บทความวิจัย

การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายแบบพหุในข้อมูลพหุระดับ: การศึกษาด้วยการจำลองข้อมูล

นวลรัตน์ ฉิมสุด ประภาศิริ รัชชประภาพรกุล* และ สิวะโชติ ศรีสุทธิยากร
ภาควิชาวิจัยและจิตวิทยาการศึกษา สาขาวิชาสถิติและสารสนเทศการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 4015 9321 อีเมล: prapasiri.r@chula.ac.th DOI: 10.14416/j.kmutnb.2024.03.009
รับเมื่อ 19 พฤษภาคม 2565 แก้ไขเมื่อ 3 สิงหาคม 2565 ตอบรับเมื่อ 25 สิงหาคม 2565 เผยแพร่ออนไลน์ 19 มีนาคม 2567
© 2024 King Mongkut's University of Technology North Bangkok. All Rights Reserved.

บทคัดย่อ

การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายแบบพหุในข้อมูลพหุระดับ: การศึกษาด้วยการจำลองข้อมูล ในการวิจัยครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุจำนวน 6 วิธี ได้แก่ วิธี Multiple Imputation Fully Conditional Specification (FCS) วิธี Random Forest (RF) และวิธี Optimal Impute (Opt.impute) ประกอบด้วยวิธี Opt.knn วิธี Opt.tree วิธี Opt.svm และวิธี Opt.cv โดยใช้การจำลองข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) ภายใต้เงื่อนไขดังนี้ 1) ประเภทการสูญหายแบ่งออกเป็น 6 รูปแบบ ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random; MCAR) การสูญหายแบบสุ่ม (Missing at Random; MAR) การสูญหายแบบไม่สุ่ม (Missing not at Random; MNAR) และประเภทของการสูญหายรูปแบบผสมรายคู่ 3 รูปแบบ ได้แก่ MCAR - MAR, MCAR - MNAR และ MAR - MNAR 2) ขนาดของตัวอย่างระดับที่หนึ่งเท่ากับ 1,000, 2,000 และ 3,000 หน่วย และขนาดตัวอย่างระดับที่สองเท่ากับ 40, 50 และ 60 หน่วย 3) อัตราการสูญหายของค่าสังเกตในตัวอย่างระดับที่หนึ่งเป็น 3 ระดับ ได้แก่ ร้อยละ 30 ร้อยละ 40 และร้อยละ 50 ตามลำดับ เมื่อพิจารณาผลการวิจัยจำแนกตามประเภทการสูญหาย 6 รูปแบบ พบว่า ข้อสูญหายรูปแบบ MCAR วิธีทดแทนค่าสูญหาย Opt.cv มีประสิทธิภาพโดยเฉลี่ยสูงที่สุด ข้อมูลสูญหายรูปแบบ MAR, MCAR - MAR และ MAR - MNAR วิธีทดแทนค่าสูญหาย Opt.svm มีประสิทธิภาพโดยเฉลี่ยสูงที่สุด ทั้งนี้เมื่อข้อสูญหายแบบ MNAR, MCAR-MNAR พบว่าวิธีทดแทนค่าสูญหาย RF มีประสิทธิภาพโดยเฉลี่ยสูงที่สุด จากการวิเคราะห์ผลการวิจัยพบว่า วิธีทดแทนค่าสูญหาย Opt.impute มีแนวโน้มให้ประสิทธิภาพโดยเฉลี่ยสูงที่สุด รองลงมาคือ วิธี RF และวิธี FCS ตามลำดับ

คำสำคัญ: การทดแทนค่าข้อมูลสูญหาย ข้อมูลพหุระดับ โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม

การอ้างอิงบทความ: นวลรัตน์ ฉิมสุด, ประภาศิริ รัชชประภาพรกุล และ สิวะโชติ ศรีสุทธิยากร, "การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายแบบพหุในข้อมูลพหุระดับ: การศึกษาด้วยการจำลองข้อมูล," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 34, ฉบับที่ 2, หน้า 1–14, เลขที่บทความ 242-126075, เม.ย.–มิ.ย. 2567.

Research Article

Comparison of the Efficiency of Multiple Imputation in Multilevel Data: A Simulation Study

Nuanrat Chimsud, Prapasiri Ratchaprapapornkul* and Siwachot Srisuttiyakorn
Department Educational Statistics, Faculty of Education Chulalongkorn University, Bangkok, Thailand

* Corresponding Author, Tel. 08 4015 9321, E–mail: prapasiri.r@chula.ac.th
DOI: 10.14416/j.kmutnb.2024.03.009
Received 19 May 2022; Revised 3 August 2022; Accepted 25 August 2022; Published online: 19 March 2024
© 2024 King Mongkut's University of Technology North Bangkok. All Rights Reserved.

Abstract

The purpose of this research was to compare the efficiency of multiple Imputation methods of multilevel missing data. Six methods of the Imputation included Multiple Imputation Fully Conditional Specification (FCS), Random Forest (RF), and four methods of Optimal Impute (Opt. impute). A simulation study was based on real-world educational data with a random coefficient model. The performance of these approaches under various conditions was investigated: 1) six types of missing data: Missing completely at random (MCAR), Missing at Random (MAR), Missing not at Random (MNAR), and three mixed types of missing data: MCAR-MAR, MCAR-MNAR and MAR-MNAR 2) the level 1 sample sizes: 1,000, 2,000, and 3,000 units and the level 2 sample sizes: 40, 50, and 60 units, 3) three missing rates of observations in the level 1 sample sizes which were three levels: 30%, 40%, and 50% respectively. The results showed that for the MCAR, Opt. cv method had the highest average efficiency; Opt. svm method was the most effective in MAR, MCAR-MAR and MAR-MNAR; and RF method was the most effective in MNAR and MCAR-MNAR. Therefore, the Opt. impute method tended to provide the highest average efficiency, followed by the RF method and the FCS method, respectively.

Keywords: Missing Data, Multilevel Data, Random Coefficients Model, Multiple Imputation, Simulation

Please cite this article as: N. Chimsud, P. Ratchaprapapornkul, and S. Srisuttiyakorn , "Comparison of the efficiency of multiple imputation in multilevel data: A simulation study," *The Journal of KMUTNB*, vol. 34, no. 2, pp. 1–14, ID. 242-126075, Apr.–Jun. 2024 (in Thai).

1. บทนำ

ถึงแม้ว่าปัจจุบันจะมีเทคโนโลยีการป้องกันไม่ให้เกิดค่าสูญหายในข้อมูลมากขึ้นหลากหลายวิธี แต่ปัญหาข้อมูลสูญหายก็ยังไม่หมดไป ซึ่งเป็นที่ทราบกันดีว่า “ค่าสูญหาย (Missing Value) ” มักเกิดขึ้นในงานวิจัยเกือบทุกสาขา ไม่ว่าจะเป็นงานวิจัยทางการแพทย์ สาธารณสุข และชีววิทยา โดยเฉพาะงานวิจัยด้านการศึกษา ที่มีการศึกษาพฤติกรรมของมนุษย์ นับเป็นปัญหาสำคัญที่ส่งผลกระทบให้ผลการวิจัยคลาดเคลื่อนไม่สามารถสะท้อนสภาพความจริงของประชากรได้ [1], [2]

เมื่อพิจารณาถึงบริบทการศึกษาของประเทศไทยพบว่า ธรรมชาติของข้อมูลถูกจัดเป็นหมวดหมู่ กลุ่มนักเรียนถูกจัดให้อยู่รวมกันเป็นห้องเรียนโดยแต่ละห้องเรียนจะถูกรวบรวมอยู่ในหมวดของโรงเรียนและในแต่ละโรงเรียน จะอยู่ภายใต้สังกัดของสำนักงานเขตพื้นที่ เป็นต้น จะสังเกตเห็นว่าลักษณะข้อมูลทางการศึกษามีโครงสร้างแบบพหุระดับ (Multilevel Data) ซึ่งในแต่ละระดับ มีความสัมพันธ์ระหว่างหน่วยข้อมูลซึ่งกันและกัน ทั้งนี้ปัญหาที่นักวิจัยมักพบคือการเก็บรวบรวมข้อมูลไม่ได้ครบตามจำนวนที่กำหนด โดยเฉพาะการเก็บข้อมูลส่วนบุคคลที่มีความอ่อนไหว (Sensitive Personal Data) เช่น รายได้ของผู้ปกครอง สถานะความเป็นอยู่ของครอบครัว ทั้งนี้หากผู้วิจัยวิเคราะห์ข้อมูลโดยไม่คำนึงถึงค่าสูญหาย อาจส่งผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูลได้ โดยทำให้สูญเสียสาระสำคัญของรายละเอียดบางอย่างไป [3], [4]

