีคุณสมบัติตามที่กำหนดหรือไม่ แต่เนื่องจากความคลาดเคลื่อนสุ่ม (ε_i) เป็นตัวแปรที่ไม่ทราบค่า จึงตรวจสอบคุณสมบัติของเรซิดวล (e_i) แทน นอกจากนี้จะต้องมีการตรวจสอบความเหมาะสมของตัวแบบ เพื่อให้ได้ตัวแบบสำหรับพยากรณ์ค่าตัวแปรตามให้มีประสิทธิภาพมากที่สุด

การตรวจสอบความเหมาะสมของตัวแบบโดยทั่วไปใช้วิธีของเดรปเปอร์และสมิทธ์ (Draper and Smith, 1981) ซึ่งจะทำได้ก็ต่อเมื่อข้อมูลมีการทำซ้ำเท่านั้น แต่ในความเป็นจริงข้อมูลที่มีการทำซ้ำมักได้จากการออกแบบการทดลอง ซึ่งถ้าเก็บข้อมูลตามสภาพที่เป็นจริงอาจจะไม่สามารถเก็บรวบรวมข้อมูลที่มีการทำซ้ำได้ จึงได้มีผู้ศึกษาหาวิธีการตรวจสอบความเหมาะสมของตัวแบบเมื่อข้อมูลไม่มีการทำซ้ำ โดยมีหลายแนวทางด้วยกัน อาทิเช่น กรีน (Green,1971) ได้สร้างแบบทดสอบสำหรับการตรวจสอบความเหมาะสมของตัวแบบที่ใช้ตัวแปรอิสระเพียง 1 ตัว แล้วจัดข้อมูลที่มีค่า X ใกล้เคียงกันให้อยู่กลุ่มเดียวกัน M กลุ่ม เพื่อใช้ประมาณความแปรปรวนของความคลาดเคลื่อนในแต่ละกลุ่มที่แบ่งนั้น แล้วเพิ่มพหุนามกำลัง q ของค่า X เข้าไปในตัวแบบของกลุ่ม และสมมติให้ตัวแบบที่มีพหุนามของค่า X เพิ่มขึ้นนี้เป็นตัวแบบที่ถูกต้อง วิธีของกรีนจะต้องมีการกำหนด M และ q ซึ่งทำได้ยาก และถ้ามีตัวแปรอิสระมากกว่า 1 ตัววิธีของกรีนต้องใช้ตัวอย่างขนาดใหญ่ ชิลลิงตัน (Shillington,1979) ได้เสนอแบบทดสอบความเหมาะสมของตัวแบบที่ได้จากการสร้างสมการเส้นถดถอย 2 เส้น แบบทดสอบดังกล่าวมีความไวต่อการจัดกลุ่ม นอกจากนี้ภายใต้ H_1 ค่าของ MSE_w มีค่ามาก อัทส์ (Utts,1982) ได้เสนอวิธีเรนโบว์โดยมีการสร้างแบบทดสอบความเหมาะสมของตัวแบบจากการเปรียบเทียบตัวแบบ 2 ตัวแบบ ตัวแบบแรกเป็นตัวแบบที่ได้จากการใช้ข้อมูลเฉพาะจุดที่มีอิทธิพลต่ำ ตัวแบบที่สองเป็นตัวแบบที่ได้จากการใช้ข้อมูลทั้งชุด นีลล์และจอห์นสัน (Neill and Johnson,1985) สร้างแบบทดสอบความเหมาะสมของตัวแบบด้วยการหาตัวประมาณค่าความแปรปรวนที่มีคุณสมบัติคงเส้นคงวา

โจลีกา (Joglekar,1989) ได้นำแบบทดสอบสำหรับการตรวจสอบความเหมาะสมของตัวแบบของชิลลิงตันมาปรับปรุงโดยการหาค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนภายในกลุ่มใหม่ $\left(\mathrm{MSE}_{W_{JG}}\right)$ ทำให้ได้แบบทดสอบที่มีความไวต่อการจัดกลุ่มน้อยลง ออกชอน (Ochshorn,1986) ได้พัฒนาวิธีการตรวจสอบความเหมาะสมของตัวแบบโดยนำวิธีการขยายตัวแบบและการแบ่งข้อมูลเป็นกลุ่มมาใช้ร่วมกัน วิธีที่ออกชอนพัฒนาขึ้นสามารถใช้กับฟังก์ชันที่ไม่ต่อเนื่องได้และสามารถใช้ได้กับตัวอย่างที่มีขนาดเล็ก คริสเตนเซน (Christensen,1991) ได้เสนอแบบทดสอบที่สามารถตรวจสอบความไม่เหมาะสมของตัวแบบที่เกิดภายในกลุ่ม (Within – Cluster lack of fit) และความไม่เหมาะสมของตัวแบบที่เกิดระหว่างกลุ่ม (Between – Cluster lack of fit) แล้วนำแบบทดสอบที่เสนอใหม่นี้เปรียบเทียบกับแบบทดสอบอื่น 5 ชนิด โดยใช้ค่ากำลังการทดสอบ (Power of the test) เป็นเกณฑ์ในการเปรียบเทียบ พบว่าไม่สามารถระบุได้ว่าแบบทดสอบใดดีที่สุด ทั้งนี้ขึ้นอยู่กับการเลือกวิธีการแบ่งกลุ่มของข้อมูล ซูและยาง (Su and Yang, 2006) ได้เสนอแบบทดสอบสำหรับการตรวจสอบความเหมาะสมของตัวแบบโดยได้มีการเพิ่มเทอมกำลังสองในการหาค่าสถิติทดสอบ มิลเลอร์และนีลล์ (Miller and Neill, 2008) ได้ศึกษาการตรวจสอบความเหมาะสมของตัวแบบโดยแบ่งข้อมูลเป็นกลุ่มให้แต่ละกลุ่มมีค่าสังเกตเท่ากับ 2, 3, 4, 5 แล้วหาค่าสถิติทดสอบ หากมีการปฏิเสธสมมติฐานว่างอย่างน้อย 1 ครั้ง จากการแบ่งข้อมูลเป็นกลุ่มโดยให้แต่ละกลุ่มมีค่าสังเกตเท่ากับ 2, 3, 4, 5 นั่นคือ ตัวแบบถดถอยเชิงเส้นตรงไม่เหมาะสมที่จะนำมาพยากรณ์ตัวแปรตาม จากการศึกษางานวิจัยเกี่ยวกับการตรวจสอบความเหมาะสมของตัวแบบกรณีไม่มีข้อมูลซ้ำ พบว่า ยังไม่มีผู้ใดวิจัยเปรียบเทียบแบบทดสอบที่เสนอโดย เดรปเปอร์และสมิทธ์ ซูและยาง มิลเลอร์และนีลล์ ดังนั้นผู้วิจัยจึงสนใจศึกษาเปรียบเทียบวิธีการตรวจสอบความเหมาะสมของตัวแบบ กรณีข้อมูลไม่มีการทำซ้ำของ 3 แบบทดสอบ ได้แก่ วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง วิธีของมิลเลอร์และนีลล์

วัตถุประสงค์ของการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์สำคัญ ดังนี้

เปรียบเทียบวิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง วิธีของมิลเลอร์และนีลล์ เพื่อการตรวจสอบความเหมาะสมของตัวแบบ กรณีข้อมูลไม่มีการทำซ้ำ

สมมติฐานการวิจัย

การวิจัยครั้งนี้มีสมมติฐานของการวิจัย ดังนี้

1. วิธีการทดสอบทั้ง 3 วิธี คือ วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง วิธีของมิลเลอร์และนีลล์ สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้

2. ในสถานการณ์ทดลองที่กำหนดให้ความคลาดเคลื่อนสุ่ม (ε) มีการแจกแจงปกติมาตรฐาน วิธีของเดรปเปอร์และสมิทธ์จะมีค่ากำลังการทดสอบสูงสุด

3. วิธีการทดสอบทั้ง 3 วิธี คือ วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง วิธีของมิลเลอร์และนีลล์ มีค่ากำลังการทดสอบเพิ่มขึ้นเมื่อสัมประสิทธิ์การถดถอยเพิ่มขึ้น

4. วิธีของเดรปเปอร์และสมิทธ์ กับ วิธีของซูและยาง ค่ากำลังการทดสอบจะลดลงเมื่อจำนวนกลุ่มเพิ่มขึ้น

ขอบเขตของการวิจัย

การวิจัยนี้เป็นการวิจัยเพื่อเปรียบเทียบวิธีการทดสอบความเหมาะสมของตัวแบบโดยมีข้อจำกัด ดังนี้

1. ขนาดตัวอย่างที่ศึกษา มีขนาด n = 15, 50 และ100

2. กำหนดให้ความคลาดเคลื่อนสุ่มมีการแจกแจงปกติมาตรฐาน นั่นคือค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ 1

3. ค่าของตัวแปรอิสระ X จะใช้ค่าสุ่มจากการแจกแจงยูนิฟอร์ม (1, 10)

4. สถิติทดสอบที่ใช้ในการเปรียบเทียบมี 3 วิธี ได้แก่ วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง วิธีของมิลเลอร์และนีลล์

5. วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง พิจารณาวิธีการแบ่งกลุ่ม 2 วิธี คือ วิธีแบ่งกลุ่มของข้อมูลโดยใช้ฟังก์ชัน cutree ซึ่งเป็นวิธีแบ่งกลุ่มของข้อมูลโดยข้อมูลที่มีค่าใกล้เคียงกันอยู่ในกลุ่มเดียวกัน ในขณะที่ต่างกลุ่มกันมีค่าแตกต่างกัน และวิธีแบ่งกลุ่มของข้อมูลให้ในแต่ละกลุ่มมีขนาดตัวอย่างเท่าๆ กัน ส่วนวิธีของมิลเลอร์และนีลล์ที่ n=15 จะกำหนด c= (7, 5, 4, 3) n=50 จะกำหนด c= (25, 17, 13, 10) n=100 จะกำหนด c = (50, 33, 25, 20)

6. วิธีของเดรปเปอร์และสมิทธ์ วิธีของซูและยาง กรณีแบ่งกลุ่มของข้อมูลโดยใช้ฟังก์ชัน cutree และแบ่งกลุ่มของข้อมูลให้ในแต่ละกลุ่มมีขนาดตัวอย่างเท่าๆ กัน มีจำนวนกลุ่มดังนี้

ตารางที่ 1.1

จำนวนกลุ่มกรณีแบ่งกลุ่มของข้อมูลโดยใช้ฟังก์ชัน cutree และกรณีแบ่งกลุ่มของข้อมูลให้ในแต่ละกลุ่มมีขนาดตัวอย่างเท่าๆ กัน

ขนาดตัวอย่าง	วิธีของเดรปเปอร์และสมิทธ์	วิธีของซูและยาง
15	c = 3, 4 ,5, 7	c = 3, 4, 5, 7
50	c = 3, 4, 5, 7, 10, 13, 17, 25	c = 3, 4, 5, 7, 10, 13, 17, 25
100	c = 3, 4, 5, 7, 10, 13, 17, 20, 25, 33, 50	c = 3, 4, 5, 7, 10, 13, 17, 20, 25, 33, 50

7. สมการถดถอยตามสมมติฐานว่าง จะเป็นสมการถดถอยที่มีความสัมพันธ์ระหว่าง X กับ Y เป็นตัวแบบเชิงเส้นตรง $Y=\beta_0+\beta_1X+\varepsilon$ โดยกำหนดค่า $\beta_0=2$ $\beta_1=0.0001, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 1.5$ และ2 โดยβ_1ในตัวแบบเชิงเส้นตรงจะมีค่าเท่ากับβ_2 ในตัวแบบพหุนามลำดับที่ 2 β_3ในตัวแบบพหุนามลำดับที่ 3 และβ_1ในตัวแบบตรีโกณมิติแบบที่ 1, 2 และตัวแบบเลขชี้กำลัง เพื่อพิจารณาความสามารถในการควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1

8. สมการถดถอยตามสมมติฐานแย้ง จะเป็นสมการถดถอยที่มีความสัมพันธ์ระหว่าง X กับ Y ในรูปแบบไม่เชิงเส้นโดยมีความชันและความโค้งของเส้นถดถอยต่างๆ กันไป ดังนี้

ตารางที่ 1.2

ตัวแบบที่ใช้ในการศึกษา

ตัวแบบที่แท้จริง	ค่าพารามิเตอร์
1.ตัวแบบพหุนามลำดับที่ 2 $Y=\beta_0+\beta_1X+\beta_2X^2+\varepsilon$	$\beta_0=\beta_1=$ 2 $\beta_2=$ 0.0001 0.001 0.01 0.05 0.1 0.5 1
2.ตัวแบบพหุนามลำดับที่ 3 $Y=\beta_0+\beta_1X+\beta_2X^2+\beta_3X^3+\varepsilon$	$\beta_0=\beta_1=$ 2 $\beta_2=$ 0.2 $\beta_3=$ 0.0001 0.001 0.01 0.05 0.1 0.5 1
3.ตัวแบบตรีโกณมิติแบบที่ 1 $Y=\beta_0+\beta_1\sin X_1+\varepsilon$	$\beta_0=$ 2 $\beta_1=$ 0.01 0.1 0.5 1 1.5 2
4.ตัวแบบตรีโกณมิติแบบที่ 2 $Y=\beta_0+\{\beta_1\cos(X_1)+\beta_1\sin(X_1)\}+\varepsilon$	$\beta_0=$ 2 $\beta_1=$ 0.01 0.1 0.5 1 1.5 2
5.ตัวแบบเลขชี้กำลัง $Y=\exp(\beta_0+\beta_1X)+\varepsilon$	$\beta_0=$ 2 $\beta_1=$ 0.01 0.1 0.5 1 1.5

โดยตัวแบบแต่ละตัวแบบจะกำหนดให้ β_0 และ β_1 ในตัวแบบพหุนามลำดับที่ 2 3 มีค่าคงที่ค่าเดียว นั่นคือมีค่าเท่ากับ 2 เพื่อศึกษาอิทธิพลของสัมประสิทธิ์การถดถอยของเทอมที่ทำให้เกิดเส้นโค้ง สำหรับ β_2 ในตัวแบบพหุนามลำดับที่ 3 จะกำหนดให้เท่ากับ 0.2 เพื่อให้กำลังการทดสอบมีค่าไม่เท่ากับ 1 ในทุกค่าของ β_3 ทุกจำนวนกลุ่ม และทุกขนาดตัวอย่าง ทำให้สามารถศึกษาอิทธิพลของ β_3 และจำนวนได้อย่างชัดเจนยิ่งขึ้น

9. ในการจำลองข้อมูลให้มีสถานการณ์ตามที่กำหนดจะทำการจำลองข้อมูลซ้ำๆ กัน 2,000 ครั้ง ในแต่ละสถานการณ์

10. กำหนดระดับนัยสำคัญของการทดสอบ เท่ากับ 0.05

ประโยชน์ที่คาดว่าจะได้รับ

การวิจัยครั้งนี้คาดว่าจะก่อให้เกิดประโยชน์ ดังนี้

1. เพื่อเป็นแนวทางให้ผู้สนใจสามารถเลือกวิธีการตรวจสอบความเหมาะสมของตัวแบบการถดถอยได้เหมาะสมในแต่ละสถานการณ์

2. สามารถนำไปใช้ทดสอบความเหมาะสมของตัวแบบสมการถดถอยกรณีข้อมูลไม่มีการทำซ้ำ

นิยามศัพท์

1. ความคลาดเคลื่อนประเภทที่ 1 (Type I error) หมายถึง ความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมติฐานว่าง (H_0) เมื่อสมมติฐานว่างเป็นจริง

2. ค่ากำลังการทดสอบ (Power of the test) หมายถึง ความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่าง (H_0) เมื่อสมมติฐานว่างเป็นเท็จ ในงานวิจัยนี้วัดโดยใช้จำนวนครั้งที่ปฏิเสธสมมติฐานว่าง เมื่อสมมติฐานว่างเป็นเท็จหารด้วยจำนวนการจำลองข้อมูลซ้ำ

3. ข้อมูลมีการทำซ้ำ หมายถึง ข้อมูลที่ค่าของ X ค่าหนึ่ง มีค่า Y หลายค่า