วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 31, ฉบับที่ 2 เม.ย.–มิ.ย. 2564
The Journal of KMUTNB., Vol. 31, No. 2, Apr.–Jun. 2021

บทความวิจัย

การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง

นลัทพร รูปหมอก* และ กมลชนก พานิชการ
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร นครปฐม

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 9912 5229 อีเมล: roopmok.n@gmail.com DOI: 10.14416/j.kmutnb.2020.11.003
รับเมื่อ 3 เมษายน 2563 แก้ไขเมื่อ 18 พฤษภาคม 2563 ตอบรับเมื่อ 20 พฤษภาคม 2563 เผยแพร่ออนไลน์ 2 พฤศจิกายน 2563
© 2021 King Mongkut's University of Technology North Bangkok. All Rights Reserved.

บทคัดย่อ

แผนแบบการทดลองแบบวัดซ้ำมีลักษณะการเก็บข้อมูลจากหน่วยตัวอย่างเดียวกัน แต่ต่างกันที่ช่วงเวลาหรือเงื่อนไขอื่น ซึ่งนิยมใช้ในงานวิจัยทางด้านการแพทย์หรือสาธารณสุข บทความนี้เสนอการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลองเมื่อสุ่มค่าข้อมูลสูญหายอย่างสุ่มสมบูรณ์ โดยประยุกต์จากวิธีการแทนที่ด้วยค่าเฉลี่ยวิธี CopyMean Trajectory วิธี CopyMean LOCF และวิธีโครงข่ายประสาทเทียม โดยใช้เกณฑ์ในการประเมินด้วยค่า MAD, RMSD และค่า Bias ซึ่งทำการทดลองทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลองโดยในชุดข้อมูลจำลองกำหนดให้ในแต่ละตัวแปรมีค่าเฉลี่ยและค่าความแปรปรวนเท่ากัน ผลการวิจัยพบว่า ในกรณีส่วนใหญ่วิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายในข้อมูลจริงและข้อมูลจำลองในกรณีไม่มีสหสัมพันธ์และสหสัมพันธ์น้อย (0, 0.3 และ 0.5) ส่วนในข้อมูลจำลองกรณีที่สหสัมพันธ์ค่อนข้างมาก (0.7 และ 0.9) วิธี CopyMean Trajectory เป็นวิธีการที่ดีที่สุดในกรณีส่วนใหญ่

คำสำคัญ: ข้อมูลสูญหาย แผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลอง แทนที่ด้วยค่าเฉลี่ย CopyMean โครงข่ายประสาทเทียม

การอ้างอิงบทความ: นลัทพร รูปหมอก และ กมลชนก พานิชการ, "การเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 31, ฉบับที่ 2, หน้า 300–308, เม.ย.–มิ.ย. 2564.

Research Article

A Comparison of Missing Data Imputation Methods in Within-subject Repeated Measure Design

Nalattaporn Roopmok* and Kamolchanok Panishkan
Department of statistics, Faculty of Science, Silpakorn University, Nakhon Pathom, Thailand

* Corresponding Author, Tel. 08 9912 5229, E–mail: roopmok.n@gmail.com
DOI: 10.14416/j.kmutnb.2020.11.003
Received 3 April 2020; Revised 18 May 2020; Accepted 20 May 2020; Published online: 2 November 2020
© 2021 King Mongkut's University of Technology North Bangkok. All Rights Reserved.

Abstract

Within-subject repeated measure design is an experimental design conducted by collecting data from the same sample unit at different times or with other conditions. It is popular in medical or public health research. This article presents a comparison of missing data imputation methods in within-subject repeated measure design when missing values are missing completely at random. The imputation methods were applied by the Mean Substitution method, CopyMean Trajectory method, CopyMean LOCF method and Artificial Neural Network method by using 3 assessment criteria such as MAD, RMSD, and Bias. All these methods were tested on both real dataset and artificial datasets when mean and variance in each variable were equally defined. The results revealed that, in the most cases, the artificial neural network method performed the best in real dataset and in artificial datasets with no correlation or low correlation (0, 0.3, and 0.5). However, in artificial datasets with high correlation (0.7 and 0.9), the CopyMean Trajectory method was the best method in the most cases.

