การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ: การประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา

น.ส.นวลรัตน์ ฉิมสุด

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต
สาขาวิชาสถิติและสารสนเทศการศึกษา ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

COMPARISON OF THE EFFICIENCY OF IMPUTATION METHODS IN MULTILEVEL DATA: APPLICATIONS TO EDUCATIONAL INEQUALITY ANALYSIS

Miss Nuanrat Chimsud

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Education in Educational Statistics and Information
Department of Educational Research and Psychology
FACULTY OF EDUCATION
Chulalongkorn University
Academic Year 2021
Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ: การประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา
โดย	น.ส.นวลรัตน์ ฉิมสุด
สาขาวิชา	สถิติและสารสนเทศการศึกษา
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.สิวะโชติ ศรีสุทธิยากร

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาครุศาสตรมหาบัณฑิต

---------- คณบดีคณะครุศาสตร์
(รองศาสตราจารย์ ดร.ศิริเดช สุชีวะ)

คณะกรรมการสอบวิทยานิพนธ์

---------- ประธานกรรมการ
(รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

---------- อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล)

---------- อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.สิวะโชติ ศรีสุทธิยากร)

---------- กรรมการ
(รองศาสตราจารย์ ดร.วรรณี แกมเกตุ)

---------- กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.สังวรณ์ งัดกระโทก)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

นวลรัตน์ ฉิมสุด : การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ: การประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา. (COMPARISON OF THE EFFICIENCY OF IMPUTATION METHODS IN MULTILEVEL DATA: APPLICATIONS TO EDUCATIONAL INEQUALITY ANALYSIS) อ.ที่ปรึกษาหลัก : อ. ดร.ประภาศิริ รัชชประภาพรกุล, อ.ที่ปรึกษาร่วม : ผศ. ดร.สิวะโชติ ศรีสุทธิยากร

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ (1) เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าข้อมูลสูญหาย 3 วิธี ได้แก่วิธี MI-FCS, วิธี RF และวิธี Opt.impute ซึ่งประกอบด้วย วิธี Opt.knn , Opt.tree, วิธี Opt.svm, และวิธี Opt.cv โดยใช้การจำลองข้อมูลและนำผลที่ได้มาประยุกต์ใช้กับข้อมูลจริง (2) เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ด้วยโมเดลพหุระดับโดยใช้ข้อมูลที่มีการทดแทนค่าสูญหาย และเปรียบเทียบผลที่ได้ กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่ไม่ได้ทดแทนค่าสูญหาย ผลการวิจัยพบว่า (1) จากการพิจารณาผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายโดยใช้การจำลองข้อมูลในภาพรวม จะพบว่าส่วนใหญ่วิธีทดแทนค่าสูญหาย Otp.impute มีแนวโน้มให้ประสิทธิภาพสูงที่สุด รองลงมาคือ วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI – FCS ตามลำดับ (2) ผู้วิจัยรวบรวมข้อมูลทุติยภูมิของนักเรียนชั้นมัธยมศึกษาปีที่ 3 จากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ปีการศึกษา 2563 จำนวน 2,109 โรงเรียนที่อยู่ในสังกัดสำนักเขตพื้นที่การศึกษามัธยมศึกษา(สพม.) นำวิธีทดแทนค่าสูญหายที่ได้จากการจำลองข้อมูลมาประยุกต์ใช้กับข้อมูลทุติยภูมิดังกล่าว ผลการวิจัย จะพบว่าสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน ส่งผลกระทบต่อผลสัมฤทธิ์ ทางการเรียนของนักเรียนระดับโรงเรียน อย่างมีนัยสำคัญทางสถิติ โดยผลกระทบที่เกิดขึ้นสะท้อนให้เห็นถึงความเหลื่อมล้ำทางการศึกษา และเมื่อเปรียบเทียบผลที่ได้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่ไม่ได้ทดแทนค่าสูญหาย แสดงให้เห็นว่าหากนำข้อมูลวิเคราะห์ผลการวิจัยโดยไม่คำนึงถึงค่าสูญหาย หรือตัดค่าสูญหายทิ้ง อาจจะส่งผลกระทบต่อการประมาณค่าพารามิเตอร์ที่แท้จริง อย่างมีนัยสำคัญทางสถิติหรือไม่สามารถอนุมานไปสู่ประชากรได้อย่างถูกต้องและแม่นยำ

สาขาวิชา	สถิติและสารสนเทศการศึกษา	ลายมือชื่อนิสิต .....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6282024627 : MAJOR EDUCATIONAL STATISTICS AND INFORMATION
KEYWORD: missing data multilevel data random coefficients model multiple imputation simulation

Nuanrat Chimsud : COMPARISON OF THE EFFICIENCY OF IMPUTATION METHODS IN MULTILEVEL DATA: APPLICATIONS TO EDUCATIONAL INEQUALITY ANALYSIS. Advisor: PRAPASIRI RATCHAPRAPAPORNKUL, Ph.D. Co-advisor: Asst. Prof. Ph.D. SIWACHOAT SRISUTTIYAKORN

The purposes of this research were to (1) compare the efficiency of multiple Imputation methods of multilevel missing data as the three methods of the Imputation included Multiple Imputation Fully Conditional Specification (MI – FCS), Random Forest (RF), and the four type of Optimal Impute (opt. impute) include Opt.knn , Opt.tree, Opt.svm and Opt.cv . A simulation study based on real-population educational data with a random coefficients model then applied the results to real data. (2) analyze the educational inequality using the multilevel analysis the using data had substituted the missing values already. Compare the results with the analysis of educational inequality that does not compensate for the missing value. The findings were as follows: (1) the Opt.impute method has the highest efficiency. This method can reduce constraints. It's flexible and increases the efficiency of multiple imputation method in multilevel data. Although types of missing data are complex and severe. followed by RF method and MI – FCS method respectively. (2) The researcher applied the secondary data of Junior High School students at the school level. The data of 2,109 schools within 42 educational clusters in the year 2020 from "The National Institute of Educational Testing Service (Public Organization) or NIETS ". Results from the study show the influence of Proportion of students whose families are underfunded and do not live with their parents school level on achievement in each education. Service

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study: Educational Statistics and Information — Student's Signature .....

Academic Year: 2021 — Advisor's Signature .....

Co-advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้จะสำเร็จลุล่วงไม่ได้หากไม่ได้รับความช่วยเหลือ ความเมตตาและดูแลความเอาใจใส่เป็นอย่างดียิ่งจาก อาจารย์ ดร.ประภาศิริ รัชชประภาพรกุล อาจารย์ที่ปรึกษาหลัก และผู้ช่วยศาสตราจารย์ ดร. สิวะโชติ ศรีสุทธิยากร อาจารย์ที่ปรึกษาร่วม ที่ได้สละเวลาอันมีค่าให้คำแนะนำในการเลือกหัวข้อ ให้คำปรึกษา ข้อคิดเห็น รวมไปถึงการแก้ไขจุดบกพร่องจนทำให้วิทยานิพนธ์เล่มนี้ มีความสมบูรณ์มากยิ่งขึ้น ขอขอบพระคุณรองศาสตราจารย์ ดร. สุชาดา บวรกิติวงศ์ อาจารย์ ดร. สุรศักดิ์ เก้าเอี้ยน และคณาจารย์ภาควิชาวิจัยและจิตวิทยาการศึกษาทุกท่าน ที่ได้มอบความรู้อันมีค่า คอยอบรมสั่งสอน และให้ความช่วยเหลือตลอดระยะเวลาการศึกษาอันจะเป็นประโยชน์ให้ผู้วิจัยสามารถนำไปปรับใช้ในการทำงานในอนาคตต่อไป

ขอขอบพระคุณ รองศาสตราจารย์ ดร. สังวรณ์ งัดกระโทก ประธานและกรรมการสอบวิทยานิพนธ์และรองศาสตราจารย์ ดร. วรรณี แกมเกตุ กรรมการสอบวิทยานิพนธ์ ที่สละเวลาอันมีค่าให้คำแนะนำและกรุณาตรวจสอบความถูกต้องวิทยานิพนธ์ฉบับนี้ รวมทั้งยังมีความเมตตาชี้แนะแนวทางการแก้ไขปัญหา ข้อบกพร่องต่าง ๆ เพื่อให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณนางสาววันวิสาข์ เสาว์สิงห์ นางสาวประกายแก้ว ไกรสงคราม นายยุทธชัย ศิลาแลง ที่ได้ร่วมแลกเปลี่ยนความคิดเห็น แชร์ประสบการณ์การเรียนรู้ และเป็นกำลังใจที่ดีตลอดมา และนางสาวณัฐธนา ยั่งยืน ที่ให้คำปรึกษาด้านการเขียนโปรแกรม คอยให้ความช่วยเหลือผู้วิจัย และขอขอบคุณพี่ ๆ ทุกท่าน จากศูนย์ทรัพย์สินทางปัญญาและบ่มเพาะวิสาหกิจ มหาวิทยาลัยธรรมศาสตร์ที่ให้คำแนะนำและคอยให้กำลังใจผู้วิจัยเสมอ

ขอขอบคุณ นายธีรเดช จ้อยประดิษฐ ที่คอยให้คำแนะนำและคำปรึกษาในด้านการเขียนรายงานและคำแนะนำอื่น ๆ ที่เป็นประโยชน์ตลอดระยะเวลาที่ผู้วิจัยศึกษา คอยเป็นกำลังใจให้คำอวยพรแก่ผู้วิจัย และเชื่อมั่นในตัวของผู้วิจัยตลอดมาจนวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง

ท้ายที่สุดนี้วิทยานิพนธ์ฉบับนี้จะสำเร็จไม่ได้หากไม่ได้รับแรงบันดาลใจและกำลังใจที่สำคัญตลอดระยะเวลาการศึกษาจาก คุณแม่ นางพัชรนันท์ ฉิมสุด คุณพี่ชายนายฉัตรชัย ขันดวง คุณพี่สาวนางอรอนงค์ ขันดวง และครอบครัว ที่ได้เลี้ยงดู อบรมสั่งสอน ให้ความรักความเอาใจใส่ให้การสนับสนุนโอกาสทางการศึกษา และขอบคุณนางสาวทิพาพันธ์ ขันดวง ที่คอยตรวจสอบคำผิด ในวิทยานิพนธ์ฉบับนี้ และหลาน ๆ ที่น่ารักทุกคน ที่คอยเป็นกำลังใจให้ผู้วิจัยตลอดระยะเวลาจนวิทยานิพนธ์สำเร็จลุล่วง

