การเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหายแบบพหุในข้อมูลพหุระดับ: การศึกษาด้วยการจำลองข้อมูล

บทคัดย่อ

วัตถุประสงค์ในการวิจัยครั้งนี้เพื่อเปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุ 6 วิธี ในข้อมูลพหุระดับ ได้แก่ วิธี Multiple Imputation Fully Conditional Specification (MI – FCS), วิธี Random Forest (RF), และ วิธี Optimal Impute (opt.impute) ประกอบด้วย opt.knn ,opt.tree, opt.svm, และ opt.cv ดำเนินการจำลองข้อมูลทางการศึกษาด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) ภายใต้เงื่อนไขดังนี้ 1) ประเภทการสูญหายแบ่งออกเป็น 6 รูปแบบ ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing completely at random หรือ MCAR) การสูญหายแบบสุ่ม (Missing at random หรือ MAR) การสูญหายแบบไม่สุ่ม (Missing not at random หรือ MNAR) และประเภทของการสูญหายรูปแบบผสมรายคู่ 3 รูปแบบ ได้แก่ MCAR - MAR, MCAR - MNAR และ MAR - MNAR 2) ขนาดตัวอย่าง และ 3) อัตราการสูญหาย โดยส่วนใหญ่ผลการวิจัยพบว่า วิธี opt.impute ให้ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (normalized root mean square error หรือ NRMSE) ต่ำที่สุด รองลงมาคือ วิธี RF และ MI – FCS ตามลำดับ ขณะที่วิธี MI – FCS ให้ค่า NRMSE ต่ำที่สุดเมื่อตัวอย่างมีขนาดเล็กและอัตราการสูญหายอยู่ในระดับต่ำที่สุดในการสูญหายรูปแบบ MCAR – MAR ดังนั้นจึงสรุปได้ว่าวิธี opt.impute ให้ประสิทธิภาพสูงที่สุด ช่วยเพิ่มประสิทธิภาพให้กับวิธีการทดแทนค่าสูญหายแบบพหุในข้อมูลพหุระดับ แม้ในประเภทการสูญหายที่มีความซับซ้อนและรุนแรง

คำสำคัญ: การทดแทนค่าข้อมูลสูญหาย, ข้อมูลพหุระดับ, โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม

Comparison of the Efficiency of Multiple Imputation in Multilevel data: A simulation study

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 (MI – FCS), Random Forest (RF), and four methods of Optimal Impute (opt.impute). A simulation study based on real-world educational data with a random coefficients model. we investigate the performance of these approaches under various conditions 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) three sample sizes and 3) three missing rates. The result showed that the opt.impute method has the lowest the normalized root mean square error (NRMSE) — among the three methods, followed by RF method, and then MI-FCS method. While MI-FCS method has the lowest NRMSE when small sample sizes and the lowest missing data percentages in MCAR – MAR. Therefore, the opt.impute method has highest efficiency. This method can reduce constraints It's flexible and increase efficiency of multiple imputation method in multilevel data. Although types of missing data are complex and severe.

Keywords: missing data, multilevel data, random coefficients model, multiple imputation, simulation

1. บทนำ

ถึงแม้ว่าปัจจุบันจะมีเทคโนโลยีการป้องกันไม่ให้เกิดค่าสูญหายในข้อมูลขึ้นหลากหลายวิธี แต่ปัญหาข้อมูลสูญหายก็ยังไม่หมดไป ซึ่งเป็นที่ทราบกันดีว่า “ ค่าสูญหาย (Missing value) ” มักเกิดขึ้นในงานวิจัยเกือบทุกสาขา ไม่ว่าจะเป็นงานวิจัยทางการแพทย์ สาธารณสุข และชีววิทยา โดยเฉพาะงานวิจัยด้านสังคมศาสตร์และพฤติกรรมศาสตร์ นับเป็นปัญหาสำคัญที่ส่งผลกระทบให้ผลการวิจัยคลาดเคลื่อนไม่สามารถสะท้อนสภาพความจริงของประชากรได้ [1], [2]

เมื่อพิจารณาถึงบริบทการศึกษาของประเทศไทยพบว่าธรรมชาติของข้อมูลถูกจัดเป็นหมวดหมู่ กลุ่มนักเรียนถูกจัดให้อยู่รวมกันเป็นห้องเรียนโดยแต่ละห้องเรียนจะถูกรวบรวมอยู่ในหมวดของโรงเรียนและในแต่ละโรงเรียนจะอยู่ภายใต้สังกัดของสำนักงานเขตพื้นที่ เป็นต้น จะสังเกตเห็นว่าลักษณะข้อมูลทางการศึกษามีโครงสร้างแบบพหุระดับ (multilevel data) ซึ่งในแต่ละระดับมีความสัมพันธ์ระหว่างหน่วยข้อมูลซึ่งกันและกัน หากผู้วิจัยวิเคราะห์ข้อมูลโดยไม่คำนึงถึงค่าสูญหาย อาจส่งผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูลได้ [3], [4]

จากสถานการณ์ข้างต้นผู้วิจัยจึงทำการศึกษาประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ โดยจากการศึกษาในอดีตที่ผ่านมามักนิยมใช้เทคนิค Single Imputation ในการทดแทนค่าสูญหาย เช่น การใช้ค่าเฉลี่ย (mean imputation) หรือการใช้ค่ามัธยฐาน (median imputation) เป็นต้น เนื่องจากเป็นวิธีที่ง่าย สะดวกและรวดเร็ว สามารถใช้โปรแกรมสำเร็จรูปในการคำนวณได้ ขณะที่เมื่อพิจารณาลักษณะของข้อมูลจริงพบว่าข้อมูลมีความซับซ้อน วิธีการทดแทนค่าสูญหายแบบดั้งเดิมอาจไม่เหมาะสมนัก [1],[2]

