

การเปรียบเทียบตัวประมาณการถดถอยเชิงเส้นพหุคูณ เมื่อข้อมูลมีปัญหาพหุสัมพันธ์และค่าปกติ

*ปัทมวดี นันทนานนท์

ในการศึกษาความสัมพันธ์ระหว่างสิ่งต่างๆ เพื่อนำไปใช้ประโยชน์ในการตัดสินใจในชีวิตประจำวันนั้นเกิดขึ้นอยู่เสมอ เช่น ความสัมพันธ์ระหว่างจำนวนสมาชิกในบ้านกับค่าไฟฟ้า รายได้เสริมกับค่าใช้จ่ายเบ็ดเตล็ด ในการดำเนินการทางธุรกิจก็เช่นเดียวกัน มีความสัมพันธ์ระหว่างปัจจัยในการดำเนินการทางธุรกิจไม่ว่าจะเป็นทางด้านบัญชีและการเงิน การผลิตและการปฏิบัติการ การตลาด การบริหารทรัพยากรมนุษย์ เช่น ความสัมพันธ์ระหว่างยอดขายกับราคาและค่าใช้จ่ายทางการตลาด ต้นทุนการผลิตกับจำนวนการผลิต เงินเดือนกับความพึงพอใจในการทำงานของพนักงาน ทักษะคติของพนักงานกับผลิตผลที่ผลิตได้ หากผู้บริหารศึกษาและเข้าใจความสัมพันธ์ของสิ่งเหล่านี้เป็นอย่างดีแล้วจะนำไปใช้เพิ่มความสามารถในการจัดการหรือในการตัดสินใจทางธุรกิจได้ขึ้นแรกเราอาจพิจารณาโดยใช้แผนภาพการกระจาย (Scatter Diagram) เพื่อพิจารณาถึงระดับและทิศทางของความสัมพันธ์ระหว่างปัจจัยที่นำมาศึกษาอย่างคร่าว ๆ รวมทั้งยังสามารถพิจารณาถึงการมีค่าผิดปกติหรือแนวโน้มของรูปแบบความสัมพันธ์ หลังจากนั้นจึงใช้การวิเคราะห์สหสัมพันธ์ (Correlation Analysis) ในการวัดระดับและทิศทางของความสัมพันธ์ระหว่างตัวแปรที่ศึกษาอย่างละเอียดขึ้น แต่ในการประยุกต์ใช้ในทางธุรกิจนั้นส่วนใหญ่เราต้องการศึกษาความสัมพันธ์ของตัวแปรต่าง ๆ เพื่อพยากรณ์หรือประมาณค่าตัวแปรตัวหนึ่งที่เราสนใจเรียกว่าตัวแปรตาม (Dependent Variable) โดยใช้ตัวแปรที่เราทราบค่าเรียกว่าตัวแปรอิสระ (Independent Variable) ในทางสถิติใช้เทคนิคที่เรียกว่าการวิเคราะห์การถดถอย (Regression Analysis) หากใช้ตัวแปรอิสระเพียงตัวเดียวในการพยากรณ์ค่าของตัวแปรตามจะเรียกว่าการวิเคราะห์การถดถอยอย่างง่าย (Simple Regression Analysis) เช่น การพยากรณ์ค่ายอดขายในปีหน้าจากงบค่าโฆษณาที่ตั้งไว้ในปีหน้า หากใช้ตัวแปรอิสระมากกว่าหนึ่งตัวในการพยากรณ์ค่าของตัวแปรตาม จะเรียกว่าการวิเคราะห์การถดถอยพหุคูณ (Multiple Regression Analysis) เช่น การพยากรณ์ต้นทุนการผลิตโดยใช้ปัจจัยที่มีอิทธิพลต่อต้นทุนการผลิตอันได้แก่ จำนวนการผลิต (Productivity) ประสิทธิภาพการผลิต (Efficiency) และทัศนคติของพนักงาน (Attitude) สำหรับผู้ที่ศึกษางานวิจัยทางด้านการวิเคราะห์การถดถอยเป็นที่ทราบกันดีว่า

*อาจารย์ประจำภาควิชาคณิตศาสตร์และสถิติ คณะศิลปศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์ : ศต.ม. (สถิติ) จุฬาลงกรณ์มหาวิทยาลัย

การใช้เทคนิคการวิเคราะห์การถดถอยจะต้องมีการประมาณค่าพารามิเตอร์ของตัวแบบ

ตัวประมาณที่นิยมใช้ประมาณค่าพารามิเตอร์ในตัวแบบคือตัวประมาณกำลังสองน้อยสุด (Least Squares Estimator) เนื่องจากคุณสมบัติของตัวประมาณนี้เป็นตัวประมาณที่ไม่เอนเอียง และในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้นทั้งหมด ตัวประมาณนี้จะให้ค่าความแปรปรวนต่ำสุดซึ่งเรียก ว่ามีคุณสมบัติ BLUE (Best Linear Unbiased Estimator) ตัวประมาณกำลังสองน้อยสุดอยู่ใน

$$\text{รูปแบบ } \underset{\sim}{\hat{\beta}}_{LS} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1} \cdot \underset{\sim}{X}'\underset{\sim}{y} \quad (1.1)$$