Keywords: Missing Values, Within-subject Repeated Measure Design, Mean Substitution, CopyMean, Artificial Neural Network

Please cite this article as: N. Roopmok and K. Panishkan, "A comparison of missing data imputation methods in within-subject repeated measure design," *The Journal of KMUTNB*, vol. 31, no. 2, pp. 300–308, Apr.–Jun. 2021 (in Thai).

1. บทนำ

ค่าข้อมูลสูญหายคือ ค่าในชุดข้อมูลที่มีค่าสังเกตขาดหายไป โดยข้อมูลสูญหายนั้น นับเป็นปัญหาสำคัญที่อาจส่งผลให้เกิดปัญหาหลายประการ เช่น ข้อมูลสูญหายทำให้กำลังการทดสอบลดลง ทำให้การประมาณค่าพารามิเตอร์เกิดความเอนเอียง ทำให้เห็นสารสนเทศจากตัวอย่างลดลง และทำให้การวิเคราะห์ข้อมูลซับซ้อนขึ้นเป็นต้น [1]

โดยทั่วไปแล้ววิธีการจัดการเกี่ยวกับค่าข้อมูลสูญหายอาจทำได้โดยการตัดค่าข้อมูลสูญหายออก หรือทำได้โดยการประมาณค่าข้อมูลสูญหาย แม้ว่าวิธีการตัดค่าข้อมูลสูญหายออกจะทำได้ง่ายกว่า แต่วิธีการนี้จะมีประสิทธิภาพหากค่าข้อมูลสูญหายนั้น ไม่มีความแตกต่างจากค่าอื่นมากนัก โดยเฉพาะอย่างยิ่งในแผนแบบการทดลองแบบวัดซ้ำภายในหน่วยทดลอง เนื่องจากข้อมูลตามเวลามีความสำคัญเป็นอย่างมากในข้อมูลวัดซ้ำการสูญหายของข้อมูลเป็นสาเหตุให้ข้อมูลที่เก็บตามเวลาไม่สมบูรณ์ และการวิเคราะห์และแปลผลจากข้อมูลอาจเกิดความผิดพลาด กำลังการทดสอบลดลง ทำให้การประมาณค่าพารามิเตอร์เกิดความเอนเอียงได้ ดังนั้นวิธีการประมาณค่าข้อมูลสูญหายจึงเป็นวิธีการที่เหมาะสมกว่าในการจัดการกับค่าข้อมูลสูญหาย ซึ่งข้อมูลแต่ละชนิดเหมาะกับวิธีการประมาณค่าข้อมูลสูญหายที่แตกต่างกัน ในงานวิจัยนี้จึงสนใจศึกษาวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลองที่เหมาะสมที่สุดในแต่ละสถานการณ์

Bingham และคณะ [2] ได้ศึกษาวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลอง โดยใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการแทนที่ด้วยค่าเฉลี่ย ซึ่งผลการวิจัยได้เปรียบเทียบค่าโมเมนต์ของชุดข้อมูลเดิมและชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายแล้ว ผลปรากฏว่าโมเมนต์ของชุดข้อมูลเดิมและชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายแล้วมีค่าใกล้เคียงกัน และโมเมนต์ของทั้งในชุดข้อมูลเดิมและชุดข้อมูลที่ประมาณค่าข้อมูลสูญหายแล้วไม่มีความแตกต่างกันตามรูปร่างของข้อมูล แต่ความผันแปรของข้อมูลจะเพิ่มขึ้นเมื่อสัดส่วนของข้อมูลสูญหายเพิ่มขึ้น

Genolini และคณะ [3] เสนอวิธีการ CopyMean สำหรับการประมาณค่าข้อมูลสูญหายทางเดียวในการศึกษาตามคาบเวลา โดยเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย 3 วิธีการ คือ วิธีการแรกการประมาณค่าข้อมูลสูญหายตามขวาง เช่น ค่าเฉลี่ยตามขวาง ค่ามัธยฐานตามขวาง ค่าสุ่มตามขวาง วิธีการที่สองการประมาณค่าข้อมูลสูญหายตามคาบเวลา เช่น ค่าเฉลี่ยตามคาบเวลา ค่ามัธยฐานตามคาบเวลา ค่าสุ่มตามคาบเวลา LOCF และการประมาณค่าในช่วงเชิงเส้น และวิธีการสุดท้ายเป็นการผสมระหว่างกระบวนการประมาณค่าตามขวาง และกระบวนการประมาณค่าตามคาบเวลา เช่น การวิเคราะห์การถดถอย และ CopyMean ในการประมาณค่าข้อมูลที่มีลักษณะการสูญหาย 3 ลักษณะ คือ MCAR, MAR และ MNAR โดยใช้เกณฑ์การประเมินประสิทธิภาพของการประมาณค่าข้อมูลสูญหายด้วยค่า *MAD, RMSD* และ *Bias* ผลการวิจัยสรุปได้ว่าวิธีการ CopyMean LOCF เป็นวิธีการที่มีประสิทธิภาพมากที่สุดในกรณีส่วนใหญ่