นวลรัตน์ ฉิมสุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค
บทคัดย่อภาษาอังกฤษ ..... ง
กิตติกรรมประกาศ ..... จ
สารบัญ ..... ฉ
สารบัญตาราง ..... ฎ
สารบัญรูปภาพ ..... ฎ
บทที่ 1 1
บทนำ 1
ความเป็นมาและความสำคัญของปัญหาวิจัย 1
คำถามวิจัย 5
วัตถุประสงค์ 5
ขอบเขตของการวิจัย 5
นิยามศัพท์เฉพาะ 8
ประโยชน์ที่ได้รับจากการวิจัย 9
ประโยชน์เชิงวิชาการ 9
ประโยชน์ในเชิงปฏิบัติ 10
บทที่ 2 11
เอกสารและงานวิจัยที่เกี่ยวข้อง 11
ตอนที่ 1 วิธีการทดแทนค่าสูญหาย 11
ความหมายของข้อมูลสูญหาย 11
ประเภทของข้อมูลสูญหาย (type of missing data) 11

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random: MCAR) 12
การสูญหายแบบสุ่ม (Missing at random: MAR).....12
การสูญหายแบบไม่สุ่ม (missing not at random: MNAR).....13
สาเหตุของการเกิดค่าสูญหาย.....13
วิธีการทดแทนค่าสูญหาย 14
วิธีการประมาณค่าสูญหายสำหรับข้อมูลพหุระดับ 21
หลักการของวิธี Multiple Imputation (MI).....26
หลักการของการประมาณค่าสูญหายด้วยวิธี Random Forest Imputation 28
หลักการของการประมาณค่าสูญหายด้วยวิธี Optimal Imputation (Opt. impute)29
เกณฑ์ที่ในการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย.....30
การหาค่าเฉลี่ยของ NRMSE, RB, และ MCSE.....30
ตอนที่ 2 ความเหลื่อมล้ำทางการศึกษา.....32
ความหมายของความเหลื่อมล้ำทางการศึกษา 32
ความสำคัญของความเหลื่อมล้ำทางการศึกษาในประเทศไทย.....32
ปัจจัยที่สะท้อนความเหลื่อมล้ำทางการศึกษา.....34
ปัจจัยด้านลักษณะของโรงเรียน.....34
ปัจจัยด้านภูมิหลังฐานะความเป็นอยู่ของครอบครัว 36
กรอบแนวคิดที่ใช้ในการวิจัย 38
บทที่ 3 39
วิธีดำเนินการวิจัย.....39
ตอนที่ 1 ตัวแปรและโมเดลที่ใช้ในการศึกษา.....40
1. ตัวแปรที่ใช้ในการวิจัย.....40
2. โมเดลที่ใช้ในการศึกษา 41
ตอนที่ 2 การศึกษาการจำลองข้อมูล.....43

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1. ข้อตกลงเบื้องต้นและเงื่อนไขที่ใช้ในการศึกษา 44
2. ขั้นตอนในการจำลองข้อมูล (Simulation procedure) 44
ขั้นตอนที่ 1 การจำลองข้อมูลสมบูรณ์ (complete data) 45
ขั้นตอนที่ 2 จำลองข้อมูลไม่สมบูรณ์ (Missing data) 46
ขั้นตอนที่ 3 วิธีทดแทนค่าสูญหาย (Imputed data) 48
ขั้นตอนที่ 4 เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย 49
3. การตรวจสอบเงื่อนไขและความถูกต้องของโปรแกรมที่ใช้ในการจำลองข้อมูล 49
3.1 การตรวจสอบลักษณะการสูญหาย 50
3.2 การตรวจสอบอัตราการสูญหายของค่าสังเกต 52
ตอนที่ 3 การประยุกต์ใช้กับข้อมูลจริง 53
บทที่ 4 55
ผลการวิเคราะห์ข้อมูล 55
ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล 64
1.1 การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย NRMSE 64
1.1.1 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ (MCAR) 64
1.1.2 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่ม (MAR) 64
1.1.3 เมื่อข้อมูลมีลักษณะการสูญหายแบบไม่สุ่ม (MNAR) 65
1.1.4 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม (MCAR - MAR) 65
1.1.5 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบไม่สุ่ม (MCAR - MNAR) 65
1.1.6 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม (MAR - MNAR) 66
1.1.7 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่มและการสูญหายแบบไม่สุ่ม (MCAR – MAR – MNAR) 66

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.2 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายด้วยค่าเฉลี่ย RB และค่าเฉลี่ย MCSE....72
1.2.1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย RB....72
1.2.2 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย MCSE72
ตอนที่ 2 ผลการวิเคราะห์สถิติพื้นฐานและการประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง 73
2.1 ผลการวิเคราะห์สถิติพื้นฐานในข้อมูลจริงระดับโรงเรียน 74
2.1.1 ผลการวิเคราะห์ผลสัมฤทธิ์ทางการเรียนของนักเรียน ระดับโรงเรียน....74
2.1.2 ผลการวิเคราะห์ข้อมูลของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ และไม่ได้อาศัยอยู่กับบิดาและมารดา....75
2.2 การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง....77
2.2.1 การพิจารณาตัวแปรและค่าสูญหายในข้อมูลที่สนใจศึกษา 77
2.2.2 การพิจารณาลักษณะของข้อมูลสูญหาย (Type of missing data) 78
2.2.3 การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง 79
ตอนที่ 3 การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูล ที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย....79
บทที่ 5....85
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ....85
สรุปผลการวิจัย....86
ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล ..86
1.1 ประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ $(\beta 1)$ ในภาพรวม 86
1.2 ความถูกต้องและความแม่นยำในการประมาณค่าพารามิเตอร์ $(\beta 1)$....86
1.3 ความผิดพลาดในการประมาณค่าคลาดเคลื่อนมาตรฐานและความคงเส้นคงวา (consistency) ของการประมาณค่าพารามิเตอร์ $(\beta 1)$....87
ตอนที่ 2 ผลการวิเคราะห์การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง....87

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตอนที่ 3 การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา 88
อภิปรายผล 89
ข้อเสนอแนะ 91
1. ข้อเสนอแนะในการนำผลการวิจัยไปใช้ 91
2. ข้อเสนอแนะในการนำการวิจัยครั้งต่อไป 92
บรรณานุกรม 93
ประวัติผู้เขียน 99

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญตาราง

ตารางที่ 2. 1 แสดงการเปรียบเทียบวิธีการประมาณค่าสูญหาย 24
ตารางที่ 3. 1 การแจกแจงของตัวแปรและค่าพารามิเตอร์ที่ใช้ในการจำลองข้อมูล.....42
ตารางที่ 3. 2 แสดงผลการตรวจสอบเงื่อนไขการสูญหายในการจำลองข้อมูล 50
ตารางที่ 4.1 แสดงค่าเฉลี่ย NRMSE ของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1).....57
ตารางที่ 4. 2 แสดงค่าเฉลี่ย RB, MCSE ของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) 67
ตารางที่ 4. 3 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน 75
ตารางที่ 4. 4 แสดงจำนวนและร้อยละของข้อมูลนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 ..75
ตารางที่ 4. 5 แสดงผลการเปรียบเทียบระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย 80

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญรูปภาพ

รูปที่ 2.1 1 หลักการของวิธี Random forest Imputation 29
รูปที่ 2.1 2 กรอบแนวคิดที่ใช้ในการวิจัย.....38
รูปที่ 3.1. 1 แผนภาพขั้นตอนการจำลองข้อมูล.....45
รูปที่ 3.1. 2 การตรวจสอบระดับอัตราการสูญหายโดยใช้การจำลองข้อมูล 52
รูปที่ 3.1. 3 การประยุกต์ใช้กับข้อมูลจริง.....53
รูปที่ 3.1. 4 การวิเคราะห์ความเหลื่อมล้ำทางการศึกษา.....54
รูปที่ 4.1. 1 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม 60
รูปที่ 4.1. 2 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม.....61
รูปที่ 4.1. 3 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม.....62
รูปที่ 4.1. 4 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม.....63
รูปที่ 4.1.5 แสดงการเปรียบเทียบลอการิทึมของค่าเฉลี่ย RB ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่มด้วยวิธีทดแทนค่าสูญหาย.....70
รูปที่ 4.1. 6 แสดงการเปรียบเทียบลอการิทึมของค่าเฉลี่ย MCSE ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่มด้วยวิธีทดแทนค่าสูญหาย.....71
รูปที่ 4.1.7 แสดงผลการวิเคราะห์องค์ประกอบหลักของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน 74
รูปที่ 4.1. 8 แสดงค่าสูญหายในสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน 76
รูปที่ 4.1. 9 แสดงความสัมพันธ์ระหว่างผลสัมฤทธิ์ทางการเรียนของนักเรียนกับสัดส่วนนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา.....76

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหาวิจัย

การวิจัยด้านการศึกษาส่วนใหญ่มีจุดมุ่งหมายเพื่อนำผลวิจัยไปประยุกต์ใช้สำหรับแก้ปัญหาทางการศึกษาที่เกิดขึ้นในสังคม ผู้วิจัยจำเป็นต้องค้นคว้าหาความรู้เพื่อทำความเข้าใจปัญหาและเก็บรวบรวมข้อมูลเพื่อตอบคำถามการวิจัยที่กำหนดไว้ อย่างไรก็ตามปัญหาส่วนใหญ่ที่นักวิจัยประสบคือกลุ่มตัวอย่างไม่ให้ความร่วมมือในการตอบคำถาม ส่งผลให้เกิดค่าสูญหายในข้อมูลที่สนศึกษา โดยค่าสูญหายอาจส่งผลกระทบในการอนุมานเชิงสถิติ เช่น อาจทำให้การประมาณค่าพารามิเตอร์เกิดความลำเอียง (biased) อำนาจในการทดสอบสมมุติฐานทางสถิติลดลง โดยทั่วไปสาเหตุของการไม่ตอบคำถามมักเกิดจากการถามข้อมูลส่วนบุคคลที่อ่อนไหว (Sensitive Personal Data) ยกตัวอย่างเช่น การสอบถามข้อมูลรายได้ผู้ปกครองของนักเรียน มีความเป็นไปได้ที่นักเรียนที่ผู้ปกครองมีรายได้ต่ำ มีแนวโน้มที่จะไม่ตอบคำถามข้อมูลส่วนบุคคลมากกว่านักเรียนที่ผู้ปกครองมีรายได้ปานกลางถึงรายได้สูง ทั้งนี้โดยทั่วไปแบ่งลักษณะการสูญหายเป็น 3 ประเภทหลัก ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (Missing Completely at Random: MCAR) การสูญหายแบบสุ่ม (Missing at Random: MAR) และการสูญหายแบบไม่สุ่ม (Missing Not at Random: MNAR) (Nugroho & Surendro, 2019; สิวะโชติ, 2009)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ปัญหาข้อมูลสูญหายเป็นปัญหาหลักที่สำคัญที่อาจส่งผลกระทบทำให้ผลการวิจัยคลาดเคลื่อนจากความจริง หากผู้วิจัยวิเคราะห์ข้อมูลโดยไม่คำนึงถึงค่าสูญหาย อาจส่งผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูลได้ โดยอาจทำให้สูญเสียสาระสำคัญของรายละเอียดบางอย่างไป อาจทำให้ได้สารสนเทศที่ได้ไม่ครบถ้วนในการตอบคำถามวิจัย ทั้งนี้จากการศึกษาของ Wood และคณะ (2004) ศึกษางานวิจัยที่ได้รับการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet และ N Engl J Med พบว่างานวิจัยร้อยละ 89 ประสบปัญหากับข้อมูลสูญหาย ขณะที่มีเพียงร้อยละ 21 เท่านั้นที่มีการจัดกระทำกับปัญหาค่าสูญหาย จากผลการศึกษานี้แสดงให้เห็นว่าการจัดการกับปัญหาข้อมูลสูญหายยังคงถูกละเลย แม้แต่ในงานวิจัยด้านการศึกษาเองก็ตาม

เมื่อพิจารณาโครงสร้างของข้อมูลทางการศึกษาพบว่าธรรมชาติของข้อมูลถูกจัดเป็นหมวดหมู่ โดยกลุ่มนักเรียนถูกจัดให้อยู่รวมกันเป็นห้องเรียน แต่ละห้องเรียนจะถูกรวบรวมอยู่ในหมวดของโรงเรียน และในแต่ละโรงเรียนจะอยู่ภายใต้สังกัดของสำนักงานเขตพื้นที่ จะสังเกตเห็นว่าลักษณะของข้อมูลมีโครงสร้างเป็นระดับแบบลดหลั่น (Hierarchical data) ซึ่งในแต่ละระดับมีความสัมพันธ์ระหว่างหน่วยของข้อมูลซึ่งกันและกัน (สังวรณ์ งัดกระโทก 2545, สิวะโชติ ศรีสุทธิยากร 2562) ทั้งนี้เมื่อพิจารณางานวิจัยด้านการศึกษา พบว่านิยมใช้แบบสอบถามในการเก็บรวบรวมข้อมูลมากกว่าวิธีอื่น เนื่องจากเป็นการเปิดโอกาสให้ผู้ตอบมีเวลาคิด มีอิสระในการตอบโดยไม่ต้องเผชิญหน้ากับผู้ถาม ช่วยให้คำตอบที่ได้มีมาตรฐานเดียวกันทำให้ง่ายต่อการวิเคราะห์ผลการวิจัย ขณะที่แบบสอบถามมีจุดอ่อน คือ ผู้ตอบไม่ตอบคำถามบ้างข้อ หรือการใช้แบบสอบถามเพียงอย่างเดียวอาจไม่เพียงพอต่อการวิเคราะห์ผล ดังนั้นมีความจึงจำเป็นต้องใช้ข้อมูลทุติยภูมิที่เก็บรวบรวมจากหน่วยงานอื่นๆ เพิ่มเติม เช่น หน่วยงานทางการศึกษาของรัฐบาล โดยจะพบว่าการเก็บข้อมูลจากหน่วยงานทางการศึกษาของรัฐมีข้อจำกัดด้านการเผยแพร่ข้อมูล โดยเฉพาะอย่างยิ่งในการวิจัยทางการศึกษาและการวิจัยทางสังคมศาสตร์และพฤติกรรมศาสตร์ ที่มีความจำเป็นต้องเก็บข้อมูลส่วนบุคคลของนักเรียน ดังนั้นจึงมีความเสี่ยงสูงที่จะพบค่าสูญหาย เนื่องจากการไม่ตอบคำถาม รวมไปถึงเกิดเหตุการณ์การนำเข้าข้อมูลผิดพลาด หรือการเก็บข้อมูลระดับโรงเรียนผ่านหน่วยงานทางการศึกษาโดยผู้กรอกข้อมูลเป็นเจ้าหน้าที่ที่ไม่ใช่เจ้าของข้อมูล เป็นต้น

เมื่อพิจารณาค่าสูญหาย (Missing value) ในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับในบริบทของงานวิจัยด้านความเหลื่อมล้ำทางการศึกษาเนื่องจากความไม่เท่าเทียมทางการศึกษามีผลอันเนื่องมาจากภูมิหลังทางเศรษฐกิจและสังคมที่แตกต่างกัน เช่น รายได้ สถานะความเป็นอยู่ จะเห็นว่ามีความจำเป็นต้องเก็บข้อมูลส่วนบุคคลที่มีความอ่อนไหว (Sensitive Personal Data) (Jo Blanden and Stephen Machin 2004, Woessmann 2014, Weiss 2017, Alberto Alesina 2018, Marks and Pokropek 2019, Howie 2020) หรือเก็บรวบรวมข้อมูลทุติยภูมิจากหน่วยงานทางการศึกษาและหน่วยงานอื่นๆ ที่เกี่ยวข้องจากหลายแหล่งที่มาโดยที่แต่ละฐานมีรูปแบบและระบบการจัดเก็บที่แตกต่างกันทำให้ยากต่อการเชื่อมโยงข้อมูลส่งผลกระทบให้เกิดค่าสูญหาย (Missing value) ในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากการพิจารณาปัญหาของค่าสูญหายที่เกิดขึ้น (Missing value) ชี้ให้เห็นว่าหากผู้วิจัยวิเคราะห์ข้อมูลโดยไม่คำนึงถึงค่าสูญหายอาจส่งผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูลทำให้สรุปผลการวิจัยคลาดเคลื่อนจากความจริง ดังนั้นผู้วิจัยจึงทบทวนเอกสารและศึกษางานวิจัยที่เกี่ยวข้องกับประสิทธิภาพของวิธีทดแทนค่าข้อมูลสูญหายเพื่อทดแทนค่าสูญหายในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ พบว่าในปัจจุบันงานวิจัยที่ผ่านมา โดยส่วนใหญ่นิยมนำเทคนิค Multiple Imputation (MI) มาใช้สำหรับทดแทนค่าสูญหาย ซึ่งวิธีทดแทนค่าสูญหายโดยใช้เทคนิค MI ในปัจจุบันสามารถนำมาประยุกต์ใช้ได้หลายวิธีด้วยกัน เช่น สร้างการประมาณค่าโดยใช้วิธี predictive mean matching , Bayesian linear regression, หรือการใช้สมการโลจิสติกส์ แต่อย่างไรก็ตามวิธีการที่ศึกษาในการวิจัยครั้งนี้ คือวิธี Fully Conditional Specification Multiple Imputation (MI-FCS) เนื่องจากเป็นวิธีที่มีความยืดหยุ่นสามารถวิเคราะห์ได้ทั้งตัวแปรต่อเนื่องและตัวแปรไม่ต่อเนื่อง โดยผู้วิเคราะห์สามารถกำหนดจำนวนชุดข้อมูลและกำหนดจำนวนรอบในการวนซ้ำได้ โดยการวนซ้ำแต่ละครั้งเพื่อให้ได้ค่าที่ดีที่สุด เป็นต้น (Audigier et al, 2018; Hughes et al, 2019; Nissen et al, 2019; สัตทัด พรประเสริฐมานิต, 2562)

นอกจากนี้ศึกษาการทดแทนค่าสูญหายด้วยวิธี Random Forest Imputation (RF) ในงานวิจัยของ Jia and Wu ให้ผลการวิจัยไม่แตกต่างกับงานวิจัยของ Nissen, Donatello, and Van Dusen ชี้ให้เห็นว่าการทดแทนค่าสูญหายด้วยวิธี RF มีประสิทธิภาพสูงสามารถทดแทนค่าสูญหายได้ดีที่สุด แม้ว่าลักษณะของข้อมูลจะมีความซับซ้อน ไม่จำเป็นต้องทราบการแจกแจงของประชากร สามารถที่จะวิเคราะห์ข้อมูลที่มีการกระจายแบบไม่ปกติได้ โดยไม่เป็นการละเมิดข้อตกลงเบื้องต้น อีกทั้งยังพบว่าจากการศึกษางานวิจัยของ Bertsimas, Pawlowski et al. (2018) จากสถาบันเทคโนโลยีแมสซาชูเซตส์ (Massachusetts Institute of Technology: MIT) ได้พัฒนาวิธีทดแทนข้อมูลสูญหาย Optimal Imputation (Opt.impute) โดยให้ความสำคัญกับการลดข้อจำกัดในอดีต โดยใช้เทคนิคการให้โปรแกรมทำงานจากการเรียนรู้ (Machine Learning) จากผลการวิจัยของ Bertsimas, Pawlowski et al. (2018) พบว่าวิธีทดแทนค่าสูญหาย Opt.impute มีประสิทธิภาพสูงที่สุด เมื่อเทียบกับวิธีการในอดีต เช่น PMM, KNN, BPCA, IKNN และ Mean เป็นต้น โดยชี้ให้เห็นว่าหากทดแทนค่าสูญหายด้วยวิธี Opt.impute จะทำให้ผลการวิจัยมีความแม่นยำเพิ่มขึ้นถึงร้อยละ 2 สามารถลดข้อผิดพลาดในการทดแทนค่าสูญหายมากถึงร้อยละ 10-15 อย่างมีนัยสำคัญทางสถิติ (Bertsimas et al.,2019)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เมื่อพิจารณาการศึกษาข้างต้น จะพบว่าการทดแทนค่าสูญหายด้วยวิธี MI - FCS และวิธี RF ซึ่งเป็นวิธีที่ได้รับความนิยมในปัจจุบัน และการทดแทนค่าสูญหายวิธี Opt.impute ซึ่งเป็นวิธีที่ได้พัฒนาขึ้นมาใหม่ ยังไม่ได้ถูกนำมาเปรียบเทียบในบริบทของข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ และทั้งนี้นอกจากการพิจารณาการประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลที่มีโครงสร้างแบบพหุระดับในเชิงเปรียบเทียบแล้ว การพิจารณาถึงสาเหตุของการเกิดค่าสูญหายยังเป็นขั้นตอนหนึ่งที่สำคัญ เนื่องจากหากผู้วิจัยทราบถึงสาเหตุและลักษณะของการสูญหายจะทำให้สามารถเพิ่มประสิทธิภาพให้กับวิธีทดแทนค่าสูญหายได้มากขึ้น โดยทั่วไปแบ่งลักษณะการเกิดค่าสูญหายได้ 3 ประเภท ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (MCAR) การสูญหายแบบสุ่ม (MAR) และการสูญหายแบบไม่สุ่ม (MNAR) โดยในทางปฏิบัตินักวิจัยไม่สามารถจำแนกได้ว่าข้อมูลมีการสูญหายประเภทใด และมีความเป็นไปได้น้อยมากที่จะเกิดการสูญหายเพียงสาเหตุเดียว และไม่สามารถอธิบายลักษณะการสูญหายจากข้อมูลที่สังเกตได้เพียงอย่างเดียว จำเป็นต้องมีการศึกษาความสัมพันธ์ของข้อมูลที่ไม่ได้สนใจศึกษาเพิ่มเติมด้วย ดังนั้นในการวิจัยครั้งนี้ผู้วิจัยจึงสนใจศึกษาลักษณะการสูญหายรูปแบบผสมเพิ่มเติมรวมทั้งหมด 7 รูปแบบ (Cro, Morris, Kenward, & Carpenter, 2020)

จากการทบทวนเอกสารที่เกี่ยวข้องกับวิธีทดแทนค่าสูญหายและการเกิดค่าสูญหายในข้อมูลทางการศึกษาที่ได้กล่าวไว้ในข้างต้น พบว่ามีประเด็นที่มีความน่าสนใจสองประเด็น ประเด็นแรกได้แก่ การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลที่มีโครงสร้างแบบพหุระดับโดยใช้การจำลองข้อมูล (Simulation data) ด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) จำนวน 3 วิธี ได้แก่ วิธี MI - FCS วิธี RF วิธี Opt.impute ภายใต้ 3 เงื่อนไขที่แตกต่างกันดังนี้ ประเภทของการสูญหาย 7 รูปแบบ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MAR, MAR-MNAR, MCAR-MAR-MNAR ขนาดตัวอย่างในระดับที่หนึ่ง 3 ขนาด ได้แก่ 1000, 2000 และ 3,000 หน่วย ขนาดตัวอย่างในระดับที่สองเท่ากับ 40, 50 และ 60 หน่วย ตามลำดับ และอัตราการสูญหายที่แตกต่างกัน 3 ระดับ เท่ากับร้อยละ 30, 40 และร้อยละ 50 โดยใช้ค่าเฉลี่ยของค่า NRMSE, ค่าเฉลี่ยของ RB และ ค่าเฉลี่ยของ MCSE เป็นเกณฑ์ในการเลือกวิธีการทดแทนค่าสูญหายที่มีประสิทธิภาพ ประเด็นที่สองคือดำเนินการนำวิธีทดแทนข้อมูลสูญหายที่เหมาะสมจากผลในประเด็นแรก มาประยุกต์ใช้กับข้อมูลจริง (Real data) โดยนำมาประยุกต์ใช้ในบริบทของความเหลื่อมล้ำทางการศึกษา โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มในการวิเคราะห์สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บิดามารดาระดับโรงเรียนซึ่งเป็นตัวแปรที่มีความอ่อนไหวที่มักเกิดค่าสูญหาย ที่ส่งผลกระทบต่อผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน โดยผลการวิจัยในครั้งนี้จะให้องค์ความรู้ทั้งในทางทฤษฎีและทางปฏิบัติ โดยนักวิจัยและผู้ที่สนใจสามารถนำผลการวิจัย ไปใช้ในการตัดสินใจเลือกวิธีทดแทนค่าข้อมูลสูญหายให้เหมาะสมกับสถานการณ์ที่ต้องการศึกษา

คำถามวิจัย

1. ประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยวิธี MI-FCS, RF และ Opt.impute มีความแตกต่างกันหรือไม่อย่างไรสำหรับการวิเคราะห์ข้อมูลพหุระดับ

2. เมื่อมีการทดแทนค่าสูญหายด้วยวิธีการที่มีประสิทธิภาพที่เหมาะสมที่สุดแล้ว ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษามีความแตกต่างกันหรือไม่อย่างไร เมื่อนำมาเทียบกับผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่ไม่ได้ทดแทนค่าสูญหาย

วัตถุประสงค์

1. เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าข้อมูลสูญหายระหว่างวิธี MI-FCS, วิธี RF และวิธี Opt.impute สำหรับการวิเคราะห์ข้อมูลพหุระดับโดยใช้การจำลองข้อมูล

2. เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษาด้วยโมเดลพหุระดับที่มีการทดแทนค่าสูญหายด้วยวิธีที่มีประสิทธิภาพและเหมาะสมกับลักษณะของข้อมูลจริงมากที่สุดที่ได้จากผลการจำลองข้อมูล และเปรียบเทียบผลที่ได้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่ไม่ได้ทดแทนค่าสูญหาย

ขอบเขตของการวิจัย

การวิจัยครั้งนี้ ผู้วิจัยมุ่งเน้นการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับโดยใช้การจำลองข้อมูล การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง และเปรียบเทียบผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่ทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย การวิจัยครั้งนี้ ช้ข้อมูลของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวน 2,109 โรงเรียน ที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) จำนวน 42 สังกัด โดยผู้วิจัยแบ่งขอบเขตของการวิจัยออกเป็น 2 ระยะ มีรายละเอียดดังนี้

ระยะที่หนึ่ง คือ การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายโดยใช้การจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล แต่ละสถานการณ์ทำซ้ำ 100 รอบ โดยมีขอบเขตของการวิจัย ดังนี้

1. โมเดลที่ใช้ศึกษา คือ โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มที่มีสองระดับ โดยที่ระดับที่หนึ่ง คือ ระดับโรงเรียน และระดับที่สอง คือ ระดับสังกัด สามารถเขียนสมการได้ ดังนี้

โมเดลระดับที่หนึ่ง: ระดับโรงเรียน $y_{ij} = \alpha_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$ (Level 1 Model)

โมเดลระดับที่หนึ่ง: ระดับสังกัด $\alpha_{0j} = \gamma_{00} + U_{0j}$, $\beta_{1j} = \gamma_{01} + U_{1j}$ (Level 2 Model)

เมื่อกำหนดให้

y_{ij}	แทน	ผลสัมฤทธิ์ทางการเรียนของนักเรียนโดยเฉลี่ยในโรงเรียนที่ i สังกัดที่ j
x_{ij}	แทน	สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา ระดับโรงเรียนที่ i และสังกัดที่ j
γ_{00}	แทน	สัมประสิทธิ์จุดตัดแกน
γ_{01}	แทน	อัตราการเปลี่ยนแปลงของสัดส่วนนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาที่มีผลต่อผลสัมฤทธิ์ทางการเรียน
$\varepsilon_{ij} \sim N(0,\sigma^2)$	แทน	ความคลาดเคลื่อนสุ่มของโมเดลในระดับโรงเรียน
$U_{ij} \sim N(0,\Sigma_u)$	แทน	ความคลาดเคลื่อนสุ่มของโมเดลในระดับระดับสังกัด

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

$\Sigma_u = \begin{bmatrix} 0.588 & 0 \\ 0 & 1.272 \end{bmatrix}$ คือ เมทริกซ์ความแปรปรวนของเวกเตอร์ความคลาดเคลื่อนสุ่มในระดับสังกัด

2. สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ และไม่ได้อาศัยอยู่กับบิดามารดา (x) กำหนดให้มีการแจกแจงแบบเบต้า (Beta Distribution) สามารถเขียนแทนได้ด้วย $x \sim \text{Beta}(\alpha,\beta)$ โดยที่ $\alpha = 0.0579, \beta = 15.278$

3. วิธีทดแทนค่าสูญหายจำนวน 3 วิธี ได้แก่ วิธี Fully Conditional Specification Multiple Imputation (MI-FCS), Random Forest Imputation (RF) และ วิธี Optimal imputation ซึ่งประกอบด้วย วิธี Opt.knn, Opt.tree, Opt.svm และ Opt.cv

4. พิจารณาประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วย ค่าเฉลี่ยของค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (Normalized Root Mean Square Error: NRMSE)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ค่าเฉลี่ยของค่าความเอนเอียงสัมพัทธ์ (Relative Biased: RB) และค่าเฉลี่ยของค่าความคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (Monte Carlo Standard Error: MCSE) ถ้าวิธีใดแสดงดังกล่าวน้อยกว่า แสดงว่ามีประสิทธิภาพในการทดแทนค่าข้อมูลสูญหายได้ดีกว่า

5. ขนาดของตัวอย่างในระดับที่หนึ่ง กำหนดให้เท่ากับ 1000, 2000, และ 3,000 หน่วย ขนาดตัวอย่างระดับที่สอง เท่ากับ 40, 50, และ 60 หน่วย ตามลำดับ

6. ลักษณะการสูญหาย 7 รูปแบบ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MNAR, MAR-MNAR, และ MCAR -MAR-MNAR

7. ระดับของอัตราการสูญหายในตัวอย่างระดับที่หนึ่ง เท่ากับ ร้อยละ 30, ร้อยละ 40 และร้อยละ 50 ตามลำดับ

เมื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล จะได้วิธีทดแทนค่าสูญหายที่เหมาะสม ภายใต้เงื่อนไขที่กำหนดข้างต้น ทั้งนี้เพื่อให้เกิดประโยชน์ในการนำไปใช้กับข้อมูลจริงในอนาคตได้ ดังนั้นการวิจัยระยะที่สองผู้วิจัยจึงดำเนินการนำวิธีทดแทนข้อมูลสูญหายมาประยุกต์ใช้กับข้อมูลจริง เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา และเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย โดยมีขอบเขตการวิจัยระยะที่ 2 ดังนี้

ระยะที่สอง คือ การวิจัยในระยะนี้ ผู้วิจัยจึงนำวิธีทดแทนค่าสูญหายที่ได้จากการจำลองข้อมูลในระยะที่ 1 มาประยุกต์ใช้กับข้อมูลจริง ซึ่งข้อมูลที่ใช้ในการวิจัยระยะนี้ คือ ข้อมูลทุติยภูมิจากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) โดยใช้ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวน 2,109 โรงเรียน ที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) ทั้งหมด 42 แห่ง โดยการวิจัยระยะนี้ ผู้วิจัยกำหนดประชากรและกลุ่มตัวอย่าง ดังนี้

1. ประชากรที่ใช้ในการวิจัย เนื่องจากการวิจัยในระยะนี้ ผู้วิจัยประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ดังนั้นผู้วิจัยจึงกำหนดประชากรโดยพิจารณารัฐธรรมนูญที่กำหนดไว้ว่ารัฐต้องจัดให้เรียนฟรี 12 ปี ตั้งแต่ระดับชั้นอนุบาลถึงระดับชั้นมัธยมศึกษาปีที่ 3 ซึ่งประชาชนทุกคนมีสิทธิได้รับการศึกษาขั้นพื้นฐานอย่างเท่าเทียม

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เมื่อพิจารณาสภาพความจริงของประเทศไทย พบว่าที่ผ่านมาอัตราการเข้าถึงการศึกษาของประชาชนภายในประเทศ มีแนวโน้มลดลงเรื่อยๆ ตามระดับการศึกษาที่สูงขึ้น (สำนักงานเลขาธิการสภาการศึกษา, 2561) จะเห็นว่าการศึกษาในประเทศไทยไม่สอดคล้องกับรัฐธรรมนูญที่กำหนดไว้ ดังนั้น การวิจัยในครั้งนี้ ผู้วิจัยจึงเลือกนักเรียนในระดับชั้นมัธยมศึกษาปีที่ 3 โดยเลือกข้อมูลของโรงเรียนที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษาในการศึกษา โดยที่ใช้ผลคะแนน O-NET แทน ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน ซึ่งเป็นตัวชี้วัดหนึ่งที่สามารถใช้สะท้อนความเหลื่อมล้ำทางการศึกษาได้อย่างเหมาะสม

2. ตัวแปรที่ใช้ในการวิจัย ประกอบด้วย ตัวแปรในระดับที่หนึ่ง คือ ระดับโรงเรียน และตัวแปรในที่สอ งคือ ระดับสังกัดเขตพื้นที่การศึกษา ซึ่งตัวแปรที่ใช้ในการวิจัยระยะที่ 2 นี้ เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษ าเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย ซึ่งประกอบด้วย

ตัวแปรอิสระที่ใช้ในการวิจัยระยะที่สอง คือ สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา ระดับโรงเรียน

ตัวแปรตามที่ใช้ในการวิจัยระยะที่สอง คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน

นิยามศัพท์เฉพาะ

วิธีทดแทนค่าสูญหาย หมายถึง การทดแทนค่าพารามิเตอร์หรือค่าตัวแปรที่ไม่สามารถเก็บรวบรวมได้จากการสำรวจ โดยเป็นวิธีการจัดกระทำให้ข้อมูลครบถ้วนสมบูรณ์ก่อนนำไปวิเคราะห์ผลการวิจัย ซึ่งในการวิจัยครั้งนี้ใช้วิธีทดแทนค่าสูญหายจำนวน 3 วิธี ได้แก่ วิธีทดแทนค่าสูญหาย MI-FCS, วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Opt.impute โดยประกอบด้วย 4 วิธี คือ วิธีทดแทนค่าสูญหาย Opt.knn , วิธีทดแทนค่าสูญหาย Opt.tree, วิธีทดแทนค่าสูญหาย Opt.svm, และวิธีทดแทนค่าสูญหาย Opt.cv ตามลำดับ

ประสิทธิภาพของวิธีทดแทนค่าสูญหาย หมายถึง ผลของกระบวนการทดแทนค่าสูญหาย โดยในการวิจัยนี้วัดประสิทธิภาพวิธีทดแทนค่าสูญหายจากเกณฑ์การวัดค่าเฉลี่ยของ NRMSE, ค่าเฉลี่ยของ RB และค่าเฉลี่ยของ MCSE โดยค่าเฉลี่ยของความเคลื่อนที่มีค่าน้อยจะเป็นเกณฑ์ที่แสดงถึงวิธีการที่มีประสิทธิภาพสูงและเหมาะสม

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ความเหลื่อมล้ำทางการศึกษา หมายถึง สภาพที่เกิดความแตกต่างของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน โดยมีสาเหตุอันเนื่องมาจากความแตกต่างของสัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียนที่แตกต่างกัน โดยการวิจัยครั้งใช้ผู้วิจัยพิจารณาความแตกต่างของสัมประสิทธิ์ความถดถอยแบบสุ่ม (β_{1j}) ของโรงเรียนในแต่ละสังกัด สพม.

สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน หมายถึง การพิจารณาตัวแปรสองตัวแปรระดับโรงเรียน คือ สัดส่วนของนักเรียนที่พักอาศัยอยู่กับบิดามารดา และสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ โดยการรวมคะแนนทั้งสองด้านเข้าด้วยกันด้วยค่าเฉลี่ยเรขาคณิต จะได้สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน

ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน หมายถึง ความรู้ความเข้าใจขั้นพื้นฐานของนักเรียนระดับโรงเรียนซึ่งการวิจัยนี้พิจารณาจากผลคะแนนจากการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (O-NET) จำนวน 4 วิชา ได้แก่ วิชาภาษาไทย วิชาภาษาอังกฤษ วิชาคณิตศาสตร์ และวิชาวิทยาศาสตร์

ประโยชน์ที่ได้รับจากการวิจัย

ประโยชน์เชิงวิชาการ

1. สามารถนำผลการวิจัยที่ได้จากการเปรียบเทียบวิธีทดแทนค่าสูญหายในการศึกษาข้อมูลที่มีโครงสร้างแบบพหุระดับครั้งนี้ มาขยายองค์ความรู้ทั้งในทางทฤษฎีและทางปฏิบัติ เพื่อนำไปประยุกต์ใช้ในข้อมูลทางการศึกษา และเป็นแนวทางในการศึกษาเปรียบเทียบวิธีทดแทนค่าสูญหายในการวิเคราะห์ข้อมูลที่มีความสัมพันธ์กันในรูปแบบอื่น ๆ หรือการสูญหายภายใต้สถานการณ์ที่แตกต่างจากงานวิจัยนี้

2. ผลการวิจัยที่ได้สามารถสะท้อนความเหลื่อมล้ำทางการศึกษาในปัจจุบันได้ในด้านภูมิหลังทางฐานะเศรษฐกิจและสังคม เป็นแนวทางในการวางแผนนโยบายและส่งเสริมการศึกษาเพื่อลดปัญหาความเหลื่อมล้ำทางการศึกษาเพื่อให้นักเรียนได้มีโอกาสทางการศึกษาที่เท่าเทียมกันมากขึ้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประโยชน์ในเชิงปฏิบัติ

1. สามารถนำผลการวิจัยไปใช้ในการตัดสินใจเลือกวิธีการทดแทนค่าข้อมูลสูญหายสำหรับข้อมูลทางการศึกษาให้เหมาะสมและสอดคล้องกับสถานการณ์ที่ต้องการศึกษาเพื่อให้ได้สารสนเทศที่ถูกต้องสอดคล้องกับข้อมูลเพื่อลดความคลาดเคลื่อนที่อาจเกิดขึ้น สามารถทดแทนค่าได้ถูกต้องแม่นยำมากขึ้นและสามารถอนุมานไปสู่ประชากรได้ดีมากยิ่งขึ้น

2. สามารถนำผลการวิจัยในครั้งนี้ไปใช้เพื่อเป็นแนวทางในการเลือกวิธีการทดแทนค่าสูญหายสำหรับข้อมูลทางการศึกษาในบริบทอื่น โดยสามารถนำไปประยุกต์ใช้โมเดลความเหลื่อมล้ำทางการศึกษาที่ประยุกต์กับปัจจัยอื่น ๆ ที่ยังไม่ได้นำมาศึกษาในครั้งนี้ เช่น ปัจจัยด้านตนเอง ปัจจัยสภาพแวดล้อมจากเพื่อน เป็นต้น และสามารถนำผลการเปรียบเทียบวิธีการทดแทนค่าสูญหายที่ได้จากผลการวิจัยครั้งนี้ไปประยุกต์ใช้พัฒนาวิธีการทดแทนค่าสูญหายในโมเดลพหุระดับที่สูงขึ้นในบริบทอื่นเพิ่มเติมต่อไปได้ในอนาคต

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาวิจัยเรื่องการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ: การประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ผู้วิจัยรวบรวมแนวคิดทฤษฎีเอกสารและงานวิจัยที่เกี่ยวข้องกับวิธีทดแทนค่าข้อมูลสูญหาย และความเหลื่อมล้ำทางการศึกษาโดยผู้วิจัยแบ่งการนำเสนอออกเป็น 2 ตอน ดังนี้ ตอนที่ 1 วิธีทดแทนค่าสูญหาย ตอนที่ 2 ความเหลื่อมล้ำทางการศึกษา โดยมีรายละเอียดดังนี้

ตอนที่ 1 วิธีการทดแทนค่าสูญหาย

ผู้วิจัยทำการศึกษาเอกสารที่เกี่ยวข้องกับวิธีการทดแทนค่าข้อมูลสูญหาย โดยแบ่งหัวข้อการนำเสนอดังนี้ ความหมายของข้อมูลสูญหาย ประเภทของข้อมูลสูญหาย สาเหตุของการเกิดค่าสูญหายวิธีทดแทนค่าสูญหาย เกณฑ์ที่ในการเปรียบเทียบวิธีการทดแทนค่าสูญหาย โดยมีรายละเอียดดังนี้

ความหมายของข้อมูลสูญหาย

ข้อมูลสูญหาย (Missing Data) หมายถึง ข้อมูลที่ไม่ครบถ้วน ข้อมูลที่ไม่สมบูรณ์หรือการไม่ปรากฏคำตอบในแบบสอบถาม เช่น ไม่ทราบคำตอบ ไม่ประสงค์ตอบ ตั้งใจปฏิเสธการตอบ ไม่มีความรู้เพียงพอ และข้อคำถามมีความยาวเกินไป เป็นต้น (Schlomer, Bauman et al. 2010, Nugroho and Surendro 2019)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประเภทของข้อมูลสูญหาย (type of missing data)

การพิจารณาประเภทของข้อมูลสูญหายเป็นขั้นตอนที่สำคัญขั้นตอนหนึ่งเนื่องจากหากทราบถึงลักษณะการสูญหายของข้อมูลจะช่วยในการพิจารณาแนวทางจัดการกับปัญหาความไม่สมบูรณ์ของข้อมูลได้อย่างเหมาะสม โดยทั่วไปจำแนกประเภทการสูญหายของข้อมูล ได้ 3 ประเภทหลัก (Roderick J. A. Little, Donald B. Rubin, 2019) ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random) การสูญหายแบบสุ่ม (Missing at random) และการสูญหายแบบไม่สุ่ม (missing not at random) ซึ่งมีรายละเอียดดังต่อไปนี้

การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random: MCAR)

การสูญหายแบบสุ่มสมบูรณ์ เป็นการสูญหายที่ลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด ข้อมูลที่สูญหายเป็นอิสระจากตัวแปรต่าง ๆ กล่าวคือข้อมูลที่สูญหายไม่มีความสัมพันธ์กับตัวแปรใด ๆ หรือค่าอื่น ๆ ทั้งที่ทราบค่าและไม่ทราบค่า การสูญหายในลักษณะนี้ไม่ขึ้นกับปัจจัยใด ๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหาย โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลที่สมบูรณ์และข้อมูลสูญหาย เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่าง ๆ ในข้อมูลที่สนใจศึกษา ตัวอย่างเช่น ข้อมูลสูญหายเกิดขึ้นเนื่องจากเครื่องมือเสียอุปกรณ์เกิดข้อบกพร่อง กลุ่มเป้าหมายที่ศึกษาล้มป่วย หรือการนำเข้าข้อมูลไม่ถูกต้องสำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นหากเกิดการสูญหายประเภทนี้สามารถตัดหน่วยตัวอย่างที่มีข้อมูลสูญหายได้ซึ่งสมการความน่าจะเป็นของการสูญหายสามารถอธิบายเพิ่มเติมได้ดังนี้ (วุฒิ สุขเจริญ 2015, Nugroho and Surendro 2019)

$$P\ (\text{Missing} \mid \text{Data Complete}) = P\ (\text{Missing})$$

การสูญหายแบบสุ่ม (Missing at random: MAR)

การสูญหายแบบสุ่ม หรืออาจเรียกว่า การสูญหายแบบสุ่มที่มีเงื่อนไข (missing conditionally random) เป็นการสูญหายที่ลักษณะของข้อมูลสูญหายไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมดแต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต กล่าวคือค่าสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในข้อมูลที่สนใจศึกษา ยกตัวอย่างเช่น หากพบว่ากลุ่มผู้ได้รับการศึกษาน้อยที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการเสพยาเสพติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพติดมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับการศึกษาซึ่งสมการความน่าจะเป็นของการสูญหายสามารถอธิบายเพิ่มเติมได้ดังนี้ (วุฒิ สุขเจริญ 2015, Nugroho and Surendro 2019)

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

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

การสูญหายแบบไม่สุ่ม (missing not at random: MNAR)

การสูญหายแบบไม่สุ่มเป็นการสูญหายที่ลักษณะของข้อมูลสูญหายไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกันรวมถึงตัวแปรตัวอื่นด้วย ตัวอย่างเช่น หากข้อมูลสูญหายของระดับรายได้ ขึ้นอยู่กับระดับรายได้ในแต่ละช่วงอายุ กล่าวคือบุคคลที่มีช่วงอายุมากที่มีรายได้ต่อเดือนอยู่ในระดับต่ำเลือกที่จะไม่ตอบคำถามรายได้ต่อเดือนมากกว่าบุคคลที่มีช่วงอายุน้อย หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใด ๆ ในข้อมูลที่สนใจศึกษาเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้นแต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในข้อมูลที่สนใจศึกษา ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกข้อมูลที่สนใจศึกษา ลักษณะข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่สามารถส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูล ก่อให้เกิดปัญญาในการวิเคราะห์ที่ร้ายแรงมากกว่า MCAR และ MAR เนื่องจากทำให้เกิดความคลาดเคลื่อนของผลงานวิจัย ซึ่งสมการความน่าจะเป็นของการสูญหายสามารถอธิบายเพิ่มเติมได้ดังนี้ (วุฒิ สุขเจริญ 2015, Nugroho and Surendro 2019)

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

ในทางปฏิบัติลักษณะของข้อมูลสูญหายประเภท MCAR มักพบไม่บ่อยนัก ข้อมูลสูญหายที่พบบ่อยคือ ข้อมูลสูญหายประเภท MAR หรือ MNAR ดังนั้นวิธีการทางสถิติต่าง ๆ ที่พัฒนาขึ้นมาเพื่อแก้ปัญหาข้อมูลสูญหาย มักดำเนินการภายใต้ข้อสมมติของ MAR หรือ MNAR เป็นส่วนใหญ่

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สาเหตุของการเกิดค่าสูญหาย

เมื่อพิจารณาจากแหล่งที่มาของการเกิดข้อมูลสามารถแบ่งการเกิดค่าสูญหายได้ 3 ประเภท รายละเอียด ดังนี้

การสูญหายที่เกิดจากกระบวนการวิจัย ได้แก่ การกำหนดประเด็นวิจัยที่กว้างมากเกินไป ส่งผลให้ผู้ตอบแบบสอบถามไม่สามารถให้ข้อมูลได้ครบทุกประเด็น การออกแบบสอบถามที่ใช้ระยะเวลาตอบมากเกินไป การใช้คำที่ไม่เหมาะสมอาจทำให้เกิดความไม่เข้าใจในคำถามทำให้ผู้ตอบแบบสอบถามละเว้นคำตอบ การกำหนดกลุ่มตัวอย่างที่กว้างเกินไป กระบวนการเก็บรวบรวมข้อมูลไม่เหมาะสมรวมไปถึงความผิดพลาดในกระบวนการนำเข้าข้อมูล

การสูญหายที่เกิดจากผู้ให้ข้อมูล ได้แก่ ผู้ให้ข้อมูลไม่เต็มใจให้ข้อมูลหรือกลุ่มตัวอย่างไม่สามารถให้ข้อมูลได้อย่างต่อเนื่อง

โดยทั่วไปแบ่งลักษณะของการค่าสูญหาย 3 รูปแบบ ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random: MCAR) การสูญหายแบบสุ่ม (Missing at random: MAR) และการสูญหายแบบไม่สุ่ม (missing not at random: MNAR)

ในการศึกษาวิจัยครั้งนี้ผู้วิจัยใช้การจำลองข้อมูลและใช้ข้อมูลทุติยภูมิ (secondary data) ที่เก็บรวบรวมจากระบบสารสนเทศจากหน่วยงานทางการศึกษาของประเทศ ดังนั้นสาเหตุของการสูญหายที่เกิดขึ้นจึงเกิดจากผู้ให้ข้อมูลกลุ่มตัวอย่างไม่สามารถให้ข้อมูลได้อย่างต่อเนื่อง ดังนั้นเพื่อความครอบคลุม ผู้วิจัยจึงทำการศึกษาโอกาสที่อาจค่าสูญหายที่เป็นไปได้ทั้งหมด 7 รูปแบบ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MAR, MAR-MNAR, MCAR-MAR-MNAR

วิธีการทดแทนค่าสูญหาย

วิธีการทดแทนค่าสูญหายเป็นกระบวนการหนึ่งในการจัดกระทำข้อมูลให้มีความพร้อม (Tidy data) ก่อนนำข้อมูลไปวิเคราะห์ผลการวิจัย การทดแทนค่าสูญหายสำหรับข้อมูลทางการศึกษาจึงเป็นประเด็นหนึ่งที่หน่วยงานทางการศึกษาควรให้ความสำคัญ เนื่องจากในสภาพความเป็นจริงกระบวนการเก็บรวบรวมข้อมูลทางการศึกษามักประสบกับปัญหาข้อมูลสูญหาย (missing data) ซึ่งอาจมีสาเหตุอันเนื่องมาจากการละเว้นคำตอบและการไม่ปรากฏคำตอบในข้อมูลที่เก็บรวบรวมได้ เช่น การสำรวจข้อมูลนักเรียนโดยลักษณะข้อคำถามที่ใช้ เป็นคำถามที่กระทบต่อความรู้สึกได้ง่าย (sensitive question) ไม่ว่าจะเป็นคำถามที่เกี่ยวกับสภาพความเป็นอยู่ของครอบครัว ลักษณะที่พักอาศัย หรือแม้กระทั้งคำถามทั่วไปทางสังคม อย่างเช่นรายได้ของผู้ปกครองต่อปี ซึ่งการถามด้วยคำถามที่กล่าวมาอาจมีแนวโน้มที่จะไม่ได้รับความร่วมมือ ในการตอบ โดยสถานการณ์ที่เกิดขึ้นนั้น ผู้วิจัยไม่สามารถควบคุมได้ หากไม่คำนึงถึงค่าสูญหายที่เกิดขึ้น อาจส่งผลกระทบให้ผลการวิเคราะห์มีความเอนเอียง (bias) มีความเสี่ยงที่จะสูญเสียเวลาในการวิเคราะห์เสียทรัพยากร เนื่องจากได้ผลการวิจัยที่คลาดเคลื่อนจากความเป็นจริง จะเห็นว่าการเลือกวิธีการจัดการกับข้อมูลสูญหายให้เหมาะสมมีความสำคัญต่อความถูกต้องแม่นยำในการสรุปผลการวิจัยเป็นอย่างยิ่ง (Horton & Lipsitz, 2001; Jiang, Josse, & Lavielle, 2020; Nissen et al., 2019; Pampaka, Hutcheson, & Williams, 2014; Pham et al., 2019; สิวะโชติ ศรีสุทธิยากร, 2552)

จากการศึกษาของ Angela M Wood (2004) ศึกษางานวิจัยที่ได้รับการตีพิมพ์ในวารสาร BMJ, JAMA, Lancet และ NEJM จำนวน 71 บทความ พบข้อมูลสูญหายปรากฏอยู่มากถึงร้อยละ 89 จากงานวิจัยทั้งหมดโดยมีเพียงร้อยละ 21 เท่านั้นที่ผู้วิจัยไม่คำนึงถึงข้อมูลสูญหาย นอกจากนี้ Nissen, Donatello et al. (2019) ศึกษางานวิจัยการศึกษาทางฟิสิกส์ (physics education research) จำนวน 28 บทความพบว่าเฉลี่ยร้อยละ 23 ไม่มีการทดแทนค่าข้อมูลที่สูญหายกล่าวคือวิเคราะห์ข้อมูลโดยไม่คำนึงถึงข้อมูลสูญหาย จากนั้นเปรียบเทียบประสิทธิภาพของผลการวิเคราะห์งานวิจัยสองประเภท ประเภทแรกคืองานวิจัยที่ไม่คำนึงถึงข้อมูลสูญหาย กับงานวิจัยที่ใช้วิธีการทดแทนค่าสูญหายด้วยวิธีต่าง ๆ ผลการวิเคราะห์พบว่าการใช้วิธีการทดแทนค่าสูญหายในการจัดกระทำข้อมูลก่อนนำไปวิเคราะห์ ให้ผลการวิเคราะห์ที่มีความคลาดเคลื่อนน้อยกว่าการไม่คำนึงถึงข้อมูลสูญหายเช่นเดียวกันกับการศึกษาของ Horton and Lipsitz (2001) และ Hughes et al. (2019) ให้ผลการวิจัยที่สอดคล้องทำนองเดียวกันว่าวิธีทดแทนค่าข้อมูลสูญหายเป็นการตรวจสอบข้อมูลเบื้องต้นที่มีความจำเป็นและเป็นขั้นตอนที่สำคัญก่อนนำข้อมูลไปวิเคราะห์ผลด้วยสถิติวิเคราะห์ เพื่อสามารถสรุปผลการวิจัยที่ถูกต้องแม่นยำมากยิ่งขึ้น

จากการศึกษาและทบทวนงานวิจัยที่เกี่ยวข้องกับวิธีการทดแทนค่าข้อมูลสูญหายพบงานวิจัยจำนวนมากที่ศึกษาประสิทธิภาพของวิธีการทดแทนค่าสูญหายในหลากหลายวิธีด้วยกัน ตั้งแต่การทดแทนค่าสูญหายโดยใช้สถิติในกลุ่มการวัดค่ากลางของข้อมูล ซึ่งเป็นที่นิยมใช้กันอย่างแพร่หลาย เช่น วิธี Mean Imputation (MEAN) เป็นการทดแทนค่าสูญหายโดยใช้วิธีการแทนที่ค่าสูญหายด้วยค่าเฉลี่ย วิธีการนี้ใช้หลักการของการหาค่าเฉลี่ย (Mean) ของข้อมูลค่าสังเกตหรือตัวแปรที่ทราบค่าและนำค่าเฉลี่ยที่ได้แทนที่ค่าสูญหาย วิธี Median Imputation (MED) เป็นการทดแทนค่าสูญหายโดยใช้การแทนที่ค่าสูญหายด้วย ค่ามัธยฐาน (Median) เป็นต้น นอกจากนี้จากการศึกษาของ Carol M. Musil (2002) นำเสนอแนวทางในการทดแทนค่าข้อมูลสูญหาย 5 วิธี ได้แก่ การลบตามรายการ (listwise deletion), การแทนที่ค่าสูญหายด้วยค่าเฉลี่ย (mean substitution), การแทนที่ค่าสูญหายด้วยสมการถดถอยอย่างง่าย (simple regression) การแทนที่ค่าสูญหายด้วยสมการถดถอย (regression imputation with error term) และการทดแทนค่าสูญหายด้วยค่าคาดหมายสูงสุด (Expectation Maximization algorithm) เปรียบเทียบการทดแทนค่าข้อมูลสูญหายด้วยวิธีดังกล่าว

ผลการวิเคราะห์พบว่าวิธีการทดแทนค่าสูญหายด้วยค่าคาดหมายสูงสุด (Expectation Maximization algorithm) มีประสิทธิภาพในการทดแทนค่าดีที่สุด แต่อย่างไรก็ตามจากการศึกษาวิธีการทดแทนค่าสูญหายข้างต้นให้ผลการวิจัยสอดคล้องกันว่าแม้จะเป็นวิธีที่ง่ายและเป็นที่รู้จักกันดี นิยมใช้กันอย่างแพร่หลายแต่มีข้อจำกัดในการนำไปใช้หลายประการ เช่น วิธี MEAN Imputation, วิธี Listwise deletion และวิธี Regression imputation เป็นวิธีการแทนที่ข้อมูลสูญหายเมื่อกลุ่มตัวอย่างมีขนาดใหญ่ และจำนวนข้อมูลสูญหายมีขนาดเล็กเท่านั้น จึงจะสามารถทดแทนค่าได้ดีภายใต้การสูญหายแบบสุ่มสมบูรณ์ (MCAR) เท่านั้น (Schlomer, Bauman et al. 2010, Vidotto, Vermunt et al. 2015, Chen and Åstebro 2016)

เมื่อพิจารณาการศึกษาความสัมพันธ์ระหว่างตัวแปรเพื่อลดมิติของตัวแปรในงานวิจัยเชิงสังคมศาสตร์นักวิจัยมักใช้เทคนิค principal component analysis (PCA) เป็นเทคนิคที่นิยมใช้กันอย่างแพร่หลายสำหรับการวิเคราะห์องค์ประกอบหลักอธิบายครั้งแรกโดย Pearson และพัฒนาโดย Hotelling มีงานวิจัยนำเทคนิค PCA ไปประยุกต์ใช้สำหรับการแทนที่ข้อมูลสูญหายโดยใช้ pcaMethods Package ในโปรแกรม R เรียกว่าเทคนิค probabilistic principal component analysis (PPCA) (Harshad Hegde 2019) โดยหลักการคือการรวมกันของเทคนิค principal component analysis (PCA) , Expectation-Maximization (EM) และ Maximum Likelihood Estimates (MLE) มีกระบวนการคือใช้ EM ในการวนซ้ำและใช้ MLE ในการทดแทนค่าข้อมูลสูญหายทำให้การทดแทนค่ามีความแม่นยำสูงขึ้น จากการศึกษาของ Harshad Hegde (2019) เปรียบเทียบเทคนิคการแทนที่ข้อมูลสูญหายสองวิธี ได้แก่ probabilistic principal component analysis (PPCA) และ multiple imputation using chained equations (MICE) โดยใช้กับข้อมูลผู้ป่วยจำนวน 41,543 คน ตัวแปรทางการแพทย์และทันตกรรมจำนวน 116 ตัวแปร ใช้ข้อมูลทุติยภูมิจากสถาบันการวิจัยซึ่งรวบรวมข้อมูลทางการแพทย์แบบบูรณาการและบันทึกสุขภาพฟันอิเล็กทรอนิกส์ (integrated medical and dental electronic health record) ทำการเปรียบเทียบสองเทคนิคด้วยการจำลองข้อมูลและใช้กับข้อมูลจริง ผลการวิเคราะห์ พบว่า PPCA ให้ผลลัพธ์ที่ดีกว่า MICE ซึ่งให้ค่าความถูกต้องโดยรวม (ความแม่นยำ: accuracy) และค่ากำลังสองเฉลี่ยความผิดพลาด: root mean square error (RMSE) ประมาณ 65% และ 0.29 ตามลำดับเมื่อเทียบกับ MICE ซึ่งให้ผลประมาณความแม่นยำ 38% พร้อม RMSE 0.83 โดยรวมการศึกษานี้แสดงให้เห็นว่าเทคนิค PPCA มีประสิทธิภาพของการแทนที่ค่าสูญหายที่สูงกว่าเทคนิค MICE ภายใต้การสูญหายในรูปแบบ MCAR

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

การศึกษางานวิจัยที่เกี่ยวข้องกับการประมาณค่าข้อมูลสูญหายได้รับการพัฒนาอย่างต่อเนื่อง เพื่อลดผลกระทบที่อาจจะเกิดขึ้นในการวิเคราะห์ผลการวิจัยซึ่งวิธีการประมาณค่าสูญหายที่นิยมและใช้กันอย่างแพร่หลายคือเทคนิค multiple imputation ซึ่งอธิบายครั้งแรกโดย Rubin (1987) และพัฒนาโดย Schafer & Graham (2002) เป็นเทคนิคในการประมาณค่าสูญหายที่มีความถูกต้องสูง เนื่องจากมีวิธีในการประมาณค่าสูญหายโดยใช้การคำนวณหลายครั้งทำการวนซ้ำเพื่อให้ได้ค่าที่ดีที่สุด จากผลการวิจัยของ Hughes, Heron et al. (2019) และ Nissen, Donatello et al. (2019) ชี้ให้เห็นว่าวิธีการประมาณค่าสูญหายด้วยวิธี Multiple Imputation ให้ผลลัพธ์ที่ถูกต้องแม่นยำและมีประสิทธิภาพทางสถิติมากกว่าการวิเคราะห์ข้อมูลโดยไม่คำนึงถึงค่าสูญหาย

ปัจจุบันมีแพคเกจในโปรแกรม R ที่เป็นที่รู้จักกันดีและเป็นที่นิยมใช้กันอย่างแพร่หลายในการประมาณค่าสูญหายไม่ว่าจะเป็น MICE Package, Amelia Package, lavaan Package, missForest Package และ pcaMethods Package เป็นต้น แพคเกจหนึ่งที่ได้รับความนิยมซึ่งใช้โดยทั่วไปในการประมาณค่าสูญหายคือการประมาณค่าสูญหายคือ MICE Package หรือการประมาณค่าสูญหายด้วยสมการลูกโซ่ (multiple imputation using chained equation analysis: MICE) เป็นที่นิยมสำหรับการประมาณค่าสูญหายเนื่องจากมีการวนซ้ำหลายครั้ง (multiple imputation) เพื่อให้ได้ค่าที่ดีที่สุด มีหลักการคือการประมาณค่าสูญหายโดยใช้สมการลูกโซ่ในการแทนที่ข้อมูลสูญหาย การแทนที่ข้อมูลสูญหายด้วยชุดข้อมูลของค่าที่เป็นไปได้มากกว่า 1 ค่า ซึ่งวิธีนี้จะมีประสิทธิภาพแม้ว่าจะมีการสูญหายมากกว่าร้อยละ 50 (Rubin 1978) ข้อมูลสูญหายถูกสร้างโมเดลขึ้นกับตัวแปรอื่นๆในชุดข้อมูลโดยต้องคำนึงถึงชนิดหรือประเภทของตัวแปรที่ถูกสร้างขึ้น โดย MICE Package สร้างการประมาณค่าโดยใช้วิธี predictive mean matching, Bayesian linear regression, logistic regression เป็นต้น (Van Buuren 2011)

วิธีหนึ่งที่ได้รับความนิยมในการประมาณค่าสูญหายด้วยเทคนิค multiple imputation คือ วิธี Multiple Imputation - Multivariate normal imputation using the EMB algorithm (MI-MVN) ซึ่งใช้ Amelia Package และ lavaan Package ในโปรแกรม R สำหรับการวิเคราะห์การประมาณค่าสูญหาย วิธีนี้เหมาะสำหรับข้อมูลที่มีการกระจาย Multivariate normal ถึงแม้ว่า MI-MVN ไม่ได้ถูกออกแบบสำหรับข้อมูลแบบเรียงอันดับ (ordinal missing data) แต่ในทางปฏิบัติถูกนำมาใช้กับข้อมูลแบบเรียงอันดับ (ordinal data) เพื่อจัดการข้อมูลสูญหาย Jia and Wu (2019) และ Enders (2015) พบว่า MI-MVN สามารถประมาณค่าได้อย่างมีประสิทธิภาพเมื่อแทนที่ค่าสูญหายประเภทมาตรวัดประมาณค่าของลิเคิร์ท (Likert) ที่ถูกรวมเข้ากับ

คะแนนมาตราส่วน นอกจากนี้วิธีการประมาณค่าสูญหายที่นิยมใช้และสามารถประมาณค่าสูญหายกับข้อมูลที่มีความซับซ้อนได้ นอกจากนี้วิธีที่ได้รับความนิยมเช่นกันคือวิธี Full information maximum likelihood (FIML) ซึ่งใช้ lavaan Package ในโปรแกรม R โดยมีหลักการพื้นฐานคือการแทนที่ข้อมูลที่สูญหายจากค่าพารามิเตอร์ประชากรโดยพิจารณาจากค่าฟังก์ชันความน่าจะเป็นสูงสุดในการประมาณค่าสูญหายซึ่งอาศัยข้อมูลที่ไม่สูญหาย ประมาณค่าพารามิเตอร์ภายใต้สมมติฐานการกระจายแบบปกติ (normality) และตัวแปรสังเกตได้อิสระต่อกัน (independent observations) เหมาะสำหรับการสูญหายแบบสุ่มสมบูรณ์ (MCAR) และการสูญหายแบบสุ่ม (MAR) เท่านั้น (Schlomer, Bauman et al. 2010, Zaninotto and Sacker 2017, Jia and Wu 2019)

แต่อย่างไรก็ตามการประมาณค่าสูญหายด้วยเทคนิค multiple imputation เช่น การประมาณค่าสูญหายด้วยสมการลูกโซ่ (MICE) และการสร้างแบบจำลองร่วม (Joint Modelling) เช่นการกำหนดให้ฟังก์ชันความหนาแน่นมีการแจกแจงปกติหลายตัวแปร (Multivariate normal) ที่กล่าวไว้ข้างต้นมีหลักการคือหาการกระจายที่เหมาะสมกับข้อมูลทั้งหมดก่อนแล้วค่อยสุ่มข้อมูลที่สูญหาย ในทางปฏิบัติโดยทั่วไปพารามิเตอร์ของแบบจำลองถูกประมาณค่าโดยใช้วิธี Expectation-Maximization (EM) และ maximize the likelihood function พบว่าวิธีการประมาณค่าสูญหายยังไม่เหมาะสมเนื่องจากการสร้างแบบจำลองร่วม (Joint Modelling) มีคุณสมบัติทางทฤษฎีที่เป็นประโยชน์ขณะที่ยังขาดความยืดหยุ่นสำหรับการประมวลผลในการวิเคราะห์ทางโปรแกรม เช่น เมื่อข้อมูลที่นำมาวิเคราะห์รวมชนิดของตัวแปรทั้งสองประเภทได้แก่ ตัวแปรต่อเนื่อง และตัวแปรไม่ต่อเนื่อง ฟังก์ชันแจกแจงปกติหลายตัวแปรไม่สามารถประมาณค่าในการสร้างแบบจำลองที่ซับซ้อนได้ (Van Buuren 2011, Singh 2016, Bertsimas, Pawlowski et al. 2018)

