การประมาณค่าแบบริดจ์เมื่อมีค่าสังเกตสูญหาย

นายวิวัฒน์ ศกลศนธิเศรษฐ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต

ภาควิชาสถิติ

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2532

ISBN 974-576-799-9

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

015978

RIDGE ESTIMATION WITH MISSING OBSERVATIONS

Mr. Wiwat Sakonsontiset

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Statistics

Department of Statistics

Graduate School

Chulalongkorn University

1989

ISBN 974-576-799-9

หัวข้อวิทยานิพนธ์ การประมาณค่าแบบริดจ์เมื่อมีค่าสังเกตสูญหาย

โดย นายวิวัฒน์ สกลสันธิเศรษฐ์

ภาควิชา สถิติ

อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. สมชัย พิศาลบุตร

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย

(ศาสตราจารย์ ดร. ถาวร วัชราภัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ มณฑา พัววิไล)

..... กรรมการ

(รองศาสตราจารย์ ดร. สมชัย พิศาลบุตร)

..... กรรมการ

(รองศาสตราจารย์ นฟรัตน์ รุ่งอุทัยศิริ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. ธีระพร วีระถาวร)

วิวัฒน์ สกลสันธิเศรษฐ์ : ความประมาณค่าแบบริดจ์เมื่อมีค่าสังเกตสูญหาย (RIDGE ESTIMATION WITH MISSING OBSERVATIONS) อ. ที่ปรึกษา : รศ. ดร. สรชัย พิศาลบุตร 126 หน้า

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีประมาณค่าพารามิเตอร์ β ในสมการถดถอย $Y = X\beta + U$ เมื่อเกิดปัญหาค่าสังเกตสูญหายและปัญหา Multicolinearity พร้อม ๆ กัน วิธีประมาณค่าที่ใช้ในการศึกษามี 5 วิธีคือ วิธี Mean-Hoerl, Kennard and Baldwin, วิธี Mean-Lawless and Wang, วิธี Regression-Hoerl, Kennard and Baldwin, วิธี Regression-Lawless and Wang และวิธี Ordinary Least Square

การศึกษาวิจัยได้ใช้เทคนิคการจำลองแบบด้วยเครื่องคอมพิวเตอร์ สร้างตัวแปรอิสระที่มีการแจกแจงแบบปกติ 5 ตัวแปร โดยกำหนดค่าสหสัมพันธ์คือ (ρ,ρ^*)= (.99,.10), (.99,.99), (.90,.10), (.90,.90) และ (.70, .30) พร้อมทั้งแปรค่าของ σ^2 ไปเป็น 5 ระดับคือ σ^2 = .01 .10 .50 1.0 และ 5.0 กำหนดให้ตัวแปรอิสระทุกตัวมีข้อมูลสูญหายโดยผันแปรไปโดยสุ่ม ตั้งแต่ 5-15% และใช้ตัวอย่างขนาด 20 และ 30 การคำนวณหาตัวประมาณค่าพารามิเตอร์ β ทั้ง 5 วิธี ได้ทำการจำลองซ้ำ ๆ กัน 15 ครั้ง การเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าแต่ละวิธีพิจารณาจากค่าผลรวมของความคลาดเคลื่อนกำลังสองเฉลี่ย

ผลการวิจัยพบว่า วิธีประมาณค่าแบบ Ordinary Least Square ยังคงใช้ได้ดีถ้าตัวแปรตาม Y มีค่าที่คล้ายคลึงกัน แต่ถ้าตัวแปรตาม Y มีค่าความแปรปรวนมาก และตัวอย่างมีขนาดใหญ่ขึ้น วิธีประมาณค่าแบบ Regression-Hoerl, Kennard and Baldwin จะเป็นวิธีประมาณค่าที่เหมาะสมกว่า อย่างไรก็ตามในสถานการณ์อื่น ๆ พบว่า โดยทั่วไปแล้ววิธีประมาณค่าแบบ Mean-Hoerl, Kennard and Baldwin เป็นวิธีประมาณค่าที่เหมาะสมที่สุด

ภาควิชา สถิติ ลายมือชื่อนิสิต วิวัฒน์ สกลสันธิเศรษฐ์

สาขาวิชา สถิติ

ปีการศึกษา 2532 ลายมือชื่ออาจารย์ที่ปรึกษา .....