Gupta และ Lam [4] ได้เสนอวิธีการโครงข่ายประสาทเทียมด้วยกระบวนการ Back-propagation เพื่อใช้ในการประมาณค่าข้อมูลสูญหายและเปรียบเทียบกับวิธีการประมาณค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และวิธีการวิเคราะห์การถดถอย จากนั้นเปรียบเทียบวิธีการจัดกลุ่มข้อมูลที่ประมาณค่าข้อมูลสูญหายแล้ว ด้วยวิธีการโครงข่ายประสาทเทียม และวิธีการวิเคราะห์การจำแนก ผลการทดลองสรุปได้ว่าวิธีการโครงข่ายประสาทเทียมเป็นวิธีการที่มีประสิทธิภาพทั้งในการจัดกลุ่มและประมาณค่าข้อมูลสูญหาย

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อหาวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำภายในหน่วยทดลองที่มีประสิทธิภาพมากที่สุดในแต่ละสถานการณ์ โดยใช้วิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพมากที่สุดในงานวิจัยก่อนหน้า นั่นคือวิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการ CopyMean และวิธีการโครงข่ายประสาทเทียม เพื่อประมาณค่าข้อมูลสูญหายทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลอง และเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายโดยใช้ค่า *MAD*, *RMSD* และค่า *Bias* เป็นเกณฑ์ในการประเมิน

2. วัสดุ อุปกรณ์และวิธีการวิจัย

กำหนดให้ตัวแปรตอบสนอง Y ถูกวัดจากช่วงเวลา k ช่วงเวลาจากตัวอย่าง n หน่วยตัวอย่าง ซึ่งประกอบด้วยค่าข้อมูลมูลสูญหาย M ค่าในชุดข้อมูลจำนวน N ชุดข้อมูล และค่าของ Y ที่เก็บจากตัวอย่างที่ i ที่ช่วงเวลาที่ j เขียนแทนด้วย y_{ij} และถ้า y_{ij} เป็นค่าข้อมูลสูญหายแล้วค่าประมาณค่าข้อมูลสูญหายจะเขียนแทนด้วย $\hat{y}_{ij}$

2.1 วิธีการประมาณค่าข้อมูลสูญหาย

2.1.1 วิธีการแทนที่ด้วยค่าเฉลี่ย (MS)

วิธีการแทนที่ด้วยค่าเฉลี่ย [2] เป็นวิธีการโดยทั่วไปที่ใช้ในการประมาณค่าข้อมูลสูญหายในแผนแบบวัดซ้ำ ประกอบด้วยส่วนประกอบ 2 ส่วน คือ การประมาณค่าตามขวางภายในหน่วยตัวอย่างซึ่งสามารถคำนวณได้จาก $\overline{Y}_{.j} = \frac{1}{n}\sum_{i} y_{ij}$ และการประมาณค่าตามคาบเวลา ซึ่งสามารถคำนวณได้จาก $\overline{I}_i = \frac{1}{t'}\sum_{i}(\overline{y}_{.j} - y_{ij})$ เมื่อ t' คือจำนวนช่วงเวลาที่ไม่ได้เป็นค่าข้อมูลสูญหายในตัวอย่างที่ i และค่าประมาณค่าข้อมูลสูญหายจะคำนวณได้จาก $\hat{Y}_{ij} = \overline{y}_{.j} - \overline{I}_i$

2.1.2 วิธีการ CopyMean

วิธีการ CopyMean [3] ประกอบด้วยส่วนประกอบ 2 ส่วน คือ การประมาณค่าตามคาบเวลาทั่วไป $\left(y_{ij}^{IM}\right)$ และการคำนวณค่าความผันแปรเฉลี่ยในช่วงเวลาที่ j ซึ่งเป็นการประมาณค่าตามขวาง ซึ่งคำนวณได้จาก $AV_j = \overline{y_{.j}} - \overline{y_{.j}^{IM}}$ เมื่อ $\overline{y_{.j}}$ และ $\overline{y_{.j}^{IM}}$ คือค่าเฉลี่ยตามขวางในช่วงเวลาที่ j ของข้อมูลเดิมและข้อมูลที่ประมาณค่าข้อมูลสูญหายแล้วตามลำดับ ค่าประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean สามารถคำนวณได้จาก $\hat{y}_{ij} = y_{ij}^{IM} + AV_j$

ทั้งนี้ วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean สามารถคำนวณได้หลายรูปแบบขึ้นกับวิธีการประมาณค่าข้อมูลสูญหายตามคาบเวลาที่เลือกใช้ ในบทความนี้เสนอวิธีการ CopyMaen Trajectory (CT) ซึ่งใช้วิธีการ Trajectory Mean เป็นวิธีการประมาณค่าข้อมูลสูญหายตามคาบเวลาซึ่งสามารถคำนวณค่า Trajectory Meanได้โดย $y_{ij}^{IM} = \overline{y_{i.}}$

และวิธีการ CopyMaen LOCF (CL) ซึ่งใช้วิธีการ LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายตามคาบเวลาคาบเวลาซึ่งสามารถคำนวณค่า LOCF ได้โดย $y_{ij}^{IM} = y_{i,j-1}$

2.1.3 วิธีการโครงข่ายประสาทเทียม (ANN)

วิธีการโครงข่ายประสาทเทียมเป็นเครื่องมือทางคอมพิวเตอร์ที่นิยมใช้อย่างแพร่หลายมีแนวคิดมาจากการทำงานของระบบประสาทของสิ่งมีชีวิต ซึ่งสามารถใช้ในงานหลากหลายประเภทเช่นการตรวจจับสิ่งของ ใช้บอกความผิดพลาดในกระบวนการต่างๆ หรือใช้ในการพยากรณ์ได้ ในบทความนี้ใช้วิธีการโครงข่ายประสาทเทียมเพื่อประมาณค่าข้อมูลสูญหาย โดยสร้างโครงข่ายประสาทเทียมจากข้อมูลที่ตัดคาบเวลาที่เป็นค่าข้อมูลสูญหายออก และกำหนด Input Layer คือตัวอย่างที่ไม่มีค่าข้อมูลสูญหาย และ Output คือตัวอย่างที่มีค่าข้อมูลสูญหาย สร้าง Hidden Layer จำนวน 2 เลเยอร์ เลเยอร์ละ 2 Node และเลือกใช้กระบวนการ Backpropagation และ Switch Function เป็น Activation Function ซึ่งเขียนได้โดย $f(x) = \frac{2x}{1+e^{-\beta x}}$ กำหนดค่า Learning Rate เท่ากับ 0.0001 จากนั้นใช้ข้อมูลจากคาบเวลาที่ตัดออกมาพยากรณ์ค่าข้อมูลสูญหาย โดยทำการประมาณค่าข้อมูลสูญหายโดยใช้โปรแกรม R ด้วยแพ็กเกจ Neuralnet บน CRAN [5]

2.2 เกณฑ์ที่ใช้ในการประเมิน

เพื่อวัดประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายใช้เกณฑ์ในการประเมิน 3 วิธี คือ

2.2.1 ค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (*MAD*)

MAD ใช้วัดความแตกต่างระหว่างค่าจริงและค่าพยากรณ์ซึ่งค่า *MAD* ต่ำ แสดงว่าค่าประมาณมีค่าใกล้เคียงกับค่าจริงค่า *MAD* สามารถคำนวณได้ดังสมการที่ (1)

$$MAD = \frac{\sum_{ij}\left|\hat{y}_{ij} - y_{ij}\right|}{N \times M} \quad (1)$$

2.2.2 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (*RMSD*)

ค่า *RMSD* ใช้วัดความแตกต่างระหว่างค่าจริงและค่าพยากรณ์ซึ่งค่า *RMSD* ต่ำ แสดงว่าค่าประมาณมีค่าใกล้เคียงกับ

