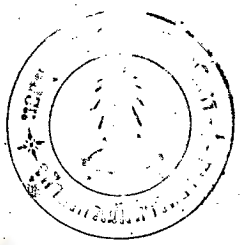
บทที่ ๑

บทนำ



๑.๑ ที่มาของปัญหา

ในการศึกษาการเปลี่ยนแปลงของตัวแปรตาม (Dependent Variable) ว่ามีผลมาจากตัวแปรอิสระ (Independent Variable) ชุดหนึ่งอย่างไรนั้น วิธีหนึ่งของการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเหล่านี้คือการศึกษาความถดถอยเชิงพหุ (Multiple Regression Analysis) ซึ่งถือว่าการใช้ตัวแปรอิสระที่เหมาะสมมากกว่าหนึ่งตัวย่อมทำให้ผลในการประมาณค่าตัวแปรตามถูกต้องมากกว่าใช้ตัวแปรอิสระเพียงตัวเดียว โดยคิดว่าตัวแปรทั้งสองประเภทนี้มีความสัมพันธ์กันในลักษณะใดลักษณะหนึ่ง ซึ่งลักษณะความสัมพันธ์ดังกล่าวอาจเป็นแบบเชิงเส้น (Linear) หรือไม่ใช่เชิงเส้น (Non-Linear) สำหรับตัวแบบ (Model) ทั่วไปของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเชิงเส้นเป็นดังนี้

$$Y = X\beta + \varepsilon$$

โดย Y เป็นเวคเตอร์ของตัวแปรตามขนาด $n \times 1$

X เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times p$

β เป็นเวคเตอร์ของสัมประสิทธิ์ความถดถอยขนาด $p \times 1$

ε เป็นเวคเตอร์ของความคลาดเคลื่อนที่เกิดขึ้นขนาด $n \times 1$

n เป็นขนาดของข้อมูลตัวอย่าง

p เป็นจำนวนสัมประสิทธิ์ความถดถอย

จากรูปแบบเชิงเส้นของตัวแปรตามข้างบนนี้วิธีที่นิยมใช้กันมากในการประมาณค่าสัมประสิทธิ์ความถดถอย คือ วิธีกำลังสองน้อยที่สุด (Least Squares Method)

ซึ่งจะได้ $\hat{\beta} = (X'X)^{-1}X'Y$ หลักการของวิธีนี้คือหาค่า $\hat{\beta}$ ในเทอมของ X และ Y ซึ่งเป็นตัวประมาณที่ไม่เอนเอียง และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ในการประมาณค่าสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุดนั้น มีสมมติฐานที่จำเป็นข้อหนึ่งกล่าวว่าตัวแปรอิสระจะต้องไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ซึ่งกรณีนี้เป็นไปได้น้อย เพราะข้อมูลหรือปัจจัยต่าง ๆ ที่นำมาศึกษาอาจมีความสัมพันธ์กันอยู่ ดังนั้นจะเห็นได้ว่าบางครั้งตัวแปรอิสระที่เราเลือกมาอธิบาย ตัวแปรตามอาจขาดความเหมาะสม คือ เกิด ill-condition ขึ้น เช่น ตัวแปรอิสระบางตัวเป็นฟังก์ชั่นของตัวแปรอิสระอื่น ๆ ลักษณะเช่นนี้อาจสรุปได้ว่าตัวแปรอิสระแต่ละตัวมีลักษณะ Collinear กัน หรือตัวแปรอิสระมีสหสัมพันธ์กัน ในกรณีที่ตัวแปรอิสระเหล่านี้มีสหสัมพันธ์สูงการประมาณค่าตัวแปรตามที่ได้จะไม่เหมาะสม และจะมีผลทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองสูง ดังนั้นในการกำหนดรูปแบบของตัวแปรตามว่ามีผลมาจากตัวแปรอิสระใดบ้างนั้นจำเป็นอย่างยิ่งที่จะต้องคำนึงถึงสิ่งต่าง ๆ เหล่านี้คือ ตัวแปรอิสระมีค่าสหสัมพันธ์สูงหรือไม่ ถ้าพบว่าตัวแปรอิสระมีค่าสหสัมพันธ์สูงก็ควรตัดตัวแปรอิสระบางตัวที่ก่อปัญหาออกจากตัวแบบก่อนที่จะทำการประมาณค่าสัมประสิทธิ์ความถดถอย เพื่อหลีกเลี่ยงข้อผิดพลาดที่อาจเกิดขึ้นในการประมาณพารามิเตอร์ β