ในการประมาณค่าพารามิเตอร์ด้วยตัวประมาณกำลังสองน้อยสุดนั้น ลักษณะของข้อมูลที่น่ามาวิเคราะห์จะต้องเป็นไปตามข้อกำหนด (Assumptions) คือความคลาดเคลื่อนสุ่ม (Random Error) $\varepsilon_i ; i = 1, \dots, n$ มีค่าเฉลี่ยเท่ากับศูนย์ ความแปรปรวนเท่ากับ σ^2 คงที่ทุกค่า i และความแปรปรวนร่วมของ $(\varepsilon_i, \varepsilon_j) ; i \neq j$ ทุกคู่เท่ากับศูนย์หรือไม่มีสหสัมพันธ์กัน (Uncorrelation) ในกรณีที่ข้อมูลไม่เป็นไปตามข้อกำหนดจะทำให้ตัวประมาณกำลังสองน้อยสุดขาดคุณสมบัติ BLUE นอกจากนี้ตัวประมาณกำลังสองน้อยสุดอาจขาดคุณสมบัติ BLUE จากสาเหตุอื่นเช่น ตัวแปรอิสระมีความสัมพันธ์กันเองที่เรียกว่ามีพหุสัมพันธ์ (Multicollinearity) และกรณีที่ข้อมูลมีค่าผิดปกติ (Outliers) จึงได้มีการคิดค้นหาตัวประมาณอื่นแทนตัวประมาณกำลังสองน้อยสุด เช่น ตัวประมาณการถดถอยที่มีความแกร่ง (Robust Regression Estimator) ตัวประมาณการถดถอยริดจ์ (Ridge Regression Estimator) และหากข้อมูลที่นำมาวิเคราะห์มีปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระและค่าผิดปกติเกิดขึ้นพร้อมกันควรจะศึกษาหาตัวประมาณอื่นแทน ได้แก่ ตัวประมาณการถดถอยริดจ์ที่มีความแกร่ง (Robust Ridge Regression Estimator)

ในการประมาณค่าพารามิเตอร์ด้วยตัวประมาณกำลังสองน้อยสุด สมมติฐานที่สำคัญข้อหนึ่งคือตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กับตัวแปรอิสระตัวอื่นซึ่งในทางปฏิบัติเกิดขึ้นได้น้อยมากและเมื่อตัวแปรอิสระเกิดพหุสัมพันธ์กันสูงจะทำให้เมทริกซ์ $X'X$ เกิดเงื่อนไขที่ไม่ดี (ill-condition) ทำให้ $|X'X|$ มีค่าเล็กลงเข้าใกล้ศูนย์ เนื่องจากเมทริกซ์ความแปรปรวนร่วมของค่าประมาณ

สมมติการถดถอยอยู่ในรูป $\text{Cov}(\underset{\sim}{\hat{\beta}}) = \sigma^2 (\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ จึงทำให้ความแปรปรวนของค่าประมาณ

สมมติการถดถอยมีค่ามาก และมีผลให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสมมติการถดถอยมีค่ามาก ดังนั้นถ้าตัวแปรอิสระมีความสัมพันธ์กันสูงอาจแก้ไขโดยการตัดตัวแปรอิสระบางตัวออกจากตัวแบบ แต่ในบางกรณีความสัมพันธ์ระหว่างตัวแปรอิสระไม่ชัดเจนทำให้การเลือกตัดตัวแปรอิสระตัวใดตัวหนึ่งออกจากตัวแบบทำได้ยากหรืออาจไม่ต้องการตัดตัวแปรอิสระตัวใดออกจากตัวแบบ เพราะถือว่าตัวแปรอิสระทุกตัวมีผลในการอธิบายตัวแปรตามได้มากพอ ๆ กัน Hoerl และ Kennard (1970) ศึกษาตัวประมาณค่าพารามิเตอร์ที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่า

ตัวประมาณกำลังสองน้อยสุดเมื่อข้อมูลเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ ตัวประมาณนี้เรียกว่าตัวประมาณการถดถอยริดจ์(Ridge Regression Estimator) อยู่ใน

$$\text{รูปแบบ } \underset{\sim}{\hat{\beta}}_{RID} = (X'X + kI)^{-1} \cdot \underset{\sim}{X}'y \quad ; k > 0 \quad (1.2)$$

เมื่อ k แทน ค่าคงที่หรือเรียกว่าเป็น พารามิเตอร์ที่เอนเอียง(Biased Parameter)