ตารางที่ 1 ค่าจริงและค่าประมาณค่าข้อมูลสูญหายในชุดข้อมูลจริง

จำนวนค่าสูญหาย	1	2		3		
ตำแหน่งค่าสูญหาย	y_{24}	y_{92}	y_{24}	y_{24}	y_{74}	y_{65}
ค่าจริง	132.82	104.71	132.82	132.82	112.51	90.56
MS	113.107	97.033	113.207	113.522	111.852	85.41
CT	87.103	99.665	87.103	87.103	85.433	93.555
CL	92.405	82.581	92.405	92.405	104.903	117.335
ANN	125.183	117.892	134.427	131.023	98.645	89.461

ตารางที่ 2 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจริง

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	19.714	388.622	19.714	13.645	221.794	13.645	8.368	133.118	8.368
CT	45.718	2090.09	45.718	25.381	1057.771	25.381	25.263	944.084	23.267
CL	40.416	1633.413	40.416	31.272	1061.553	31.272	24.933	802.729	7.083
ANN	**7.637**	**58.327**	**7.637**	**7.394**	**88.173**	**7.394**	**5.587**	**65.559**	**5.587**

ค่าจริงเช่นเดียวกับค่า *MAD* แต่ค่า *RMSD* ไม่สามารถวัดทิศทางของค่าพยากรณ์ได้ และมีความไวต่อค่านอกกลุ่ม วิธีการประเมินนี้จึงไม่เหมาะสมกับข้อมูลที่มีค่านอกกลุ่ม ค่า *RMSD* สามารถคำนวณได้ดังสมการที่ (2)

$$RMSD = \sqrt{\frac{\sum_{ij}(\hat{y}_{ij} - y_{ij})^2}{N \times M}} \quad (2)$$

2.2.3 ค่าความเอนเอียง (*Bias*)

ค่า *Bias* เป็นวิธีการวัดค่าความคลาดเคลื่อนในการพยากรณ์โดยพิจารณาทิศทางของข้อมูลร่วมด้วย ค่าความเอนเอียงจะเป็นวิธีการประเมินที่ไม่เหมาะสมหากค่าความเอนเอียงของข้อมูลแต่ละค่ามีค่ามากแต่มีทิศทางตรงกันข้ามกัน ดังนั้นจึงควรพิจารณาร่วมกับค่า *MAD* และ *RMSD* ซึ่งค่า *Bias* สามารถคำนวณได้ดังสมการที่ (3)

$$Bias = \frac{\sum_{ij}(\hat{y}_{ij} - y_{ij})}{N \times M} \quad (3)$$

3. ผลการทดลอง

จากข้อมูลทั้งในชุดข้อมูลจริง และชุดข้อมูลจำลองสุ่มค่าข้อมูลสูญหายจำนวน 1, 2 และ 3 ค่าแบบ MCAR (Missing Completely at Random) ในแต่ละชุดข้อมูล และประมาณค่าข้อมูลสูญหาย ผลการทดลองที่ได้คือค่าสัมประสิทธิ์สหสัมพันธ์มีอิทธิพลต่อประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหาย แต่จำนวนค่าข้อมูลสูญหาย ไม่มีผลต่อประสิทธิภาพ โดยพิจารณาจากวิธีการประมาณค่าข้อมูลสูญหายที่ทำให้ค่าประเมินมีค่าน้อยที่สุดจะเป็นวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพมากที่สุดในแต่ละเกณฑ์การประเมิน และจะแสดงเป็นตัวหนาในตาราง

3.1 ผลการวิจัยโดยใช้ชุดข้อมูลจริง

ในชุดข้อมูลจริงใช้ข้อมูลชุด Sky Drive [6] ข้อมูลชุดนี้เก็บค่าอัตราการเต้นของหัวใจจากนักบินจำนวน 11 คน ซึ่งเป็นเพศชายจำนวน 8 คน และเพศหญิงจำนวน 3 คน มีช่วงอายุระหว่าง 18–40 ปี โดยวัดอัตราการเต้นของหัวใจจากเครื่อง Polar F6 Heart Rate Monitor ในช่วงเวลา 5 ช่วง

คือ ก่อนขึ้นบิน ขณะกำลังขึ้นบิน เมื่อเครื่องบินเหนือพื้นดิน 1524 เมตร เมื่อเครื่องบินเหนือพื้นดิน 3048 เมตร และเมื่อเครื่องบินกำลังลงจอด

จากตารางที่ 1 และ 2 แสดงผลการประมาณค่าข้อมูลสูญหายและค่าประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจริงตามลำดับ ในชุดข้อมูลจริงวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในทุกจำนวณค่าข้อมูลสูญหายด้วยเกณฑ์ในการประเมินทั้ง 3 เกณฑ์

