บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในงานวิจัยด้านต่าง ๆ ไม่ว่าจะเป็นด้านวิทยาศาสตร์ หรือสังคมศาสตร์นั้น หากต้องการคาดคะเนหรือพยากรณ์ค่าของตัวแปรตาม(dependent variable)ที่เราสนใจ ว่าขึ้นอยู่กับกลุ่มของตัวแปรอิสระ(independent variables) อย่างไร โดยทั่วไปจะใช้เทคนิคที่เรียกว่าการวิเคราะห์การถดถอยพหุคูณ(Multiple Regression Analysis) และนิยมประมาณค่าพารามิเตอร์หรือสัมประสิทธิ์การถดถอยของตัวแบบ ด้วยตัวประมาณกำลังสองน้อยสุด(Least Squares Estimators) เนื่องจากตัวประมาณดังกล่าวจะมีคุณสมบัติ BLUE (Best Linear Unbiased Estimator) กล่าวคือ จะเป็นตัวประมาณที่ไม่เอนเอียง และในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้นทั้งหมด ตัวประมาณกำลังสองน้อยสุดจะให้ค่าความแปรปรวนต่ำสุด

พิจารณา ตัวแบบการถดถอยพหุคูณเชิงเส้น ซึ่งมีรูปแบบดังต่อไปนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \quad ; i = 1,2,...,n$$

หรือเขียนในรูปเมทริกซ์ ได้ดังนี้

$$\underset{\sim}{y} = X\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}$ แทนเวกเตอร์ของตัวแปรตามที่มีขนาด (n x 1)

X แทนเมทริกซ์ของตัวแปรอิสระที่มีขนาด(n x q)

$\underset{\sim}{\beta}$ แทนเวกเตอร์ของสัมประสิทธิ์ถดถอยที่มีขนาด (q x 1)

$\underset{\sim}{\varepsilon}$ แทนเวกเตอร์ของความคลาดเคลื่อนที่มีขนาด (n x 1)

n แทนขนาดตัวอย่าง

p แทนจำนวนตัวแปรอิสระในตัวแบบ

และ q แทนจำนวนพารามิเตอร์ในตัวแบบ = p + 1

ได้กล่าวไปแล้วว่า ในการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้นพหุคูณนิยมประมาณค่าด้วยตัวประมาณกำลังสองน้อยสุด ซึ่งมีรูปแบบตัวประมาณดังต่อไปนี้

$$\underset{\sim}{\hat{\beta}}_{LS} = (X'X)^{-1}.X'\underset{\sim}{y} \qquad (1.1)$$

แต่ในการประมาณค่าพารามิเตอร์ ด้วยตัวประมาณกำลังสองน้อยสุดที่จะมีคุณสมบัติ BLUE นั้น ลักษณะของข้อมูลจะต้องเป็นไปตามข้อสมมติ (Assumptions) คือ ความคลาดเคลื่อนสุ่ม $\varepsilon_i ; i = 1,...,n$ มีค่าเฉลี่ยเท่ากับศูนย์ ความแปรปรวนเท่ากับ σ^2 คงที่ทุกค่า i และความแปรปรวนร่วมของ $(\varepsilon_i, \varepsilon_j) ; i \neq j$ ทุกคู่เท่ากับศูนย์หรือไม่มีสหสัมพันธ์กัน(uncorrelation) ดังนั้น ในกรณีที่ข้อมูลไม่เป็นไปตามข้อสมมติดังกล่าว การใช้ตัวประมาณกำลังสองน้อยสุดจะทำให้ตัวประมาณพารามิเตอร์ขาดคุณสมบัติ BLUE นอกจากนี้ ตัวประมาณกำลังสองน้อยสุดอาจขาดคุณสมบัติ BLUE อันเนื่องจากสาเหตุอื่นได้เช่นกัน เช่น กรณีที่ตัวแปรอิสระมีความสัมพันธ์กัน ที่เรียกว่า มีพหุสัมพันธ์(Multicollinearity) และกรณีที่ข้อมูลมีค่าผิดปกติ(Outliers) เป็นต้น ดังนั้น จึงได้มีการคิดค้นหาตัวประมาณอื่นแทนตัวประมาณกำลังสองน้อยสุด เมื่อเกิดกรณีที่ไม่เหมาะสมที่จะใช้ตัวประมาณกำลังสองน้อยสุด ตัวประมาณอื่น ๆ เช่น ตัวประมาณการถดถอยที่มีความแกร่ง(Robust Regression Estimator) และตัวประมาณการถดถอยริดจ์ (Ridge Regression Estimator) เป็นต้น ในความเป็นจริง ข้อมูลในด้านต่าง ๆ ที่นำมาวิเคราะห์นั้น อาจมีลักษณะของข้อมูลไม่เป็นไปตามข้อสมมติ และอาจเกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระและมีค่าผิดปกติเกิดขึ้นอีกด้วย ซึ่งหากเป็นเช่นนี้ ควรจะศึกษาหาตัวประมาณอื่น ๆ แทน เช่น ตัวประมาณการถดถอยริดจ์ที่มีความแกร่ง (Robust Ridge Regression Estimator) เป็นต้น

ในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณด้วยตัวประมาณกำลังสองน้อยสุดตามสมการ (1.1) นั้น สมมติฐานที่สำคัญข้อหนึ่ง คือ ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กับตัวแปรอิสระตัวอื่น ซึ่งในทางปฏิบัติเกิดขึ้นได้น้อยมาก และเมื่อตัวแปรอิสระเกิดพหุสัมพันธ์กันสูงจะทำให้เมทริกซ์ $X'X$ เกิดเงื่อนไขที่ไม่ดี(ill – condition) นั่นคือ ทำให้ $|X'X|$ มีค่าเล็กลงเข้าใกล้ศูนย์ เนื่องจาก เมทริกซ์ความแปรปรวนร่วมของค่าประมาณสัมประสิทธิ์การถดถอยอยู่ในรูป $\text{Var}(\underset{\sim}{\hat{\beta}}) = \sigma^2(X'X)^{-1}$ จึงส่งผลให้ความแปรปรวนของค่าประมาณสัมประสิทธิ์การถดถอยมีค่ามาก และมีผลทำให้ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Squares Error) ของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณมีค่ามาก ดังนั้น ถ้าตัว

แปรอิสระมีความสัมพันธ์กันสูง เราอาจแก้ไขโดย การตัดตัวแปรอิสระบางตัวออกจากตัวแบบ แต่ในบางกรณี ความสัมพันธ์ระหว่างตัวแปรอิสระไม่ชัดเจน ทำให้การตัดตัวแปรอิสระตัวใดตัวหนึ่งออกจากตัวแบบทำได้ยาก หรืออาจไม่ต้องการตัดตัวแปรอิสระตัวใดออกจากตัวแบบ เพราะถือว่า ตัวแปรอิสระทุกตัวมีผลในการอธิบายตัวแปรตามได้มากพอ ๆ กัน

Hoerl และ Kennard(1970) ได้ศึกษาหาตัวประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยเชิงเส้นพหุคูณ ที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าตัวประมาณกำลังสองน้อยสุด เมื่อข้อมูลเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยตัวประมาณนี้ สร้างจากหลักการนำค่าคงที่ค่าหนึ่งที่มากกว่าศูนย์(k) มาบวกกับสมาชิกทุกตัวในแนวทแยงมุมของเมทริกซ์ XX เนื่องจาก เมทริกซ์ความแปรปรวนร่วมของตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณอยู่ในรูป $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ ดังนั้น การนำค่าคงที่ที่มากกว่าศูนย์(k) มาบวกกับสมาชิกทุกตัวในแนวทแยงมุมของเมทริกซ์ XX จะทำให้อินเวอร์สของเมทริกซ์ดังกล่าวมีค่าลดลง ซึ่งจะทำให้ได้ตัวประมาณที่มีค่าความแปรปรวนต่ำกว่าตัวประมาณกำลังสองน้อยสุด แต่จะมีความเอนเอียง ดังนั้น จึงต้องพิจารณาประสิทธิภาพของตัวประมาณโดยใช้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ตัวประมาณดังกล่าวเรียกว่า ตัวประมาณการถดถอยริดจ์(Ridge Regression Estimator) ซึ่งมีรูปแบบของตัวประมาณดังนี้

$$\underset{\sim RID}{\hat{\beta}} = (X'X + kI)^{-1}.X'y \qquad ;k > 0 \tag{1.2}$$

ในทางปฏิบัติ นอกจากข้อมูลของตัวแปรอิสระอาจจะมีสหสัมพันธ์กัน ซึ่งอาจก่อให้เกิดผลกระทบต่อประสิทธิภาพของตัวประมาณแล้ว ลักษณะของข้อมูลที่นำมาใช้ในการวิเคราะห์การถดถอยนั้น อาจมีค่าสังเกตบางค่าที่มีค่าสูงหรือต่ำกว่าค่าสังเกตส่วนใหญ่ของข้อมูลโดยปกติ ซึ่งค่าสังเกตดังกล่าว เรียกว่า ค่าผิดปกติ(Outliers) สาเหตุการเกิดข้อมูลผิดปกติอาจเกิดจากความคลาดเคลื่อนจากการวัดค่า เช่น การบันทึกข้อมูลที่ไม่ถูกต้องของผู้บันทึกข้อมูล การใช้เครื่องมือวัดค่าที่มีคุณภาพต่ำ เป็นต้น หรืออาจเกิดจากเหตุการณ์ผิดปกติที่ไม่สามารถคาดการณ์ล่วงหน้าได้ เช่น อุบัติเหตุ หรือ อุทกภัยต่าง ๆ หรืออาจเกิดจากการที่มีบางหน่วยตัวอย่างของข้อมูลไม่ได้มาจากการแจกแจงเดียวกันกับการแจกแจงของประชากรที่ศึกษา เป็นต้น การที่ข้อมูลมีค่าผิดปกติเกิดขึ้นนั้นอาจส่งผลให้การประมาณค่าสัมประสิทธิ์การถดถอยด้วยตัวประมาณกำลังสองน้อยสุดไม่เหมาะสมได้ เนื่องจาก สมการการถดถอยพหุคูณที่ได้ หรือเส้นสมการถดถอยจะถูกปรับทิศทางไปตามข้อมูลที่ผิดปกติ และอาจทำให้ค่าประมาณของความแปรปรวนของ

ความคลาดเคลื่อนมีค่าสูงขึ้นกว่าปกติ ดังนั้น ค่าประมาณความแปรปรวนของ $\hat{\beta}$ จึงมีค่าสูงขึ้นด้วย และเมื่อนำไปทดสอบสมมติฐานของพารามิเตอร์ ก็จะทำให้ผลการทดสอบผิดพลาดได้

ดังนั้น หากนำข้อมูลที่ใช้ในการวิเคราะห์การถดถอยมาตรวจสอบค่าผิดปกติแล้ว พบว่าข้อมูลมีค่าผิดปกติ ผู้วิเคราะห์ไม่ควรตัดค่าผิดปกตินั้นออกทันที แต่ควรพิจารณาถึงสาเหตุของการเกิดค่าผิดปกติดังกล่าว เช่น หากเกิดจากการบันทึกข้อมูลผิดพลาด วิธีการแก้ไขอาจทำโดยการตัดข้อมูลที่มีค่าผิดปกติออกไป หรืออาจทำการปรับค่าให้เหมาะสม จากนั้น จึงนำข้อมูลที่เหลือมาวิเคราะห์การถดถอยต่อไป โดยใช้ตัวประมาณกำลังสองน้อยสุดในการประมาณค่าพารามิเตอร์ แต่หากไม่สามารถอธิบายถึงสาเหตุของการเกิดค่าผิดปกติได้ อาจแก้ไขโดยการแปลงข้อมูล(Data Transformation) เพื่อลดอิทธิพลของข้อมูลที่มีค่าผิดปกติลง และนอกจากวิธีการแปลงข้อมูลแล้ว มีนักสถิติหลายท่านได้คิดค้นตัวประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นไว้เป็นจำนวนมาก โดยมีหลักการคือ พยายามลดอิทธิพลของข้อมูลที่เป็นค่าผิดปกติลง โดยไม่มีการตัดค่าผิดปกติดังกล่าวทิ้ง หรือเป็นการให้น้ำหนักของค่าสังเกตที่เป็นค่าผิดปกติน้อยกว่าค่าสังเกตที่เป็นข้อมูลส่วนใหญ่ ซึ่งตัวประมาณที่สร้างจากหลักการดังกล่าว เรียกว่า ตัวประมาณการถดถอยที่มีความแกร่ง(Robust Regression Estimator) ซึ่งการวิจัยในครั้งนี้ จะทำการศึกษาโดยใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด(Least Absolute Value Estimator) ซึ่งตัวประมาณการถดถอยดังกล่าว จะหาค่าประมาณของพารามิเตอร์ $\hat{\beta}$ ที่ทำให้ผลรวมของค่าสัมบูรณ์ของความคลาดเคลื่อนมีค่าต่ำสุด กล่าวคือ

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \left| y_i - \underset{\sim}{x_i}' \underset{\sim}{\beta} \right| \tag{1.3}$$

การวิเคราะห์การถดถอยเชิงเส้นพหุคูณนั้น ลักษณะของข้อมูลอาจมีพหุสัมพันธ์ระหว่างตัวแปรอิสระและมีค่าผิดปกติในตัวแปรตามขึ้นพร้อมกัน ในกรณีเช่นนี้ ควรจะศึกษาหาตัวประมาณอื่น ๆ แทน โดยตัวประมาณค่าพารามิเตอร์ที่เหมาะสมน่าจะเป็นตัวประมาณที่ได้จากการผสมผสานระหว่างตัวประมาณที่ใช้ในการแก้ปัญหาพหุสัมพันธ์และตัวประมาณที่ใช้ในการแก้ปัญหาที่มีค่าผิดปกติเกิดขึ้น หรือกล่าวว่า เป็นตัวประมาณที่ได้จากการผสมผสานระหว่างตัวประมาณการถดถอยริดจ์ และตัวประมาณการถดถอยที่มีความแกร่ง ซึ่งตัวประมาณดังกล่าวเรียกว่า ตัวประมาณการถดถอยริดจ์ที่มีความแกร่ง(Robust Ridge Regression Estimator)

และในการวิจัยครั้งนี้ สนใจที่จะศึกษาตัวประมาณที่เสนอโดย Pfaffenberger และ Dielman (1984) ซึ่งเป็นตัวประมาณที่สร้างจากตัวประมาณการถดถอยริดจ์จากสมการที่ (1.2) และตัวประมาณค่าสัมบูรณ์น้อยสุดที่สร้างจากสมการที่ (1.3) ซึ่งเรียกว่า ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(Ridge Least Absolute Value Estimator) ซึ่งมีรูปแบบตัวประมาณดังนี้

$$\underset{\sim RLAV}{\hat{\beta}} = \left(X'X + k^{*}I\right)^{-1}X'\underset{\sim}{y} \qquad (1.4)$$

