

ดวงภรณ์ โปทาวี 2552: การประมาณค่าข้อมูลสูญหายในการวัดซ้ำด้วยวิธีมาร์คอฟเชนมอนติคาร์โลและวิธีคอปูลาส์ ปรินญาวิทยาสาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ
ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ลีลี อิงศรีสว่าง,
Ph.D. 99 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณค่าสูญหายในข้อมูลที่มีการวัดซ้ำด้วยวิธี Markov Chain Monte Carlo (MCMC) และวิธี Copulas โดยเปรียบเทียบกับวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ใช้ข้อมูลที่ได้จากการจำลองสถานการณ์ที่มีการวัดซ้ำ 3 ครั้ง ด้วยเทคนิคมอนติคาร์โล ภายใต้เงื่อนไขที่กำหนดคือ 1) ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร มีโครงสร้างเมตริกซ์ความสัมพันธ์ 2 แบบ คือ เมตริกซ์ความสัมพันธ์แบบ Compound Symmetry (CS) หรือ แบบ Autoregressive (AR) 2) กำหนดระดับความสัมพันธ์ระหว่างค่าวัดซ้ำมี 3 ระดับ คือ ระดับต่ำ ($\rho = 0.3$) ระดับปานกลาง ($\rho = 0.5$) และ ระดับสูง ($\rho = 0.7$) 3) ขนาดตัวอย่างที่ศึกษา 3 ขนาดคือ 30, 70 และ 100 และ 4) กำหนดให้ตำแหน่งการวัดซ้ำครั้งสุดท้ายมีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม โดยมีระดับการสูญหายของข้อมูล 5%, 10%, 20% และ 30% ตามลำดับ ทำให้มีสถานการณ์ที่เป็นไปได้ทั้งหมด 72 สถานการณ์ สำหรับการประมาณค่าสูญหายด้วยวิธี MCMC กำหนดจำนวนการแทนค่าข้อมูลสูญหายแต่ละค่า 5 ครั้ง ทำการจำลองแต่ละสถานการณ์ 1,000 ครั้ง ด้วยโปรแกรม SAS และวัดประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) วิธีการประมาณค่าสูญหายที่ให้ค่า MSE ต่ำกว่า ถือว่าเป็นวิธีที่ดีกว่า

ผลการจำลองทุกสถานการณ์ของการประมาณค่าข้อมูลสูญหาย พบว่า วิธี Copulas มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC โดยให้ค่า MSE ต่ำสุด สำหรับวิธี MCMC มีประสิทธิภาพดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย เฉพาะในกรณีโครงสร้างเมตริกซ์ความสัมพันธ์แบบ Autoregressive ส่วนผลที่ได้จากการนำวิธีประมาณค่าสูญหายไปประยุกต์กับข้อมูลจริงจำนวน 2 ชุด คือ 1) ข้อมูลรอบเฮอร์ของผู้เข้าร่วมโครงการ “นพรัตน์ไร้พุง มุ่งสุขภาพดี” จากการวัดรอบเฮอร์ 4 ครั้ง และ 2) ข้อมูลปริมาณน้ำฝนรายเดือนของเดือน มิถุนายน – สิงหาคม 2550 ของกรมอุตุนิยมวิทยาภาคเหนือ พบว่า วิธี Copulas มีประสิทธิภาพดีกว่า วิธี MCMC และ วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ซึ่งสอดคล้องกับผลที่ได้จาก ข้อมูลจำลองสถานการณ์

Duangporn Potawee 2009: The Estimation of Missing Data in Repeated Measurements Using Markov Chain Monte Carlo and Copulas Methods. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Lily Ingsisawang, Ph.D. 99 pages.

The objectives of this research were to study two data imputation methods: Markov Chain Monte Carlo (MCMC) and Copulas, and compare these two methods with the simple form of mean imputation in repeated measures data. The datasets used in this study were simulated by the Monte Carlo technique. Each unit was measured on three occasions under the following conditions: 1) data had multivariate normal distribution with two types of correlation structures: Compound Symmetry (CS) and Autoregressive (AR), 2) the level of inter-correlation among repeated observations are determined as low level (0.3), middle level (0.5) and high level (0.7), 3) the sample sizes used are: 30, 70, and 100 units, and 4) the last measurement was assigned to be missing at random (MAR) by rates of missing data of 5%, 10%, 20% and 30%. These gave rise to a total of 72 possible situations. For the MCMC method, the number of imputations was set to be five. Each defined situation was repeated 1,000 times by SAS program, and the performance of each imputation method was evaluated using mean squared error (MSE). The lower MSE was used to indicate the more effective imputation method.

The results from all situations in the simulation study showed that the Copulas method was the most effective method, yielding the lowest value of MSE. The MCMC method was more effective than the mean imputation when the repeated data had correlation structure as AR. In addition, we applied the above three imputation methods with two datasets in practice: 1) the "Nopparat without Fat, but Healthy" project where the data on waist circumference of participants were repeat measured four times during November 21, 2008 – March 3, 2009; and 2) the data on monthly rainfall in the northern part of Thailand where the Department of Meteorology reported the volume of rainfall by month and by rain station in the north region during June - August 2009. The results also indicated that the Copulas was the most effective method, which was consistent with the simulation data.