บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการวิเคราะห์ข้อมูลทางสถิติผู้วิจัยอาจจะประสบกับปัญหาซึ่งข้อมูล หรือค่าสังเกตมีค่าบางค่าสูงหรือต่ำมาก หรือเป็นค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกับค่าสังเกตส่วนใหญ่ โดยที่ในกรณีหลังนี้จะพบมากในข้อมูลที่ได้จากการวางแผนการทดลองผิดพลาด ข้อมูลผิดปกติดังกล่าวเรียกว่า"outlier" ซึ่งแอนคอมบ์(Anscombe, 1960) กล่าวว่าสาเหตุของการเกิดข้อมูลผิดปกติมีสาเหตุ 3 ประการคือ

ประการแรก เกิดจากความผันแปรที่มีอยู่ในประชากรที่ศึกษา(inherent variability) ซึ่งเป็นความผันแปรที่ไม่สามารถจะหลีกเลี่ยงได้แม้จะมีการควบคุมการวัดหรือการปฏิบัติการเป็นอย่างดี ความคลาดเคลื่อนนี้ยังคงอยู่แก้ไขไม่ได้ นอกจากจะเปลี่ยนประชากรหรือวัตถุประสงค์ในการศึกษา

ประการที่สอง ความคลาดเคลื่อนที่เกิดจากการวัด(meansurement error)เป็นความคลาดเคลื่อนที่เกิดจากการบันทึกข้อมูล หรือเครื่องมือเครื่องใช้ในการวัดมีคุณภาพต่ำความคลาดเคลื่อนนี้อาจแก้ไข้ให้หมดไปได้

ประการที่สาม ความคลาดเคลื่อนที่เกิดจากการปฏิบัติการ(execution error) อาทิเช่น การลงรหัส การเจาะบัตร เป็นต้น

ในทางปฏิบัติจะพบลักษณะข้อมูลผิดปกติดังกล่าวในการศึกษาทางด้านการแพทย์ทางด้านชีววิทยา ทางด้านผลผลิตเกษตรกรรม และอุตสาหกรรม

เบคแมน และคุ๊ก(Beckman and Cook, 1983) ได้สรุปความหมายของค่าผิดปกติออกเป็น 2 ลักษณะดังนี้

ลักษณะแรก ค่าสังเกตมีค่าสูงหรือต่ำมาก(extreme) หรือเป็นค่าที่เบี่ยงเบน(deviation)ไปจากกลุ่มค่าสังเกตที่ศึกษา ค่าผิดปกติลักษณะนี้เรียกว่า "discordant observation"

ลักษณะที่สอง ค่าสังเกตที่มีลักษณะการแจกแจงแตกต่างจากการแจกแจงของประชากรที่สนใจศึกษา ค่าผิดปกติลักษณะนี้เรียกว่า"contaminate observation"

ดังนั้นข้อมูลผิดปกติจะมีความหมายตามลักษณะทั้งสองดังที่กล่าวมาแล้วข้างต้น

วิธีวิเคราะห์ความถดถอยเชิงเส้นเป็นวิธีวิเคราะห์ข้อมูลเชิงสถิติ ซึ่งได้ถูกนำไปใช้ในงานวิจัยทางด้านสังคมศาสตร์ และวิทยาศาสตร์ ซึ่งบางครั้งผู้วิจัยจะประสบกับปัญหาข้อมูลผิดปกติ ซึ่งจะทำการวิเคราะห์ข้อมูลที่ได้มาโดยที่มีข้อมูลผิดปกติอยู่บางค่าก็จะทำให้ผลการวิเคราะห์ความถดถอยเกิดความผิดพลาดมากขึ้น ทั้งนี้เพราะว่าข้อมูลผิดปกติมีผลต่อการประมาณค่าสัมประสิทธิ์การถดถอย(β)ซึ่งทำให้สมการการถดถอยที่ได้เบี่ยงเบนไปในทิศทาง หรือตำแหน่งของค่าผิดปกติ ดังนั้นการศึกษาวิธีตรวจสอบค่าผิดปกติในการวิเคราะห์ความถดถอยจะช่วยให้ผู้วิจัยสามารถแก้ปัญหาดังกล่าวได้