จากสถานการณ์ข้างต้นผู้วิจัยศึกษาประสิทธิภาพของวิธีทดแทนค่าสูญหายพบว่า ในปัจจุบันนิยมนำเทคนิค Multiple Imputation (MI) มาทดแทนค่าสูญหายในข้อมูลที่มีโครงสร้างแบบพหุระดับเนื่องจากเทคนิคนี้มีหลักการคือใช้การคำนวณหลายครั้งทำการวนซ้ำเพื่อให้ได้ค่าที่ดีที่สุด อย่างไรก็ตามเทคนิค MI มีหลายวิธี [2], [5]–[9] วิธีการที่ศึกษาในงานวิจัยนี้คือ Multiple Imputation Fully Conditional Specification (FCS) [10]–[12] นอกจากนี้ศึกษางานวิจัยของ Jia และ Wu [13] ให้ผลการวิจัยไม่แตกต่างกับงานวิจัยของ Nissen และคณะ [2] และ Kokla และคณะ [14] แสดงให้เห็นว่าวิธีทดแทนค่าสูญหาย Random Forest (RF) มีประสิทธิภาพสูงเช่นกันเมื่อเปรียบเทียบกับวิธีในอดีต ยิ่งกว่านั้น Bertsimas และคณะ [15] พัฒนาวิธี Optimal Impute (Opt.impute) ด้วยการให้โปรแกรมทำงานจากการเรียนรู้ (Machine Learning) สามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง ช่วยเพิ่มประสิทธิภาพให้กับวิธีทดแทนค่าสูญหายได้

จากผลการศึกษาชี้ให้เห็นว่าการดำเนินการทดแทนค่าสูญหายเป็นขั้นตอนที่สำคัญ โดยเฉพาะอย่างยิ่งในงานวิจัยทางการศึกษาที่มีความจำเป็นต้องเก็บข้อมูลส่วนบุคคลที่มีความอ่อนไหวมักพบค่าสูญหายเนื่องจากได้รับข้อมูลไม่ครบถ้วนตามจำนวนที่กำหนด ดังนั้นการวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีทดแทนค่าสูญหายแบบพหุของการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่ม (β) เมื่อค่าสูญหายเกิดขึ้นที่ตัวแปรอิสระระหว่างเก็บรวบรวมข้อมูลในการวิจัยทางการศึกษา โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) จำนวน 6 วิธี ได้แก่ วิธี FCS วิธี RF และ วิธี Opt.impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ วิธี Opt.knn วิธี Opt.tree วิธี Opt.svm และวิธี Opt.cv โดยใช้การจำลองข้อมูลด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม ประกอบด้วยเงื่อนไข ดังนี้ ประเภทของการสูญหาย 6 ประเภท กำหนดขนาดตัวอย่างสองระดับ เมื่อให้ขนาดของตัวอย่างระดับที่หนึ่งเท่ากับ 1,000 2,000 และ 3,000 หน่วย และขนาดตัวอย่างระดับที่สองเท่ากับ 40 50 และ 60 หน่วย ตามลำดับทั้งนี้กำหนดอัตราการสูญหายเป็น 3 ระดับ ได้แก่ ร้อยละ 30 40 และ 50 ตามลำดับ ผลการวิจัยในครั้งนี้จะให้องค์ความรู้ ซึ่งสามารถนำไปใช้ในการตัดสินใจเลือกวิธีทดแทนค่าข้อมูลสูญหายให้เหมาะสมกับสถานการณ์ที่ต้องการศึกษา เพื่อให้ได้สารสนเทศที่ถูกต้อง ลดความคลาดเคลื่อนที่อาจเกิดขึ้น และสามารถอนุมานไปสู่ประชากรได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

2. วัสดุ อุปกรณ์และวิธีการวิจัย

การวิจัยครั้งนี้ เป็นการจำลองข้อมูลทางการศึกษา โดย

ใช้เทคนิคมอนติคาร์โล แต่ละสถานการณ์ทำซ้ำ 100 รอบ เพื่อเปรียบเทียบวิธีทดแทนค่าสูญหายแบบพหุของ การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่มเมื่อค่าสูญหายเกิดขึ้นที่ตัวแปรอิสระระหว่างเก็บรวบรวมข้อมูลในการวิจัยทางการศึกษา โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม โดยมีวิธีการดำเนินการวิจัยดังนี้

2.1 ขอบเขตการศึกษา

การวิจัยครั้งนี้ ผู้วิจัยจำลองข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ โดยกำหนดค่าพารามิเตอร์ในการจำลองข้อมูลทางการศึกษา จากข้อมูลของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 ในโรงเรียนที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.)

เมื่อพิจารณาโครงสร้างของข้อมูลโรงเรียนในสังกัด สพม. ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวน 42 สังกัดนั้นพบว่า ในแต่ละสังกัดมีจำนวนโรงเรียนโดยเฉลี่ยสังกัดละ 23 โรงเรียน จากการวิเคราะห์ข้อมูลดังกล่าว ผู้วิจัยพบว่าข้อมูลมีการสูญหายแบบรายโรงเรียนโดยเฉลี่ยมากกว่าร้อยละ 30 ซึ่งอยู่ในระดับที่สูงมาก โดยข้อมูลสูญหายส่วนใหญ่พบในนักเรียนที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำและนักเรียนที่ไม่ได้พักอาศัยอยู่กับบิดาและมารดา เนื่องจากเป็นข้อมูลส่วนบุคคลที่กระทบต่อความรู้สึกได้ง่าย ดังนั้นจึงมีความเสี่ยงสูงที่ผู้วิจัยอาจไม่ได้รับข้อมูลที่เพียงพอสำหรับการนำข้อมูลมาวิเคราะห์ผลการวิจัย

นอกจากนี้หากพิจารณาบริบทของการศึกษาไทยพบว่าตัวชี้วัดที่สำคัญในการวัดคุณภาพทางการศึกษา คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียน จากเหตุผลที่กล่าวมานั้น ผู้วิจัยจึงกำหนด ตัวแปรอิสระ คือ สัดส่วนของนักเรียนที่ไม่ได้อาศัยกับบิดาและมารดา โดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำและตัวแปรตาม คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียน สำหรับการจำลองข้อมูลในครั้งนี้

ดังนั้นการเปรียบเทียบวิธีทดแทนค่าสูญหายแบบพหุในการวิจัยทางการศึกษาโดยใช้การจำลองข้อมูลในครั้งนี้ ผู้วิจัยจึงกำหนดเงื่อนไขในการจำลอง ดังนี้ ตัวแปรระดับที่หนึ่งคือระดับโรงเรียนมีขนาดเท่ากับ 1,000 2,000 และ 3,000 หน่วย และตัวแปรระดับสองคือระดับสังกัดมีขนาดเท่ากับ 40 50 และ 60 หน่วย ตามลำดับ โดยกำหนดค่าสูญหายที่ตัวแปรอิสระ ได้แก่ สัดส่วนของนักเรียนในโรงเรียนที่ไม่ได้อาศัยอยู่กับกับบิดาและมารดาโดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำ (ระดับที่หนึ่ง) มีอัตราการสูญหายที่ตัวแปรของข้อมูลในแต่ละกรณีที่แตกต่างกัน จำแนกตามประเภทของการสูญหาย 6 ประเภท โดยในแต่ละประเภทมีอัตราสูญหายเท่ากับร้อยละ 30 40 และ 50 ตามลำดับ โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม รายละเอียดดังสมการที่ (1)–(3)

โมเดลระดับโรงเรียน

$$y_{ij} = \alpha_{0j} + \beta_{1j}(x)_{ij} + \varepsilon_{ij} \quad (1)$$

โมเดลระดับสังกัด

$$\beta_{1j} = \gamma_{01} + U_{1j}, \quad (2)$$

$$\alpha_{0j} = \gamma_{00} + U_{0j} \quad (3)$$

เมื่อกำหนดให้

y_{ij} คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียน โดยเฉลี่ยของโรงเรียนที่ i และสังกัดที่ j

x_{ij} คือ สัดส่วนของนักเรียนที่ไม่ได้อาศัยกับบิดา และมารดาโดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำในโรงเรียนที่ i และสังกัดที่ j

γ_{00} คือ สัมประสิทธิ์จุดตัดแกน

γ_{01} คือ สัดส่วนของนักเรียนในโรงเรียนที่ไม่ได้อาศัยกับบิดาและมารดา โดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำ

$\varepsilon_{ij} \sim N(0, \sigma^2)$ คือ ความคลาดเคลื่อนสุ่มของโมเดลระดับโรงเรียน

$U_{ij} \sim \mathrm{N}(0, \sum u)$ คือ ความคลาดเคลื่อนสุ่มของโมเดลระดับสังกัด

โดยที่ $\gamma_{00} = -1.502 \text{ x } e^{-19}$, $U_{0j} \sim N(0,0.588)$ และ $\gamma_{01} = -1.02$, $U_{1j} \sim N(0,1.272)$ และ $\Sigma_u = \begin{bmatrix} 0.588 & 0 \\ 0 & 1.272 \end{bmatrix}$

คือ เมทริกซ์ความแปรปรวนของเวกเตอร์ความคลาดเคลื่อนสุ่มในระดับสังกัด

เมื่อพิจารณาประเภทการสูญหายของข้อมูล (Type of Missing Data) โดยจำแนกประเภทการสูญหายของข้อมูลได้ 3 ประเภทหลัก [1], [16] ซึ่งมีรายละเอียดดังต่อไปนี้

การสูญหายแบบสุ่มสมบูรณ์ (MCAR) เป็นการสูญหายที่ลักษณะของข้อมูลสูญหายเกิดขึ้นอย่างสุ่มสมบูรณ์ข้อมูลที่สูญหายเป็นอิสระหรือไม่มีความสัมพันธ์กับตัวแปรสังเกตได้ทั้งที่ทราบค่าและไม่ทราบค่า [1], [16] ซึ่งสามารถอธิบายได้ด้วยสมการที่ (4)

$$P\,(missing \mid complete\ data) = P\,(missing) \qquad (4)$$

การสูญหายแบบสุ่ม (MAR) อาจเรียกว่าการสูญหายแบบสุ่มที่มีเงื่อนไข เป็นการสูญหายที่ลักษณะของข้อมูลที่สูญหายเกิดขึ้นอย่างสุ่มภายในบางส่วน หรือบางกลุ่มของค่าสังเกตหรือค่าที่สูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในข้อมูลที่สนใจศึกษา [1], [16] ซึ่งสามารถอธิบายได้ด้วยสมการที่ (5)

$$P(missing \mid complete\ data) = P(missing \mid observed\ data) \qquad (5)$$

การสูญหายแบบไม่สุ่ม (MNAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม โดยที่ค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกันหรือตัวแปรตัวอื่นภายนอกข้อมูลที่สนใจศึกษา [1], [16] ซึ่งสามารถอธิบายได้ด้วยสมการที่ (6)

$$P(missing \mid complete\ data) \neq P(missing \mid observed\ data) \qquad (6)$$

ในทางปฏิบัติผู้วิจัยไม่สามารถทราบได้อย่างแน่ชัดว่าการสูญหายเกิดขึ้นในรูปแบบใด และข้อมูลจริงมีความเป็นไปได้น้อยมากที่จะเกิดการสูญหายเพียงสาเหตุเดียว ดังนั้นผู้วิจัยจึงสนใจศึกษาประเภทของการสูญหายรูปแบบผสมรายคู่เพิ่มเติมจากที่กล่าวไว้ข้างต้น รวมทั้งหมด 6 รูปแบบ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MNAR และ MAR-MNAR ตามลำดับ

วิธีทดแทนค่าสูญหาย (Methods) ซึ่งมีรายละเอียดดังนี้

1) หลักการทดแทนค่าสูญหายด้วยวิธี FCS

การทดแทนค่าสูญหายด้วยวิธี FCS ใช้หลักการทดแทนค่าสูญหายด้วยเทคนิค MI กล่าวคือการทดแทนค่าสูญหายด้วยชุดข้อมูลของค่าที่เป็นไปได้มากกว่า 1 ค่า ใช้การคำนวณหลายครั้ง เพื่อให้ได้ค่าที่ดีที่สุด ทั้งนี้การทดแทนค่าสูญหายด้วยเทคนิค MI มีหลายวิธีด้วยกัน เช่น สร้างการประมาณค่าโดยใช้วิธี Predictive Mean Matching, Bayesian Linear Regression, Logistic ในงานวิจัยนี้เลือกใช้ การกำหนดค่าด้วย วิธี FCS มีหลักการคือใช้สมการทดแทนค่าสูญหายในลักษณะวนซ้ำ อันดับแรกสร้างสมการทำนายค่าสูญหายตัวแปรแรกด้วยการกำหนดแบบจำลองการทดแทนค่าสูญหาย โดยให้ตัวแปรที่สูญหายแต่ละตัวเป็นตัวแปรตาม (y) หลังจากนั้นวนไปทำนายตัวแปรที่สอง (อิงตัวแปรแรกที่ทดแทนค่าสูญหายเรียบร้อยแล้ว) ต่อมาวนไปยังตัวแปรที่สามและไปจนครบตัวแปรสุดท้าย แล้วจึงวนกลับมายังตัวแปรที่หนึ่งทำเช่นนี้ไปไปเรื่อย ๆ จนครบจำนวนรอบที่กำหนด โดยผู้วิเคราะห์จะสามารถกำหนดจำนวนชุดข้อมูลและจำนวนรอบในการวนได้ [10], [11] โดยวิธีนี้ใช้ “mice” Package ในโปรแกรม R สำหรับการวิเคราะห์วิธีทดแทนค่าสูญหาย

2) หลักการทดแทนค่าสูญหายด้วยวิธี RF

แนวคิดของวิธี Random Forest Imputation คือการรวมกันของทฤษฎี Bagging (Bootstrap Aggregation) ซึ่งเป็นพื้นฐานของ Random Forest Classification โดยมีหลักการคือการสร้างโมเดลจากการใช้แผนภาพต้นไม้ตัดสินใจ (Decision Tree) หลายโมเดลในการวิเคราะห์แต่ละครั้ง มีกระบวนการโดยในแต่ละโมเดลจะได้รับชุดข้อมูล ที่แตกต่างกันซึ่งเป็นซับเซตของชุดข้อมูลทั้งหมด และขณะการทำนายค่าข้อมูลสูญหายจะกำหนดให้แต่ละแผนภาพต้นไม้ทำนายค่าข้อมูลสูญหายในแต่ละโมเดล โดยแต่ละโมเดลเป็นอิสระต่อกัน หลังจากนั้นคำนวณผลการทำนายค่าข้อมูลสูญหายด้วยการโหวตผลลัพธ์ (Vote Output) ซึ่งผลลัพธ์ที่ถูกเลือกโดยแผนภาพต้นไม้ตัดสินใจมากที่สุด โดยวิธีนี้ใช้ “missForest” Package ในโปรแกรม R สำหรับวิเคราะห์การทดแทนค่า

สูญหาย ดังนั้นผลลัพธ์ที่ได้จากการวิเคราะห์จะมีความถูกต้องแม่นยำและมีประสิทธิภาพที่สูง

3) หลักการทดแทนค่าสูญหายด้วยวิธี Opt.impute

วิธี Opt.impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ Opt.knn, Opt.svm, Opt.tree และ Opt.cv หลักการคือการรวมแนวคิดการทดแทนค่าสูญหาย จำนวน 3 วิธี [16] ได้แก่ K-Nearest Neighbors based (Opt.knn), Support Vector Machines based (Opt.svm) และ Decision Tree based (Opt.tree) โดยสร้างการแทนที่ค่าสูญหายหลายครั้งด้วยเทคนิค MI ซึ่งจะช่วยประมาณค่าได้ดีกว่าการใส่แบบจำลองกับชุดข้อมูลเพียงครั้งเดียว โดยปรับค่าในจุดข้อมูลสูญหายและข้อมูลทั้งหมดพร้อมกันเพื่อทดแทนค่าสูญหายให้เหมาะสม สามารถใช้ได้ทั้งตัวแปรต่อเนื่องและไม่ต่อเนื่อง นอกจากนี้พิจารณาวิธีการผสมสามวิธี ได้แก่ Opt.cv (optimal cross-validated) ซึ่งคำนวณจาก Opt.knn, Opt.svm และ Opt.tree ตามลำดับ โดยวิธีนี้ใช้โปรแกรมจูเลีย Julia และ Interface to 'Interpretable AI' Modules หรือ “iai” Package (AI software) ในโปรแกรม R สำหรับวิเคราะห์การทดแทนค่าสูญหาย

เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย จากการทำซ้ำ 100 รอบ ของแต่ละสถานการณ์สำหรับการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่ม เมื่อ Mean คือ ค่าเฉลี่ยตามการจำลอง 100 รอบ สามารถพิจารณาจากค่าเฉลี่ยของค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (Normalized Root Mean Square Error; NRMSE) และค่าเฉลี่ยของค่าความเอนเอียงสัมพัทธ์ (Relative Biased; RB) เมื่อกำหนดให้

$N_{1\ell}$ คือ ขนาดของตัวอย่างระดับที่หนึ่งหรือจำนวนโรงเรียน โดยที่ N_{11} = 1,000, N_{13} = 2,000, N_{13} = 3,000

$N_{2\ell}$ คือ ขนาดของตัวอย่างระดับที่สอง หรือจำนวนสังกัด โดยที่ N_{21} = 40, N_{22} = 50, N_{23} = 60

θ_j คือ ค่าประมาณพารามิเตอร์ β จากข้อมูลจริงของสังกัดที่ j

$\hat{\theta}_{ij}$ คือ ค่าประมาณพารามิเตอร์ β ของโรงเรียนที่ i สังกัดที่ j

$\overline{\hat{\theta}_j}$ คือ ค่าเฉลี่ยของการประมาณพารามิเตอร์ β จำนวน $N_{2\ell}$

$var(\theta_j)$ คือ ค่าความแปรปรวนของค่าประมาณพารามิเตอร์ β

จากข้อมูลจริงของสังกัดที่ j เมื่อ j คือ สังกัดที่ j โดยที่ j = 1,2,3,....,$N_{2\ell}$

N คือ การจำลองรอบที่ N โดยที่ N = 1,2,3,....,100 โดยสามารถคำนวณได้จากสมการที่ (7)–(8) ดังนี้

$$NRMSE\left(\overline{\hat{\theta}_j}\right)=\sqrt{\frac{mean\left(\sum_{j=1}^{N_2}(\theta_j-\hat{\theta}_j)^2\right)^1}{var\ (\theta_j)}} \quad (7)$$

$$RB\left(\overline{\hat{\theta}_j}\right)=\frac{1}{\theta_j}\sum_{N=1}^{100}\left(\frac{\sum_{j=1}^{N_2}(\theta_j-\hat{\theta}_j)^2}{100}\right) \quad (8)$$

2.2 ขั้นตอนการวิจัย

2.2.1 ขั้นที่ 1 ผู้วิจัยจำลองข้อมูลสมบูรณ์ทางการศึกษาที่มีโครงสร้างแบบพหุระดับเมื่อระดับที่หนึ่งคือระดับโรงเรียนและระดับที่สองคือระดับสังกัด โดยกำหนดค่าพารามิเตอร์จากข้อมูลจริงของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 ในโรงเรียนที่อยู่ในสังกัด สพม. ในการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โลและสร้างโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม จากสมการที่ (1)–(3) กำหนดตัวแปรอิสระ คือ สัดส่วนของนักเรียนในโรงเรียนที่ไม่ได้อาศัยอยู่กับบิดาและมารดา โดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำ (x) และกำหนดให้ตัวแปรตาม คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียน โดยเฉลี่ยระดับโรงเรียน (y)

ผู้วิจัยกำหนดการแจกแจงของค่าพารามิเตอร์ตามลักษณะของข้อมูลจริงข้างต้น ในการสร้างแบบจำลองครั้งนี้ โดยกำหนดให้สัดส่วนของนักเรียนในโรงเรียนที่ไม่ได้อาศัยอยู่กับบิดาและมารดา โดยที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำ (x) มีการแจกแจงแบบเบต้า (Beta Distribution) สามารถ

เขียนแทนได้ด้วย $x \sim \text{Beta}(\alpha, \beta)$ เมื่อกำหนดให้ α = 0.0579, β = 15.278 สร้างโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม โดยนำค่าพารามิเตอร์ที่สร้างขึ้นมาคำนวณหาค่าผลสัมฤทธิ์ทางการเรียน (y_{ij}) ของนักเรียนของโรงเรียนที่ i สังกัดที่ j ในสมการที่ (1)–(3)

2.2.2 ขั้นที่ 2 นำชุดข้อมูลที่ได้ จากการจำลองขั้นที่ 1 ดำเนินการตามรูปแบบการสูญหายที่ตัวแปรในแต่ละกรณีที่แตกต่างกัน 6 รูปแบบ มีรายละเอียดดังนี้

กรณีที่ 1: การสูญหายแบบสุ่มสมบูรณ์ (MCAR)

กำหนดให้ x สูญหายแบบสุ่มสมบูรณ์ โดยที่ค่า x ที่สูญหาย ไม่มีความสัมพันธ์ ไม่มีเกี่ยวข้อง หรือไม่ได้ขึ้นอยู่กับค่า x และ y ที่เก็บรวบรวมได้ เช่น สาเหตุของการไม่ตอบคำถามของนักเรียนมาจาก นักเรียนลืมตอบ นักเรียนล้มป่วย หรือคุณครูในโรงเรียนนำเข้าข้อมูลผิดพลาด เป็นต้น จะเห็นว่าสาเหตุของการไม่ตอบไม่มีความเกี่ยวข้องกับคำถามที่เก็บรวบรวมได้ สามารถอธิบายได้ ด้วยสมการที่ (9)

$$P((x)\ Missing \mid complete\ data) = P((x,y)\ Missing) \tag{9}$$

กรณีที่ 2: การสูญหายแบบสุ่ม (MAR)

กำหนดให้ ค่า x สูญหายอย่างสุ่ม หรืออาจเรียกว่าสูญหายแบบมีเงื่อนไข เมื่อให้ค่าสูญหาย x มีความสัมพันธ์หรือขึ้นอยู่กับค่า y ที่เก็บรวบรวมได้ เช่น สาเหตุของการไม่ตอบคำถามด้านความสัมพันธ์ภายในครอบครัวของนักเรียนขึ้นอยู่กับผลสัมฤทธิ์ทางการเรียนของนักเรียน โดยนักเรียนที่มีผลสัมฤทธิ์ทางการเรียนต่ำบางคนจะไม่ตอบคำถาม การพักอาศัยอยู่กับกับบิดาและมารดาและฐานะทางเศรษฐกิจของครอบครัว เป็นต้น สามารถอธิบายได้ด้วยสมการที่ (10)

$$P(x\ missing \mid complete\ data) = P(x\ missing \mid y\ observed) \tag{10}$$

กรณีที่ 3: การสูญหายแบบไม่สุ่ม (MNAR)

กำหนดให้ ค่า x ที่สูญหายมีความสัมพันธ์หรือขึ้นอยู่กับ x ที่สูญหายเอง เช่น กลุ่มนักเรียนที่ไม่ได้อาศัยอยู่กับบิดาและมารดาและครอบครัวมีฐานะทางเศรษฐกิจต่ำเป็นกลุ่มนักเรียนที่มีแนวโน้มจะไม่ให้ข้อมูลเนื่องจากไม่ได้อาศัยอยู่กับบิดาและมารดาและครอบครัวมีฐานะทางเศรษฐกิจต่ำ ทั้งนี้จะเห็นว่าสถานการณ์การสูญหายในลักษณะนี้ มีโอกาสหรือความน่าจะเป็นของค่าสูญหายขึ้นอยู่กับค่าของตัวแปรที่มีค่าสูญหายเองซึ่งเป็นข้อมูลที่ผู้วิจัยไม่ทราบค่า โดยสามารถอธิบายได้ด้วยสมการที่ (11)

$$P(x\ missing \mid complete\ data) \neq P(x\ missing \mid (x,y)\ observed) \tag{11}$$

กรณีที่ 4: การสูญหายแบบผสมรายคู่ MCAR–MAR

นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการให้มีการสุ่มการสูญหายแบบ MCAR ตามกรณีที่ 1 หลังจากนั้นนำชุดข้อมูลที่ได้มาสุ่มการสูญหายแบบ MAR ตามกรณีที่ 2 จะได้การสูญหายแบบผสมรายคู่ MCAR - MAR

กรณีที่ 5: การสูญหายแบบผสมรายคู่ MCAR–MNAR

นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการกำหนดให้มีการสุ่มการสูญหายแบบ MCAR ตามกรณีที่ 1 หลังจากนั้นนำชุดข้อมูลที่ได้ สุ่มการสูญหายแบบ MNAR ตามกรณีที่ 3 จะได้การสูญหายแบบผสมรายคู่ MCAR–MNAR

