

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ข้อมูล (data) มีความสำคัญมากในยุคปัจจุบัน โดยเฉพาะอย่างยิ่งในการประมวลข้อมูลให้เป็นสารสนเทศ (information) ไม่ว่าจะเป็นหน่วยงาน หรือองค์กรใดๆ ก็มีการเก็บข้อมูล เพื่อนำไปทดลอง วิเคราะห์วางแผนหรือตัดสินใจให้เกิดประโยชน์สูงสุด ซึ่งข้อมูลเหล่านี้จะต้องอาศัยวิธีการทางสถิติมาเป็นเครื่องมือในการวิเคราะห์ผล โดยในขั้นต้น อาจต้องทราบก่อนว่าลักษณะการกระจายของข้อมูลเป็นอย่างไร ข้อมูลมีการแจกแจงแบบไหน ก่อนที่จะนำข้อมูลไปทดสอบสมมติฐานหรือวิเคราะห์ผลต่อไป

ในการวิเคราะห์ข้อมูลทางสถิติ มักมีข้อสมมติเบื้องต้นว่าประชากรมีการแจกแจงแบบปกติ เช่น ในการทดสอบ t หรือการทดสอบ F นอกจากนั้น การวิเคราะห์การถดถอย ก็ยังมีเงื่อนไขที่ว่าความคลาดเคลื่อนมีการแจกแจงแบบปกติ โดยถ้าข้อมูลที่นำมาวิเคราะห์ไม่เป็นไปตามเงื่อนไขที่กำหนดไว้ ก็อาจส่งผลต่อการวิเคราะห์ข้อมูล ทำให้ผลการวิเคราะห์ไม่ถูกต้อง และอาจทำให้เกิดความเสียหายเมื่อนำผลจากการวิเคราะห์ไปใช้ ดังนั้น การตรวจสอบข้อมูลว่ามีลักษณะเหมาะสมและสอดคล้องกับเงื่อนไขเบื้องต้นจึงเป็นเรื่องสำคัญ

การแจกแจงแบบปกติ ถือเป็นการแจกแจงที่สำคัญมาก นอกจากเป็นเงื่อนไขในการวิเคราะห์ทางสถิติส่วนใหญ่แล้ว เหตุการณ์ที่เกิดขึ้นโดยธรรมชาติส่วนใหญ่ก็มีลักษณะใกล้เคียงการแจกแจงแบบปกติ คือ ข้อมูลจะมีรูปร่างลักษณะเป็นรูปประฆังคว่ำ ข้อมูลส่วนมากจะอยู่ตรงกลางรอบๆ ค่าเฉลี่ย และค่อยๆ ลดน้อยลงไปด้วยสมมาตรทั้งทางซ้ายและขวา จึงทำให้การแจกแจงแบบปกติเป็นการแจกแจงที่นิยมใช้กันอย่างแพร่หลาย อีกหนึ่งการแจกแจงแบบต่อเนื่องที่น่าสนใจคือการแจกแจงแบบโลจิสติก เป็นการแจกแจงที่มีการนำไปประยุกต์ในหลายด้าน เช่น ด้านการตลาด ด้านชีววิทยา ด้านเทคโนโลยี ด้านพลังงาน ที่สำคัญคือใช้สำหรับตัวแบบการเจริญเติบโต (Growth Models) การแจกแจงแบบโลจิสติก มีรูปร่างลักษณะเป็นรูปประฆังคว่ำ ที่มีความสมมาตร ของเขตเป็นได้ตั้งแต่ $-\infty$ ถึง ∞ ซึ่งมีความใกล้เคียงกับการแจกแจงแบบปกติมาก โดยต่างกันเพียงส่วนหางที่หนักกว่า และมีความโด่งกว่าเล็กน้อยเท่านั้น

เนื่องจากการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติกมีความใกล้เคียงกันมาก โดยเฉพาะในกรณีที่ข้อมูลถูกเก็บมาเพียงบางส่วน หรือข้อมูลมีจำนวนน้อย ทำให้เราไม่สามารถแยกได้ว่าข้อมูลมีการแจกแจงแบบปกติ หรือมีการแจกแจงแบบโลจิสติก ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาว่า คุณสมบัติโดยทั่วไปของการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติกเหมือนหรือ

แตกต่างกันอย่างไร ถ้านำข้อมูลที่มีการแจกแจงแบบโลจิสติกมาทำการวิเคราะห์ข้อมูลทางสถิติที่มีเงื่อนไขเบื้องต้นว่าประชากรมีการแจกแจงแบบปกติ จะส่งผลกระทบอย่างไร และจะส่งผลในกรณีใดบ้าง และทำการทดสอบว่าตัวสถิติใดที่มีประสิทธิภาพในการคัดกรองข้อมูลที่มีการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติ

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษา และเปรียบเทียบประสิทธิภาพตัวสถิติสำหรับการคัดกรองการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติ
2. เพื่อศึกษาผลกระทบในการใช้วิธีการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก
3. เพื่อสรุป และหาข้อเสนอแนะในการวิเคราะห์ทางสถิติเมื่อข้อมูลมาจากการแจกแจงแบบ โลจิสติก

ขอบเขตของการวิจัย

ในการวิจัยครั้งนี้กระทำภายใต้ขอบเขตดังนี้

1. การแจกแจงที่นำมาใช้ในการศึกษา มี 2 การแจกแจง คือ การแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติก
2. ระดับนัยสำคัญที่ใช้ในการทดสอบสมมติฐาน คือ 0.01, 0.05, 0.10
3. ขนาดตัวอย่างที่ใช้ในศึกษาเท่ากับ 10, 25, 50 และ 100
4. การศึกษาครั้งนี้ใช้โปรแกรม R เวอร์ชัน 2.9.2 โดยการจำลองข้อมูลในแต่ละสถานการณ์จะกระทำซ้ำ 5000 รอบ

ข้อตกลงเบื้องต้น

ในการวิจัยครั้งนี้ แบ่งส่วนที่สนใจศึกษาออกเป็น 3 ส่วน โดยกำหนด ดังนี้

ส่วนที่ (1) การเปรียบเทียบประสิทธิภาพของตัวสถิติที่ใช้สำหรับคัดกรองการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติ ใช้ตัวสถิติทดสอบ ดังนี้

- ตัวสถิติ Kolmogorov-Smirnov
- ตัวสถิติ Shapiro Wilk
- ตัวสถิติ Anderson Darling

- ตัวสถิติ Lilliefors
- ตัวสถิติ Cramer Von Mises
- ตัวสถิติ Chi-square

โดยในส่วนนี้สนใจที่จะศึกษาว่าตัวสถิติทั้ง 6 ที่นำมาศึกษา ตัวสถิติใดสามารถคัดกรองการแจกแจงแบบโลจิสติกออกจากการแจกแจงแบบปกติได้ดีที่สุด โดยพิจารณาความสามารถในการควบคุมความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 และกำลังการทดสอบ

ส่วนที่ (2) การศึกษา และเปรียบเทียบคุณสมบัติทั่วไปของการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติก ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1

- ทดสอบว่าการแจกแจงแบบโลจิสติกยกกำลังสองมีการแจกแจงโคสแควร์หรือไม่
- ทดสอบว่าข้อมูลที่มีการแจกแจงแบบโลจิสติกเมื่อนำค่าเฉลี่ยหารด้วยส่วนเบี่ยงเบนมาตรฐานส่วนรากที่สองของจำนวนข้อมูลมีการแจกแจงแบบทีหรือไม่
- ทดสอบว่าข้อมูลที่มีการแจกแจงแบบโลจิสติกเมื่อนำค่าความแปรปรวนของตัวอย่าง 2 ชุด มาหารกัน มีการแจกแจงแบบเอฟหรือไม่

ในส่วนนี้สนใจศึกษาว่าถ้าแปลงข้อมูลที่มีการแจกแจงแบบโลจิสติกไปเป็นการแจกแจงแบบอื่นๆ จะสามารถคัดกรองจากการแจกแจงแบบปกติได้หรือไม่

ส่วนที่ (3) การศึกษาผลกระทบของการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก

สมการการวิเคราะห์การถดถอยที่ใช้ คือ

- $Y = \beta_0 + \beta_1 X + \varepsilon$

โดย $\beta_0 = 1, \beta_1 = 0.5, 1, 2$

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

โดย $\beta_0 = 1, \beta_1 = 1, 2, \beta_2 = 1$

X_1, X_2 มีการแจกแจงแบบ $U(-100, 100)$, $N(0, 1111.56)$ และ $Exp(0.023)$

กรณี X_1, X_2 ไม่อิสระ กำหนดให้มีความสัมพันธ์กัน 10%, 30% และ 50%

ε มีการแจกแจงแบบ $L(0, \sqrt{3}/\pi)$ ¹

¹ $L(0, \sqrt{3}/\pi)$ คือ การแจกแจงแบบโลจิสติกที่มีค่าเฉลี่ย = 0 ความแปรปรวน = 1

ในส่วนนี้สนใจศึกษาว่า เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก การประมาณค่าสัมประสิทธิ์การถดถอย (ภายใต้เงื่อนไข ϵ มีการแจกแจงแบบปกติ) จะให้ค่าที่ถูกต้องหรือไม่ โดยใช้ความสามารถในการควบคุมความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 ของการทดสอบสมมติฐาน ไม่สามารถหาคำจำกัดการทดสอบได้เนื่องจากสมมติฐานยังเป็นสมมติฐานประกอบ และสนใจศึกษาว่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีการแจกแจงแบบปกติหลายตัวแปรหรือไม่ โดยใช้ค่า p-value

หมายเหตุ เพื่อความสะดวกในการกล่าวถึงครั้งต่อไป ขอกำหนดสัญลักษณ์ดังนี้

$N(\mu, \sigma^2)$ แทน การแจกแจงแบบปกติที่มี location parameter $= \mu$, scale parameter $= \sigma^2$
ซึ่งมีค่าเฉลี่ย μ ความแปรปรวน σ^2

$L(\mu^*, \sigma^*)$ แทน การแจกแจงแบบโลจิสติกที่มี location parameter $= \mu^*$, scale parameter $= \sigma^*$ ซึ่งมีค่าเฉลี่ย μ^* ความแปรปรวน $(\pi\sigma^*)^2/3$

$U(a, b)$ แทน การแจกแจงแบบยูนิฟอร์มที่มี $\min = a$, $\max = b$
ซึ่งมีค่าเฉลี่ย $(a+b)/2$ ความแปรปรวน $(b-a)^2/12$

$Exp(\theta)$ แทน การแจกแจงแบบเอกซ์โปเนนเชียล ที่มี scale parameter $= \theta$
ซึ่งมีค่าเฉลี่ย θ ความแปรปรวน θ^2

เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการพิจารณา คือ ความสามารถในการควบคุมความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 (Type I Error) ด้วยเกณฑ์การพิจารณาของแบรดเลย์ (Bradley 1978 : อ้างถึงใน ศิริรัตน์ วงศ์ประกรณ์กุล 2539 : 35) ตัวสถิติทดสอบจะควบคุมค่าความคลาดเคลื่อนประเภทที่ 1 ได้ ก็ต่อเมื่อความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 จากผลการทดลองต้องอยู่ในช่วง $[0.5\alpha, 1.5\alpha]$ ซึ่งแบ่งออกเป็น 3 กรณี คือ

กรณี 1 ที่ระดับนัยสำคัญ 0.01 ต้องอยู่ในช่วง $[0.005, 0.015]$

กรณี 2 ที่ระดับนัยสำคัญ 0.05 ต้องอยู่ในช่วง $[0.025, 0.075]$

กรณี 2 ที่ระดับนัยสำคัญ 0.10 ต้องอยู่ในช่วง $[0.05, 0.15]$

ในกรณีที่ค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 จากผลการทดลอง อยู่ในขอบเขตที่ระบุสำหรับแต่ละเกณฑ์ที่กำหนดก็จะถือว่า ตัวสถิติทดสอบนั้นมีค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 เท่ากับ α ที่กำหนด และสามารถควบคุมค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 ได้ ถ้าค่าความน่าจะเป็น

ของการเกิดความคลาดเคลื่อนประเภทที่ 1 จากผลการทดลองของตัวสถิติทดสอบใดอยู่นอกขอบเขตที่ระบุ จะถือว่าตัวสถิติทดสอบนั้นไม่สามารถควบคุมค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 ได้

