การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ

นางสาวพรศิริ หมื่นไชยศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต

ภาควิชาสถิติ

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2529

ISBN 974-566-475-8

A COMPARISON OF MISSING VALUES ESTIMATION METHODS IN MULTIVARIATE ANALYSIS

Miss Pornsiri Muenchaisri

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Statistics

Graduate School

Chulalongkorn University

1986

ISBN 974-566-475-8

หัวข้อวิทยานิพนธ์ การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ

โดย นางสาว พรศิริ หมื่นไชยศรี

ภาควิชา สถิติ

อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. สรชัย พิศาลบุตร

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....

(รองศาสตราจารย์ ดร. สรชัย พิศาลบุตร)

รักษาการในตำแหน่งรองคณบดีฝ่ายวิชาการ

ปฏิบัติราชการแทนรักษาการในตำแหน่งคณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. สุชาดา กีระนันทน์)

..... กรรมการ

(รองศาสตราจารย์ ดร. สรชัย พิศาลบุตร)

..... กรรมการ

(รองศาสตราจารย์ กัลยา ครองแก้ว)

..... กรรมการ

(อาจารย์ ดร. สุพล ดุรงค์วัฒนา)

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

หัวข้อวิทยานิพนธ์ การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ

ชื่อนิสิต นางสาว พรศิริ หมื่นไชยศรี

อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. สัมฤทัย พิศาลบุตร

ภาควิชา สถิติ

ปีการศึกษา 2528

บทคัดย่อ

ในการวิเคราะห์ตัวแปรพหุ หากมีปัญหาเกี่ยวกับข้อมูลบางตัวสูญหายจะทำให้ไม่สามารถวิเคราะห์ข้อมูลได้ วิธีการแก้ปัญหาอย่างหนึ่งก็คือ ตัดค่าสังเกตชุดนั้นทิ้งไป แต่การแก้ปัญหาโดยวิธีนี้จะมีผลทำให้จำนวนค่าสังเกตน้อยลง และสูญเสียรายละเอียดของข้อมูลบางตัวไป วิธีการแก้ปัญหาอีกวิธีหนึ่งก็คือ ต้องทำการประมาณค่าสูญหายนั้น แต่เนื่องจากวิธีการประมาณค่าสูญหายมีหลายวิธีซึ่งแต่ละวิธีต่างก็มีข้อดีและข้อเสียแตกต่างกันไป ดังนั้นการวิจัยนี้จึงสนใจเปรียบเทียบวิธีการประมาณค่าสูญหายที่นิยมใช้กันทั่วไป 4 วิธีคือ วิธีที่ใช้ค่าเฉลี่ย วิธีวิเคราะห์ความถดถอยพหุเชิงเส้น วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นตัดแปลง และวิธีวิเคราะห์ส่วนประกอบหลัก โดยใช้ค่าความคลาดเคลื่อนเฉลี่ยเป็นเกณฑ์ในการเปรียบเทียบสถานการณ์ต่าง ๆ ซึ่งจำลองการทดลองขึ้นโดยใช้เทคนิคมอนติคาร์โล แต่ละสถานการณ์ต่างกันขึ้นอยู่กับขนาดตัวอย่าง $n = 30$ 50 70 100 200 จำนวนตัวแปร $p = 3$ 5 7 10 และขนาดความสัมพันธ์ระหว่างตัวแปร $\rho = 0.1$ 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 แต่เนื่องจากบางช่วงของการทดลอง ถ้า ρ มีค่าแตกต่างกัน มีผลการเรียงลำดับของค่าความคลาดเคลื่อนเฉลี่ยเหมือนกัน ดังนั้นผู้วิจัยจึงไม่ได้นำสถานการณ์เหล่านั้นมาเสนอในวิทยานิพนธ์ แต่จะเสนอสถานการณ์ที่แตกต่างกันเพียง 106 สถานการณ์เท่านั้น

จากการวิจัยที่ระดับนัยสำคัญ 0.05 พบว่าวิธีการประมาณค่าสูญหายทั้ง 4 วิธีให้ค่าความคลาดเคลื่อนเฉลี่ยไม่แตกต่างกันอย่างมีนัยสำคัญ ดังนั้นจึงกล่าวได้ว่า ไม่ว่าจะเป็นสถานการณ์ใดก็ตามที่กำหนดเหล่านี้ ถ้าหากมีข้อมูลสูญหายเกิดขึ้น สามารถเลือกวิธีการประมาณค่าสูญหายวิธีใดก็ได้ใน 4 วิธีนี้ แต่วิธีการประมาณค่าสูญหายที่ง่ายที่สุดและใช้เวลาในการประมวล