3.2 ผลการวิจัยโดยใช้ชุดข้อมูลจำลอง

ในชุดข้อมูลจำลอง ได้จำลองข้อมูลจากการแจกแจงปรกติพหุแปรซึ่งมีฟังก์ชันความน่าจะเป็นดังสมการที่ (4)

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) \quad (4)$$

เมื่อ μ คือ เวกเตอร์ค่าเฉลี่ย

Σ คือ เมทริกซ์ความแปรปรวนความแปรปรวนร่วม

$\boldsymbol{V}^{1/2} = \sqrt{\boldsymbol{diag}(\Sigma)} = \boldsymbol{\sigma}$

$\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{\rho}\boldsymbol{V}^{1/2}$

โดยกำหนด k = 4, n = 5 และพารามิเตอร์ $\mu_{\tilde{A}}$ = 20; i = 1, 2, 3, 4, σ_i^2 = 25, i = 1, 2, 3, 4 เท่ากันทั้ง 5 ชุดข้อมูล แต่กำหนดสัมประสิทธิ์สหสัมพันธ์แตกต่างกันคือ ρ = 0, 0.3, 0.5, 0.7, 0.9 ตามลำดับ และจำลองชุดข้อมูลสถานการณ์ละ 1,000 รอบ

ผลการวิจัยในชุดข้อมูลจำลองที่ 1 เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0 พบว่า เมื่อใช้ค่า *MAD* และ *RMSD* เป็นเกณฑ์ในการประเมินวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในทุกจำนวนค่าข้อมูลสูญหาย แต่เมื่อพิจารณาจากค่า Bias วิธีการแทนที่ด้วยค่าเฉลี่ยเป็นวิธีการที่ดีที่สุดเมื่อสุ่มค่าข้อมูลสูญหาย 1 และ 3 ค่า และวิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดเมื่อสุ่มค่าข้อมูลสูญหาย 2 ค่า ดังแสดงในตารางที่ 3

ผลการวิจัยในชุดข้อมูลจำลองที่ 2 เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.3 พบว่า วิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในทุกจำนวนค่าข้อมูลสูญหาย ในทุกเกณฑ์การประเมินยกเว้นในกรณีที่ใช้ค่า *Bias* เป็นเกณฑ์เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในกรณีนี้ ดังแสดงในตารางที่ 4

ตารางที่ 3 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 1 (ρ = 0)

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	4.775	36.097	**0.288**	5.535	49.038	0.941	5.153	40.221	**0.16**
CT	4.764	34.689	0.458	5.043	40.791	0.605	4.688	33.301	0.255
CL	5.657	49.701	0.316	5.369	43.816	**0.552**	5.454	47.389	0.277
ANN	**4.131**	**26.606**	2.033	**4.397**	**28.685**	1.412	**4.191**	**27.328**	1.225

ตารางที่ 4 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 2 (ρ = 0.3)

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	2.407	9.35	0.071	2.777	11.8	0.33	3.73	15.228	0.697
CT	2.23	7.938	**0.023**	2.645	10.737	0.273	3.49	13.094	0.653
CL	2.638	10.44	0.152	2.848	12.795	0.126	4.23	18.269	0.784
ANN	**1.03**	**1.517**	0.057	**1.459**	**3.107**	**0.056**	**2.364**	**6.853**	**0.203**

เมื่อพิจารณาในชุดข้อมูลจำลองที่ 3 เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.5 พบว่า วิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในกรณีส่วนใหญ่ยกเว้นในกรณีที่ใช้ค่า Bias เป็นเกณฑ์ เมื่อสุ่มค่าข้อมูลสูญหาย 1 ค่า และ 3 ค่า วิธีการ CopyMean LOCF และ CopyMean Trajectory จะเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดตามลำดับ ดังแสดงในตารางที่ 5

ผลการวิจัยในชุดข้อมูลจำลองที่ 4 เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.7 พบว่า เมื่อใช้ค่า *MAD* และ *RMSD* เป็นเกณฑ์ในการประเมินวิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในทุกจำนวนค่าข้อมูลสูญหาย แต่เมื่อพิจารณาจากค่า *Bias* วิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการโครงข่ายประสาทเทียม และวิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดเมื่อสุ่มค่าข้อมูลสูญหาย 1, 2 และ 3 ค่าตามลำดับดังแสดงในตารางที่ 6