ในกรณีที่ตัวแปรอิสระมีค่าสหสัมพันธ์กันสูง ลักษณะดังกล่าวนี้เรียกว่าพหุสัมพันธ์ในตัวแปรอิสระ (Multicollinearity) และเราไม่ต้องการตัดตัวแปรอิสระตัวใดตัวหนึ่งออกไปจากตัวแบบ เพราะถือว่าตัวแปรอิสระทุกตัวที่เลือกเข้ามาในตัวแบบนี้ ต่างมีความสำคัญและมีผลต่อการเปลี่ยนแปลงของตัวแปรตามมากพอสมควร ดังนั้นเมื่อประมาณค่าสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุดจะมีผลทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองสูงขึ้นด้วย ฉะนั้นในการประมาณค่าสัมประสิทธิ์ความถดถอยก็น่าจะพิจารณาค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่เกิดขึ้นด้วย หากพบว่ามีวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยใดที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุดก็น่าที่จะนำวิธีนั้นมาใช้ในการประมาณค่าสัมประสิทธิ์ความถดถอย ซึ่งเมื่อประมาณปี พ.ศ. ๒๕๑๓

Hoerl และ Kennard ได้ศึกษาพบว่าวิธีหนึ่งของการประมาณค่าสัมประสิทธิ์ความถดถอยที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุดคือ วิธีริดจ์ รีเกรสชั่น[๑] (Ridge Regression) วิธีนี้มีข้อดีคือไม่จำเป็นต้องตัดตัวแปรอิสระออกจากตัวแบบ ถึงแม้ว่าจะเกิดพหุสัมพันธ์ในตัวแปรอิสระ และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด

อย่างไรก็ดีการประมาณค่าสัมประสิทธิ์ความถดถอยด้วยวิธี ริดจ์ รีเกรสชั่น จะให้ตัวประมาณที่เอนเอียง (biased estimator)

หลักการของวิธี ริดจ์ รีเกรสชั่น นอกจากลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้น้อยลง ในบรรดากลุ่มของตัวประมาณค่าที่เอนเอียง และไม่เอนเอียงแล้วยังเป็นการปรับค่าประมาณพารามิเตอร์ β ของตัวแปรอิสระใหม่ ซึ่ง Hoerl และ Kennard ได้เสนอแนะวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยของริดจ์ รีเกรสชั่น โดยพิจารณาจากวิธีกำลังสองน้อยที่สุด ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเป็นฟังก์ชั่นของ $(X'X)^{-1}$ ฉนั้น การที่จะลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้ต่ำลงจึงต้องมุ่งไปที่การลดค่า $(X'X)^{-1}$ ให้ต่ำลง วิธีการของริดจ์ รีเกรสชั่น อาจสามารถปรับค่า $(X'X)^{-1}$ ให้ต่ำลงได้ โดยการบวกค่าคงที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุม (Diagonal) ของเมทริกซ์ $(X'X)^{-1}$ ค่าที่ปรับใหม่ของเมทริกซ์ $(X'X)^{-1}$ คือ เมทริกซ์ $(X'X + KI)^{-1}$ เมื่อ $K > 0$ จึงน่าที่จะได้ศึกษาเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยระหว่างวิธีกำลังสองน้อยที่สุด และวิธีริดจ์ รีเกรสชั่น ว่าให้ผลแตกต่างกันเช่นไร และแต่ละวิธีจะใช้เหมาะสมกับข้อมูลประเภทใดบ้าง

---

[๑]Arthur E. Hoerl and Robert W. Kennard, "Ridge Regression : Biased Estimation for Nonorthogonal Problem," Technometrics (February 1970) : p. 55-66.