I แทน เมทริกซ์เอกลักษณ์ (Identity Matrix) ที่มีขนาด $(q \times q)$

ตัวประมาณการถดถอยริดจ์เป็นตัวประมาณที่ไม่เอนเอียง และสร้างจากการนำค่าคงที่ค่าหนึ่ง ที่มากกว่าศูนย์(k) มาบวกกับสมาชิกทุกตัวในแนวเส้นทแยงมุมของเมทริกซ์ $X'X$ ทำให้ $(X'X)^{-1}$ และเมทริกซ์ความแปรปรวนร่วมของตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณ

$(\underset{\sim}{\hat{\beta}}) = \sigma^2 (X'X)^{-1}$ มีค่าลดลง ทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณนี้ต่ำกว่าตัว

ประมาณกำลังสองน้อยสุดได้ โดยขึ้นอยู่กับทางเลือกค่า k ที่เหมาะสมด้วย ในทางปฏิบัติเราไม่ทราบค่า k ที่เหมาะสม มีนักวิจัยเสนอวิธีการหาค่า k ที่เหมาะสมไว้หลายวิธี เริ่มแรก Hoerl และ Kennard(1970) ใช้วิธีพิจารณาจากกราฟ (Graphical Technique) หรือเรียกว่า Ridge Trace แต่เป็นวิธีการที่ค่อนข้างยุ่งยาก จึงมีนักวิจัยคิดวิธีการหาค่า k โดยไม่ใช้กราฟไว้หลายวิธี ซึ่ง Gibbons(1981) ทำการวิจัยเพื่อเปรียบเทียบวิธีการในการหาค่า k ที่เหมาะสม พบว่าโดยภาพรวมวิธีการเลือกค่า k ที่ดีที่สุดคือวิธีการประมาณค่า k ที่คิดค้นโดย Hoerl, Kennard และ Bladwin(1975) แทนด้วย k_{HKB} และอยู่ใน

$$\text{รูปแบบ } k_{HKB} = \frac{qs^2}{\underset{\sim}{\hat{\beta}}_{LS}' \underset{\sim}{\hat{\beta}}_{LS}} \quad (1.3)$$

$$\text{เมื่อ } s^2 = \frac{\left(\underset{\sim}{y} - \underset{\sim}{X} \underset{\sim}{\hat{\beta}}_{LS} \right) \left(\underset{\sim}{y} - \underset{\sim}{X} \underset{\sim}{\hat{\beta}}_{LS} \right)'}{n - q} \text{ เป็นตัวประมาณค่าพารามิเตอร์ } \sigma^2$$

ลักษณะของข้อมูลที่นำมาวิเคราะห์อาจมีค่าสังเกตบางค่าที่มีค่าสูงหรือต่ำกว่าค่าสังเกตส่วนใหญ่ของข้อมูลโดยปกติ เรียกว่าค่าผิดปกติหรือค่านอกกลุ่ม (Outliers) สาเหตุการเกิดข้อมูลผิดปกติอาจเกิดจากความคลาดเคลื่อนจากการวัดค่า เช่น การบันทึกข้อมูลที่ไม่ถูกต้องของผู้บันทึกข้อมูล การใช้เครื่องมือวัดค่าที่มีคุณภาพต่ำ หรืออาจเกิดจากการเกิดค่าผิดปกติที่ไม่สามารถคาดการณ์ล่วงหน้าได้ เช่น อุบัติเหตุหรืออุทกภัย หรืออาจเกิดจากการที่มีบางหน่วยตัวอย่างของข้อมูลไม่ได้มาจากการแจกแจงเดียวกันกับการแจกแจงของประชากรที่ศึกษา การที่มีข้อมูลมีค่าผิดปกติเกิดขึ้นอาจส่งผลให้การประมาณค่าพารามิเตอร์ด้วยตัวประมาณกำลังสองน้อยสุดไม่เหมาะสม เนื่องจากสมการถดถอยเชิงเส้นพหุคูณจะถูกปรับ