และนอกจากนี้ ผู้วิจัยยังสนใจที่จะศึกษาตัวประมาณริดจ์ที่มีการถ่วงน้ำหนัก (Weighted Ridge Estimator) ซึ่งเสนอโดย Askin และ Montgomery(1980) ซึ่งเป็นตัวประมาณที่สร้างจากตัวประมาณริดจ์กับตัวประมาณกำลังสองน้อยสุดแบบถ่วงน้ำหนัก ซึ่งมีรูปแบบตัวประมาณดังนี้

$$\underset{\sim WRID}{\hat{\beta}} = \left(X'WX + kI\right)^{-1}X'W\underset{\sim}{y} \qquad (1.5)$$

ซึ่งการวิจัยในครั้งนี้ จะเลือกใช้ค่าของน้ำหนัก $w_{ii} = \frac{1}{|e_i|}$ ซึ่งเสนอโดย Pfaffenberger และ Dielman(1990)

เมื่อ e_i แทน ส่วนเหลือ(Residual)จากการใช้ตัวประมาณค่าสัมบูรณ์น้อยสุดในการประมาณค่าตัวแปรตาม y หรือกล่าวว่า เป็นผลต่างระหว่างค่าจริงของตัวแปรตาม y กับค่าประมาณของตัวแปรตาม($\hat{y}$) ซึ่งประมาณค่าโดยใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด

ดังนั้น ในการวิจัยครั้งนี้ ผู้วิจัยสนใจที่จะศึกษาในกรณีที่ข้อมูลมีปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระและ/หรือมีค่าผิดปกติในตัวแปรตาม โดยตัวประมาณการถดถอยที่จะนำมาศึกษาในครั้งนี้มีจำนวน 5 ตัว คือ ตัวประมาณกำลังสองน้อยสุด (Least Squares Estimator) ตัวประมาณค่าสัมบูรณ์น้อยสุด (Least Absolute Value Estimator) ตัวประมาณการถดถอยริดจ์ (Ridge Regression Estimator) ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(Ridge Least Absolute Value Estimator) และ ตัวประมาณริดจ์ที่มีการถ่วงน้ำหนัก (Weighted Ridge Estimator) ภายใต้สถานการณ์ที่ข้อมูลมีพหุสัมพันธ์ระหว่างตัวแปรอิสระและ/หรือมีค่าผิดปกติในตัวแปรตามในระดับต่าง ๆ แล้วทำการเปรียบเทียบประสิทธิภาพของตัวประมาณทั้ง 5 ตัว โดยใช้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง(Root Mean Squares Error : RMSE) ซึ่งตัวประมาณใดที่ให้ค่า RMSE ต่ำสุด จะถือว่าเป็นตัวประมาณที่มีประสิทธิภาพ

สูงสุดในแต่ละสถานการณ์ และข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลองด้วยเทคนิคการจำลองมอนติคาร์โล(Monte Carlo Simulation Technique) โดยทำการทดลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของตัวประมาณการถดถอยโดยใช้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณการถดถอย เมื่อข้อมูลมีพหุสัมพันธ์ระหว่างตัวแปรอิสระและ/หรือมีค่าผิดปกติในตัวแปรตาม โดยตัวประมาณการถดถอยที่นำมาใช้ในการวิจัยในครั้งนี้ มีจำนวน 5 ตัว ดังต่อไปนี้

1. ตัวประมาณกำลังสองน้อยสุด(Least Squares Estimator : LS)
2. ตัวประมาณค่าสัมบูรณ์น้อยสุด(Least Absolute Value Estimator : LAV)
3. ตัวประมาณริดจ์(Ridge Estimator : RID)
4. ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(Ridge Least Absolute Value Estimator : RLAV)
5. ตัวประมาณริดจ์ที่มีการถ่วงน้ำหนัก(Weighted Ridge Estimator : WRID)