เมื่อพิจารณาในชุดข้อมูลจำลองที่ 5 เมื่อกำหนดค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.9 พบว่า วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในกรณีที่ใช้ค่า *MAD* และ *RMSD* เป็นเกณฑ์ในการประเมิน ส่วนในกรณีที่ใช้ค่า *Bias* เป็นเกณฑ์เมื่อสุ่มค่าข้อมูลสูญหาย 1, 2 และ 3 ค่า วิธีการแทนที่ด้วยค่าเฉลี่ย วิธีการโครงข่ายประสาทเทียม และวิธีการ CopyMean LOCF จะเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดตามลำดับ ดังแสดงในตารางที่ 7

ตารางที่ 5 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 3 (ρ = 0.5)

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	4.145	24.433	0.191	4.023	24.32	0.236	4.288	26.499	0.493
CT	4.054	22.346	0.132	3.939	23.311	0.19	3.925	22.183	**0.491**
CL	4.438	27.168	**0.025**	4.426	29.981	0.228	4.236	26.148	0.572
ANN	**2.672**	**15.858**	0.227	**3.224**	**21.065**	0.129	**3.496**	**21.288**	1.757

ตารางที่ 6 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 4 (ρ = 0.7)

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	2.627	10.361	**0.399**	2.623	10.399	0.221	3.967	17.045	0.156
CT	**2.583**	**10.024**	0.517	**2.411**	**9.146**	0.319	**3.549**	**13.712**	0.067
CL	3.077	13.757	0.54	2.647	11.188	0.316	4.259	19.498	**0.062**
ANN	3.518	18.368	0.552	3.833	22.683	**0.056**	5.701	33.921	0.38

ตารางที่ 7 ผลการประเมินวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลจำลองที่ 5 (ρ = 0.9)

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
MS	1.621	3.926	0.059	1.709	4.613	0.066	1.413	3.275	0.244
CT	**1.524**	**3.427**	0.072	**1.596**	**4.076**	0.04	**1.305**	**2.796**	0.165
CL	1.704	4.648	**0.005**	1.78	4.872	**0.027**	1.5	3.633	**0.134**
ANN	3.28	24.06	0.21	3.969	31.147	0.331	3.716	27.563	0.411

ตารางที่ 8 วิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในชุดข้อมูลจำลอง

วิธีการประมาณค่า	สูญหาย 1 ค่า			สูญหาย 2 ค่า			สูญหาย 3 ค่า		
	MAD	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*	*MAD*	*RMSD*	*|Bias|*
ชุดที่ 1 (ρ = 0.0)	ANN	ANN	MS	ANN	ANN	CL	ANN	ANN	MS
ชุดที่ 2 (ρ = 0.3)	ANN	ANN	CT	ANN	ANN	ANN	ANN	ANN	ANN
ชุดที่ 3 (ρ = 0.5)	ANN	ANN	CL	ANN	ANN	ANN	ANN	ANN	CT
ชุดที่ 4 (ρ = 0.7)	CT	CT	MS	CT	CT	ANN	CT	CT	CL
ชุดที่ 5 (ρ = 0.9)	CT	CT	CL	CT	CT	CL	CT	CT	CL

จากตารางที่ 8 แสดงวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในแต่ละเกณฑ์การประเมินจำแนกตามจำนวณค่าข้อมูลสูญหายพบว่า วิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในชุดข้อมูลที่ 1, 2 และ 3 ที่ไม่มีค่าสัมประสิทธิ์สหสัมพันธ์และที่มีค่าสัมประสิทธิ์สหสัมพันธ์น้อย (0, 0.3, 0.5) แต่ในชุดข้อมูลที่ 4 และ 5 ที่มีค่าสัมประสิทธิ์สหสัมพันธ์มาก (0.7, 0.9) วิธีการ CopyMean Trajectory จะเป็นวิธีการที่ดีที่สุดในการประมาณค่าข้อมูลสูญหายในทุกจำนวณของค่าข้อมูลสูญหาย