ทิศทางไปตามข้อมูลที่ผิดปกติ และอาจทำให้ค่าประมาณของความแปรปรวนและความคลาดเคลื่อนมีค่าสูงขึ้น เมื่อนำไปทดสอบสมมติฐานของค่าพารามิเตอร์จะทำให้ผลการทดสอบผิดพลาดได้ เพราะค่าประมาณความแปรปรวนของค่าพารามิเตอร์มีค่าสูงขึ้นด้วย อย่างไรก็ตามค่าผิดปกติบางค่าอาจให้ข้อมูลที่ เป็นประโยชน์ต่อการหาสมการถดถอยเชิงเส้นที่เหมาะสมได้ ดังนั้นหากนำข้อมูลที่ใช้ในการวิเคราะห์ มาตรวจสอบค่าผิดปกติแล้วพบว่าข้อมูลมีค่าผิดปกติและมีอิทธิพลต่อการประมาณค่าสัมประสิทธิ์การ ถดถอยรวมอยู่ ผู้วิเคราะห์ไม่ควรตัดค่าผิดปกตินั้นออกทันทีแต่ควรพิจารณาถึงสาเหตุของการเกิดค่าผิด ปกติดังกล่าว เช่น หากเกิดจากการบันทึกข้อมูลผิดพลาด วิธีการแก้ไขอาจทำได้โดยการตัดข้อมูลที่มีค่าผิด ปกติทิ้งไปหรืออาจทำการปรับค่าให้เหมาะสม จากนั้นจึงนำข้อมูลที่เหลือมาวิเคราะห์การถดถอยต่อไป โดยใช้ตัวประมาณกำลังสองน้อยสุดในการประมาณค่าพารามิเตอร์ แต่หากค่าผิดปกติถูกบันทึกมาอย่าง ถูกต้องและทราบถึงสาเหตุของการเกิดค่าผิดปกติ เช่น เกิดจากรูปแบบของสมการถดถอยที่ใช้ไม่เหมาะ สม ผู้วิเคราะห์ควรทำการปรับปรุงตัวแบบถดถอยให้มีความเหมาะสมมากขึ้น แต่หากไม่สามารถอธิบาย ถึงสาเหตุของการเกิดค่าผิดปกติได้ อาจแก้ไขโดยการแปลงข้อมูล(Data Transformation)เพื่อลดอิทธิพล ของข้อมูลที่มีค่าผิดปกติลง และนอกจากวิธีการแปลงข้อมูลแล้วมีนักสถิติหลายท่านได้คิดค้นตัว ประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้น โดยมีหลักการคือพยายามลดอิทธิพลของข้อมูลที่มีค่า ผิดปกติลงโดยไม่มีการตัดค่าผิดปกติทิ้ง หรือเป็นการให้น้ำหนักของค่าสังเกตที่เป็นค่าผิดปกติน้อยกว่า ค่าสังเกตที่เป็นข้อมูลส่วนใหญ่ เรียกว่าตัวประมาณการถดถอยที่มีความแกร่ง(Robust Regression Estimator) ศึกษาโดยใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด(Least Absolute Value Estimator) ซึ่งเป็นตัว ประมาณที่ทำให้ผลรวมของค่าสัมบูรณ์ของความคลาดเคลื่อนมีค่าต่ำสุด กล่าวคือ

$$\min_{\beta} \sum_{i=1}^n |y_i - x_i' \beta| \quad (1.4)$$

และใช้เทคนิคโปรแกรมเชิงเส้น(Linear Programming Technique) ในการประมาณค่า พารามิเตอร์ที่ทำให้ผลรวมของค่าสัมบูรณ์ของความคลาดเคลื่อนมีค่าต่ำสุด และใช้วิธีการซิมเพล็กซ์ (Simplex Method) ในการหาค่าประมาณพารามิเตอร์ในตัวแบบ

ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ หากข้อมูลเกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปร อิสระและค่าผิดปกติในตัวแปรตามขึ้นพร้อมกัน ควรจะศึกษาหาตัวประมาณอื่นแทนโดยตัวประมาณค่า พารามิเตอร์ที่เหมาะสมจะเป็นตัวประมาณที่ได้จากการผสมผสานระหว่างตัวประมาณที่ใช้ในการแก้ ปัญหาพหุสัมพันธ์และปัญหาที่มีค่าผิดปกติ หรือเป็นตัวประมาณที่ได้จากการผสมผสานระหว่างตัว ประมาณการถดถอยริดจ์และตัวประมาณการถดถอยที่มีความแกร่ง เรียกว่าตัวประมาณการถดถอยริดจ์ ที่มีความแกร่ง(Robust Ridge Regression Estimator) จะศึกษาตัวประมาณการถดถอยริดจ์ที่มีความแกร่ง ที่เสนอโดย Pfaffenberger และ Dielman(1984) ซึ่งเป็นตัวประมาณที่สร้างจากตัวประมาณการถดถอย ริดจ์จากสมการ(1.2) และตัวประมาณค่าสัมบูรณ์น้อยสุดที่สร้างจากหลักการจากสมการ(1.4) เรียกว่าตัว ประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(Ridge Least Absolute Value Estimator) มีรูปแบบตัวประมาณดังนี้

$$\hat{\beta}_{\sim RLAV} = (X'X + k^*I)^{-1} X'y \quad (1.5)$$

โดยค่า k^* ได้จากวิธีการประมาณค่า k ที่คิดค้นโดย Hoerl, Kennard และ Bladwin (1975) แต่อิงพื้นฐานจากตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) แทนการอิงพื้นฐานจากตัวประมาณกำลังสองน้อยสุด (LS) ค่า k^* อยู่ในรูป

$$k^* = \frac{q s_{LAV}^2}{\hat{\beta}_{\sim LAV}' \hat{\beta}_{\sim LAV}} \quad (1.6)$$

$$\text{เมื่อ } s_{LAV}^2 = \frac{\left(y - X \hat{\beta}_{\sim LAV} \right)' \left(y - X \hat{\beta}_{\sim LAV} \right)}{n - q} \quad \text{เป็นตัวประมาณค่าพารามิเตอร์ } \sigma^2$$