ทอดมูลกรณ์ สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

WIWAT SAKONSONTISET : RIDGE ESTIMATION WITH MISSING OBSERVATIONS. THESIS ADVISOR. ASSO. PROF. SORACHAI BHISALBUTRA, Ph.D. 126 PP.

The objectives of this research is to make use of methods in estimating parameter β , in the regression model $Y = X\beta + U$ when problem of missing observations and multicolinearity is encountered. Five methods for estimating parameter were Mean-Hoerl, Kennard and Badwin, Mean-Lawless and Wang, Regression-Hoerl, Kennard and Baldwin, Regression-Lawless and Wang and Ordinary Least Square.

This study made use of computer in generating data by simulation technique, simulate poeudo normal of five independent variable, which correlation coefficients were $(\rho,\rho^*) = (.99,.10)$, $(.99,.99)$,$(.90,.10)$, $(.90,.90)$ and $(.70,.30)$, and varying the value of $\sigma^2 = .01$, .10, .50, 1.0 and 5.0 respectively. Every independent variable assigned to be missing randomly, and varying from 5 to 10 percent. The sample of size 20 and 30 were used Simulation in each case was repeated 15 times and each time calculate for each estimator. In order to compare these five methods the Total Mean Square Errors were used.

The results of this research were the following if dependent variables Y are highly homogeneous, the Ordinary Least Square method would be a worthwhile candidate, but when large sample and variance of dependent variables Y are maximum, the Regression-Hoerl, Kennard and Baldwin method tend to be the best estimation method, however, the Mean-Hoerl, Kennard and Baldwin method is more appropriate than another four methods in other cases.

ภาควิชา สถิติ ลายมือชื่อนิสิต วิวัฒน์ ศกลสนธิเศรษฐ์

สาขาวิชา สถิติ

ปีการศึกษา 2532 ลายมือชื่ออาจารย์ที่ปรึกษา .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ รองศาสตราจารย์ ดร. สรชัย พิศาลบุตร อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำ และข้อคิดเห็นต่าง ๆ ของการวิจัยมาด้วยดีตลอด จึงขอกราบขอบพระคุณอาจารย์มา ณ ที่นี้ด้วย

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา ซึ่งสนับสนุนในด้านการศึกษา และให้กำลังใจแก่ผู้วิจัยเสมอมา และขอกราบขอบพระคุณ อาจารย์ทุกท่านที่ประสิทธิ์ประสาทวิชาความรู้ให้โดยตลอด จนสำเร็จการศึกษา

วิวัฒน์ สกลสันธิเศรษฐ์

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง

บทคัดย่อภาษาอังกฤษ ..... จ

กิตติกรรมประกาศ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญรูป ..... ฎ

บทที่ 1 บทนำ 1

1.1 ความเป็นมาของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 3

1.3 สมมติฐานของการวิจัย 4

1.4 ขอบเขตของการวิจัย 4

1.5 ประโยชน์ที่คาดว่าจะได้รับ 5

บทที่ 2 ทฤษฎีที่เกี่ยวข้องในการวิจัย 6

2.1 Ridge Regression 6

2.2 การประมาณค่าสังเกตที่สูญหาย 26

บทที่ 3 วิธีดำเนินการวิจัย 28

3.1 วิธีการวิจัย 28

3.2 สรุปวิธีประมาณค่าพารามิเตอร์ 29

3.3 การสร้างข้อมูลของตัวแปรอิสระ 32

3.4 การสร้างข้อมูลของตัวแปรตาม 32

3.5 ขั้นตอนการประมาณค่าสังเกตที่สูญหาย 33

3.6 วิธีวิเคราะห์ข้อมูล 35

3.7 โปรแกรมที่ใช้ในการวิจัย 36

สารบัญ (ต่อ)

หน้า

บทที่ 4 สรุปผลการวิจัยและข้อเสนอแนะ 38

4.1 ผลการวิจัย 38

4.2 สรุปผลการวิจัยและข้อเสนอแนะ 73

บรรณานุกรม 75

ภาคผนวก 77

ประวัติผู้เขียน 126

สารบัญตาราง

หน้า

ตารางที่

1 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงมาก $[(\rho,\rho^*) = (.99,.99)]$ และขนาดตัวอย่างเท่ากับ 20 41