เมื่อพิจารณาค่าประเมินเมื่อกำหนดค่าข้อมูลสูญหายจำนวน 1, 2 และ 3 ค่าตามลำดับพบว่าค่าประเมินด้วยเกณฑ์การประเมินทั้ง 3 วิธี มีค่าใกล้เคียงกัน เมื่อพิจารณาค่าประเมินในกรณีที่สุ่มค่าข้อมูลสูญหายต่างกันจากชุดข้อมูลเดียวกัน แต่เมื่อพิจารณาค่าประเมินในกรณีที่ค่าสหสัมพันธ์ต่างกัน พบว่า เมื่อค่าสหสัมพันธ์มีค่ามากขึ้นแล้วค่าประเมินด้วยเกณฑ์การประเมินทั้ง 3 วิธี มีแนวโน้มที่ลดลงเมื่อค่าสหสัมพันธ์เพิ่มขึ้น

4. อภิปรายผลและสรุป

ชุดข้อมูลจำลองที่ 1–3 (กรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์คือ 0, 0.3, 0.5) ผลการทดลองคือ วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการโครงข่ายประสาทเทียมเป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดในกรณีส่วนใหญ่ซึ่งสอดคล้องกับในงานวิจัยของ Gupta และ Lam [4]

ในงานวิจัยของ Genolini และคณะ [3] วิธีการ CopyMean LOCF เป็นวิธีการประมาณค่าข้อมูลสูญหายที่เหมาะสมที่สุดในการศึกษาข้อมูลตามคาบเวลา ในงานวิจัยนี้เมื่อพิจารณาในชุดข้อมูลจำลองที่ 4–5 (กรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์คือ 0.7 และ 0.9) วิธีการ CopyMean Trajectory เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุด ซึ่งทั้งในงานวิจัยนี้ในกรณีที่กำหนดค่าสัมประสิทธิ์สหสัมพันธ์มาก และในงานวิจัยของ Genolini และคณะ [3] วิธีการ CopyMean เป็นวิธีการประมาณค่าข้อมูลสูญหายที่ดีที่สุดเช่นเดียวกัน แต่แตกต่างกันที่วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตามคาบเวลา ซึ่งอาจเกิดจากในงานวิจัยนี้ได้กำหนดค่าความแปรปรวนของทุกตัวแปรมีค่าเท่ากัน และขนาดตัวอย่างในงานวิจัยนี้มีขนาดเล็กกว่าในงานวิจัยของ Genolini และคณะ [3] ในการประมาณค่าข้อมูลสูญหายด้วยวิธีการ CopyMean ที่เลือกใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีการของการศึกษาตามคาบเวลาด้วยวิธีการ Trajectory Mean จึงเหมาะสมกว่าวิธีการ LOCF

ในงานวิจัยครั้งนี้ได้กำหนดความแปรปรวนเท่ากันในทุกตัวแปรและกำหนดขนาดตัวอย่าง n = 5 ซึ่งตัวอย่างมีขนาดเล็ก ในการศึกษาครั้งต่อไปควรศึกษาในกรณีที่กำหนดพารามิเตอร์ความแปรปรวนไม่เท่ากัน และควรศึกษาในขนาดตัวอย่างอื่นเพิ่มเติม

เอกสารอ้างอิง

[1] H. Kang, “The prevention and handling of the missing data,” *The Korean Society of Anesthesiologists*, vol. 64, no. 5, pp. 402–406, 2013.

[2] C. R. Bingham, M. Stemmler, A. C. Peterson, and J. A. Graber, "Imputing missing data values in repeated measurement within-subjects designs," *Methods of Psychological Research Online*, vol. 3, no. 2, pp. 131–155, 1998.

[3] C. Genolini, A. Lacombe, R. cochard, and F. Subtil, "CopyMean: A new method to predict monotone missing values in longitudinal studies," *Computer Methods and Programs in Biomedicine*, vol. 132, pp. 29–44, 2016.

[4] A. Gupta and M. S. Lam, "Estimating missing values using neural networks," *The Journal of the Operational Research Society*, vol. 41, pp. 229–238, 1996.

[5] S. Fritsch, F. Guenther, M. N. Wright, M. Suling, and S. M. Mueller. (2019, February). Package 'neuralnet', GitHub, Inc., [Online]. Available: https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf

[6] K. I. Singley, B. D. Hale, and D. Russell, "Heart rate, anxiety, and hardiness in novice (Tandem) and experienced (Solo) skydivers," *Journal of Sport Behavior*, vol. 35, no. 4, pp. 453–469, 2012.