จากการศึกษาในปัจจุบันนิยมนำเทคนิค Multiple Imputation (MI) มาทดแทนค่าสูญหายในข้อมูลที่มีโครงสร้างแบบพหุระดับ (multilevel data) เนื่องจากเทคนิคนี้มีหลักการคือใช้การคำนวณหลายครั้งทำการวนซ้ำเพื่อให้ได้ค่าที่ดีที่สุด อย่างไรก็ตามเทคนิค MI มีหลายวิธี [2],[5],[6],[7],[8],[9] วิธีการที่ศึกษาในงานวิจัยนี้คือ Multiple Imputation Fully Conditional Specification (MI - FCS) [10],[11] นอกจากนี้ศึกษางานวิจัยของ Jia and Wu [12] ให้ผลการวิจัยไม่แตกต่างกับงานวิจัยของ Nissen, Donatello, and Van Dusen [2] แสดงให้เห็นว่าการทดแทนค่าข้อมูลสูญหายด้วยวิธี Random Forest (RF) มีประสิทธิภาพสูงเช่นกันเมื่อเปรียบเทียบกับวิธีดั้งเดิมในอดีต [13] ยิ่งกว่านั้น Bertsimas, Pawlowski, and Zhuo et al. [14] พัฒนาวิธี Optimal Impute (opt.impute) ด้วยการให้โปรแกรมทำงานจากการเรียนรู้ (Machine Learning) ผลการวิจัยให้การประมาณค่าสูญหายที่มีประสิทธิภาพสูงแม้อัตราการสูญหายมาก สามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง ช่วยเพิ่มประสิทธิภาพให้กับวิธีทดแทนค่าสูญหายแบบดั้งเดิมได้

จากผลการศึกษาในข้างต้นให้แสดงให้เห็นว่าการดำเนินการทดแทนค่าสูญหายแบบพหุเป็นขั้นตอนที่สำคัญที่นักวิจัยไม่ควรละเลย งานวิจัยนี้จึงเปรียบเทียบวิธีทดแทนค่าสูญหายแบบพหุในข้อมูลที่มีโครงสร้างแบบพหุระดับ (multilevel data) จำนวน 6 วิธี ได้แก่ 1) วิธี MI – FCS, 2) วิธี RF , และ 3) วิธี opt.impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ opt.knn, opt.tree, opt.svm, และ opt.cv โดยใช้การจำลองข้อมูลด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) ครอบคลุม 3 เงื่อนไข คือ ประเภทของการสูญหาย 6 รูปแบบ ขนาดตัวอย่าง และอัตราการสูญหายที่แตกต่างกัน ผลการวิจัยในครั้งนี้ จะให้องค์ความรู้ทั้งในทางทฤษฎีและทางปฏิบัติเพื่อนำไปประยุกต์ใช้ในข้อมูลทางการศึกษาสามารถนำผลการวิจัยไปใช้ในการตัดสินใจเลือกวิธีทดแทนค่าข้อมูลสูญหายให้เหมาะสมกับสถานการณ์ที่ต้องการศึกษา เพื่อให้ได้สารสนเทศที่ถูกต้องลดความคลาดเคลื่อนที่อาจเกิดขึ้น และสามารถอนุมานไปสู่ประชากรได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

2. วัสดุ อุปกรณ์ และวิธีการวิจัย

การวิจัยครั้งนี้ เป็นการจำลองข้อมูลทางการศึกษาด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) โดยมีวิธีการดำเนินการวิจัยดังนี้

2.1 ขอบเขตการศึกษา

การวิจัยครั้งนี้กำหนดให้ประชากรเป็นข้อมูลที่ได้จากการจำลองข้อมูลพหุระดับ จากงานวิจัยเรื่อง "ความเหลื่อมล้ำทางการศึกษาและปัจจัยที่ส่งผล: การวิเคราะห์พหุระดับร่วมกับดัชนีเอ็นโทรปีนัยทั่วไปที่อิงมัธยฐาน" [4] ซึ่งลักษณะโครงสร้างของข้อมูลพหุระดับที่จำลองขึ้นมีความสอดคล้องกับธรรมชาติของข้อมูลทางการศึกษา สามารถเขียนสมการได้ ดังนี้

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

เมื่อ y_{ij} *แทน ผลสัมฤทธิ์ทางการเรียนของนักเรียน โดยเฉลี่ย ของโรงเรียนที่* i *และสังกัดที่* j

x_{ij} *แทน สัดส่วนของนักเรียนที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำในโรงเรียนที่* i *และสังกัดที่* j

โมเดลระดับสังกัด $\beta_{1j} = \gamma_{01} + U_{1j}$, $\alpha_{0j} = \gamma_{00} + U_{0j}$

เมื่อ γ_{00} *แทน สัมประสิทธิ์จุดตัดแกน*

γ_{01} *แทน สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำ*

$\varepsilon_{ij} \sim N(0,\sigma 2)$ *และ* $U_{ij} \sim N(0,\sigma 2)$

แทน ความคลาดเคลื่อนสุ่มของโมเดลในระดับโรงเรียนและระดับสังกัด

โดยที่ $\gamma_{00} = 0.4425,\ U_{0j} \sim N(0,\ 0.0616)$
$\gamma_{01} = -0.1479,\ U_{1j} \sim N(0,\ 0.0692)$

$$\sigma 2 = \begin{bmatrix} 0.0616 & 0 \\ 0 & 0.0692 \end{bmatrix}$$

ทั้งนี้ กำหนดขนาดตัวอย่าง 3 ขนาด คือ 50 100 และ 300 โรงเรียน อัตราการสูญหายที่ตัวแปรของข้อมูลในแต่ละกรณีที่แตกต่างกัน จำแนกตามประเภทของการสูญหาย 6 รูปแบบ โดยในแต่ละรูปแบบมีอัตราสูญหายเท่ากับร้อยละ 20 ร้อยละ 30 และร้อยละ 50 ตามลำดับ

ประเภทการสูญหายของข้อมูล (Type of Missing Data) โดยจำแนกประเภทการสูญหายของข้อมูลได้ 3 ประเภทหลัก [1],[15] ซึ่งมีรายละเอียดดังต่อไปนี้