2 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงมาก $[(\rho,\rho^*) = (.99,.10)]$ และขนาดตัวอย่างเท่ากับ 20 43

3 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงมาก $[(\rho,\rho^*) = (.99,.10)]$ และขนาดตัวอย่างเท่ากับ 30 45

4 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงมาก $[(\rho,\rho^*) = (.90,.10)]$ และขนาดตัวอย่างเท่ากับ 30 47

5 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงปานกลาง $[(\rho,\rho^*) = (.90,.90)]$ และขนาดตัวอย่างเท่ากับ 20 51

6 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรสุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงปานกลาง $[(\rho,\rho^*) = (.90,.90)]$ และขนาดตัวอย่างเท่ากับ 20 53

สารบัญตาราง

หน้า

ตารางที่

7 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรลุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงปานกลาง $[(\rho,\rho^*) = (.90,.10)]$ และขนาดตัวอย่างเท่ากับ 30 55

8 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรลุ่ม U เมื่อมีปัญหา Multicolinearity รุนแรงปานกลาง $[(\rho,\rho^*) = (.90,.90)]$ และขนาดตัวอย่างเท่ากับ 30 57

9 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรลุ่ม U เมื่อมีปัญหา Multicolinearity ไม่รุนแรง $[(\rho,\rho^*) = (.70,.30)]$ และขนาดตัวอย่างเท่ากับ 20 60

10 แสดงการเปรียบเทียบวิธีประมาณค่าจำแนกตามความแปรปรวนของตัวแปรลุ่ม U เมื่อมีปัญหา Multicolinearity ไม่รุนแรง $[(\rho,\rho^*) = (.70,.30)]$ และขนาดตัวอย่างเท่ากับ 30 62

11 แสดงค่าสัมประสิทธิ์ของการผันแปร (coefficient of variation) ของการทดลองซ้ำ 15 ครั้ง 65

12 แสดงค่าสัมประสิทธิ์ของการผันแปร (coefficient of variation) ของการทดลองซ้ำ 25 ครั้ง 67

13 แสดงค่าสัมประสิทธิ์ของการผันแปร (coefficient of variation) ของการทดลองซ้ำ 35 ครั้ง 69

14 แสดงค่าสัมประสิทธิ์ของการผันแปร (coefficient of variation) ของการทดลองซ้ำ 40 ครั้ง 71

สารบัญรูป

รูปที่		หน้า
1	ผังงานของโปรแกรมที่ใช้ในการวิจัย	37
2	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.99,.99)]$ และขนาดตัวอย่างเท่ากับ 20	42
3	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.99,.10)]$ และขนาดตัวอย่างเท่ากับ 20	44
4	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.99,.99)]$ และขนาดตัวอย่างเท่ากับ 30	46
5	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.99,.10)]$ และขนาดตัวอย่างเท่ากับ 30	48
6	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho, \rho^*) = (.99,.10)]$ และขนาดตัวอย่างเท่ากับ 20	52
7	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.90,.10)]$ และขนาดตัวอย่างเท่ากับ 20	54
8	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.90,.10)]$ และขนาดตัวอย่างเท่ากับ 30	56
9	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.90,.90)]$ และขนาดตัวอย่างเท่ากับ 30	58
10	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.70,.30)]$ และขนาดตัวอย่างเท่ากับ 20	61
11	กราฟแสดงค่า TMSE ในกรณีที่ $[(\rho,\rho^*) = (.70,.30)]$ และขนาดตัวอย่างเท่ากับ 30	63
12	ผังงานโดยสรุปในการนำผลการวิจัยไปใช้ในทางปฏิบัติ	74

คำอธิบายสัญญลักษณ์และคำย่อ

n คือ ขนาดตัวอย่าง

ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_1 กับ X_2, X_1 กับ X_3 และ X_2 กับ X_3

ρ^* คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_4 กับ X_5

TMSE คือ ค่า Total Mean Square Error

Mean-HKB คือ วิธีประมาณค่าแบบ Mean-Hoerl, Kennard and Baldwin

Mean-LW คือ วิธีประมาณค่าแบบ Mean-Lawless and Wang

Regression-HKB คือ วิธีประมาณค่าแบบ Regression-Hoerl, Kennard and Beldwin

Regression-LW คือ วิธีประมาณค่าแบบ Regression-Lawless and Wang

OLS คือ วิธีประมาณค่าแบบ Ordinary Least Square