กรณีที่ 6: การสูญหายแบบผสมรายคู่ MAR–MNAR

นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการ กำหนดให้มีการสุ่มการสูญหายแบบ MAR ตามกรณีที่ 2 หลังจากนั้นนำชุดข้อมูลที่ได้สุ่มการสูญหายแบบ MNARตามกรณีที่ 3 จะได้การสูญหายแบบผสมรายคู่ MCAR-MNAR

เมื่อดำเนินการจำลองข้อมูลสูญหายแยกตามประเภทของการสูญหาย 6 ประเภท ในข้างต้นเรียบร้อยแล้ว เพื่อให้ข้อมูลสูญหายเป็นไปตามเงื่อนไขที่กำหนด ผู้วิจัยจึงตรวจสอบเงื่อนไขและความถูกต้องของโปรแกรมที่ใช้ในการจำลองข้อมูล ภายใต้ประเภทของการสูญหายหลัก 3 รูปแบบ มีรายละเอียดดังนี้

กรณีที่ 1: เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MCAR โดยใช้ Little's Test of

MCAR ในการทดสอบสมมติฐาน ดังนี้

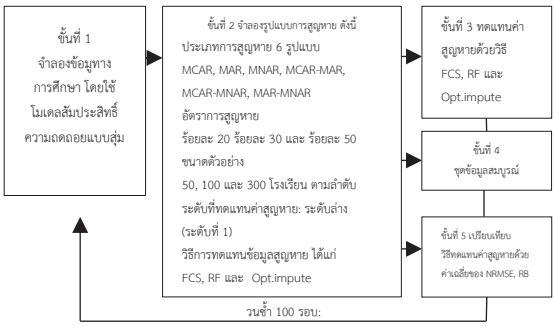
H_0 : การสูญหายในการจำลองข้อมูลเป็นแบบ MCAR

H_1 : การสูญหายในการจำลองข้อมูลไม่เป็นแบบ MCAR

ผลการทดสอบพบว่า p-value = 0.838 จะเห็นว่า ค่า p-value > 0.05 ดังนั้นจึงสามารถสรุปได้ว่า การสูญหายที่เกิดขึ้นในตัวแปร x ในการจำลองข้อมูล สถานการณ์นี้ เป็นไปตามเงื่อนไขที่กำหนด

กรณีที่ 2: เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MAR ผู้วิจัยดำเนินการแปลงตัวแปร x ที่สูญหายให้มีค่าเท่ากับ 1 และค่ารวบรวมได้ในตัวแปร x ให้เท่ากับ 0 หลังจากนั้นทดสอบความสัมพันธ์ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสูญหายในตัวแปร x กับค่ารวบรวมได้ในตัวแปร y พบว่า มีความสัมพันธ์ในทิศทางบวกเท่ากับ 0.3202 ดังนั้นจึงสามารถสรุปได้ว่าเกิดการสูญหายแบบมีเงื่อนไข หรือการสูญหายแบ MAR ดังนั้นการจำลองข้อมูลในสถานการณ์นี้เป็นไปตามเงื่อนไขที่กำหนด

กรณีที่ 3: เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MNAR ด้วยการทดสอบสมมติฐานโดยใช้ Little's Test of MCAR ผลทดสอบพบว่า ปฏิเสธสมมติฐานคือข้อมูลที่จำลองขึ้นมีลักษณะ การสูญหายแบบ MCAR อย่างมีนัยสำคัญทางสถิติ (p-value = 0.000) กล่าวคือข้อมูลที่จำลองขึ้นไม่ได้เกิดขึ้นอย่างสุ่มสมบูรณ์ เมื่อพิสูจน์ได้แล้วว่าการจำลองข้อมูลสูญหายไม่ได้เป็น MCAR อย่างไรก็ตามการสูญหายแบบ MNAR เป็นการสูญหายที่มีสาเหตุมาจากตัวแปรที่เราไม่ทราบค่า หรือไม่มีข้อมูลให้นำมาวิเคราะห์ได้ ดังนั้นในทางปฏิบัติจึงทำได้เพียงตรวจสอบจากข้อมูลที่เก็บรวบรวมได้ว่ามีแนวโน้มที่จะมีการสูญหายแบบ MAR หรือไม่ หากผลการตรวจสอบอธิบายการสูญหายแบบ MAR ได้น้อย จะทำให้สามารถสรุปได้ว่าข้อมูลมีแนวโน้มที่จะเป็น MNAR มากขึ้น โดยพิจารณาความสัมพันธ์ของข้อมูลที่รวบรวมได้กับข้อมูลสูญหาย ในตัวแปร x พบว่า ค่าสูญหายที่เกิดขึ้นมีความสัมพันธ์กับค่าที่รวบรวมได้ y เพียง 0.002 ดังนั้นสรุปได้ว่ามีแนวโน้มที่ข้อมูลจำลองในกรณีนี้เป็นไปตามเงื่อนไขการสูญหายแบบ MNAR



รูปที่ 1 ขั้นตอนการจำลองข้อมูล

ทั้งนี้ในส่วนของการตรวจสอบข้อมูลสูญหายแบบผสม 3 รูปแบบ ได้แก่ MCAR–MAR, MCAR–MNAR และ MAR–MNAR ผู้วิจัยมีหลักการตรวจสอบเช่นเดียวกับประเภทของการสูญหายหลัก 3 รูปแบบ ดังตารางที่ 1

ตารางที่ 1 ผลการตรวจสอบเงื่อนไขการสูญหายในการจำลองข้อมูล

เงื่อนไข	ค่าสูญหาย	ค่ารวบรวมได้	สถิติทดสอบ
MCAR	x	x และ y	Little's MCAR test (p-value = 0.838)
MAR	x	y	Correlation test (ρ_{xy} = 0.3202)
MNAR	x	x และ y	Little's MCAR test (p-value = 0.000) and Correlation test (ρ_{xy} = 0.0002)

2.2.3 ขั้นที่ 3 ทดแทนค่าสูญหายด้วยวิธี FCS วิธี RF และวิธี Opt.impute โดยที่ในวิธี Opt.impute ประกอบด้วยวิธี Opt.knn, Opt.tree, Opt.svm และ Opt.cv หลังจากนั้นจะได้ชุดข้อมูลสมบูรณ์ในขั้นที่ 4 นำมาเปรียบเทียบประสิทธิภาพด้วยค่า NRMSE และ RB ในขั้นที่ 5 ตามลำดับ การศึกษาด้วยการจำลองข้อมูลในครั้งนี้ ประกอบด้วย 5 ขั้นตอนที่สำคัญ โดยสามารถสรุปรายละเอียด ดังรูปที่ 1

ตารางที่ 2 ค่าเฉลี่ยของ NRMSE และ RB ในการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่มในการวิจัยทางการศึกษา