นอกจากนี้ยังศึกษาตัวประมาณการถดถอยริดจ์ที่มีความแข็งแกร่งอีกตัวหนึ่งที่เสนอโดย Askin และ Montgomery (1980) ซึ่งเป็นตัวประมาณที่สร้างจากตัวประมาณริดจ์กับตัวประมาณกำลังสองน้อยสุดแบบถ่วงน้ำหนัก เรียกว่าตัวประมาณริดจ์ที่มีการถ่วงน้ำหนัก (Weighted Ridge Estimator) รูปแบบตัวประมาณเป็นดังนี้

$$\hat{\beta}_{\sim WRID} = (X'WX + kI)^{-1} X'W y \quad (1.7)$$

เมื่อ W แทน เมทริกซ์แยงมุมที่มี w_{ii} เป็นสมาชิกในแนวเส้นทแยงมุม

w_{ii} แทน ค่าของน้ำหนักของแต่ละค่าสังเกต

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

แทน ส่วนเหลือจากการใช้ตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) ในการประมาณค่าพารามิเตอร์ และใช้วิธีการประมาณค่า k ที่คิดค้นโดย Hoerl, Kennard และ Bladwin (1975) แทนด้วย k_{HKB} และอยู่ในรูปแบบสมการ(1.3)

จากงานวิจัยผู้เขียนจะศึกษาในกรณีที่มีข้อมูลเกิดปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระและมีค่าผิดปกติในตัวแปรตาม ตัวประมาณการถดถอยเชิงเส้นพหุคูณที่ศึกษามี 5 ตัว คือ ตัวประมาณกำลังสองน้อยสุด (Least Squares Estimator : LS) จากสมการ(1.1) ตัวประมาณการถดถอยริดจ์ (Ridge Regression Estimator : RID) จากสมการ(1.2) โดยใช้ค่า k_{HKB} จากสมการ (1.3) ตัวประมาณค่าสัมบูรณ์น้อยสุด (Least Absolute Value Estimator : LAV) จากสมการ (1.4) ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด (Ridge

Least Absolute Value Estimator : RLAV) จากสมการ (1.5) โดยใช้ค่า k^* จากสมการ (1.6) และตัวประมาณ ริดจ์ที่มีการถ่วงน้ำหนัก (Weighted Ridge Estimator : WRID) จากสมการ (1.7) โดยใช้ค่า k_{HKB} จากสมการ (1.3) โดยทดลองภายใต้สถานการณ์ที่ข้อมูลมีพหุสัมพันธ์ระหว่างตัวแปรอิสระและมีค่าผิดปกติในตัวแปรตามในระดับต่างๆ ทำการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าพารามิเตอร์ทั้ง 5 ตัว โดยใช้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Squares Error : RMSE) ตัวประมาณที่ให้ค่า RMSE ต่ำสุดจะถือว่าเป็นตัวประมาณที่มีประสิทธิภาพสูงสุดในแต่ละสถานการณ์ ข้อมูลที่ใช้ในการทดลองได้จากการจำลองด้วยเทคนิคการจำลองมอนติคาร์โล (Monte Carlo Simulation Technique) ทดลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์

ขอบเขตของการวิจัย

1. ตัวแบบการถดถอยเชิงเส้นพหุคูณอยู่ในรูปแบบ

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

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

เมื่อ $\underset{\sim}{y}$	แทน เวกเตอร์ของตัวแปรตามที่มีขนาด (n x 1)
$\underset{\sim}{X}$	แทน เมทริกซ์ของตัวแปรอิสระที่มีขนาด (n x q)
$\underset{\sim}{\beta}$	แทน เวกเตอร์ของสัมประสิทธิ์ถดถอยที่มีขนาด (q x 1)
$\underset{\sim}{\varepsilon}$	แทน เวกเตอร์ของความคลาดเคลื่อนที่มีขนาด (n x 1)
n	แทน ขนาดตัวอย่าง

และ q แทน จำนวนพารามิเตอร์ที่ต้องการประมาณค่าในตัวแบบเท่ากับ 4

โดยมีข้อกำหนด (Assumptions) ว่า rank ของเมทริกซ์ $\underset{\sim}{X}$ เท่ากับ q (Full Rank) ; (q < n) ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่เป็นอิสระกันและมีการแจกแจงเดียวกันที่มีค่าเฉลี่ย $E(\underset{\sim}{\varepsilon}) = 0$ และ

เมทริกซ์ความแปรปรวนร่วม $E(\underset{\sim}{\varepsilon} \underset{\sim}{\varepsilon}') = \sigma^2 I$ ซึ่งศึกษาในกรณีที่ $\underset{\sim}{\beta} = 1$

2. การแจกแจงของความคลาดเคลื่อนสุ่ม แบ่งออกเป็น 2 การแจกแจง ดังนี้

2.1) การแจกแจงปกติ (Normal Distribution)