1.3 สมมติฐานของการวิจัย

สมมติฐานของการวิจัย ในการเลือกใช้ตัวประมาณพารามิเตอร์ของตัวแบบเชิงเส้นพหุคูณเมื่อข้อมูลมีค่าผิดปกติในตัวแปรตาม และ/หรือ มีพหุสัมพันธ์ระหว่างตัวแปรอิสระในระดับต่าง ๆ เป็นดังนี้

1. เมื่อตัวแปรอิสระมีพหุสัมพันธ์ ตัวประมาณริดจ์ (RID) จะให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ในการประมาณค่าพารามิเตอร์ของตัวแบบต่ำกว่าตัวประมาณอื่น ๆ

2. เมื่อตัวแปรตามมีค่าผิดปกติ ตัวประมาณค่าสัมบูรณ์น้อยสุด(LAV) จะให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ในการประมาณค่าพารามิเตอร์ของตัวแบบต่ำกว่าตัวประมาณอื่น ๆ

3. เมื่อตัวแปรตามมีค่าผิดปกติและมีพหุสัมพันธ์ระหว่างตัวแปรอิสระ ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(RLAV) และตัวประมาณริดจ์ที่มีการถ่วงน้ำหนัก(WRID) จะให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ในการประมาณค่าพารามิเตอร์ของตัวแบบต่ำกว่าตัวประมาณอื่น ๆ

1.4 ขอบเขตของการวิจัย

1. ในการวิจัยครั้งนี้ ตัวแบบการถดถอยเชิงเส้นพหุคูณอยู่ในรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad ; i = 1,2,...,n$$

หรือเขียนในรูปเมทริกซ์ ได้ดังนี้

$$\underset{\sim}{y} = X\,\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}$ แทนเวกเตอร์ของตัวแปรตามที่มีขนาด (n x 1)

X แทนเมทริกซ์ของตัวแปรอิสระที่มีขนาด(n x q)

$\underset{\sim}{\beta}$ แทนเวกเตอร์ของสัมประสิทธิ์ถดถอยที่มีขนาด (q x 1)

$\underset{\sim}{\varepsilon}$ แทนเวกเตอร์ของความคลาดเคลื่อนที่มีขนาด (n x 1)

n แทนขนาดตัวอย่าง

และ q แทนจำนวนพารามิเตอร์ในตัวแบบ = 4

โดยมีข้อกำหนดว่า rank ของเมทริกซ์ X เท่ากับ q ; (q < n) ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่เป็นอิสระต่อกันและมีการแจกแจงเดียวกัน ที่มีค่าเฉลี่ย $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ และ เมทริกซ์ความแปรปรวน - ความแปรปรวนร่วม $E\left(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}'\right) = \sigma^2 I$

2. การแจกแจงของความคลาดเคลื่อนสุ่ม แบ่งออกเป็น 2 การแจกแจง ดังนี้

2.1 การแจกแจงปกติ(Normal Distribution)

$$f(\varepsilon)=\frac{1}{\sqrt{2\pi\sigma^2}}.\exp\left\{\frac{-1}{2\sigma^2}(\varepsilon-\mu)^2\right\} \qquad ;-\infty<\mu<\infty,\ \sigma^2>0$$

เมื่อ ε แทน ความคลาดเคลื่อนสุ่ม

μ แทน ค่าเฉลี่ยของความคลาดเคลื่อนสุ่ม

σ^2 แทน ความแปรปรวนของความคลาดเคลื่อนสุ่ม

ในงานวิจัยครั้งนี้ จะศึกษาในกรณีที่ μ = 0 และ σ^2 = 3

2.2 การแจกแจงปกติปลอมปน(Contaminated Normal Distribution)

$$f(\varepsilon)=(1-P)N(\mu,\sigma^2)+PN(\mu,C^2\sigma^2)$$

เมื่อ P แทน สัดส่วนการปลอมปน(Proportion of Contamination)

และ C แทน สเกลแฟคเตอร์(Scale Factor)