เงื่อนไข				เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่ม (β)											
				NRMSE						RB					
Type of Missing	N_1	N_2	M	FCS	RF	Opt. knn	Opt. tree	Opt. svm	Opt.cv	FCS	RF	Opt. knn	Opt. tree	Opt. svm	Opt.cv
MCAR	1,000	40	30	0.12	0.06	0.07	0.06	0.06	**0.05**	−0.08	−0.37	−0.25	−0.09	**−0.06**	−0.12
			40	0.13	0.06	**0.01**	0.09	0.09	0.07	−0.50	−0.13	−0.40	0.04	**−0.01**	−0.17
			50	0.14	0.10	0.08	0.11	0.11	**0.01**	−4.18	1.55	−0.55	−0.42	**−0.04**	−0.26
	2,000	50	30	0.12	0.09	0.09	0.07	0.07	**0.06**	**−0.04**	−0.28	−0.39	−0.08	**−0.04**	−1.69
			40	0.14	**0.06**	0.10	0.12	0.09	0.08	−0.56	−0.18	−0.32	**−0.12**	−0.14	−0.20
			50	0.17	**0.09**	0.11	0.11	0.11	**0.09**	−0.39	**−0.30**	−0.44	0.31	−0.37	**0.30**
	3,000	60	30	0.15	0.11	0.09	0.07	0.07	**0.06**	**−0.01**	−0.22	−0.32	−0.11	−0.42	−0.36
			40	0.12	0.09	0.10	0.09	**0.01**	0.09	**0.27**	−0.42	−0.31	−0.32	**−0.27**	−0.76
			50	0.14	**0.07**	0.10	0.10	0.11	0.09	−0.69	0.56	−0.55	−0.34	**−0.14**	−0.48
MAR	1,000	40	30	0.15	0.10	0.12	**0.09**	**0.09**	**0.09**	−0.40	−0.22	**−0.11**	−0.27	−0.46	−0.27
			40	0.18	0.13	0.15	0.12	**0.11**	0.12	−0.49	−0.35	0.95	**−0.05**	**−0.05**	**−0.05**
			50	0.19	0.14	0.15	0.13	**0.12**	**0.13**	−0.50	−0.39	−0.27	**−0.25**	−0.66	**−0.25**
	2,000	50	30	0.14	0.10	0.11	0.11	**0.09**	0.10	−0.54	−0.33	−0.25	−0.26	**−0.20**	−0.26
			40	0.15	0.15	0.12	0.12	**0.10**	0.11	−0.34	−0.40	−0.43	−0.34	**−0.20**	−0.33
			50	0.17	0.15	0.13	0.12	**0.10**	0.12	−0.60	−0.35	−0.58	−0.26	**−0.19**	−0.26
	3,000	60	30	0.11	0.07	0.09	**0.08**	**0.08**	**0.08**	−0.40	−0.15	0.41	−0.21	**−0.12**	−0.21
			40	0.13	0.11	0.11	**0.08**	**0.08**	**0.08**	−0.59	−0.38	−0.46	−0.31	**−0.26**	−0.30
			50	0.16	0.12	0.12	0.10	**0.08**	0.10	−0.66	0.45	−0.53	−0.45	**−0.30**	−0.44
MNAR	1,000	40	30	0.12	0.05	0.08	**0.05**	0.09	0.07	−0.26	−0.05	−0.12	−0.21	−0.20	−0.19
			40	0.14	**0.07**	0.09	**0.07**	0.11	0.10	−0.57	**−0.18**	−0.30	−0.40	−0.44	−0.42
			50	0.19	**0.09**	0.10	**0.09**	**0.09**	**0.09**	0.56	**0.31**	−0.29	−0.89	−0.81	−0.83
	2,000	50	30	0.13	**0.05**	0.07	0.07	0.07	0.07	3.00	−0.06	−0.01	−0.04	**−0.05**	−0.08
			40	0.13	**0.07**	0.08	0.08	0.09	0.08	−0.11	−0.14	−0.19	−0.15	**−0.08**	−0.12
			50	0.16	**0.09**	**0.09**	**0.09**	**0.09**	**0.09**	0.51	**0.10**	−0.48	−0.65	−0.72	−0.67
	3,000	60	30	0.12	**0.06**	0.08	0.08	0.09	0.08	−0.27	−0.26	−0.29	−0.23	−0.23	**−0.22**
			40	0.14	**0.07**	0.08	0.08	0.08	**0.07**	0.15	−0.25	−0.15	**−0.02**	−0.14	−0.22
			50	0.15	0.08	0.10	0.09	0.10	0.10	−1.75	−0.48	**−0.12**	−0.32	−0.29	−0.34
MCAR-MAR	1,000	40	30	0.16	0.10	0.10	0.07	**0.06**	0.07	−1.54	−1.05	**−0.32**	−0.33	**−0.32**	**−0.32**
			40	0.18	0.13	0.12	0.10	**0.07**	0.08	−1.48	1.47	**−0.23**	0.95	1.39	1.39
			50	0.22	0.22	0.14	**0.09**	0.11	0.11	−1.66	−0.69	−0.67	**−0.25**	−0.50	−0.26
	2,000	50	30	0.12	0.08	0.10	0.09	0.08	**0.07**	−0.61	−0.14	0.01	0.04	**0.02**	−0.11
			40	0.17	**0.10**	0.11	0.11	0.11	0.11	−2.21	−0.69	−0.67	0.67	**−0.37**	−0.67
			50	0.17	**0.12**	0.13	**0.12**	**0.12**	**0.12**	0.12	−0.21	−1.23	−0.67	**0.02**	−0.67
	3,000	60	30	0.10	**0.05**	0.06	**0.05**	**0.05**	**0.05**	−0.96	**−0.07**	**−0.07**	−0.28	−0.27	−0.28
			40	0.13	0.07	0.08	**0.06**	0.07	**0.06**	**−0.19**	−0.34	−0.29	−0.37	−0.30	−0.37
			50	0.17	0.12	**0.10**	0.12	**0.10**	**0.10**	−0.75	−0.44	−0.43	−0.42	−0.27	−0.43

ตารางที่ 2 ค่าเฉลี่ยของ NRMSE และ RB ในการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่มในการวิจัยทางการศึกษา (ต่อ)

เงื่อนไข				เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่ม (β)											
				NRMSE						RB					
Type of Missing	N_1	N_2	M	FCS	RF	Opt. knn	Opt. tree	Opt. svm	Opt.cv	FCS	RF	Opt. knn	Opt. tree	Opt. svm	Opt.cv
MCAR-MNAR	1,000	40	30	0.14	0.09	0.09	0.07	0.04	**0.01**	–0.09	–0.13	–0.37	–0.14	**–0.07**	0.14
			40	0.16	0.11	0.10	**0.09**	**0.09**	0.10	–0.61	–0.45	–0.22	–0.28	**–0.20**	–0.28
			50	0.20	0.13	0.11	0.11	**0.08**	0.09	**–0.25**	–0.41	–0.59	–0.82	–0.72	–0.82
	2,000	50	30	0.15	**0.08**	0.09	**0.08**	0.09	**0.08**	6.96	1.33	**0.08**	–0.16	–0.16	–0.16
			40	0.16	**0.09**	0.11	0.12	0.11	0.12	–0.58	**–0.18**	–0.30	–0.34	–0.29	–0.34
			50	0.17	**0.09**	0.12	0.11	0.12	0.11	–0.58	**–0.14**	–0.42	–0.36	–0.32	–0.36
	3,000	60	30	0.12	0.05	**0.02**	0.05	0.05	0.05	0.64	0.93	–0.23	–0.22	–0.21	–0.22
			40	0.12	**0.06**	0.07	0.07	0.07	0.07	–0.86	**–0.10**	1.39	0.88	0.66	0.88
			50	0.17	**0.09**	**0.09**	**0.09**	**0.09**	**0.09**	–0.40	**–0.15**	–0.19	–0.32	–0.23	–0.32
MAR-MNAR	1,000	40	30	0.14	0.06	0.11	0.10	0.05	0.09	–0.84	–0.25	**–0.30**	–0.16	–0.21	–0.24
			40	0.18	0.12	0.12	0.13	**0.08**	0.10	1.11	–1.71	–0.13	–1.18	**–0.03**	–1.98
			50	0.19	0.11	0.13	0.12	**0.09**	0.11	1.79	0.13	–0.52	–0.79	**–0.07**	0.47
	2,000	50	30	0.13	0.09	0.10	0.09	**0.02**	0.07	–0.62	–0.22	–0.20	–0.06	**–0.01**	–0.12
			40	0.14	0.10	0.11	0.11	**0.08**	0.10	–0.56	**–0.27**	–0.37	–0.44	–0.44	–0.42
			50	0.16	**0.12**	0.13	0.13	0.13	**0.12**	–1.52	–0.29	–0.77	–0.57	–0.56	**–0.50**
	3,000	60	30	0.12	0.08	**0.07**	0.09	0.09	0.08	–0.15	–0.25	**–0.03**	–0.34	–0.37	–0.32
			40	0.15	**0.10**	**0.10**	0.11	0.11	0.11	1.29	–0.10	**–0.04**	–0.47	–0.98	–1.24
			50	0.16	0.19	**0.09**	0.10	0.10	**0.09**	0.12	0.01	**–0.01**	–0.57	–0.51	–0.04

หมายเหตุ: N_1 และ N_2 แทน ขนาดตัวอย่างระดับที่หนึ่งและขนาดตัวอย่างระดับที่สอง, M แทน อัตราสูญหายในระดับที่หนึ่ง, ตัวหนา หมายถึงค่าเฉลี่ยของ NRMSE และ RB ต่ำที่สุด

2.3 การวิเคราะห์ข้อมูล

ผู้วิจัยวิเคราะห์ข้อมูลโดยใช้โปรแกรม R version 4.2.1 ในการจำลองข้อมูลและโปรแกรม Julia 1.7.2 เพื่อเปรียบเทียบประสิทธิภาพวิธีทดแทนค่าสูญหายแบบพหุ 6 วิธี ได้แก่ วิธี FCS, วิธี RF และวิธี Opt.impute ประกอบด้วย วิธี Opt. knn, วิธี Opt.tree, วิธี Opt.svm และวิธี Opt.cv ใช้เกณฑ์ค่าเฉลี่ยของค่า NRMSE เพื่อทดสอบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุในภาพรวม และใช้ค่า RB เพื่อตรวจสอบความเอนของวิธีทดแทนค่าสูญหายแบบพหุ จากการทำซ้ำ 100 รอบ ทั้งนี้เปรียบเทียบค่าดังกล่าว ด้วยการพิจารณาค่าที่น้อยที่สุดจะมีประสิทธิภาพสูงที่สุด