สำหรับวิธีตรวจสอบค่าผิดปกติได้มีผู้ศึกษาดังนี้ มิกกี้, ดันน์ และคลาชก์ (Mickey, Dunn and Clark, 1967)ได้ใช้วิธีการวิเคราะห์ความถดถอยแบบขั้นตอน (stepwise reqression) และการเพิ่มตัวแปรหุ่น(dummy variable)เข้าไปในสมการความถดถอยเพื่อทำการแยกค่าผิดปกติ แต่วิธีนี้ไม่เหมาะสมในกรณีที่มีค่าผิดปกติมากกว่าหนึ่งค่า ต่อมาทิตเจน, มัวร์ และเบคแมน(Tietjen, Moore and Beckman , 1973)ได้เสนอวิธีตรวจสอบค่าผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย พร้อมกับเสนอตัวสถิติทดสอบ R_n ซึ่งต่อมาเพรสค็อต(Prescott, 1975) ได้นำตัวสถิติทดสอบ R_n มาใช้ตรวจสอบค่าผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นพหุ (multiple linear regression) พร้อมกับแสดงว่าตัวสถิติทดสอบ R_n มีการแจกแจงแบบเอฟ เมอวิน จี มาราซิงห์(Mervyn G. Marasinghe, 1985) ได้เสนอวิธีการวิเคราะห์แบบหลายขั้นตอน(multistage procedure) และตัวสถิติ F_{k*} เพื่อใช้ตรวจสอบค่าผิดปกติหลายค่าในการวิเคราะห์ความถดถอยเชิงเส้น บุญสม(2532) ได้ศึกษาวิธีตรวจสอบค่าผิดปกติในสมการการถดถอยเชิงเส้นพหุ โดยศึกษาเปรียบเทียบวิธีของเดนนิสคุ๊ก(Dennis Cook, 1977) วิธีของแอนดูร์ และเพรตจิบอน(Andrew and Pregibon, 1978) และวิธีของจีแบร์รี(G. Barrie Wetherill, 1986)ซึ่งศึกษาในกรณีที่การแจกแจงของความผิดพลาดมี 2 แบบ คือสเกลคอนทามิเนตนอร์มอล และโลเคชันคอนทามิเนตนอร์มอล ผลการศึกษาปรากฏว่าวิธีของจีแบร์รีมีความสามารถในการควบคุมความคลาดเคลื่อนประเภทที่ 1 ได้ดี

ผู้วิจัยจึงสนใจที่จะเปรียบเทียบวิธีตรวจสอบข้อมูลผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้น โดยใช้วิธีของทิตเจน, มัวร์ และเบคแมน (Tietjen, Moore and Beckman, 1973(TMB)) วิธีของเมอวิน จี มาราซิงห์(Mervyn .G. Marasinghe ,1985(M)) และวิธีของจีแบร์รี(G. Barrie Wetherill, 1986(GB))

1.2 วัตถุประสงค์ของการวิจัย

ต้องการศึกษาเปรียบเทียบตัวสถิติที่ใช้ตรวจสอบข้อมูลผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย เมื่อความผิดพลาดมีการแจกแจงแบบหางยาวกว่าการแจกแจงปกติ และมีการแจกแจงแบบเบ้ ของวิธีการตรวจสอบ 3 วิธีการ คือ

1 วิธีการของทิตเจน, มัวร์ และเบคแมน (TMB)

2 วิธีการของเมอวิน จี มาราซิงห์ (M)

3 วิธีการของจีแบร์รี (GB)**

โดยพิจารณาจากความสามารถในการควบคุมความน่าจะเป็นของความคลาดเคลื่อนแบบที่ 1 และอำนาจการทดสอบ เพื่อหาวิธีการที่เหมาะสมในการตรวจสอบข้อมูลผิดปกติ

1.3 สมมติฐานของการวิจัย

วิธีการของเมอวิน จี มาราซิงห์ (M) เป็นวิธีที่มีอำนาจการทดสอบสูงสุดในกรณีที่มีค่าผิดปกติหลายค่าเนื่องจากวิธีดังกล่าวเป็นวิธีที่เหมาะสมกับข้อมูลที่มีค่าผิดปกติหลายค่า

1.4 ข้อตกลงเบื้องต้น

1 ค่าความผิดพลาด(ε_i) มีการแจกแจงแบบเดียวกัน(ยกเว้นกรณีค่าผิดปกติค่าความผิดพลาดอาจจะมีการปลอมปนของการแจกแจงอื่นได้) และเป็นอิสระต่อกัน

2 กรณีที่มีข้อมูลผิดปกติมากกว่าหนึ่งค่า จะถือว่าไม่เกิดเหตุการณ์ที่ค่าผิดปกติค่าหนึ่งมีผลต่อค่าผิดปกติอีกค่าหนึ่ง หรือมีผลต่อค่าสังเกตค่าอื่น ๆ ทำให้การทดสอบให้ผลผิดพลาด (masking effect)

3 การสร้างค่าผิดปกติ จะกำหนดตำแหน่งของค่าผิดปกติเพื่อหาค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ

4 การประมาณสัมประสิทธิ์การถดถอย(β) ใช้วิธีกำลังสองน้อยที่สุด(least square method)

** วิธีการของจีแบร์รี(GB)ในงานวิจัยครั้งนี้หมายถึงวิธีการของจีแบร์รี เวทเธอร์ริว (G . Barrie Wetherill, 1986)

1.5 ขอบเขตของการวิจัย

1 งานวิจัยครั้งนี้จะศึกษากรณีการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย

2 ขนาดตัวอย่างที่ศึกษา n = 20, 50 และ 100

3 จำนวนค่าผิดปกติที่ศึกษา k=0, 1, 2 และ 3 ทุกขนาดตัวอย่างที่ศึกษา

4 กำหนดระดับนัยสำคัญของการทดสอบ α = 0.05 และ 0.01

5 การแจกแจงของความผิดพลาดที่สนใจศึกษา มีดังนี้

5.1 การแจกแจงแบบหางยาวกว่าการแจกแจงปกติ (long tailed distribution)จะใช้การแจกแจงดังต่อไปนี้

5.1.1 การแจกแจงแบบสเกลคอนทามิเนตนอร์มอล จะศึกษาในกรณีที่ค่าสเกลแฟคเตอร์(c)= 3, 5 และ 10 และเปอร์เซนต์การปลอมปน(p)= 5%, 10% และ 25% ตามลำดับ

5.1.2 การแจกแจงแบบโลเคชันคอนทามิเนตนอร์มอล จะศึกษาในกรณีที่ค่าโลเคชันแฟคเตอร์(a) = 3, 5 และ 15 และเปอร์เซนต์การปลอมปน(p) = 5%, 10% และ 25% ตามลำดับ

5.1.3 การแจกแจงแบบที จะศึกษาเฉพาะกรณีที่ขนาดตัวอย่าง (n)= 20 ณ ระดับความเป็นอิสระ(d.f)= 18 ทั้งนี้เพราะเมื่อขนาดตัวอย่างสูงขึ้นการแจกแจงแบบทีจะเข้าใกล้การแจกแจงปกติและทำให้ค่าเฉลี่ยของความผิดพลาดกำลังสอง(mean square error) ต่ำกว่าความเป็นจริง

5.2 การแจกแจงแบบเบ้(skewed distribution) จะใช้การแจกแจงดังต่อไปนี้

5.2.1 การแจกแจงแบบลอกนอร์มอล จะศึกษาในกรณีที่ μ = 0, σ^2 = 0.1, 0.3, 0.5 และ 0.7 ตามลำดับ

5.2.2 การแจกแจงแบบแกมมา จะศึกษาในกรณีที่ α = 1, 2, 3 และ 10 เมื่อ β = 1

5.2.3 การแจกแจงแบบไวบูลล์ จะศึกษาในกรณีที่ α = 1, 2, 3 และ 10 เมื่อ β = 1

6 ในงานวิจัยครั้งนี้จะทำการสร้างแบบจำลองข้อมูลให้มีสถานการณ์ตามที่ต้องการ โดยใช้วิธีมอนติคาร์โลซิมูเลชันด้วยเครื่อง IBM 370/3031 ณ สถาบันบริการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย โดยเขียนโปรแกรมด้วยภาษา FORTRAN และทำการทดลองซ้ำ 500 ครั้งในแต่ละสถานการณ์

1.6 คำจำกัดความ

1 outlier หมายถึงค่าสังเกตที่มีค่ามากหรือน้อยกว่าค่าสังเกตอื่น ๆ หรือค่าสังเกตที่ไม่ได้มาจากประชากรเดียวกันกับค่าสังเกตอื่น ๆ

2 masking effect หมายถึง เหตุการณ์ที่ค่าผิดปกติค่าหนึ่ง มีผลต่อค่าผิดปกติอีกค่าหนึ่ง หรือมีผลต่อค่าสังเกตค่าอื่น ๆ ทำให้การทดสอบค่าผิดปกติให้ผลคลาดเคลื่อน

3 ความคลาดเคลื่อนประเภทที่ 1 (α) หมายถึงความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมติฐานว่าง(H_0) เมื่อสมมติฐานว่างจริง

4 ความคลาดเคลื่อนประเภทที่ 2 (β) หมายถึงความคลาดเคลื่อนที่เกิดจากการยอมรับสมมติฐานว่าง(H_0) เมื่อสมมติฐานว่างนั้นไม่จริง

5 อำนาจการทดสอบ ($1-\beta$) หมายถึงความน่าจะเป็นที่จะปฏิเสธสมมติฐานว่าง(H_0) เมื่อสมมติฐานว่างไม่จริง

6 ความแกร่งของการทดสอบ (robustness) หมายถึงคุณสมบัติของการทดสอบที่ไม่ไวต่อการเปลี่ยนแปลงของปัจจัยอื่นที่ไม่ใช่ปัจจัยที่ต้องการทดสอบ เช่น การฝ่าฝืนข้อตกลงเบื้องต้นของการทดสอบนั้น ซึ่งสิ่งที่ใช้พิจารณาความแกร่งของการทดสอบคือ ค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 (probability of type I error)

1.7 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางให้ผู้สนใจสามารถเลือกวิธีการตรวจสอบข้อมูลผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นอย่างง่าย เมื่อการแจกแจงของความผิดพลาดเป็นแบบเบ้และแบบหางยาวกว่าการแจกแจงปกติ