ตัวสถิติทดสอบที่สามารถควบคุมค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 ได้ จะนำมาหาค่ากำลังการทดสอบต่อไป เพื่อเปรียบเทียบว่าตัวสถิติทดสอบใดให้ค่ากำลังการทดสอบสูงสุดถือว่ามีประสิทธิภาพดีที่สุด

คำจำกัดความที่ใช้ในการวิจัย

1. ความคลาดเคลื่อนประเภทที่ 1 (Type I Error) เป็นความคลาดเคลื่อนที่เกิดจากการปฏิเสธสมมติฐานว่าง ทั้งที่สมมติฐานว่างนั้นเป็นจริง และมักจะเรียกความน่าจะเป็นของความคลาดเคลื่อนประเภทนี้ว่า ระดับนัยสำคัญ (Level of Significance) ใช้สัญลักษณ์ α แทนความคลาดเคลื่อนประเภทนี้ โดยที่ $\alpha = P(\text{ปฏิเสธ } H_0 \text{ เมื่อ } H_0 \text{ จริง})$

2. ความคลาดเคลื่อนประเภทที่ 2 (Type II Error) เป็นความคลาดเคลื่อนที่เกิดจากการยอมรับสมมติฐานว่าง ทั้งที่สมมติฐานว่างนั้นไม่เป็นจริง ความน่าจะเป็นที่จะยอมรับสมมติฐานว่างที่ไม่เป็นจริงนี้จะใช้สัญลักษณ์แทนด้วย β โดยที่ $\beta = P(\text{ยอมรับ } H_0 \text{ เมื่อ } H_0 \text{ ไม่เป็นจริง})$

3. กำลังการทดสอบ (Power of test) คือ ความน่าจะเป็นของการปฏิเสธสมมติฐานว่างเมื่อสมมติฐานว่างไม่เป็นจริง ดังนั้น กำลังการทดสอบทางสถิติ ก็คือ ความน่าจะเป็นในการตัดสินใจที่ถูกต้อง ใช้สัญลักษณ์แทนด้วย $1-\beta$

ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับในงานวิจัยครั้งนี้ คือ

1. สามารถเลือกใช้ตัวสถิติสำหรับการคัดกรองการແຈ້ງແຈ້ງแบบโลจิสติกจากการແຈ້ງແຈ້ງแบบปกติในกรณีต่างๆ ได้
2. สามารถหาข้อสรุปได้ว่ากรณีใดที่ต้องระมัดระวังเมื่อความคลาดเคลื่อนมีการແຈ້ງແຈ້ງแบบโลจิสติกในการวิเคราะห์การถดถอยภายใต้เงื่อนไขการແຈ້ງແຈ້ງแบบปกติ
3. เพื่อเป็นแนวทางในการศึกษาวิจัยต่อไป

วิธีดำเนินการวิจัย

1. การทดสอบความสามารถของตัวสถิติที่ใช้คัดกรองข้อมูลที่มีการແຈ້ງແຈ້ງแบบโลจิสติกจากการແຈ້ງແຈ້ງแบบปกติ