3. ผลการทดลอง

ผู้วิจัยมุ่งศึกษาประสิทธิภาพของวิธีการทดแทนค่าสูญหายด้วยการจำลองข้อมูลทางการศึกษาโดยใช้ข้อมูล ทางการศึกษาที่มีลักษณะพหุระดับด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มภายใต้ประเภทของการสูญหาย ขนาดตัวอย่าง และอัตราการสูญหายที่แตกต่างกัน แบ่งการนำเสนอตามเกณฑ์ที่ใช้เปรียบเทียบประสิทธิภาพวิธีการทดแทนสูญหายแบบพหุ 6 วิธี ได้แก่ วิธี FCS วิธี RF วิธี Opt.knn วิธี Opt.tree วิธี Opt.svm และวิธี Opt.cv โดยมีรายละเอียดดังนี้

3.1 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุด้วยค่าเฉลี่ยของ NRMSE

ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุด้วยค่าเฉลี่ยของ NRMSE จากตารางที่ 2 ในภาพรวมพบว่า ส่วนใหญ่วิธีทดแทนค่าสูญหาย Opt.impute ให้ค่าเฉลี่ยของ NRMSE ต่ำที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF ขณะที่วิธีทดแทน ค่าสูญหาย FCS มีค่าเฉลี่ยของ NRMSE ค่อนข้างสูงที่สุด

เมื่อพิจารณาค่าเฉลี่ยของ NRMSE จำแนกตามประเภทของการสูญหายรายเงื่อนไข พบว่า

เมื่อพิจารณาการสูญหายแบบ MCAR พบว่า วิธีทดแทนค่าสูญหาย Opt.cv มีแนวโน้มให้ค่าเฉลี่ยของ NRMSE ต่ำที่สุด เมื่อตัวอย่างระดับที่หนึ่งเท่ากับ 1,000 และตัวอย่างระดับที่สองเท่ากับ 40 โดยมีค่าเฉลี่ยของ NRMSE อยู่ระหว่างช่วง 0.01–0.09 ทั้งนี้จะสังเกตได้ว่าค่าเฉลี่ยของ NRMSE มีแนวโน้มสูงขึ้นเมื่ออัตราการสูญหายเพิ่มขึ้นสูงขึ้น

ทั้งนี้หากพิจารณาการสูญหายแบบ MAR พบว่า วิธีทดแทนค่าสูญหาย Opt.svm มีแนวโน้มให้ค่าเฉลี่ยของ NRMSE ต่ำที่สุด โดยมีค่าเฉลี่ยของ NRMSE อยู่ระหว่างช่วง 0.08–0.12 รองลงมาคือพบว่าวิธีทดแทนค่าสูญหาย Opt.cv และวิธีทดแทนค่าสูญหาย Opt.tree ตามลำดับ ทั้งนี้จะสังเกตได้ว่าค่าเฉลี่ยของ NRMSE

นอกจากนี้พิจารณาการสูญหายแบบ MNAR พบว่า วิธีทดแทนค่าสูญหาย RF มีค่าเฉลี่ยของ NRMSE ต่ำที่สุด ซึ่งอยู่ระหว่างช่วง 0.05–0.09 รองลงคือวิธีทดแทนค่าสูญหาย Opt.tree โดยมีค่าเฉลี่ยของ NRMSE อยู่ระหว่างช่วง 0.05–0.09 ขณะที่พบว่า วิธีทดแทนค่าสูญหาย Opt.knn, Opt.svm และ Opt.cv ให้ค่าเฉลี่ยของ NRMSE ไม่แตกต่างกัน

เมื่อลักษณะการสูญหายเกิดขึ้นแบบผสมรายคู่ MCAR–MAR ในภาพรวมพบว่า วิธีทดแทนค่าสูญหาย Opt.impute ให้ค่าเฉลี่ย NRMSE ค่อนข้างต่ำที่สุดอยู่ในช่วง 0.05–0.15 โดยเมื่อพิจารณาค่าเฉลี่ยของ NRMSE รายเงื่อนไขจะสังเกตได้ว่าค่าเฉลี่ย NRMSE มีแนวโน้มสูงขึ้นเมื่อขนาดของกลุ่มตัวอย่างและอัตราการสูญหายเพิ่มขึ้น

พิจารณาลักษณะการสูญหายเกิดขึ้นแบบผสมรายคู่ MCAR–MNAR พบว่า วิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย NRMSE ต่ำที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย Opt.tree และ Opt.svm ซึ่งให้ค่าเฉลี่ยของ NRMSE ไม่แตกต่างกัน นอกจากนี้เมื่อลักษณะการสูญหายเกิดขึ้นแบบผสมรายคู่ MAR–MNAR พบว่า วิธีทดแทนค่าสูญหาย Opt.impute ให้ค่าเฉลี่ย NRMSE ต่ำที่สุด โดยมีค่าเฉลี่ยของ NRMSE อยู่ในช่วง 0.05–0.13 รองมาคือวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย FCS ตามลำดับ

3.2 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุด้วยค่าเฉลี่ยของ RB

จากการพิจารณาผลการวิเคราะห์ประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ยของ RB จะสังเกตเห็นว่าค่าเฉลี่ย RB มีค่าใกล้เคียงกันเมื่อพิจารณาภายในเงื่อนไขเดียวกัน ขณะที่หากพิจารณารายเงื่อนไขพบว่า ค่าเฉลี่ย RB จะเข้าใกล้ศูนย์เมื่ออัตราการสูญหายอยู่ในระดับต่ำ โดยที่ วิธีทดแทนค่าสูญหาย Opt.impute และวิธีทดแทนค่าสูญหาย RF จะเข้าใกล้ศูนย์มากกว่าวิธีทดแทนค่าสูญหาย FCS แสดงให้เห็นว่ามีค่าใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริงมากกว่า ยกตัวอย่างเช่น เมื่อพิจารณาการสูญหายแบบ MAR พบว่าโดยส่วนใหญ่ วิธีทดแทนค่าสูญหาย Opt.svm มีแนวโน้มให้ค่าเฉลี่ยของ RB ต่ำที่สุด โดยมีค่า RB อยู่ในช่วง –0.05 ถึง –0.66

เช่นเดียวกันเมื่อลักษณะการสูญหายเกิดขึ้นแบบผสมรายคู่ MAR - MNAR ในขนาดของตัวอย่างระดับที่หนึ่งเท่ากับ 1,000 หน่วย โดยที่ขนาดของตัวอย่างระดับที่สองเท่ากับ 40 หน่วย จะสังเกตได้ว่าวิธีทดแทนค่าสูญหาย Opt.impute มีค่า RB อยู่ในช่วง –1.98–0.47 และวิธีทดแทนค่าสูญหาย RF มีค่าเฉลี่ย RB อยู่ในช่วง –1.71–0.13 ซึ่งจะเห็นว่าค่าเฉลี่ย RB เข้าใกล้ศูนย์มากกว่าวิธีทดแทนค่าสูญหาย FCS ซึ่งมีค่าเฉลี่ย RB อยู่ในช่วง –1.52–1.79 เป็นต้น

จากการเปรียบเทียบวิธีทดแทนค่าสูญหายแบบพหุของการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่มจำนวน 3 วิธี ได้แก่ วิธีทดแทนค่าสูญหาย FCS วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Otp.impute

ประกอบด้วยวิธี Otp.knn วิธี Otp.tree วิธี Otp.svm และวิธี Otp.cv จากตารางที่ 2 ทำให้สามารถสรุปได้ว่าปัจจัยที่สำคัญที่ผู้วิจัยไม่ควรมองข้ามในกระบวนการทดแทนค่าสูญหายคือลักษณะของการสูญหาย (Type of Missing) ขนาดของกลุ่มตัวอย่าง (Simple Size) และอัตราค่าสูญหาย (Percentage of Missing Rate) ของข้อมูลที่สนใจศึกษา จากการพิจารณาในภาพรวม จะสังเกตเห็นได้ชัดว่าวิธีทดแทนค่าสูญหาย Otp.impute มีประสิทธิภาพสูงที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI–FCS ตามลำดับ

4. อภิปรายผลและสรุป

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุจำนวน 6 วิธี ในการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์ความถดถอยสุ่มสำหรับการวิจัยทางการศึกษา ภายใต้เงื่อนไขที่แตกต่างกันคือประเภทของการสูญหาย 6 รูปแบบ คือ MCAR, MAR, MNAR, MCAR–MAR, MCAR–MNAR และ MAR–MNAR อัตราการสูญหายระดับที่หนึ่งเท่ากับร้อยละ 30 40 และ 50 โดยกำหนดขนาดตัวอย่างระดับที่หนึ่งเท่ากับ 1,000 2,000 3,000 หน่วย เมื่อขนาดตัวอย่างระดับที่สองเท่ากับ 40 50 60 หน่วย วนซ้ำ 100 รอบ ตามลำดับ โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม

เมื่อพิจารณาภาพรวมของผลการวิเคราะห์มีประเด็นที่น่าสนใจและนำไปสู่การอภิปรายผลและสรุป โดยมีรายละเอียดดังนี้

4.1 จากการเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหายแบบพหุ

จำนวน 6 วิธี ได้แก่ วิธีทดแทนค่าสูญหาย FCS วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Opt.impute ประกอบด้วยวิธี Opt.knn, Opt.tree, Opt.svm และ Opt.cv พบว่า วิธีทดแทนค่าสูญหาย Opt.impute มีประสิทธิภาพสูงที่สุดแม้ในสถานการณ์ที่อัตราการสูญหายสูงมากถึงร้อยละ 50 เมื่อเทียบกับวิธีทดแทนค่าสูญหาย RF และวิธี FCS นอกจากนี้จะสังเกตได้ว่า เมื่อขนาดของตัวอย่างเพิ่มขึ้น ส่งผลให้ค่าเฉลี่ยของ NRMSE และ RB มีแนวโน้มลดลง

จากผลการวิเคราะห์ชี้ให้เห็นว่าสามารถนำวิธีทดแทนค่าสูญหาย Opt.impute มาประยุกต์ใช้กับข้อมูลที่มีโครงสร้างแบบพหุระดับได้อย่างสมเหตุสมผล กล่าวคือ จากการศึกษางานวิจัยของ Bertsimas และคณะ [15] พบว่าวิธีทดแทนค่าสูญหาย Opt.impute ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อวิเคราะห์และเรียนรู้รูปแบบการสูญหายโดยใช้การฝึก (Train) เมื่อขนาดตัวอย่างเพิ่มสูงขึ้น ทำให้กระบวนการเรียนรู้ (Algorithm) รูปแบบการสูญหายมีมากขึ้น จากหลักการดังกล่าว จึงทำให้การทดแทนค่าสูญหายด้วยวิธี Opt.impute มีความแม่นยำและมีความน่าเชื่อถือมากขึ้นเมื่อขนาดตัวอย่างเพิ่มสูงขึ้นส่งผลให้การประมาณค่าพารามิเตอร์ในโมเดลมีความตรงและเข้าใกล้พารามิเตอร์ค่าจริงมากขึ้น อย่างไรก็ตามหากไม่นำปัจจัยด้านการสูญหายมาพิจารณาจะพบว่า เมื่อขนาดตัวอย่างเพิ่มสูงขึ้นยังคงช่วยให้การประมาณค่าพารามิเตอร์ดีขึ้นเช่นกัน [17]

4.2 จากผลการเปรียบเทียบประสิทธิภาพวิธีทดแทนค่าสูญหายแบบพหุในการสูญหายแบบ MCAR

พบว่า วิธีทดแทนค่าสูญหาย Opt.cv มีประสิทธิภาพสูงที่สุด รองลงมา คือ วิธีทดแทนค่าสูญหาย RF ตามลำดับ โดยที่วิธีทดแทนค่าสูญหาย Opt.cv มีค่าเฉลี่ยของ NRMSE อยู่ระหว่างช่วง 0.01–0.09 ขณะที่วิธีทดแทนค่าสูญหาย RF มีค่าเฉลี่ยของ NRMSE อยู่ระหว่าง 0.06–0.11 ตามลำดับ

จากการพิจารณาผลการวิเคราะห์ดังกล่าวข้างต้น จะเห็นว่าวิธีทดแทนค่าสูญหายแบบพหุ Opt.cv มีประสิทธิภาพสูงที่สุด เมื่อพบการสูญหายแบบ MCAR แต่ในทางปฏิบัติผู้วิจัยมีความเห็นว่าสามารถเลือกใช้วิธี RF ทดแทนได้ เนื่องจากเมื่อพิจารณาหลักการของวิธี Opt.cv พบว่าวิธี Opt.cv รวมการคำนวณ Opt.knn, Opt.svm และ Opt.tree เพื่อเลือกค่าที่ดีที่สุด จึงทำให้ใช้ระยะเวลาในการคำนวณที่ค่อนข้างนานเมื่อเทียบกับ วิธีทดแทนค่าสูญหาย RF ซึ่งอาจ

ทำให้ผู้วิจัยเสียเวลาวิเคราะห์ผลการวิจัยทั้งที่วิธี Opt.cv และวิธี RF ให้ประสิทธิภาพ ไม่แตกต่างกันมาก ดังนั้นเมื่อเกิดการสูญหายแบบ MCAR ควรเลือกใช้วิธี RF แทนวิธี Opt.cv จึงจะมีความเหมาะสมมากกว่า

5. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการอนุญาตให้ใช้ AI Software โดยไม่มีค่าใช้จ่ายจากสถาบันเทคโนโลยีแมสซาชูเซตส์ (Massachusetts Institute of Technology; MIT) ประเทศสหรัฐอเมริกา และขอขอบคุณสถาบันทดสอบการศึกษาแห่งชาติ (สทศ.) ที่อนุเคราะห์ข้อมูลทุติยภูมิในการวิจัยครั้งนี้

เอกสารอ้างอิง

[1] S. V. Buuren, *Flexible Imputation of Missing Data*. New York: Chapman and Hall/CRC, 2018, pp.3-18

[2] J. Nissen, R. Donatello, and B. V. Dusen, "Missing data and bias in physics education research: A case for using multiple imputation," *Physical Review Physics Education Research*, vol. 15, no. 2, 2019.

[3] S. Ngudratoke, "The principles of multilevel path analysis, and multilevel latent variable growth curve model: Muthen-based approach," *Journal of Research Methodology*, vol. 15, no. 1, pp. 85–104, 2002 (in Thai).

[4] S. Srisuttiyakorn, "Educational inequality and its factors: Multilevel analysis integrated with median-based class of generalized entropy inequality Index," *Journal of Research Methodology*, vol. 32, no. 3, pp. 356–386, 2019 (in Thai).

[5] A. C. Black, O. Harel, and D. B. McCoach, "Missing data techniques for multilevel data: Implications of model misspecification," *Journal of Applied Statistics*, vol. 38, no. 9, pp. 1845–1865, 2011.

[6] H. Nugroho and K. Surendro, "Missing data problem in predictive analytics," in *Proceedings ICSCA*, 2019, pp.95–100.

[7] G. L. Schlomer, L. Bauman, and N. A. Card, "Best practices for missing data management in counseling psychology," *Journal of Couns Psychol*, vol. 57, no. 1, pp. 1–10, 2010.

[8] S. V. Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Journal of Statistical Methods in Medical Research*, vol. 16, no. 3, pp. 195–197, 2007.

[9] S. V. Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Journal of Statistical Software*, vol. 45, no. 3, 2011.

[10] S. V. Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Journal of Statistical Methods in Medical Research*, vol. 16, no. 3, pp. 195–197, 2007.

[11] V. Audigier, I. R. White, S. Jolani, T. Debray, M. Quartagno, J. Carpenter, S. V. Buuren, and M. Resche-Rigon, "Multiple imputation for multilevel data with continuous and variables," *Statistical Science*, vol. 33, no. 2, pp. 160–183, 2018.

[12] S. Pornprasertmani, "Missing data handling (Multilevel Modeling)," Ph.D. dissertation, Faculty of Psychology, Chulalongkorn University, Thailand, 2019 (in Thai).

[13] F. Jia, and W. Wu, "Evaluating methods for handling missing ordinal data in structural

equation modeling," *Behav Res Methods*, vol. 51, no. 5, pp. 2337–2355, 2019.

[14] M. Kokla, J. Viranen, M. Kolehmainen, J. Paananen, and K Hanhineva, "Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study," *BMC Bioinformatics*, 2019.

[15] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From predictive methods to missing data imputation: An optimization approach," *Journal of Machine Learning Research 18*, pp. 1–39, 2018.

[16] S. Srisuttiyakorn, "Missing data analysis," *Journal of Education*, vol. 52, no. 1, pp. 217–223, 2019 (in Thai).

[17] J. Lorah and A. Womac, "Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling," *Behavior Research Methods*, vol. 51, pp. 440–450, 2019.