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

เมื่อ μ แทน ค่าเฉลี่ยของความคลาดเคลื่อนสุ่ม ซึ่งศึกษาในกรณีที่
 σ^2 แทน ความแปรปรวนของความคลาดเคลื่อนสุ่ม ซึ่งศึกษาในกรณีที่

2.2) การแจกแจงปกติปน (Contaminated Normal Distribution)

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

เมื่อ P แทน สัดส่วนการปน (Percent of Contamination)
 C แทน สเกลแฟกเตอร์ (Scale Factor)

ซึ่งศึกษาในกรณีที่ $\mu = 0$, $\sigma^2 = 3^a$, $P = 5\%, 8\%, 10\%, 15\%$ และระดับค่าผิดปกติ 3 ระดับคือเล็กน้อย ปานกลาง และรุนแรง ใช้เกณฑ์กำหนดขนาดค่าผิดปกติด้วย Box Plot จะได้ว่า เมื่อ C อยู่ระหว่าง 4-6 ตัวแปรตามจะมีค่าผิดปกติระดับเล็กน้อยถึงปานกลาง ถ้าค่า C อยู่ระหว่าง 12-13 ตัวแปรตามจะมีค่าผิดปกติในระดับรุนแรง ซึ่งศึกษาโดยให้ค่า $C = 4$ เมื่อตัวแปรตามมีค่าผิดปกติระดับเล็กน้อย ให้ $C = 6$ เมื่อตัวแปรตามมีค่าผิดปกติระดับปานกลาง และให้ $C = 13$ เมื่อตัวแปรตามมีค่าผิดปกติระดับรุนแรง

3. จำนวนตัวแปรอิสระที่ใช้ในการศึกษาเท่ากับ 3 ตัวแปร และศึกษาในกรณีมีพหุสัมพันธ์ระหว่างตัวแปรอิสระเพียง 1 คู่ (x_1 กับ x_2) ระดับสหสัมพันธ์ (ρ) ระหว่างตัวแปรอิสระคู่ดังกล่าว 7 ระดับคือ 0.1, 0.3, 0.5, 0.7, 0.9, 0.95 และ 0.99 ส่วนตัวแปรอิสระอีก 1 ตัวที่เหลือ (x_3) ไม่มีสหสัมพันธ์กับตัวแปรอิสระตัวอื่นๆ ซึ่งศึกษาในกรณีที่ตัวแปรอิสระ มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 5 การสร้างตัวแปรอิสระ x_1 กับ x_2 ให้มีสหสัมพันธ์กันใช้วิธีการจำลองของ Wilchern และ Churchill (1978) สูตรในการคำนวณเป็นดังนี้

$$X_{ij} = (1 - \rho^2)^{1/2} \cdot Z_{ij} + \rho Z_{i4} \quad ; i = 1, 2, \dots, n \quad , j = 1, 2, 3$$

เมื่อ $Z_{i1}, \dots, Z_{i4}$ แทน ตัวเลขสุ่มที่มีการแจกแจงปกติมาตรฐานที่เป็นอิสระต่อกัน (Independent Standard Normal Random Number)

ρ แทน ค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ x_1 กับ x_2

4. ขนาดตัวอย่าง (n) จำนวน 6 ระดับคือ 20, 30, 35, 40, 50 และ 60

5. เกณฑ์ที่ใช้ในการพิจารณา เปรียบเทียบประสิทธิภาพของตัวประมาณพารามิเตอร์แต่ละตัว โดยใช้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Squares Error : RMSE) ของตัวประมาณการถดถอยพหุคูณเชิงเส้น สูตรในการคำนวณอยู่ในรูป

$$RMSE_i = \sqrt{\sum_{j=1}^{1,000} \frac{(\beta_{ij} - \hat{\beta}_{ij})^2}{1,000}} \quad ; i = 1, 2, 3, 4 \quad , j = 1, 2, \dots, 1000$$

เมื่อ β_{ij} แทน ค่าจริงของพารามิเตอร์ในสมการถดถอยตัวที่ i ในการจำลองรอบที่ j

$\hat{\beta}_{ij}$ แทน ค่าประมาณของพารามิเตอร์ในสมการถดถอยตัวที่ i ในการจำลองรอบที่ j

$RMSE_i$ แทน ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณค่าพารามิเตอร์ตัวที่ i ในสมการถดถอย

หมายถึง เมื่อทดลองใช้ ค่าอื่นจะให้ผลการเปรียบเทียบตัวประมาณไม่แตกต่างกัน

ตารางที่ 1 แสดงตัวประมาณค่าพารามิเตอร์ที่ควรเลือกใช้ในแต่ละสถานการณ์ เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ (ไม่มีค่าผิดปกติ) และการแจกแจงปกติปลอมปน (มีค่าผิดปกติ) โดยจำแนกตามระดับขนาดตัวอย่าง (n) ระดับค่าผิดปกติในตัวแปรตาม สัดส่วนการปลอมปน และระดับสหสัมพันธ์ของตัวแปรอิสระ x_1 กับ x_2 (ρ)