- 1.1 ศึกษาค้นคว้าเอกสาร และข้อมูลเกี่ยวกับตัวสถิติที่ใช้คัดกรองข้อมูลที่มีการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติ
- 1.2 สร้างข้อมูลที่มีการแจกแจงและขนาดตัวอย่าง (n_j) ที่กำหนด
- 1.3 คำนวณค่าตัวสถิติทั้ง 6 ตัวสถิติ และทดสอบว่าข้อมูลมีการแจกแจงแบบปกติหรือไม่
- 1.4 ทำซ้ำข้อ 1.2 – 1.3 จำนวน 5,000 รอบ
- 1.5 หาค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 และหาค่ากำลังการทดสอบ
2. การศึกษา และเปรียบเทียบคุณสมบัติทั่วไปของการแจกแจงแบบปกติ และการแจกแจงแบบโลจิสติก ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1
 - 2.1 ศึกษาข้อมูลเกี่ยวกับคุณสมบัติทั่วไปของการแจกแจงแบบปกติมาตรฐาน (ค่าเฉลี่ย 0 และความแปรปรวน 1)
 - 2.2 ทำการทดสอบว่าการแจกแจงแบบโลจิสติกมีคุณสมบัติเช่นเดียวกับการแจกแจงแบบปกติหรือไม่
 - 2.2.1 ทดสอบความสัมพันธ์ของการแจกแจงแบบโลจิสติกกับการแจกแจงแบบไคสแควร์
 - 2.2.1.1 สร้างข้อมูลที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ และขนาดตัวอย่าง (n_j) ที่กำหนด
 - 2.2.1.2 คำนวณผลรวมของข้อมูลยกกำลังสอง $\left(\sum_{i=1}^n x_i^2\right)$ และทดสอบว่าข้อมูลมีการแจกแจงแบบไคสแควร์ที่ $df = n_j$ หรือไม่
 - 2.2.2 ทดสอบความสัมพันธ์ของการแจกแจงแบบโลจิสติกกับการแจกแจงแบบที
 - 2.2.2.1 สร้างข้อมูลที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ และขนาดตัวอย่าง (n_j) ที่กำหนด
 - 2.2.2.2 คำนวณค่า $\frac{\bar{X}}{s/\sqrt{n_j}}$ และทดสอบว่าข้อมูลมีการแจกแจงแบบที ที่ $df = n_j - 1$ หรือไม่
 - 2.2.3 ทดสอบความสัมพันธ์ของการแจกแจงแบบโลจิสติกกับการแจกแจงแบบเอฟ
กรณีข้อมูลเริ่มต้น มี 2 ชุด
 - 2.2.3.1 สร้างข้อมูลที่มีการแจกแจง $N(0,1)$ และ $L(0, \sqrt{3}/\pi)$ อย่างละ 2 ชุด กำหนดเป็น ชุด A, B, C และ D ตามลำดับ ให้มีจำนวนตามขนาดตัวอย่าง (n_j) ที่กำหนด

2.2.3.2 คำนวณค่า $\frac{s_A^2}{s_D^2}$, $\frac{s_C^2}{s_B^2}$, $\frac{s_C^2}{s_D^2}$ และทดสอบว่าข้อมูลมีการแจกแจงแบบเอฟ

ที่ $df1 = n_1 - 1$, $df2 = n_2 - 1$ หรือไม่

2.2.4 ทดสอบความสัมพันธ์ของการแจกแจงแบบโลจิสติกกับการแจกแจงแบบเอฟ
กรณีข้อมูลเริ่มต้น มี 1 ชุด

2.2.4.1 สร้างข้อมูลที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ และขนาดตัวอย่าง (n_i) ที่กำหนด

2.2.4.2 นำข้อมูลที่สร้างขึ้นมาแบ่งเป็น 2 ชุดเท่า ๆ กัน และคำนวณค่า $\frac{s_1^2}{s_2^2}$

2.2.4.3 ทำซ้ำข้อ 2.2.4.2 จำนวน n_i รอบ

2.2.4.4 นำข้อมูล $\frac{s_1^2}{s_2^2}$ ที่เก็บไว้ n_i ค่า มาทดสอบว่ามีการแจกแจงแบบเอฟ ที่

$df1 = n_1 - 1$, $df2 = n_2 - 1$ หรือไม่

2.3 ทำซ้ำข้อ 2.2.1 – 2.2.4 กรณีละ 5,000 รอบ

2.4 หาค่ากำลังการทดสอบ

3. การศึกษาผลกระทบของการประมาณค่าสัมประสิทธิ์การถดถอย เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก

3.1 ศึกษาค้นคว้าเกี่ยวกับวิธีการประมาณค่าสัมประสิทธิ์การถดถอย

3.2 ทดลองประมาณค่าสัมประสิทธิ์การถดถอยในกรณีต่างๆ โดยความคลาดเคลื่อนได้จากการสร้างข้อมูลโดยตรง

3.2.1 กรณีการถดถอยอย่างง่าย

3.2.1.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ และสร้างตัวแปรอิสระ (X) ให้มีการแจกแจงและขนาดตัวอย่าง (n_i) ที่กำหนด

3.2.1.2 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X + \varepsilon$

3.2.1.3 ประมาณค่า β_0 , β_1 ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0 , b_1 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.2.2 กรณีการถดถอยเชิงพหุ ตัวแปรอิสระ (X_1, X_2) ไม่มีความสัมพันธ์กัน

3.2.2.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ และสร้างตัวแปรอิสระ (X_1, X_2) ให้มีการแจกแจงและขนาดตัวอย่าง (n_i) ที่กำหนด