ผลน้อยที่สุดคือวิธีค่าเฉลี่ย ซึ่งเป็นวิธีการประมาณค่าสูญหายที่จะทำให้ได้ค่าความคลาดเคลื่อนเฉลี่ยไม่แตกต่างไปจากการใช้วิธีการประมาณอีก 3 วิธีที่เหลือ แต่อย่างไรก็ตาม ถ้าพิจารณาให้ละเอียดในแต่ละสถานการณ์ เมื่อเปรียบเทียบค่าความคลาดเคลื่อนเฉลี่ยแล้วพบว่ามีค่าแตกต่างกัน แม้ว่าจะไม่แตกต่างกันอย่างมีนัยสำคัญก็ตาม แต่ในการประมาณต่าง ๆ ผู้วิจัยต้องพยายามทำให้ค่าความคลาดเคลื่อนเฉลี่ยมีค่าน้อยที่สุด อาจกล่าวได้ว่า ถ้า p = 3วิธีที่ใช้ค่าเฉลี่ยจะดีที่สุดเมื่อ ρ = 0.1 วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นตัดแปลง จะดีที่สุดเมื่อ ρ = 0.2 - 0.7วิธีวิเคราะห์ส่วนประกอบหลัก จะดีที่สุดเมื่อ ρ = 0.9ถ้า p = 5 วิธีที่ใช้ค่าเฉลี่ย จะดีที่สุดเมื่อ ρ = 0.1 - 0.2 วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นตัดแปลง จะดีที่สุดเมื่อ ρ = 0.3 วิธีวิเคราะห์ส่วนประกอบหลักจะดีที่สุดเมื่อ ρ = 0.5 - 0.9 ถ้า p = 7 วิธีที่ใช้ค่าเฉลี่ยจะดีที่สุดเมื่อ ρ = 0.1 - 0.2 วิธีวิเคราะห์ความถดถอยพหุเชิงเส้นตัดแปลงจะดีที่สุดเมื่อ ρ = 0.3 - 0.4 วิธีวิเคราะห์ส่วนประกอบหลักจะดีที่สุดเมื่อ ρ = 0.5 - 0.8 ถ้า p = 10 วิธีวิเคราะห์ส่วนประกอบหลักจะดีที่สุดเมื่อ ρ = 0.2 - 0.5

Thesis Title	A comparison of missing values estimation methods in multivariate analysis.
Name	Miss Pornsiri Muenchaisri
Thesis Advisor	Associate Professor Sorachai Bhisalbutra , Ph.D.
Department	Statistics
Academic year	1985

ABSTRACT

The purpose of this study is to investigate the four well known missing value estimation methods in multivariate analysis namely, 1) Mean 2) Multiple Linear Regression 3) Modified Multiple Linear Regression 4) Principal Component, using mean square errors as means of comparison.

The data for each experiment were obtained through simulation using the Monte Carlo technique. The computer program was designed to calculate the mean square error for each methods in different situations with varying sample size n = 30 50 70 100 200 number of variables p = 3 5 7 10 and correlation coefficient ρ = 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 However, some intervals of ρ provide the same ranking results of mean square error. These situations are omitted and thus only 106 different situations are presented in the thesis.

The result of this study shows that, at 5% level of significance, mean square errors of the four methods are not significantly different. So if there exists the missing value problem, any one of these methods can be used to estimate the missing value. Nevertheless, the easiest method which also uses least processing time is the first method, Mean.

In addition, attempt to obtain the smallest mean square error is made by considering for each situation which method has the smallest mean square error. The results are up to n , p and ρ. Conclusively, if the number of variables are three, Mean is the best when ρ = 0.1, Modified multiple linear regression is the best when ρ = 0.2 - 0.7 and Principal component is the best when ρ = 0.9. If the number of variables are five, Mean is the best when ρ = 0.1 - 0.2 Modified multiple linear regression is the best when ρ = 0.3 and Principal component is the best when ρ = 0.5 - 0.9. If the number of variables are seven Mean is the best when ρ = 0.1 - 0.2, Modified multiple linear regression is the best when ρ = 0.3 - 0.4 and Principal component is the best when ρ = 0.5 - 0.8. If the number of variables are ten, Principal component is the best when ρ = 0.2 - 0.5.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณาของรองศาสตราจารย์ ดร. สรชัย พิศาลบุตร คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ที่ให้คำแนะนำ ปรึกษา ตลอดจนแก้ไขข้อบกพร่องต่าง ๆ เป็นอย่างดียิ่ง ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณท่านอาจารย์ทุก ๆ ท่านที่ได้ประสิทธิประสาทวิชาความรู้ให้แก่ผู้วิจัยมาโดยตลอด

ขอขอบคุณเจ้าหน้าที่สถาบันบริการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ช่วยประมวลผลงานวิจัยตั้งแต่ต้นจนแล้วเสร็จ

ขอขอบคุณ เพื่อน ๆ และพี่ ๆ ทุกคนที่ให้ความช่วยเหลือ และคำแนะนำต่าง ๆ

สุดท้ายนี้ ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ และพี่สาวที่ช่วยส่งเสริมและสนับสนุนการเรียนของผู้วิจัยตลอดมา

พรศิริ หมื่นไชยศรี

สารบัญ

หน้า

สารบัญตาราง

หน้า

หน้า

หน้า

หน้า