ขนาด ตัวอย่าง	ระดับค่า ผิดปกติ	สัดส่วน ปลอมปน	ระดับสหสัมพันธ์ของตัวแปรอิสระ x_1 กับ $x_2 (\rho)$						
			0.1	0.3	0.5	0.7	0.9	0.95	0.99
20	ไม่มี	0%	LS	LS	RID	RID	RID	RID	RID
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	เล็กน้อย	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	ปานกลาง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
30	ไม่มี	0%	LS	LS	RID	RID	RID	RID	RID
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	เล็กน้อย	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	ปานกลาง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
35	ไม่มี	0%	LS	LS	LS	RID	RID	RID	RID
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	เล็กน้อย	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	ปานกลาง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	15%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		5%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	RLAV	RLAV	RLAV	RLAV	RLAV

ขนาด ตัวอย่าง	ระดับค่า ผิดปกติ	สัดส่วน ปลอม	ระดับสหสัมพันธ์ของตัวแปรอิสระ x_1 กับ x_2 (ρ)						
			0.1	0.3	0.5	0.7	0.9	0.95	0.99
40	ไม่มี	0%	LS	LS	LS	RID	RID	RID	RID
	เล็กน้อย	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	ปาน กลาง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
50	ไม่มี	0%	LS	LS	LS	RID	RID	RID	RID
	เล็กน้อย	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	ปาน กลาง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
60	ไม่มี	0%	LS	LS	LS	RID	RID	RID	RID
	เล็กน้อย	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	ปาน กลาง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
	รุนแรง	5%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		8%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		10%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV
		15%	LAV	LAV	LAV	RLAV	RLAV	RLAV	RLAV

ผลการทดลองจากตารางที่ 1 แสดงตัวประมาณค่าพารามิเตอร์ที่ให้ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด (RMSE) ในแต่ละสถานการณ์ เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ (ไม่มีค่าผิดปกติ) และการแจกแจงปกติปลอมปน (มีค่าผิดปกติ) โดยจำแนกตามระดับขนาดตัวอย่าง (n) ระดับค่าผิดปกติในตัวแปรตาม สัดส่วนการปลอมปน และระดับสหสัมพันธ์ของตัวแปรอิสระ กับ x_1 กับ x_2 (ρ) สรุปผลการทดลองได้ดังนี้

1. เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ (ไม่มีค่าผิดปกติในตัวแปรตาม)

1.1) เมื่อระดับสหสัมพันธ์ในตัวแปรอิสระมีค่าน้อยกว่าหรือเท่ากับ 0.5

ตัวประมาณกำลังสองน้อยสุด (LS) จะให้ค่า RMSE ต่ำที่สุดในทุกระดับขนาดตัวอย่าง ยกเว้นในกรณีที่ระดับสหสัมพันธ์เท่ากับ 0.5 และขนาดตัวอย่างเท่ากับ 20 และ 30 ตัวประมาณการถดถอยริดจ์ (RID) จะให้ค่า RMSE ต่ำที่สุด

1.2) เมื่อระดับสหสัมพันธ์ในตัวแปรอิสระมีค่ามากกว่า 0.5

ตัวประมาณการถดถอยริดจ์ (RID) จะให้ค่า RMSE ต่ำที่สุดในทุกระดับขนาดตัวอย่าง

2. เมื่อความคลาดเคลื่อนมีการแจกแจงปกติปลอมปน (มีค่าผิดปกติในตัวแปรตาม)

2.1) เมื่อระดับสหสัมพันธ์ในตัวแปรอิสระมีค่าน้อยกว่าหรือเท่ากับ 0.5

ตัวประมาณค่าสัมบูรณ์น้อยสุด (LAV) จะให้ค่า RMSE ต่ำที่สุดในทุกระดับค่าผิดปกติในตัวแปรตาม ทุกระดับของสัดส่วนการปลอมปน และทุกระดับขนาดตัวอย่าง ยกเว้นในกรณีที่ระดับสหสัมพันธ์เท่ากับ 0.5 และขนาดตัวอย่างเท่ากับ 20, 30 และ 35 ในทุกระดับค่าผิดปกติในตัวแปรตามและทุกระดับของสัดส่วนการปลอมปน ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด (RLAV) จะให้ค่า RMSE ต่ำที่สุด

2.2) เมื่อระดับสหสัมพันธ์ในตัวแปรอิสระมีค่ามากกว่า 0.5

ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด (RLAV) จะให้ค่า RMSE ต่ำที่สุดในทุกระดับค่าผิดปกติในตัวแปรตาม ทุกระดับของสัดส่วนการปลอมปน และทุกระดับขนาดตัวอย่าง

3. ผลกระทบจากปัจจัยต่างๆ ที่มีต่อค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) พบว่าค่า RMSE จะแปรผันตามกับระดับค่าผิดปกติในตัวแปรตาม สัดส่วนการปลอมปน และระดับสหสัมพันธ์ในตัวแปรอิสระ แต่จะแปรผกผันกับระดับขนาดตัวอย่าง เมื่อกำหนดให้ปัจจัยอื่นๆ คงที่

การที่จะนำผลการศึกษาที่ได้ในครั้งนี้นำไปประยุกต์ใช้กับข้อมูลจริงในการประมาณค่าหรือพยากรณ์ค่าของตัวแปรตามที่ต้องการ มีหลักในการปฏิบัติดังนี้

1. ผลการวิจัยจะนำไปใช้กับข้อมูลจริงได้ในกรณีที่ความคลาดเคลื่อนสุ่มมีการแจกแจงปกติหรือการแจกแจงปกติปลอมปนเท่านั้น จึงควรทดสอบการแจกแจงของความคลาดเคลื่อนสุ่มก่อนว่ามีการแจกแจงปกติหรือไม่ โดยสามารถทดสอบการแจกแจงของตัวแปรตามแทนได้ หากความคลาดเคลื่อนสุ่มมีการแจกแจงปกติตัวแปรตามก็จะมีการแจกแจงปกติเช่นกัน การทดสอบว่าตัวแปรตามมีการแจกแจงปกติหรือไม่ ใช้สถิติทดสอบ Kolmogorov-Smirnov Test (K-S Test) หรือ Shapiro-Wilk Test

2. ข้อมูลที่จะนำมาวิเคราะห์จะต้องมีจำนวนตัวแปรอิสระ 3 ตัวแปร และมีสหสัมพันธ์ระหว่างตัวแปรอิสระเพียง 1 คู่ โดยตัวแปรอิสระอีกตัวหนึ่งจะต้องไม่มีสหสัมพันธ์กับตัวแปรอิสระตัวอื่นๆ ที่เหลือ

3. การทดสอบว่าระดับพหุสัมพันธ์ของตัวแปรอิสระอยู่ในระดับใด ทดสอบโดยใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) หรืออาจพิจารณาควบคู่กับแผนภาพการกระจาย (Scatter Diagram) ของตัวแปรอิสระแต่ละคู่ว่ามีความสัมพันธ์กันหรือไม่

4. การทดสอบว่ามีค่าผิดปกติในตัวแปรตามหรือไม่ มีค่าผิดปกติอยู่ในระดับใด และมีสัดส่วนของค่าผิดปกติเท่าไรในข้อมูลที่จะนำมาวิเคราะห์ ทดสอบโดยใช้ Box Plot หรืออาจพิจารณาจากแผนภาพการกระจาย (Scatter Diagram) ของข้อมูลของตัวแปรตามว่ามีค่าสังเกตที่มีค่าสูงหรือต่ำกว่าค่าสังเกตส่วนใหญ่ของข้อมูลในชุดเดียวกันหรือไม่

การเลือกใช้ตัวประมาณค่าพารามิเตอร์ที่เหมาะสมจะสามารถพยากรณ์ค่าของตัวแปรตามที่น่าสนใจได้ถูกต้องมากยิ่งขึ้น และสรุปการเลือกใช้ตัวประมาณในสถานการณ์ต่าง ๆ ดังตาราง

ระดับค่าผิดปกติ	สัดส่วนการปลอมปน	ขนาดตัวอย่าง	ระดับสหสัมพันธ์ในตัวแปรอิสระ	ตัวประมาณค่าพารามิเตอร์ที่ควรเลือกใช้
ไม่มีค่าผิดปกติ	0 %	น้อยกว่า 35	น้อยกว่า 0.4	ตัวประมาณค่าสองนัยสุด(LS)
			มากกว่าหรือเท่ากับ 0.4	ตัวประมาณริดจ์(RID)
		35 ถึง 60	น้อยกว่าหรือเท่ากับ 0.6	ตัวประมาณค่าสองนัยสุด(LS)
			มากกว่า 0.6	ตัวประมาณริดจ์(RID)
ทุกระดับ	ทุกระดับ	น้อยกว่า 40	น้อยกว่า 0.4	ตัวประมาณค่าสัมบูรณ์น้อยสุด(LAV)
			มากกว่าหรือเท่ากับ 0.4	ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(RLAV)
		40 ถึง 60	น้อยกว่าหรือเท่ากับ 0.6	ตัวประมาณค่าสัมบูรณ์น้อยสุด(LAV)
			มากกว่า 0.6	ตัวประมาณริดจ์ที่มีค่าสัมบูรณ์น้อยสุด(RLAV)

Bibliography

- Hoerl, A.E. and Kennard, R.W. (2000). Ridge regression : Biased Estimation for Nonorthogonal Problems. **TECHNOMETRICS**, 42, PART 1, pp.69-82.
- Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975). Ridge regression : some simulations. **Communications in Statistics**, 4, pp. 105-123.
- Pfaffenberger, R.C. and Dielman, T.E. (1990). **Robust Regression : Analysis and Applications**. New York : Marcel Dekker.
- Rousseeuw, Peter J. and Leroy, Annick M. (2003). **Robust Regression and Outlier Detection**. New Jersey : John Wiley & Sons.