3.2.2.2 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

3.2.2.3 ประมาณค่า $\beta_0, \beta_1, \beta_2$ ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0, b_1, b_2 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.2.3 กรณีการถดถอยเชิงพหุ ตัวแปรอิสระ (X_1, X_2) มีความสัมพันธ์กัน

3.2.3.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$

3.2.3.2 สร้าง Z_1, Z_2, Z_3 ให้มีการแจกแจง และพารามิเตอร์ที่ทำให้ X_1, X_2 สัมพันธ์กัน 10%, 30% และ 50% ขนาดตัวอย่าง (n_j) เมื่อกำหนด $X_1 = Z_1 + Z_2$ และ $X_2 = Z_2 + Z_3$

3.2.3.3 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

3.2.3.4 ประมาณค่า $\beta_0, \beta_1, \beta_2$ ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0, b_1, b_2 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.3 ทดลองประมาณค่าสัมประสิทธิ์การถดถอยในกรณีต่างๆ โดยคัดกรองความคลาดเคลื่อนจากการแจกแจงแบบปกติก่อนนำไปใช้

3.3.1 กรณีการถดถอยอย่างง่าย

3.3.1.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ นำไปคัดกรองโดยใช้ตัวสถิติ Shapiro Wilk และคัดกรองจากความสัมพันธ์กับตัวสถิติเอฟ ก่อนนำไปใช้

3.3.1.2 สร้างตัวแปรอิสระ (X) ให้มีการแจกแจงและขนาดตัวอย่าง (n_j) ที่กำหนด

3.3.1.3 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X + \varepsilon$

3.3.1.4 ประมาณค่า β_0, β_1 ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0, b_1 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.3.2 กรณีการถดถอยเชิงพหุ ตัวแปรอิสระ (X_1, X_2) ไม่มีความสัมพันธ์กัน

3.3.2.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ นำไปคัดกรองโดยใช้ตัวสถิติ Shapiro Wilk และคัดกรองจากความสัมพันธ์กับตัวสถิติเอฟ ก่อนนำไปใช้

3.3.2.2 สร้างตัวแปรอิสระ (X_1, X_2) ให้มีการแจกแจงและขนาดตัวอย่าง (n_j) ที่กำหนด

3.3.2.3 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

3.3.2.4 ประเมินค่า $\beta_0, \beta_1, \beta_2$ ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0, b_1, b_2 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.3.3 กรณีการถดถอยเชิงพหุ ตัวแปรอิสระ (X_1, X_2) มีความสัมพันธ์กัน

3.3.3.1 สร้างความคลาดเคลื่อนที่มีการแจกแจง $L(0, \sqrt{3}/\pi)$ นำไปคัดกรองโดยใช้ ตัวสถิติ Shapiro Wilk และคัดกรองจากความสัมพันธ์กับตัวสถิติเอฟ ก่อนนำไปใช้

3.3.3.2 สร้าง z_1, z_2, z_3 ให้มีการแจกแจง และพารามิเตอร์ที่ทำให้ X_1, X_2 สัมพันธ์กัน 10%, 30% และ 50% ขนาดตัวอย่าง (n) เมื่อกำหนด $X_1 = z_1 + z_2$ และ $X_2 = z_2 + z_3$

3.3.3.3 คำนวณค่าตัวแปรตาม (Y) จากสมการ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

3.3.3.4 ประเมินค่า $\beta_0, \beta_1, \beta_2$ ด้วยวิธี OLS แล้วแทนด้วยสัญลักษณ์ b_0, b_1, b_2 ตามลำดับ ทำการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ เท่ากับค่าสัมประสิทธิ์การถดถอยเริ่มต้นที่กำหนดไว้หรือไม่

3.4 ทำซ้ำข้อ 3.2.1 – 3.2.3 และ 3.3.1 – 3.3.3 จำนวน 5,000 รอบ

3.5 หาค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1

3.6 ทำการทดสอบว่า b_1 ที่เก็บค่าไว้มีการแจกแจงแบบปกติหลายตัวแปรหรือไม่

โดยทดสอบว่าผลรวมเชิงเส้น $C_1' \underline{\beta}$ มีการแจกแจงแบบปกติหรือไม่ เมื่อ

$C_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, C_2 = \begin{bmatrix} 1 \\ 10 \end{bmatrix}$ และ $C_3 = \begin{bmatrix} 10 \\ 1 \end{bmatrix}$ สำหรับกรณีการถดถอยอย่างง่าย

โดยทดสอบว่าผลรวมเชิงเส้น $C_1' \underline{\beta}$ มีการแจกแจงแบบปกติหรือไม่ เมื่อ

$C_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, C_2 = \begin{bmatrix} 0.1 \\ 1 \\ 10 \end{bmatrix}$ และ $C_3 = \begin{bmatrix} 10 \\ 1 \\ 0.1 \end{bmatrix}$ สำหรับกรณีการถดถอยเชิงพหุ ทั้งกรณีตัวแปรอิสระ

(X_1, X_2) ไม่มีความสัมพันธ์กัน และมีความสัมพันธ์กัน

ลำดับขั้นตอนในการเสนอผลการวิจัย

การวิจัยครั้งนี้ นำเสนอผลการวิจัยโดยเสนอค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 ค่ากำลังการทดสอบ และค่า p-value ตามแต่ละกรณี นำเสนอในรูปแบบกราฟ และตาราง โดยการนำเสนอผลการวิจัยแบ่งออกเป็น 3 ส่วน ดังนี้

ส่วนที่ (1) นำเสนอผลการเปรียบเทียบประสิทธิภาพของตัวสถิติทดสอบที่ใช้ในการคัดกรองการแจกแจงแบบโลจิสติกจากการแจกแจงแบบปกติของ 6 ตัวสถิติ คือ ตัวสถิติ Kolmogorov-Smirnov, ตัวสถิติ Shapiro Wilk, ตัวสถิติ Anderson Darling, ตัวสถิติ Lilliefors, ตัวสถิติ Cramer Von Mises และตัวสถิติ Chi-square นำเสนอค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 และค่ากำลังการทดสอบ

ส่วนที่ (2) นำเสนอผลการศึกษาความสัมพันธ์ของการแจกแจงแบบโลจิสติกที่มีค่าเฉลี่ย 0 และความแปรปรวน 1 กับการแจกแจงแบบโคสเคอร์ว การแจกแจงแบบที และการแจกแจงแบบเอฟ เพื่อที่จะทดสอบว่า ตัวสถิติที่ใช้ในการทดสอบการแจกแจงแบบดังกล่าวสามารถที่จะคัดกรองการแจกแจงซึ่งมาจากการแจกแจงแบบโลจิสติกจากการแจกแจงดังกล่าว (X มีการแจกแจงแบบโลจิสติกแล้ว $1. \left(\sum_{i=1}^n x_i^2 \right)$ มีการแจกแจงแบบโคสเคอร์วหรือไม่ $2. \frac{\bar{X}}{s/\sqrt{n_j}}$ มีการแจกแจงแบบทีหรือไม่ $3. \frac{s_1^2}{s_2^2}$ มีการแจกแจงแบบเอฟหรือไม่) อย่างมีนัยสำคัญหรือไม่ โดยนำเสนอค่ากำลังการทดสอบ

ส่วนที่ (3) นำเสนอผลกระทบของการประมาณค่าสัมประสิทธิ์การถดถอย ทั้งกรณีการถดถอยอย่างง่าย การถดถอยเชิงพหุเมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน และการถดถอยเชิงพหุเมื่อตัวแปรอิสระมีความสัมพันธ์กัน เมื่อความคลาดเคลื่อนมีการแจกแจงแบบโลจิสติก โดยนำเสนอค่าความน่าจะเป็นของการเกิดความคลาดเคลื่อนประเภทที่ 1 สำหรับการทดสอบว่าค่าสัมประสิทธิ์การถดถอยที่ประมาณได้ให้ค่าสอดคล้องกับค่าจริงหรือไม่ (ไม่สามารถหาลำดับการทดสอบได้ไม่สามารถหาลำดับการทดสอบได้เนื่องจากสมมติฐานแย้งเป็นสมมติฐานประกอบ) และค่า p-value สำหรับการทดสอบว่าสัมประสิทธิ์การถดถอยที่ประมาณได้มีการแจกแจงแบบปกติหลายตัวแปรหรือไม่