ในงานวิจัยครั้งนี้ จะศึกษาในกรณีที่ μ = 0 , σ^2 = 3 , P = 0.05 , 0.08 , 0.10 และ 0.15 และมีระดับค่าผิดปกติ 3 ระดับ คือ ระดับเล็กน้อย ปานกลาง และรุนแรง โดยใช้เกณฑ์การกำหนดขนาดค่าผิดปกติด้วย Box Plot จะได้ว่า

เมื่อค่า C = 4 - 6 ตัวแปรตามจะมีค่าผิดปกติระดับเล็กน้อย - ปานกลาง

และ เมื่อค่า C = 12 - 13 ตัวแปรตามจะมีค่าผิดปกติระดับรุนแรง

ดังนั้น ในงานวิจัยครั้งนี้ เลือกกำหนดค่า C = 4 เมื่อตัวแปรตามมีค่าผิดปกติระดับเล็กน้อย C = 6 เมื่อตัวแปรตามมีค่าผิดปกติระดับปานกลาง และ เลือกกำหนดค่า C = 13 เมื่อตัวแปรตามมีค่าผิดปกติระดับรุนแรง ซึ่งได้กล่าวรายละเอียดการใช้เกณฑ์การกำหนดขนาดค่าผิดปกติด้วย Box Plot ไว้ในบทที่ 3 (หน้า 24)

3. จำนวนตัวแปรอิสระที่ใช้ในการวิจัยเท่ากับ 3 ตัวแปร และกำหนดระดับสหสัมพันธ์ระหว่างตัวแปรอิสระ x_1 กับ x_2 (ρ) จำนวน 7 ระดับ คือ 0.1 , 0.3 , 0.5 , 0.7 , 0.9 , 0.95 และ 0.99

4. ขนาดตัวอย่าง(n) มีจำนวน 6 ระดับ คือ 20 , 30 , 35 , 40 , 50 และ 60

5. ในการวิจัยครั้งนี้ ทำการจำลองข้อมูลโดยใช้เทคนิคการจำลองมอนติคาร์โล (Monte Carlo Simulation Technique) โดยทำการจำลองจำนวน 1,000 ครั้ง ในแต่ละสถานการณ์

1.5 เกณฑ์ที่ใช้ในการพิจารณา

ในการวิจัยครั้งนี้ จะทำการเปรียบเทียบประสิทธิภาพของตัวประมาณพารามิเตอร์แต่ละตัว โดยการเปรียบเทียบค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณการถดถอยพหุคูณเชิงเส้น(Root Mean Squares Error : RMSE) โดยมีสูตรในการคำนวณดังต่อไปนี้

$$RMSE_i = \sqrt{\sum_{j=1}^{1,000} \frac{\left(\beta_{ij} - \hat{\beta}_{ij}\right)^2}{1,000}} \quad , i = 0,1,2,3$$

เมื่อ β_{ij} แทนค่าจริงของพารามิเตอร์ตัวที่ i ในสมการถดถอย ในการจำลองรอบที่ j

$\hat{\beta}_{ij}$ แทนค่าประมาณของพารามิเตอร์ตัวที่ i ในสมการถดถอย ในการจำลองรอบที่ j

$RMSE_i$ แทนค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณพารามิเตอร์ตัวที่ i ในสมการถดถอย

1.6 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการตัดสินใจเลือกใช้ตัวประมาณพารามิเตอร์ที่มีประสิทธิภาพ ในตัวแบบการถดถอยพหุคูณเชิงเส้น ในกรณีที่ข้อมูลมีพหุสัมพันธ์ระหว่างตัวแปรอิสระและ/หรือมีค่าผิดปกติในตัวแปรตาม และเป็นแนวทางสำหรับการวิจัยเพิ่มเติม ในการเลือกใช้ตัวประมาณอื่น ๆ หรือการแจกแจงของความคลาดเคลื่อนในรูปแบบอื่น ๆ