การสูญหายแบบสุ่มสมบูรณ์ (MCAR) เป็นการสูญหายที่ลักษณะของข้อมูลสูญหายเกิดขึ้นอย่างสุ่มสมบูรณ์ ข้อมูลที่สูญหายเป็นอิสระหรือไม่มีความสัมพันธ์กับตัวแปรสังเกตได้ทั้งที่ทราบค่าและไม่ทราบค่า [1],[15] ซึ่งสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P(\text{missing} \mid \text{complete data}) = P(\text{missing})$$

การสูญหายแบบสุ่ม (MAR) อาจเรียกว่าการสูญหายแบบสุ่มที่มีเงื่อนไข เป็นการสูญหายที่ลักษณะของข้อมูลที่สูญหายเกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกตหรือค่าที่สูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในข้อมูลที่สนใจศึกษา [1],[15] ซึ่งสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P(\text{missing} \mid \text{complete data}) = P(\text{missing} \mid \text{observed data})$$

การสูญหายแบบไม่สุ่ม (MNAR) เป็นลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม โดยที่ค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกันหรือตัวแปรตัวอื่นภายนอกข้อมูลที่สนใจศึกษา [1],[15] ซึ่งสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P(\text{missing} \mid \text{complete data}) \neq P(\text{missing} \mid \text{observed data})$$

ในทางปฏิบัติผู้วิจัยไม่สามารถทราบได้อย่างแน่ชัดว่าการสูญหายเกิดขึ้นในรูปแบบใด และข้อมูลจริงมีความเป็นไปได้น้อยมากที่จะเกิดการสูญหายเพียงสาเหตุเดียว [16] ดังนั้นผู้วิจัยจึงสนใจศึกษาประเภทของการสูญหายรูปแบบผสมรายคู่เพิ่มเติมจากที่กล่าวไว้ข้างต้น รวมทั้งหมด 6 รูปแบบ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MNAR และMAR-MNAR ตามลำดับ

วิธีทดแทนค่าสูญหาย (Methods) ซึ่งมีรายละเอียดดังนี้

1) หลักการทดแทนค่าสูญหายด้วยวิธี MI – FCS

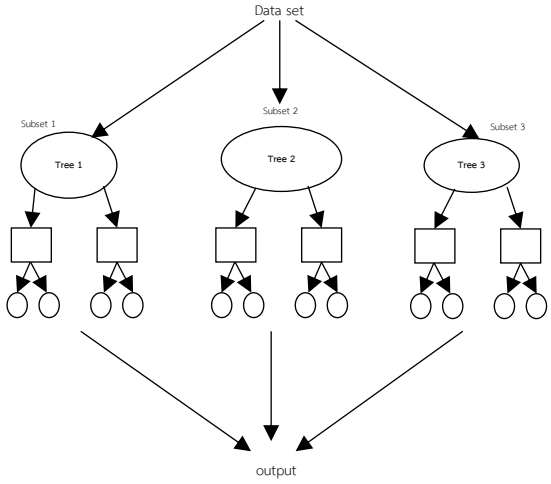
การทดแทนค่าสูญหายด้วยวิธี MI – FCS ใช้หลักการทดแทนค่าสูญหายด้วยเทคนิค MI กล่าวคือการทดแทนค่าสูญหายด้วยชุดข้อมูลของค่าที่เป็นไปได้มากกว่า 1 ค่า ใช้การคำนวณหลายครั้ง เพื่อให้ได้ค่าที่ดีที่สุด ทั้งนี้การทดแทนค่าสูญหายด้วยเทคนิค MI มีหลายวิธีด้วยกัน เช่น สร้างการประมาณค่าโดยใช้วิธี predictive mean matching , Bayesian linear regression, logistic ในงานวิจัยนี้เลือกใช้ การกำหนดค่าด้วย วิธี MI - FCS มีหลักการคือใช้สมการทดแทนค่าสูญหายในลักษณะวนซ้ำ มีขั้นตอนดังนี้ อันดับแรกสร้างสมการทำนายค่าสูญหายตัวแปรแรกด้วยการกำหนดแบบจำลองการทดแทนค่าสูญหาย โดยให้ตัวแปรที่สูญหายแต่ละตัวเป็นตัวแปรตาม (y) หลังจากนั้นวนไปทำนายตัวแปรที่สอง (อิงตัวแปรแรกที่ทดแทนค่าสูญหายเรียบร้อยแล้ว) ต่อมาวนไปยังตัวแปรที่สามและไปจนครบตัวแปรสุดท้าย แล้วจึงวนกลับมายังตัวแปรที่หนึ่ง ทำเช่นนี้ไปไปเรื่อย ๆ จนครบจำนวนรอบที่กำหนด โดยผู้วิเคราะห์จะสามารถกำหนดจำนวนชุดข้อมูลและจำนวนรอบในการวนได้ [10],[11] โดยวิธีนี้ใช้ “mice” Package ในโปรแกรม R สำหรับวิเคราะห์การทดแทนค่าสูญหาย ซึ่งรายละเอียดดังนี้

$$y_{11} = \beta_{0(11)} + \beta_{1(11)}(x_1) + \beta_{2(11)}(x_2) + \ldots \beta_{n(11)}(x_n)$$
$$\ldots$$
$$\ldots$$
$$\ldots$$
$$y_{ij} = \beta_{0(ij)} + \beta_{1(ij)}(x_1) + \beta_{2(ij)}(x_2) + \ldots \beta_{n(ij)}(x_n)$$

เมื่อ β_{ij} *แทน ค่าประมาณการทดแทนข้อมูลสูญหาย (Imputed value) ในรอบที่* i *ชุดข้อมูลที่* j

2) หลักการทดแทนค่าสูญหายด้วยวิธี RF

แนวคิดของวิธี Random forest Imputation คือ การรวมกันของทฤษฎี Bagging (Bootstrap Aggregation) ซึ่งเป็นพื้นฐานของ Random Forest Classification โดยมีหลักการคือการสร้างโมเดลจากการใช้แผนภาพต้นไม้ตัดสินใจ (decision tree) หลายโมเดลในการวิเคราะห์แต่ละครั้ง มีกระบวนการโดยในแต่ละโมเดลจะได้รับชุดข้อมูลที่แตกต่างกันซึ่งเป็นซับเซตของชุดข้อมูลทั้งหมด และขณะการทำนายค่าข้อมูลสูญหายจะกำหนดให้แต่ละแผนภาพต้นไม้ทำนายค่าข้อมูลสูญหายในแต่ละโมเดล โดยแต่ละโมเดลเป็นอิสระต่อกัน หลังจากนั้นคำนวณผลการทำนายค่าข้อมูลสูญหายด้วยการโหวตผลลัพธ์ (vote output) ซึ่งผลลัพธ์ที่ถูกเลือกโดยแผนภาพต้นไม้ตัดสินใจมากที่สุดปรากฏดังรูปที่ 1 โดยวิธีนี้ใช้ “missForest” Package ในโปรแกรม R สำหรับวิเคราะห์การทดแทนค่าสูญหาย ดังนั้นผลลัพธ์ที่ได้จากการวิเคราะห์จะมีความถูกต้องแม่นยำและมีประสิทธิภาพที่สูง



รูปที่ 1 หลักการทดแทนค่าสูญหายด้วยวิธี RF

3) หลักการทดแทนค่าสูญหายด้วยวิธี opt. impute

วิธี opt. impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ opt.knn, opt.svm, opt.tree และ opt.cv หลักการคือการรวมแนวคิดการทดแทนค่าสูญหาย จำนวน 3 วิธี [14] ได้แก่ K-Nearest Neighbors based (opt.knn), Support Vector Machines based (opt.svm) และ Decision Tree based (opt.tree) โดยสร้างการแทนที่ค่าสูญหายหลายครั้งด้วยเทคนิค MI ซึ่งจะช่วยประมาณค่าได้ดีกว่าการใส่แบบจำลองกับชุดข้อมูลเพียงครั้งเดียว

โดยปรับค่าในจุดข้อมูลสูญหายและข้อมูลทั้งหมดพร้อมกัน เพื่อทดแทนค่าสูญหายให้เหมาะสม สามารถใช้ได้ทั้งตัวแปรต่อเนื่องและไม่ต่อเนื่อง นอกจากนี้พิจารณาวิธีการผสมสองวิธี ได้แก่ opt.cv (optimal cross - validated) ซึ่งเลือกวิธีที่ดีที่สุดจาก opt.knn, opt.svm และ opt.tree ตามลำดับ โดยวิธีนี้ใช้โปรแกรมจูเลีย Julia และ Interface to 'Interpretable AI' Modules หรือ “iai” Package (AI software) ในโปรแกรม R สำหรับวิเคราะห์การทดแทนค่าสูญหาย

ประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุ ซึ่งเกณฑ์ที่ใช้มีรายละเอียดดังนี้

ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (normalized root mean square error: NRMSE) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพเพื่อทดสอบประสิทธิภาพของชุดข้อมูลและตรวจสอบความน่าเชื่อถือของโมเดล โดยสามารถคำนวณได้จากสมการดังต่อไปนี้

$$NRMSE = \sqrt{\frac{mean\left((X^{comp} - X^{imp})^2\right)}{var\left(X^{comp}\right)}}$$

เมื่อ X^{comp} *แทน ค่าของข้อมูลจริงที่สังเกตได้*

X^{imp} *แทน ค่าของข้อมูลที่ได้จากการทดแทนค่าสูญหาย*

$var\left(X^{comp}\right)$ *แทน ค่าความแปรปรวนของค่าข้อมูลจริงที่สังเกตได้*

ค่าความเอนเอียงสัมพัทธ์ (Relative Biased: RB) เป็นเกณฑ์ที่ใช้สำหรับวัดความเอนเอียงกลุ่มของวิธีการทดแทนค่าสูญหายแบบพหุ โดยมีสูตรในการคำนวณดังนี้

$$RB\left(\hat{\theta}i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\theta}i-\theta}{\theta}\right) \quad ; \quad i = 1,2,3, \ldots\ldots, 100$$

เมื่อ $\hat{\theta}i$ แทน *ค่าประมาณพารามิเตอร์ที่ได้จากการจำลองจำนวน* 100 *รอบ*

θi แทน *ค่าประมาณพารามิเตอร์รอบที่* i

n แทน *ขนาดตัวอย่าง*

2.2 ขั้นตอนการวิจัย

2.2.1 ขั้นที่ 1 ผู้วิจัยจำลองข้อมูลความเหลื่อมล้ำทางการศึกษา โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) ที่มีลักษณะข้อมูล แบบ 2 ระดับ ได้แก่ ระดับโรงเรียน และระดับสังกัด [4],[17] โดยที่ตัวแปรอิสระ คือ สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำระดับโรงเรียน (x) ตัวแปรตาม คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียนโดยเฉลี่ย ระดับโรงเรียน (y)

2.2.2 ขั้นที่ 2 นำชุดข้อมูลที่ได้ จากการจำลองขั้นที่ 1 ดำเนินการตามรูปแบบการสูญหายที่ตัวแปรในแต่ละกรณีที่แตกต่างกัน 6 รูปแบบ มีรายละเอียดดังนี้

กรณีที่ 1: การสูญหายแบบสุ่มสมบูรณ์ (MCAR)

กำหนดให้ x และ y สูญหาย โดยที่ค่า x และ y ที่สูญหาย ไม่มีความสัมพันธ์กับค่า x และ y ที่สังเกตได้ โดยสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P\left((x,y)_{Missing} \mid complete\ data\right) = P\left((x,y)_{Missing}\right)$$

กรณีที่ 2: การสูญหายแบบสุ่ม (MAR)

กำหนดให้ ค่า x ที่สูญหายมีความสัมพันธ์กับค่า y ที่สังเกตได้ โดยสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P\left(x_{missing} \mid complete\ data\right) = P\left(x_{missing} \mid y_{observed}\right)$$

กรณีที่ 3: การสูญหายแบบไม่สุ่ม (MNAR)

กำหนดให้ ค่า y ที่สูญหายมีความสัมพันธ์กับ y ด้วยกันเอง โดยสามารถอธิบายได้ด้วยสมการความน่าจะเป็น ดังนี้

$$P\left(y_{missing} \mid complete\ data\right) \neq P\left(y_{missing} \mid (x,y)_{observed}\right)$$

กรณีที่ 4: การสูญหายแบบผสมรายคู่ MCAR-MAR

นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการโดยกำหนดให้มีการสุ่มการสูญหายแบบสุ่มสมบูรณ์ (MCAR) ตามกรณีที่ 1 หลังจากนั้น นำชุดข้อมูลที่ได้มาสุ่มการสูญหายแบบสุ่ม (MAR) ตามกรณีที่ 2 จะได้การสูญหายแบบผสมรายคู่ MCAR-MAR

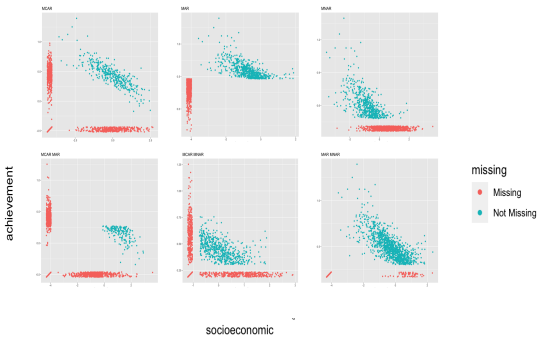
กรณีที่ 5: การสูญหายแบบผสมรายคู่ MCAR-MNAR

นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการโดยกำหนดให้มีการสุ่มการสูญหายแบบสุ่มสมบูรณ์ (MCAR) ตามกรณีที่ 1 หลังจากนั้นนำชุดข้อมูลที่ได้สุ่มการสูญหายแบบไม่สุ่ม (MNAR) ตามกรณีที่ 3 จะได้การสูญหายแบบผสมรายคู่ MCAR-MNAR

กรณีที่ 6: การสูญหายแบบผสมรายคู่ MAR-MNAR

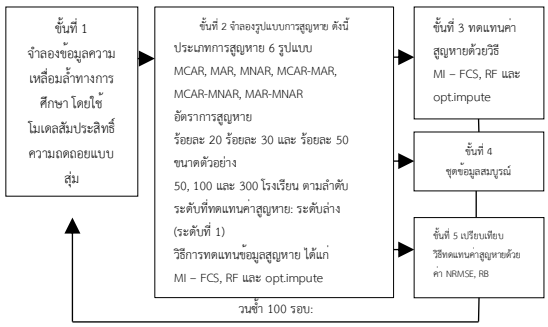
นำชุดข้อมูลที่ได้จากการจำลองขั้นที่ 1 มาดำเนินการโดยกำหนดให้มีการสุ่มการสูญหายแบบสุ่ม (MAR) ตามกรณีที่ 2 หลังจากนั้นนำชุดข้อมูลที่ได้สุ่มการสูญหายแบบไม่สุ่ม (MNAR) ตามกรณีที่ 3 จะได้การสูญหายแบบผสมรายคู่ MCAR-MNAR

ทั้งนี้เมื่อดำเนินการจำลองข้อมูลแยกตามประเภทของการสูญหาย 6 รูปแบบ ที่ได้กล่าวมาข้างต้น ตามขั้นที่ 1 ถึง 2 จะได้แผนภาพการกระจายของความสัมพันธ์ระหว่างตัวแปรสูญหายกับตัวแปรสังเกตได้ที่มีความแตกต่างกันปรากฏดังรูปที่ 2 ดังนี้



รูปที่ 2 ความสัมพันธ์ของตัวแปรสูญหายกับตัวแปรสังเกตได้จำแนกตามประเภทของการสูญหาย

2.2.3 ขั้นที่ 3 ทดแทนค่าสูญหายด้วยวิธี MI – FCS, วิธี RF และ วิธี 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 ปรากฏดังรูปที่ 3



รูปที่ 3 ขั้นตอนการจำลองข้อมูล

2.3 การวิเคราะห์ข้อมูล

ผู้วิจัยวิเคราะห์ข้อมูลโดยใช้โปรแกรม R version 4.1.3 และโปรแกรม Julia 1.7.2 ในการจำลองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพวิธีทดแทนค่าสูญหายแบบพหุ 6 วิธี ได้แก่ วิธี MI-FCS, วิธี RF และวิธี otp.impute ประกอบด้วย otp.knn, otp.tree, otp.svm และ otp.cv ใช้เกณฑ์ค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (NRMSE) และทดสอบความเชื่อมั่นของวิธีการทดแทนค่าสูญหายแบบพหุ โดยใช้ค่าความเอนเอียงสัมพัทธ์ (RB) ทั้งนี้เปรียบเทียบค่าดังกล่าวด้วยการพิจารณาค่าที่น้อยที่สุดจะมีประสิทธิภาพสูงที่สุด นอกจากนี้ ใช้การนับจำนวนวิธีทดแทนค่าสูญหายแบบพหุซึ่งมีประสิทธิภาพสูงที่สุด (Count of wins) ด้วย NRMSE ที่มีค่าต่ำที่สุดโดยใช้ตารางและแผนภาพนำเสนอผลการทดลอง (data visualization)

3. ผลการทดลอง

ผู้วิจัยมุ่งศึกษาประสิทธิภาพของวิธีการทดแทนค่าสูญหายด้วยการจำลองข้อมูลทางการศึกษาโดยใช้ข้อมูลความเหลื่อมล้ำทางการศึกษาที่มีลักษณะพหุระดับด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) [4] ภายใต้ประเภทของการสูญหาย ขนาดตัวอย่าง และอัตราการสูญหายที่แตกต่างกัน แบ่งการนำเสนอตามเกณฑ์ที่ใช้เปรียบเทียบ

ประสิทธิภาพวิธีการทดแทนสูญหายแบบพหุ 6 วิธี ได้แก่ MI-FCS, RF otp.knn, otp.tree, otp.svm และ otp.cv โดยมีรายละเอียดดังนี้

3.1 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุด้วย NRMSE

เมื่อพิจารณาภาพรวมจากรูปที่ 4 พบว่าวิธีทดแทนค่าสูญหายด้วยวิธี MI – FCS มีค่า NRMSE ค่อนข้างสูงที่สุด รองลงมาเป็นวิธี RF และวิธี opt.impute ให้ค่า NRMSE ต่ำที่สุด นอกจากนี้จะสังเกตเห็นได้ชัดว่าขนาดตัวอย่างและอัตราการสูญหาย ส่งผลกระทบต่อค่า NRMSE ในทิศทางบวก ซึ่งมีแนวโน้มว่า NRMSE จะมีค่าสูงขึ้นเมื่อขนาดตัวอย่างและอัตราการสูญหายเพิ่มสูงขึ้น

พิจารณาการสูญหายแบบผสมรายคู่รูปแบบ MCAR-MAR เมื่อขนาดตัวอย่างเท่ากับ 50 และอัตราการสูญหายเท่ากับ 20 พบว่าการทดแทนค่าสูญหายแบบพหุด้วยวิธี MI – FCS ให้ค่า NRMSE ต่ำที่สุด ในขณะที่อัตราการสูญหายเพิ่มสูงขึ้น วิธี opt.impute ให้ค่า NRMSE ที่ต่ำกว่าวิธี MI – FCS อย่างเห็นได้ชัด ปรากฏดังรูปที่ 4 สำหรับการพิจารณาการสูญหายแบบผสมรายคู่รูปแบบ MCAR-MNAR เมื่อขนาดตัวอย่างเท่ากับ 100 อัตราการสูญหายเท่ากับ 20 และร้อยละ 30 ตามลำดับ พบว่าการทดแทนค่าสูญหายแบบพหุด้วยวิธี RF และวิธี MI – FCS ให้ค่า NRMSE ค่อนข้างต่ำ ขณะที่เมื่ออัตราการสูญหายเพิ่มสูงขึ้น วิธี opt.impute ให้ค่า NRMSEที่ต่ำกว่าอย่างเห็นได้ชัดโดยสังเกตได้จากรูปที่ 4

เมื่อเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหายแบบพหุโดยใช้การนับจำนวนวิธีทดแทนค่าสูญหายแบบพหุซึ่งมีประสิทธิภาพสูงที่สุด (Count of wins) ด้วยค่า NRMSE ต่ำที่สุด (Count of wins) ภาพรวมส่วนใหญ่พบว่า opt.impute ถูกจัดอยู่ในอันดับสูง (NRMSE ต่ำ) เมื่อพิจารณาจำแนกตามประเภทของรูปแบบการสูญหาย MCAR, MAR, MNAR, MCAR - MAR จะสังเกตได้ชัดว่าวิธี opt.cv จัดอยู่ในอันดับสูงที่สุด (NRMSE ต่ำที่สุด) ขณะที่วิธี MI – FCS ถูกจัดอยู่ในอันดับสุดท้ายทุกกรณี (NRMSE สูงสุด) ทั้งนี้หากพิจารณาการสูญหายในรูปแบบ MCAR – MNAR พบว่าวิธี RF จะถูกจัดอยู่ในอันดับสูงกว่า opt.impute นอกจากนี้เมื่อมีการสูญหายในรูปแบบMAR – MNAR วิธี opt.knn จัดอยู่ในอันดับสูงที่สุดปรากฏดังตารางที่ 1 และสามารถสังเกตได้จากรูปที่ 5

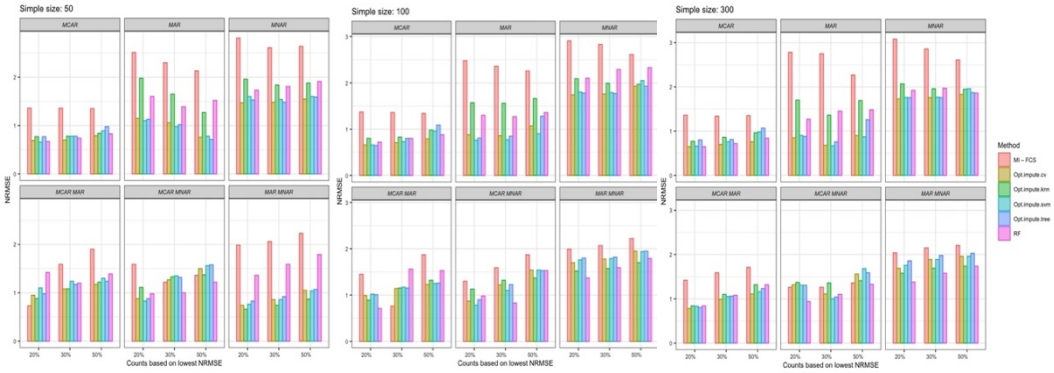
3.2 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายแบบพหุด้วย RB

ผลการวิเคราะห์ประสิทธิภาพและความเชื่อมั่นวิธีการทดแทนค่าสูญหายแบบพหุจากการเปรียบเทียบด้วยค่าความเอนเอียงสัมพัทธ์ (RB) ในภาพรวมจะสังเกตเห็นว่า RB มีค่าใกล้เคียงกัน ขณะที่หากพิจารณาจำแนกตามขนาดตัวอย่าง พบว่าค่า RB จะเข้าใกล้ศูนย์ เมื่อตัวอย่างมีขนาดเท่ากับ 50 และ 100 เมื่อเทียบกับตัวอย่างขนาดเท่ากับ 300 แสดงให้เห็นว่าการทดแทนค่าสูญหายในตัวอย่างขนาดเท่ากับ 50 และ 100 ในข้อมูลพหุระดับด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) ใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริงอย่างเห็นได้ชัด ปรากฏดังรูปที่ 6

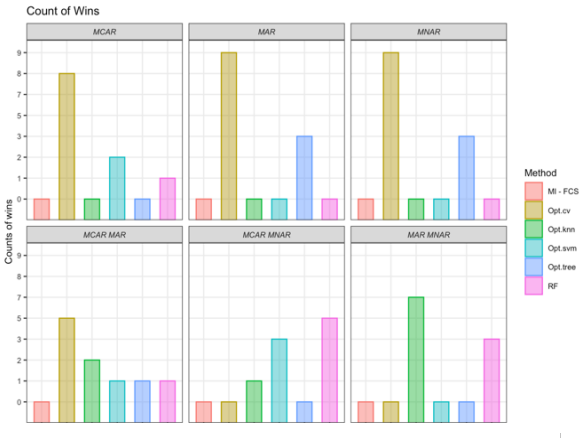
จากผลการทดลองข้างต้นทำให้สามารถสรุปได้ว่าขนาดตัวอย่างและอัตราการสูญหายส่งผลกระทบต่อประสิทธิภาพของวิธีการทดแทนค่าสูญหายกล่าวคือเมื่อขนาดตัวอย่างมากกว่าหรือเท่ากับ 50 ขึ้นไป และอัตราการสูญหายมากกว่าหรือเท่ากับร้อยละ 20 วิธี opt.impute เป็นวิธีการทดแทนค่าสูญหายแบบพหุที่ให้ประสิทธิภาพสูง โดยเฉพาะอย่างยิ่งเมื่อข้อมูลสูญหายอยู่ในรูปแบบผสมรายคู่ และ MNAR และในทางตรงกันข้ามหากขนาดตัวอย่างน้อยกว่าหรือเท่ากับ 50 และอัตราการสูญหายต่ำกว่าหรือเท่ากับร้อยละ 20 วิธี RF และวิธี MI – FCS จะให้ประสิทธิภาพที่สูงมากกว่าวิธี opt.impute โดยเฉพาะเมื่อข้อมูลสูญหายอยู่ในรูปแบบผสมรายคู่ MCAR-MAR และ MCAR-MNAR ดังนั้นจะเห็นว่าขนาดตัวอย่าง อัตราการสูญหายรวมถึงประเภทของการสูญหายมีผลต่อประสิทธิภาพวิธีการทดแทนค่าสูญหายแบบพหุสำหรับข้อมูลทางการศึกษาที่มีลักษณะแบบพหุระดับ

ตารางที่ 1 การนับจำนวนครั้งที่ค่า NRMSE ต่ำที่สุด (Count of wins)

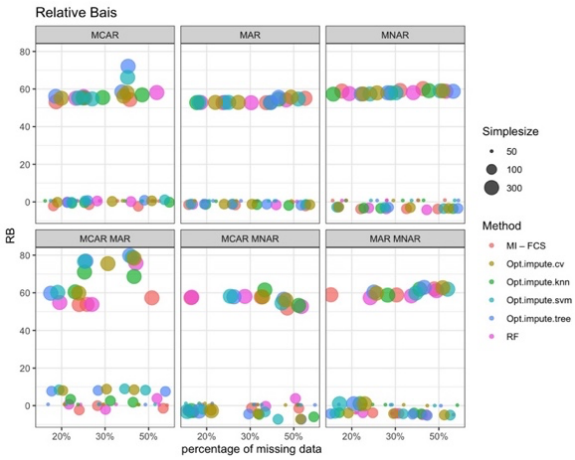
Type	Methods					
	MI - FCS	RF	Opt.impute			
			Opt.knn	Opt.tree	Opt.svm	Opt.cv
MCAR	0	1	0	0	2	8
MAR	0	0	0	3	6	0
MNAR	0	0	0	3	0	9
MCAR - MAR	0	1	2	1	1	5
MCAR - MNAR	0	5	1	0	3	0
MAR - MNAR	0	3	7	0	0	0



รูปที่ 4 เปรียบเทียบประสิทธิภาพวิธีทดแทนค่าสูญหายแบบพหุด้วย NRMSE



รูปที่ 5 เปรียบเทียบการนับจำนวนวิธีทดแทนค่าสูญหายแบบพหุที่มีประสิทธิภาพสูงที่สุด



รูปที่ 6 เปรียบเทียบประสิทธิภาพวิธีทดแทนค่าสูญหายแบบพหุด้วย RB

4. อภิปรายผลและสรุป

เมื่อพิจารณาภาพรวมของผลการทดลองมีประเด็นที่น่าสนใจและนำไปสู่การอภิปรายผลและสรุป โดยมีรายละเอียดดังนี้

4.1 จากการเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหายแบบพหุด้วยค่า NRMSE จากการจำลองข้อมูลภายใต้รูปแบบการสูญหาย 6 รูปแบบ โดยรวมจะเห็นได้ชัดว่าส่วนใหญ่การทดแทนค่าสูญหายแบบพหุด้วยวิธี opt.impute มีประสิทธิภาพสูงที่สุดแม้ในสถานการณ์ที่มีการสูญหายแบบซับซ้อนและรุนแรง เช่น การสูญหายในรูปแบบผสม หรือ MNAR และอัตราการสูญหายที่สูงถึง ร้อยละ 50 ซึ่งสอดคล้องกับงานวิจัยของ Bertsimas, Pawlowski, and Zhuo [15] พบว่า otp.impute ให้ผลการทดแทนค่าสูญหายที่มีประสิทธิภาพสูงแม้อัตราการสูญหายมาก สามารถลดข้อจำกัดในวิธีการแบบดั้งเดิมและยังช่วยเพิ่มประสิทธิภาพให้กับวิธีการทดแทนค่าสูญหายแบบพหุ ได้อย่างถูกต้องและเหมาะสมมากยิ่งขึ้น

4.2 เมื่อพิจารณาประสิทธิภาพของวิธี MI-FCS พบว่าหากพิจารณาการสูญหายแบบผสมรายคู่รูปแบบ MCAR-MAR เมื่อขนาดตัวอย่างเท่ากับ 50 และอัตราการสูญหายเท่ากับร้อยละ 20 จะสังเกตเห็นว่าการทดแทนค่าสูญหายแบบพหุด้วยวิธี MI – FCS ให้ค่า NRMSE ต่ำที่สุด ซึ่งไม่แตกต่างจากงานวิจัยของ Audigier et al [10] และ สัตทัด พรประเสริฐมานิต [11] ทดสอบพบว่าวิธี MI - FCS เป็นวิธีการที่ง่าย สามารถประมาณค่าข้อมูลพหุระดับได้ แต่เมื่อพิจารณาโดยภาพรวมพบว่าส่วนใหญ่การทดแทนค่าสูญหายแบบพหุด้วยวิธี MI-FCS มีค่า NRMSE สูงมากกว่าวิธี RF และ opt.impute ที่เป็นเช่นนี้อาจเป็นเพราะในงานวิจัยนี้ กำหนดให้ข้อมูลมีอัตราการสูญหายที่สูงมากถึงร้อยละ 50 ดังนั้นจึงอาจสรุปได้ว่าวิธี MI – FCS น่าจะเหมาะกับตัวอย่างขนาดเล็กและอัตราการสูญหายที่ต่ำกว่าร้อยละ 20 เป็นต้น

3.4 จากการพิจารณาค่า RB เป็นที่น่าสังเกตว่าตัวอย่างขนาด 50 และ 100 ให้ค่า RB ที่ต่ำกว่าตัวอย่างขนาด 300 แสดงให้เห็นว่าการประมาณค่าตัวอย่างขนาดเล็กใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริง ในขณะที่งานวิจัยของ Zimmerman et al. [18] และชนินันท์ พฤกษ์ประมูล และคณะ [19] พบว่าเมื่อสุ่มตัวอย่างขนาดใหญ่จะให้ค่าใกล้เคียงพารามิเตอร์มากกว่าสุ่มตัวอย่างขนาดที่เล็กกว่า เนื่องจากงานวิจัยนี้เป็นการศึกษาข้อมูลที่มีโครงสร้างแบบพหุระดับซึ่งแตกต่างจากงานวิจัยดังที่ได้กล่าวมา เมื่อพิจารณาโครงสร้างของข้อมูลและลักษณะการสูญหายในงานวิจัยนี้พบว่าข้อมูลมีการสูญหายในระดับที่หนึ่งซึ่งข้อมูลที่สูญหายมีความสัมพันธ์และส่งผลกระทบต่อข้อมูลระดับที่สองจึงอาจจะส่งผลกระทบทำให้ค่า RB ไม่ได้ขึ้นอยู่กับขนาดของตัวอย่างเพียงอย่างเดียว ทั้งนี้อาจขึ้นอยู่กับโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) ที่นำมาทดสอบในงานวิจัยนี้ด้วยซึ่งไม่แตกต่างจากงานวิจัยของ Lorah and Womac [8]

4.3 ถึงแม้ว่าวิธีการทดแทนค่าสูญหายแบบพหุด้วยวิธี otp.cv จะมีประสิทธิภาพสูงที่สุด (NRMSE ต่ำที่สุด) เมื่อมีการสูญหายรูปแบบ MNAR แต่ในทางปฏิบัติผู้วิจัยมีความเห็นว่าสามารถเลือกใช้วิธี opt.tree แทนได้เนื่องจากวิธี otp.cv ใช้ระยะเวลาในการคำนวณที่ค่อนข้างมากกว่า opt.tree โดยที่ให้ค่า RB ที่ไม่แตกต่างกันมากนัก

5. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับความช่วยเหลือในการอนุญาตให้ใช้ AI software โดยไม่มีค่าใช้จ่าย จากสถาบันเทคโนโลยีแมสซาชูเซตส์ (Massachusetts Institute of Technology หรือ MIT) ประเทศสหรัฐอเมริกา

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, July 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, January - April 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, September-December 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, November 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, June 2007.

[9] S. V. Buuren, "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Journal of Statistical Software,* vol. 45, no. 3, December 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, June 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," *Statist. Sci*, vol. 33, no. 2, pp. 160 - 183, May 2018.

[11] S. Pornprasertmani, "Missing Data Handling (Multilevel Modeling)," Ph.D. dissertation, Faculty of Psychology, Chulalongkorn University, Thailand, 2019 (In Thai)

[12] 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.

[13] 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.

[14] 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.

[15] S. Srisuttiyakorn, “Missing Data Analysis,” *Journal of Education*, vol. 52, no. 1, pp. 217-223, January-March 2019. (In Thai)

[16] 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, January 2019.

[17] N. Bannakan, “EDUCATIONAL INEQUALITY STUDENT IN OFFICE OF THE BASIC EDUCATION COMMISSION,” *Journal of Research Methodology*, vol. 14, no.2, pp. 217-223, January-March 2019. (In Thai)

[17] Napapach Bannakan. (2019). EDUCATIONAL INEQUALITY STUDENT IN OFFICE OF THE BASIC EDUCATION COMMISSION. Journal of Research Methodology, *Vol.* 14*, No.* 2*, July - December* 2019. (In Thai)

[18] D. W. Zimmerman, B. D. Zumbo, and R. H. Williams, “Bias in estimation and hypothesis testing of correlation,” *Psicológica,* vol. 24, no. 1, pp. 133-158, 2003.

[19] C. Pruekpramool. “The Principles of Multilevel Path Analysis, Multilevel Factor Analysis, and Multilevel Latent Variable Growth Curve Model: Muthen-Based Approach Efficiency of Pearson, Spearman and Kendall’s Correlation Coefficients When Data is Non-normal Distributed,” *Journal of Research Methodology*, Vol. 15, No. 2, (In Thai)