บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการวิเคราะห์ทางสถิติ สิ่งที่สำคัญประการหนึ่ง คือ ลักษณะของข้อมูลที่มีความถูกต้อง แม่นยำ ผลการวิเคราะห์ย่อมมีประสิทธิภาพและเชื่อถือได้ ในทางปฏิบัติการเก็บรวบรวมข้อมูลเพื่อใช้ในการวิเคราะห์ บางครั้งข้อมูลที่ได้มาไม่เป็นไปตามสภาวะการณ์ที่ศึกษาหรือควบคุมอยู่ เช่น การศึกษาทางด้านชีววิทยา ทางด้านการแพทย์ ลักษณะเช่นนี้ทำให้ข้อมูลบางค่าแตกต่างไปจากข้อมูลอื่นมาก บางค่ามีค่าสูงมาก บางค่ามีค่าต่ำมาก (outlier) ความแตกต่างที่เกิดขึ้นมีสาเหตุสำคัญสามประการ (Ascomebe F.J.:1960) ประการแรกเกิดจากความผันแปรที่มีอยู่แล้วในประชากรที่ศึกษา (Inherent variability) เป็นความผันแปรที่ไม่สามารถจะหลีกเลี่ยงได้ แม้จะมีการควบคุมการวัด การปฏิบัติการอื่น ๆ อย่างดี ความคลาดเคลื่อนนี้ยังคงอยู่แก้ไขไม่ได้ นอกจากจะเปลี่ยนประชากรหรือวัตถุประสงค์ในการศึกษา ประการที่สอง ความคลาดเคลื่อนที่เกิดจากการวัด (measurement error) เป็นความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูล หรือเครื่องมือเครื่องใช้ในการวัด มีคุณภาพต่ำ ความคลาดเคลื่อนนี้อาจแก้ไขหรือตัดทิ้งได้ ประการสุดท้ายความคลาดเคลื่อนที่เกิดจากการปฏิบัติการ (execution error) เช่น การลงรหัส การเจาะบัตร เป็นต้น

การศึกษาสมการการถดถอยเชิงเส้นพหุ

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (1.1)$$

โดยที่ $\underset{\sim}{Y}$ เป็นเวคเตอร์ของตัวแปรตามที่มีขนาด nxl

X เป็นเมตริกซ์ของตัวแปรอิสระคงที่ที่มีขนาด nxp และมี rank = p

$\underset{\sim}{\beta}$ เป็นเวคเตอร์สัมประสิทธิ์การถดถอย เป็นพารามิเตอร์ ที่ไม่ทราบค่า แทนความชันของเส้นถดถอย

$\underset{\sim}{\varepsilon}$ เป็นความคลาดเคลื่อนที่มีขนาด nxl

ในการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด (Least square) เป็นวิธีที่ให้ตัวประมาณ (estimator) ที่มีคุณสมบัติเป็น BLUE (Best Linear Unbiased Estimator) เมื่อค่าสังเกต Y_i มีค่าผิดปกติ* ปนอยู่ จะเนื่องมาจากสาเหตุใดก็ตาม มีผลทำให้ลักษณะการแจกแจงของความคลาดเคลื่อน (error) ไม่เป็นไปตามข้อตกลงเบื้องต้น คือ ไม่เป็นการแจกแจงแบบปกติ เช่น อาจเป็นการแจกแจงแบบปกติปลอมปน การแจกแจงที่มีหางยาว (long tails) หรือมีการแจกแจงไปทางหางมาก (heavy tails) (Hawkin D.M.:1980) ซึ่งมีผลต่อสัมประสิทธิ์การถดถอย สมการการถดถอยจะเบี่ยงเบนไปในทิศทาง หรือตำแหน่งของค่าผิดปกติ การพิจารณาตัดค่าสังเกตที่ผิดปกติควรระมัดระวัง เพราะอาจทำให้ขาดสารสนเทศ (information) บางอย่าง ถ้าค่าผิดปกตินั้นเป็นกุญแจสำคัญในการวิเคราะห์ข้อมูลนั้น ดังนั้นในการศึกษาวิธีการตรวจสอบค่าผิดปกติ ทำให้เข้าใจและรู้โครงสร้างของข้อมูลได้ดียิ่งขึ้น