๑.๒ วัตถุประสงค์ของการวิจัย

๑. ศึกษาการประมาณค่าพารามิเตอร์ในสมการถดถอยโดยวิธีริดจ์ รีเกรสชั่น

๒. เพื่อเปรียบเทียบการประมาณค่าพารามิเตอร์ในสมการถดถอยระหว่างวิธีกำลังสองน้อยที่สุดและวิธีริดจ์ รีเกรสชั่น

๑.๓ ขอบเขตของการวิจัย

๑. ในการศึกษาครั้งนี้จะทำการเปรียบเทียบการประมาณค่าสัมประสิทธิ์ความถดถอยระหว่างวิธีกำลังสองน้อยที่สุด และวิธีริดจ์ รีเกรสชั่น โดยพิจารณาค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเป็นเกณฑ์

๒. ข้อมูลที่ใช้เป็นข้อมูลทุติยภูมิ (Secondary data) โดยเก็บรวบรวมจากแหล่งต่าง ๆ ที่เกี่ยวข้องและเอกสารต่าง ๆ ที่มีอยู่ โดยข้อมูลแต่ละชุดอยู่ระหว่าง ๙ - ๑๕ ปี และในการศึกษาครั้งนี้รูปแบบจะอยู่ในลักษณะเชิงเส้นซึ่งจะใช้ข้อมูล ๕ ชุด ในการศึกษาเปรียบเทียบการประมาณค่าสัมประสิทธิ์ความถดถอยทั้ง ๒ วิธีดังกล่าว

๓. ข้อมูลทั้ง ๕ ชุดที่นำมาศึกษาครั้งนี้ได้เลือกเอาข้อมูลชุดที่ตัวแปรอิสระมีค่าสหสัมพันธ์อยู่ระหว่าง ๐.๗๐ - ๐.๙๙ เป็นส่วนมาก

๑.๔ ข้อตกลงเบื้องต้น

๑. ในการกำหนดตัวแบบสำหรับข้อมูลทั้ง ๕ ชุดตัวแบบที่ใช้ศึกษาอยู่ในลักษณะเชิงเส้นดังนี้

$$Y = X\beta + \varepsilon$$

โดย $\varepsilon \sim N(0 , \sigma^2 I_n)$, $E(\varepsilon'\varepsilon) \neq 0$

๒. ในการวิเคราะห์ค่าสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุดนั้นจะใช้วิธีที่มีตัวแปรอิสระทุกตัวในตัวแบบ และวิธี Step - Wise Multiple Regression

ส่วนวิธีการวิเคราะห์สัมประสิทธิ์ความถดถอยโดยวิธีริดจ์ รีเกรสชั่น จะใช้วิธี Ordinary Ridge Regression

๑.๕ สมมติฐานในการวิจัย

สมมติฐานในการวิจัยกรณีที่ตัวแปรอิสระมีค่าสหสัมพันธ์สูงเป็นดังนี้

การประมาณค่าสัมประสิทธิ์ความถดถอยโดยวิธีริดจ์ รีเกรสชั่น ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด

๑.๖ ประโยชน์ที่คาดว่าจะได้รับ

๑. การวิจัยนี้จะเป็นแนวทางให้นักวิจัยเข้าใจวิธีวิเคราะห์ความถดถอยพหุเชิงเส้นโดยวิธี ริดจ์ รีเกรสชั่น ได้โดยง่าย

๒. เป็นแนวทางในการตัดสินใจว่าจะใช้วิธีใดในการวิเคราะห์ข้อมูลที่มีปัญหาแต่ละประเภท

๓. จากโปรแกรมคอมพิวเตอร์สำหรับหาค่าตัวประมาณ $\hat{\beta}^*$ ซึ่งแสดงไว้ในภาคผนวกสามารถนำไปใช้ประมาณค่าได้สะดวกขึ้น

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย