บทที่ 1

บทนำ

1.1 ความเป็นมาของปัญหา

ในการวิเคราะห์ความถดถอยนั้น ปัญหาที่ผู้วิเคราะห์ต้องเผชิญในทางปฏิบัติ 2 ประการที่มีความสำคัญไม่ยิ่งหย่อนไปกว่ากันก็คือ

1. ปัญหาข้อมูลสูญหาย
2. ปัญหา Multicolinearity

สำหรับปัญหาข้อมูลสูญหาย ผู้วิเคราะห์จำเป็นต้องคัดตัวอย่างที่ไม่สมบูรณ์ทิ้งไป จะนำไปใช้ร่วมเป็นข้อมูลสำหรับการวิเคราะห์มิได้ผลกระทบที่ได้รับจากวิธีปฏิบัติดังกล่าวก็คือ ขนาดตัวอย่างจะมีค่าเล็กกว่าเดิม และถ้ามีจำนวนตัวอย่างที่ไม่สมบูรณ์ที่ต้องคัดทิ้งเป็นจำนวนมาก เช่น ในกรณีของอนุกรมเวลา ขนาดตัวอย่างจะเล็กลง จนกระทั่งมีจำนวนใกล้เคียงกับจำนวนตัวแปรอิสระ ผลที่ตามมาก็คือ MSE (MSE = SSE/n-p = $\hat{\sigma}^2$) มีค่าสูงมาก มีผลให้สมการประมาณค่าขาดความน่าเชื่อถือ

สำหรับปัญหานี้ได้มีผู้ทำการศึกษาไว้หลายท่าน เอฟิฟีและอีลาช่อฟ (A.A. Afifi and R.M. Elashoff) ทำการศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการประมาณค่าข้อมูลที่สูญหายไปในกรณีของสมการถดถอยเชิงเส้นอย่างง่ายรวม 11 วิธี ต่อมา เคเลเจียน (H.H. Kelejian) ได้ศึกษาปัญหาเดียวกัน โดยนำวิธีการไปใช้กับสมการถดถอยพหุคูณ นอกจากนี้ก็ยังมีผู้ที่ได้ทำการศึกษาปัญหาเดียวกันนี้อีกหลายครั้ง ซึ่งในบรรดาวิธีประมาณค่าสังเกตที่สูญหายที่มีอยู่หลายวิธีนั้น วิธีที่ได้รับความนิยมนำไปใช้ในการประมาณค่าสังเกตที่สูญหายอยู่เสมอ มีอยู่ด้วยกัน 2 วิธี คือ วิธีแทนด้วยค่าเฉลี่ย (Mean Substitution) และวิธีสมการถดถอย (Regression)

ในกรณีที่ตัวแปรอิสระมีสหสัมพันธ์ค่อนข้างสูง กล่าวคือ $r_{st} \to 1$; $s \neq t$: $s,t = 1, ..., P$ ซึ่งแสดงว่าตัวแปร X_s และ X_t มีความสัมพันธ์ต่อกันในปริมาณที่ค่อนข้างสูงจนอาจสามารถใช้แทนกันได้ ซึ่งเป็นปัญหาที่เรียกว่า Multicolinearity แล้ว ผู้วิเคราะห์จำเป็นต้องหาทางแก้ปัญหา Multicolinearity ดังกล่าวพร้อมกันไปกับการแก้ปัญหาข้อมูลสูญหาย

ปัญหา Multicolinearity คือ ปัญหาที่เกิดขึ้นจากสถานการณ์ที่ตัวแปรอิสระมิได้เป็นอิสระต่อกัน หมายความว่า เวกเตอร์ใด ๆ ของเมตริกซ์ X อาจเกิดขึ้นจากการประกอบกันเชิงเส้นของเวกเตอร์อื่น ๆ (Linear Combination) ผลสะท้อนที่ปรากฏแก่งานวิเคราะห์สมการถดถอยก็คือ $\det(X'X) \rightarrow 0$ (Near Singularity) ซึ่งจะมีผลให้สมาชิกของเมตริกซ์ $X'X$ มีค่าสูงมาก และ $\hat{V}(\hat{\beta}_i) = \sigma^2 (X'X)^{-1}_{ii}$ มีค่าสูง เมื่อ $(X'X)^{-1}_{ii}$ คือ สมาชิกลำดับที่ (i,i) ของเมตริกซ์ $(X'X)^{-1}$ งานวิเคราะห์สมการถดถอยจึงขาดความน่าเชื่อถือ และในกรณีที่ดีเทอร์มิแนนต์ของเมตริกซ์ $X'X$ มีค่าเท่ากับ 0 กล่าวคือ $\det(X'X) = 0$ ซึ่งเป็นกรณีของปัญหา Multicolinearity อย่างสมบูรณ์ (Perfect Multicolinearity) จะพบว่า $(X'X)^{-1}$ ไม่ปรากฏค่า การวิเคราะห์สมการถดถอยจะล้มเหลวอย่างสิ้นเชิง

Ridge Regression เป็นวิธีประมาณค่าพารามิเตอร์สำหรับแก้ปัญหา Multicolinearity โดยตรง โดยมีหลักการที่พัฒนาขึ้นมาจากความคิดที่ว่าเราสามารถแปลงรูปเมตริกซ์ขนาด P x Pใด ๆ ให้เป็น Diagonal matrix Λ โดยที่ $\Lambda = \text{diag.} (\lambda_1, \lambda_2, \ldots, \lambda_p)$ เมื่อ λ_i , $i = 1,2,\ldots, P$ คือ Eigen value ของเมตริกซ์ดังกล่าว และ $\Lambda^{-1} = \text{diag.} (\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_p^{-1})$ และถ้าเมตริกซ์ใดมีธรรมชาติใกล้เคียงกับ Singular matrix มาก λ_{min} จะมีค่าใกล้ 0 อันมีผลให้ $\lambda_{min}^{-1} \rightarrow \infty$ และ $\det \Lambda \rightarrow 0$ ถ้าเมตริกซ์ขนาด P x P ดังกล่าวคือเมตริกซ์ $(X'X)$ เราจะไม่อาจคำนวณหา $(X'X)^{-1}$ และ $\hat{\beta}$ เมื่อ $\hat{\beta} = (X'X)^{-1} X'Y$ ได้เลยถ้า $\lambda_{min} \rightarrow 0$ ดังนั้น หลักการทั่วไปของ Ridge Regression ก็คือ การเพิ่มค่าให้แก่ λ_i แต่ละตัวเท่า ๆ กัน (วิธีนี้คือ Simple Ridge Regression คือเปลี่ยน $\Lambda = \text{diag.} (\lambda_1, \lambda_2, \ldots, \lambda_p)$ เป็น $\Lambda = \text{diag.} (\lambda_1 + k, \lambda_2 + k, \ldots, \lambda_p + k)$ โดยที่ $0 \leqslant k \leqslant 1$) ผลที่ตามมาก็คือ λ_{min} มีค่าสูงขึ้นซึ่งจะไม่เป็นอุปสรรคในการหา $(X'X)^{-1}$และ $\hat{\beta}$ อีกต่อไป และโดยทั่วไปเรานิยมเพิ่มค่าให้แก่ $(X'X)_{ii}$ คือเปลี่ยนจาก $(X'X)$ เป็น $(X'X + kI_p)$

สังเกตว่า $\hat{\beta}_{LS} = (X'X)^{-1} X'Y = WY$ เมื่อ W คือ เมตริกซ์ขนาด P x n ของตัวถ่วงน้ำหนัก เมื่อเราเปลี่ยนตัวถ่วงน้ำหนักจาก $W = (X'X)^{-1} X'$ เป็น $W^* = (X'X + kI_p)^{-1} X$ ผลที่ตามมาก็คือ $\hat{\beta}^*(k) = (X'X + kI_p)^{-1} X'Y$ เป็น Biased estimator แต่ $\hat{\beta}^*(k)$ ให้ค่า MSE ต่ำกว่า $\hat{\beta}_{LS}$

จากการศึกษาของวิชเช่อร์นและเช่อร์ชีล (Wichern, D.W. and Churchill, G.A., 1978) กิบบอน (Gibbons, D.G. 1981) แมคโดนาล์ดและกาลานู (Mcdonald, G.C. and Galarnesu, D.I., 1975) พบว่าในจำนวนวิธีประมาณค่า k มากกว่า 10 วิธีนั้น วิธีที่ดีที่สุดและเป็นที่นิยมใช้กันมีอยู่ 2 วิธีคือ

1. วิธี Hoerl, Kennard and Baldwin (HKB)
2. วิธี Lawless and Wang (LW)

ปัญหาที่พบก็คือ การผสมผสานเทคนิคสำหรับแก้ปัญหาข้อมูลสูญหายกับเทคนิคการแก้ปัญหา Multicolinearity เข้าด้วยกันจะให้ผลดีเพียงใด สำหรับปัญหาดังกล่าวนี้ ได้เคยมีการศึกษากันมาแล้ว แต่การศึกษาในครั้งนั้นเป็นการศึกษาที่มีขอบเขตค่อนข้างจะจำกัดในการนำไปใช้ในเชิงปฏิบัติ เพราะเป็นการศึกษาปัญหาในกรณีที่เกิด Multicolinearity ในสถานการณ์ที่มีเฉพาะตัวแปรอิสระเพียงตัวเดียวเท่านั้น คือตัวแปรอิสระ X_1 ที่เกิดปัญหาข้อมูลสูญหาย ผลการศึกษาดังกล่าวนี้ แม้จะช่วยแก้ปัญหาในกรณีที่เกิดปัญหา Multicolinearity พร้อมกันไปกับการเกิดปัญหาข้อมูลสูญหายในระดับหนึ่ง แต่ในทางปฏิบัติแล้ว ผู้วิเคราะห์มักจะประสบกับกรณีที่ตัวแปรอิสระที่จะนำมาใช้ในการวิเคราะห์สมการถดถอยเกิดปัญหา Multicolinearity พร้อมกันนั้นก็เกิดปัญหาข้อมูลสูญหายในหลายตัวแปรอิสระ ซึ่งเมื่อเกิดปัญหาดังกล่าวนี้ ผู้วิเคราะห์ก็ไม่สามารถที่จะนำผลการศึกษาที่มีอยู่เดิมมาใช้ในการแก้ปัญหาได้

ดังนั้นในการศึกษาครั้งนี้ จึงมุ่งศึกษาในกรณีที่เกิดปัญหา Multicolinearity ในสถานการณ์ที่ตัวแปรอิสระหลายตัวเกิดปัญหาข้อมูลสูญหาย เพื่อหาวิธีการที่เหมาะสมในการแก้ปัญหาดังกล่าว โดยใช้วิธีการดังต่อไปนี้

1. วิธี Mean-Hoerl, Kennard and Baldwin
2. วิธี Mean-Lawless and Wang
3. วิธี Regression-Hoerl, Kennard and Baldwin
4. วิธี Regression-Lawless and Wang
5. วิธี Ordinary Least Square

1.2 <u>วัตถุประสงค์ของการวิจัย</u>

เพื่อเปรียบเทียบวิธีประมาณค่าทั้ง 5 วิธี เมื่อตัวแปรอิสระมีข้อมูลบางส่วนสูญหาย และตัวแปรอิสระบางคู่หรือทุกคู่มีสหสัมพันธ์ต่อกันค่อนข้างสูง

1.3 สมมติฐานของการวิจัย

ในการวิจัยครั้งนี้ วิธี Regression-Hoerl, Kennard and Baldwin จะเป็นวิธีที่เหมาะสมที่สุดในขณะที่วิธี Ordinary Least Square จะเป็นวิธีที่เชื่อถือได้น้อยที่สุด

1.4 ขอบเขตของการวิจัย

1. ตัวแปรอิสระที่ใช้ในการศึกษามี 5 ตัว คือ X_1 , X_2, ..., X_5

2. ขนาดตัวอย่างกำหนดไว้เป็น 2 ระดับ คือ $n = 20,30$ โดยมุ่งศึกษาเฉพาะกรณีตัวอย่างขนาดเล็กเท่านั้น

3. การสูญหายของข้อมูลกำหนดให้ปรากฏขึ้นโดยสุ่ม และกำหนดให้ตัวแปรอิสระทุกตัวมีอัตราการสูญหายของข้อมูลแตกต่างกันไปโดยสุ่มตั้งแต่ 5-15%

4. $X_i \sim N(0,1)$ และ $U \sim N(0, \sigma^2)$ โดยที่ $\sigma^2 = .01, .10, .50, 1.0$ 5.0

5. กำหนดค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ เพื่อบังคับให้ตัวแปรอิสระเกิด Multicolinearity ในระดับที่ต้องการ เป็น 5 ระดับด้วยกัน โดยจัดแบ่งตามความรุนแรงของปัญหา Multicolinearity ดังนี้

(1) $(\rho, \rho^*) = (.99,.99),(.99,.10)$ แสดงว่ามีปัญหา Multicolinearity มาก

(2) $(\rho, \rho^*) = (.90,.90),(.90,.10)$ แสดงว่ามีปัญหา Multicolinearity ปานกลาง

(3) $(\rho, \rho^*) = (.70,.30)$ แสดงว่ามีปัญหา Multicolinearity น้อย

โดยที่ ρ คือ สหสัมพันธ์ระหว่าง X_1 กับ X_2 , X_1 กับ X_3 และ X_2 กับ X_3

ρ^* คือ สหสัมพันธ์ระหว่าง X_4 กับ X_5

6. ข้อมูลของตัวแปรสร้างขึ้นโดยสุ่มในแต่ละลักษณะจำนวน 15 ครั้ง

1.5 <u>ประโยชน์ที่คาดว่าจะได้รับ</u>

ทำให้สามารถเลือกใช้เทคนิคการประมาณค่าพารามิเตอร์ในสมการถดถอยพหุคูณ เมื่อตัวแปรอิสระมีข้อมูลสูญหายไปบางส่วนและตัวแปรอิสระบางคู่มีสหสัมพันธ์ค่อนข้างสูงได้อย่างเหมาะสม