การศึกษา วิธีตรวจสอบค่าผิดปกติ ได้มีผู้ศึกษาไว้หลายท่าน เช่น การศึกษาของแอนโคมบ์ เอฟ เจ (Ancombe F.J. 1960:123-147) มิกกี้ ดันน์และคลาร์ก (Mickey Dunn and Clark 1967 1967:105-111) ได้ใช้วิธีการ stepwise regression method และการบวกตัวแปรหุ่น (dummy variable) เข้าไปเพื่อทำการแยกค่าผิดปกติ แต่วิธีการนี้ไม่เหมาะสม ในกรณีที่ค่าผิดปกติมีมากกว่าหนึ่งค่า เจนเทิลแมน และวิลค์ (Gentleman and Wilk, 1975:387-410) ได้เสนอวิธีการแยกค่าผิดปกติ กรณีที่มีมากกว่าหนึ่ง โดยการพิจารณาเซตย่อยของค่าสังเกต $\binom{n}{k}$ ว่ากลุ่มใดมีผลบวกกำลังสองของค่าความคลาดเคลื่อน (Q_k) มากที่สุด แสดงว่าค่าสังเกตกลุ่มนั้นเป็นค่าผิดปกติ ซึ่งต้องเสียค่าใช้จ่ายในการคำนวณสูง เบอร์นาร์ด โรสเนอร์ (Bernard Rosner, 1975:221-227 ได้ศึกษาเปรียบเทียบวิธีการตรวจสอบค่าผิดปกติ 4 วิธี คือ Extreme Studentized Deviate (ESD) Studentized Range (STR) Kurtosis (KUR) R-Statistic (RST) ปรากฏว่าวิธี ESD เป็นวิธีตรวจสอบที่ดีที่สุด

---

*ค่าผิดปกติ คือ ค่าสังเกตที่มีค่ามากหรือน้อยกว่าค่าสังเกตอื่น ๆ อย่างผิดปกติ

อาร์เดนนิล์ คุ๊ก (R. DENNIS COOK) ได้เล่นอวิธีการตรวจล่อบค่าผิดปกติในล่มการถดถอยเชิงเล้น ในปี 1977 มีแนวความคิดว่าค่าผิดปกติย่อมมีอิทธิพลต่อสัมประสิทธิ์การถดถอยโดยตรง จึงดูการเปลี่ยนแปลงของสัมประสิทธิ์การถดถอย เมื่อมีค่าผิดปกติกับไม่นำค่าผิดปกติมาพิจารณา

แอนดูร์และเพรตจิบอน (Andrew and Pregibon) ได้มีแนวความคิดว่าค่าผิดปกติมีอิทธิพลต่อค่าความคลาดเคลื่อน และดีเทอร์มินัลของผลคูณของเมตริกซ์ตัวแปรตาม จีแบร์รี เว็ทเทอร์ล (G.Barrie Wetheril) ในปี 1986 ได้เล่นอวิธีการตรวจล่อบค่าผิดปกติ โดยพิจารณาจากค่าความคลาดเคลื่อนมาตรฐานเป็นตัวทดล่อบค่าผิดปกติ จากแนวความคิดทั้งล่ามนี้น่าล่นใจที่จะศึกษาต่อไปว่า วิธีการทดล่อบทั้งล่าม วิธีใดจะมีอำนาจการทดล่อบค่าผิดปกติที่สูง

ในการศึกษาครั้งนี้ ผู้วิจัยจะทำการศึกษาอำนาจของการทดล่อบ และความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของทั้ง 3 วิธี โดยวิธีการตรวจล่อบแบบซีเควนเชียล (Sequential) ตรวจล่อบค่าผิดปกติทีละหนึ่ง แล้วตัดค่าผิดปกติออก เพื่อตรวจล่อบค่าผิดปกติต่อไป ขนาดตัวอย่างจะลดลงทีละหนึ่ง โดยศึกษาภายใต้แบบจำลองที่ล่ร้างขึ้น จากเทคนิคมอนติคาร์โลซิมูเลชั่น (Monte Carlo Simulation Technique) ซึ่งเป็นเทคนิคที่จะทำให้ได้ผลล่รุปจากล่ภาพการที่เป็นการทดลอง ภายใต้ขนาดตัวอย่าง ลักษณะการแจกแจงความคลาดเคลื่อน ค่าเฉลี่ย และความแปรปรวนตามที่กำหนดได้

1.2 วัตถุประล่งค์ของการวิจัย

ศึกษาเปรียบเทียบล่ถิติที่ใช้ตรวจล่อบค่าลังเกตที่มีอิทธิพลและค่าผิดปกติ ในล่มการการถดถอยเชิงเล้นพหุ ของการทดล่อบ 3 วิธี คือ

1.2.1 การทดล่อบโดยวิธี จีแบร์รี

1.2.2 การทดล่อบวิธีคุ๊ก

1.2.3 การทดล่อบวิธีแอนดูร์และเพรตจิบอน

โดยศึกษาอำนาจการทดล่อบ และความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1

1.3 สมมติฐานของการวิจัย

วิธีการตรวจสอบของคุ๊กเป็นวิธีที่มีอำนาจการทดสอบสูงสุด

1.4 ข้อตกลงเบื้องต้น

1.4.1 ศึกษาตัวแปรอิสระ X เป็นค่าคงที่และเป็นอิสระกันมี rank = p

1.4.2 ค่าความคลาดเคลื่อน (R_i) มีการแจกแจงแบบเดียวกัน (ยกเว้นค่าผิดปกติ) และอิสระกัน

1.4.3 ค่าสังเกตที่มีอิทธิพล (Influential observation) หมายถึง ค่าสังเกตที่มีผลกระทบต่อสาระสำคัญในการวิเคราะห์ข้อมูล เช่น มีผลต่อสัมประสิทธิ์การถดถอย (β) สมการการถดถอย ($\hat{y}$) และความคลาดเคลื่อน (Residual:R_i)

1.4.4 ในการประมาณค่าสัมประสิทธิ์การถดถอย ใช้วิธีกำลังสองน้อยที่สุด (Least square method)

1.5 ขอบเขตของการวิจัย

1.5.1 ศึกษาอำนาจการทดสอบ และความคลาดเคลื่อนประเภทที่ 1 ในการตรวจสอบค่าสังเกตที่มีอิทธิพลและค่าผิดปกติ ในสมการการถดถอยเชิงเส้นพหุ วิธี จีแบร์รี วิธีคุ๊ก วิธีแฮนดูร์และเพรตจิบอน เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติปลอมปน

1.5.2 ในการสร้างค่าผิดปกติ สร้างจากการแจกแจงแบบปกติ ที่มีค่าเฉลี่ย เท่ากับ 0 และความแปรปรวน เท่ากับ 1 โดยการสร้างลักษณะการแจกแจงแบบปกติปลอมปน 2 ลักษณะคือ สเกลคอนทามิเนต (Scale-contaminated normal distribution) สร้างโดยการเปลี่ยนค่าความแปรปรวนเป็น 9 16 25 ตามลำดับ โลเคชั่นคอนทามิเนต (Location-contaminated normal distribution) สร้างโดยการเปลี่ยนค่าเฉลี่ยของความคลาดเคลื่อน จาก 0 เป็น 4 6 15 ตามลำดับ การสร้างค่าผิดปกติจะกำหนดตำแหน่งของค่าผิดปกติ เพื่อหาอำนาจการทดสอบและความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1

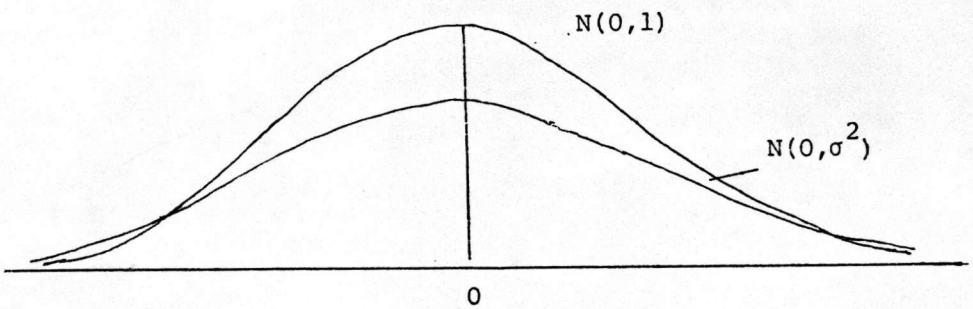
1.5.3 การศึกษาทุกรูปแบบ ศึกษาโดยการกำหนดค่า X เป็นค่าคงที่

1.5.4 ศึกษาในกรณีที่จำนวนตัวแปรอิสระ เป็น 2 4 6 8 10 ขนาดตัวอย่าง 20 30 50 70 จำนวนค่าผิดปกติเป็น 1 2

1.5.5 กำหนดระดับนัยสำคัญ (α) 0.01 0.05 0.10

1.5.6 ในการศึกษาการวิจัยครั้งนี้ จำลองการทดลองขึ้นโดยใช้เทคนิคมอนติคาร์-โลซิมูเลชั่น จากเครื่องคอมพิวเตอร์ IBM 370/3031 ซึ่งจะศึกษาเมื่อค่าความคลาดเคลื่อนเป็นแบบปกติปลอมปน 2 ลักษณะ คือ สเกลคอนทามิเนต โลเคชั่นคอนทามิเนต รูปแบบการแจกแจงเป็นดังนี้

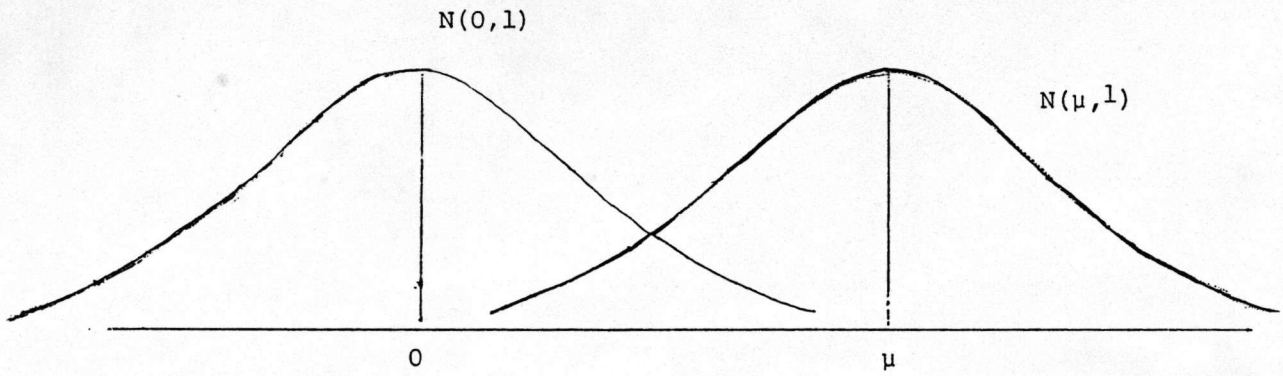
1.5.6.1 สเกลคอนทามิเนต (Scale - contaminated normal disribution)



ลักษณะการแจกแจงแบบปกติปลอมปนที่พิจารณาในวิทยานิพนธ์นี้เป็นการแจกแจงที่แปลงมาจากการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเป็น 1 ซึ่งมีฟังก์ชั่นการเปลี่ยนแปลงดังนี้

$$F = (N - N_1)\, N(0,1) + N_1\, N(0, \sigma^2)$$

1.5.6.2 โลเคชั่นคอนทามิเนต (Location-contaminated normal distribution)



ฟังก์ชั่นการเปลี่ยนแปลงเป็นดังนี้

$$F = (N-N_1)N(0,1) + N_1\ N(\mu,1)$$

เมื่อ N เป็นจำนวนขนาดตัวอย่าง

N_1 เป็นจำนวนค่าผิดปกติ

1.5.7 การจำลองการทดลอง จะกระทำซ้ำกัน 100 ครั้ง ในแต่ละการทดลอง

1.6 คำจำกัดความ

1.6.1 ความคลาดเคลื่อนประเภทที่ 1 (Type I error) เป็นความผิดพลาดที่เกิดจากการปฏิเสธสมมติฐาน H_o เมื่อสมมติฐาน H_o ถูก

1.6.2 ความคลาดเคลื่อนประเภทที่ 2 (Type II error) เป็นความผิดพลาดที่เกิดจากการยอมรับสมมติฐาน H_o เมื่อสมมติฐาน H_o ผิด

1.6.3 อำนาจของการทดสอบ (Power of the test) คือ ความน่าจะเป็นที่จะเป็นปฏิเสธสมมติฐาน H_o เมื่อสมมติฐาน H_o ผิด

1.7 ประโยชน์ของการวิจัย

1.7.1 ทำให้ผู้วิจัยเข้าใจโครงสร้างของปัญหา และรู้ข้อบกพร่องของข้อมูล เพื่อเป็นแนวทางในการแก้ปัญหา และหาทางป้องกันมิให้เกิดความผิดพลาดในการทดลองและเก็บรวบรวมข้อมูล

1.7.2 ทำให้ผู้ใช้สามารถเลือกวิธีทดสอบที่เหมาะสม ในกรณีที่มีค่าผิดปกติในสมการการถดถอยเชิงเส้นพหุ