ขณะที่ในทางตรงกันข้ามการศึกษาแบบจำลองของ Van Buuren (2007) จะพบว่าการสร้างแบบจำลองร่วมการกำหนดค่าเงื่อนไขแบบสมบูรณ์ (Fully Conditional Specification: FCS) มีความหยืดหยุ่นโดยสามารถวิเคราะห์ได้ทั้งตัวแปรต่อเนื่องและตัวแปรไม่ต่อเนื่องวิธีการคือหาสมการทำนายค่าสูญหายของตัวแปรแรกจากตัวแปรที่สังเกตได้ในข้อมูล แล้ววนกลับไปทำนายตัวแปรตัวที่สองโดยอาศัยตัวแปรตัวแรกที่แทนค่าสูญหายแล้ว หลังจากนั้นทำซ้ำในตัวแปรตัวถัดไปจนตัวแปรสุดท้ายแล้ววนกลับตัวแปรตัวที่หนึ่งไปเรื่อย ๆ สามารถกำหนดจำนวนชุดข้อมูลและจำนวนรอบในการวนซ้ำได้โดยการวนซ้ำแต่ละครั้ง แต่มีข้อเสียคือวิธีการนี้ยังไม่มีทฤษฎีทางสถิติอย่างสมบูรณ์มาสนับสนุน (Bertsimas, Pawlowski et al. 2018, สันทัด พรประเสริฐมานิต 2562)

ตัวอย่างในการกำหนดเงื่อนไขขึ้นอยู่กับการถดถอยที่ได้รับความนิยม เช่น วิธี Predictive Mean Matching (PMM) หลักการคือเป็นการประมาณค่าสูญหายเป็นตัวอย่างที่สุ่มมาจากชุดของค่าสังเกตที่ใกล้กับการทำนายการถดถอย กล่าวคือสามารถหาค่าคาดหวังของข้อมูลสูญหายได้จากข้อมูลที่ทราบค่าจากนั้นทำการแทนที่หรือประมาณค่าข้อมูลที่สูญหายด้วยข้อมูลที่ทราบค่าที่มีค่าคาดหวังใกล้เคียงกับค่าคาดหวังของข้อมูลที่เกิดการสูญหายมากที่สุด (Van Buuren 2011, อุษณีย์ วงษ์อำมาตย์ 2555, วราพร ลิ่มชูเชื้อ 2556)

นอกจากนี้เมื่อพบความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างตัวแปร (non-linear relationship between the variables) การประมาณค่าสูญหายที่ได้กล่าวมาข้างต้นอาจให้การประมาณค่าที่ไม่เหมาะสม (Bertsimas, Pawlowski et al. 2018) วิธีการสำหรับสถิติไม่อิงพารามิเตอร์ (Non-parametric) ที่นิยมใช้กันคือวิธี K-Nearest Neighbor Imputation (KNN) มีหลักการ คือเป็นวิธีทดแทนค่าสูญหายด้วยค่าสังเกตที่ทราบค่าที่มีลักษณะคล้ายคลึงกันโดยวิธี KNN จะประมาณค่าโดยใช้ค่าใกล้สุดโดยพิจารณาเลือกหน่วยตัวอย่างจากชุดข้อมูลที่ทราบค่า ที่มีลักษณะคล้ายคลึงกับหน่วยตัวอย่างที่สูญหายมากที่สุดโดยพิจารณาเลือกหน่วยตัวอย่าง K ตัว จากข้อมูลจากนั้นแทนที่ค่าเฉลี่ยของหน่วยตัวอย่างที่คล้ายกัน โดยวิธี KNN พิจารณาค่าระยะห่างความคลึงกันจากระยะทางยุคลิค(ภัทฐิดา นิลภัทรฉัตร 2559) แต่วิธี KNN มีข้อเสียคืออาจเกิดข้อผิดพลาดในการคำนวณค่าหากข้อมูลมีจำนวนมากและมีความซับซ้อนมากไปและวิธีนี้เหมาะกับเฉพาะข้อมูลประเภทนามบัญญัติ (Nominal) เท่านั้น เช่น ข้อมูลเพศชาย หญิง อาชีพ เป็นต้น

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

นอกจากนี้จากงานวิจัยของ Jia and Wu (2019) ศึกษาการประมาณวิธี Multiple Imputation Random Forest (RF) เป็นวิธีการที่ไม่อิงพารามิเตอร์ (non-parametric) หลักการประมาณค่าข้อมูลสูญหายคือเป็นประมาณค่าที่มีสูญหายสุ่มโดยวิธีนี้สามารถใช้ได้กับข้อมูลต่อเนื่องและข้อมูลไม่ต่อเนื่องการประมาณค่าสูญหายด้วยวิธีนี้มีข้อดีคือไม่มีข้อตกลงเบื้องต้นที่เกี่ยวข้องกับประชากร ใช้เมื่อไม่ทราบการแจกแจงของประชากรจึงมีความหยืดหยุ่นกล่าวคือข้อมูลมีการกระจายแบบใดก็ได้จึงสามารถที่จะวิเคราะห์ข้อมูลที่มีการกระจายแบบไม่ปกติและความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างตัวแปรที่ไม่สามารถกำหนดพารามิเตอร์ได้ และจากการศึกษาของ Doov et al. (2014) เปรียบเทียบวิธีการประมาณ MI-RF กับ MI-LOGIT ในการวิเคราะห์สมการถดถอยโลจิสติก พบว่า MI-RF สามารถประมาณค่าข้อมูลสูญหายได้มีประสิทธิภาพ แต่อย่างไรก็ตามวิธี MI-RF ยังไม่ได้นำมาเปรียบเทียบในบริบทของข้อมูลโครงสร้างพหุระดับ (multilevel model)

จากงานวิจัยของ Doov et al. (2014) และJia and Wu (2019) ให้ผลวิจัยสอดคล้องกับงานวิจัยของ Nissen, Donatello et al. (2019) พบว่าวิธีการแทนที่ข้อมูลสูญหายด้วย RF มีประสิทธิภาพสูงแม้ข้อมูลมีความซับซ้อน และอัตราการสูญหายที่สูง โดย Nissen, Donatello et al. (2019) เปรียบเทียบวิธีการประมาณค่าสูญหายจำนวน 9 วิธี ได้แก่ ZERO, MEAN, minimum value (MIN), half mini- mum (1/2 MIN), Singular Value Decomposition (SVD), Probabilistic Principal Component Analysis (PPCA), Baye-sian Principal Component Analysis (BPCA), Random Forest (RF), และ K-Nearest Neighbors (KNN) ภายใต้การสูญหาย 7 รูปแบบ ได้แก่ MCAR, MAR, MNAR และอีก 4 กลไกคือ MCAR-MAR, MCAR-MNAR, MAR-MNAR และ MCAR-MAR-MNAR ด้วยอัตราการสูญหายที่แตกต่างกัน 4 ระดับ ได้แก่ ร้อยละ 5, 10, 20 และ 30 ตามลำดับ โดยการจำลองข้อมูลทางชีววิทยาด้วยการสร้างข้อมูล 2 ชุดแยกกัน ผลการวิจัยพบว่าวิธี RF สามารถประมาณค่าสูญหายได้ดีที่สุดในทุกกรณี และวิธี 1/2 MIN มีประสิทธิภาพต่ำในทุกกรณียกเว้นเมื่อข้อมูลมีการสูญหายแบบ MNAR วิธี 1/2 MIN ให้การประมาณค่าที่มีประสิทธิภาพมากที่สุด

ทั้งนี้จากการศึกษางานวิจัยของ Bertsimas, Pawlowski et al. (2018) ศึกษาการประมาณค่าสูญหายด้วยเทคนิค Opt.impute ซึ่งรวมกรอบแนวคิดการประมาณค่าสูญหายไว้ 3 วิธี คือ K- nearest neighbors, support vector machines, และ decision tree based method นำมาใช้สำหรับเทคนิค multiple Imputation เพื่อเพิ่มความหยืดหยุ่นและสะดวกในทางปฏิบัติมากขึ้นสามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง ผู้วิจัยนำเทคนิคดังกล่าวมาเปรียบเทียบกับ 5 วิธี ได้แก่ mean impute, K-nearest neighbors, iterative knn, Bayesian PCA, and predictive-mean matching เพื่อแสดงให้เห็นถึงประสิทธิภาพของวิธี opt.imputeโดยใช้ชุดข้อมูล 84 ชุดจาก UCI Machine Learning Repository ผลการวิเคราะห์แสดงให้เห็นว่า Opt.impute ให้ผลการประมาณค่าสูญหายที่มีประสิทธิภาพมากกว่าทุกวิธีที่นำมาเปรียบเทียบอย่างมีนัยสำคัญทางสถิติและวิธีการนี้ให้ความสำคัญกับการเพิ่มความหลากหลายในการใช้งานเพื่อลดข้อจำกัดที่ผ่านมาปรับปรุงวิธีการให้สอดคล้องกับการใช้งานจริง ทันสมัยในทุกสถานการณ์ที่อาจเกิดขึ้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วิธีการประมาณค่าสูญหายสำหรับข้อมูลพหุระดับ

เมื่อผู้วิจัยพิจารณางานวิจัยทางการศึกษาพบว่าธรรมชาติของตัวแปรมีลักษณะเป็นลำดับชั้น สังเกตได้จากการบริหารจัดการทางการศึกษามีลักษณะเป็นลำดับชั้นจากส่วนกลางลงไปยังเขตพื้นที่ โรงเรียน และนักเรียน ยกตัวอย่างเช่น ตัวแปรระดับนักเรียน ประกอบด้วย คุณลักษณะของนักเรียน สิ่งแวดล้อมทางบ้าน ผลสัมฤทธิ์ทางการเรียน เป็นต้น ตัวแปรระดับห้องเรียน ประกอบด้วย คุณลักษณะของครูผู้สอนสภาพการเรียนการสอน เป็นต้น และตัวแปรระดับโรงเรียน ประกอบด้วย คุณลักษณะผู้บริหารการจัดการเรียนการสอน นโยบายของโรงเรียน เป็นต้น โดยตัวแปรในระดับล่างหรือระดับที่สองจะได้รับอิทธิพลร่วมของตัวแปรที่อยู่ระดับสูงกว่าหรือระดับที่หนึ่ง เช่น ผลสัมฤทธิ์ทางการเรียนของนักเรียน ซึ่งเป็นตัวแปรระดับนักเรียนได้รับอิทธิพลจากตัวแปรระดับสูงกว่าคือตัวแปรห้องเรียน ได้แก่สภาพการเรียน การสอนของครู คุณลักษณะของครู เป็นต้น

จะเห็นได้ว่าลักษณะของข้อมูลทางการศึกษามีโครงสร้างแบบพหุระดับ (Multilevel data) มีความสัมพันธ์ระหว่างหน่วยข้อมูลซึ่งกันและกัน หากผู้วิจัยวิเคราะห์ข้อมูลพหุเคราะห์ระดับโดยไม่คำนึงถึงค่าสูญหาย (Missing data) อาจทำให้เกิดผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูลและส่งผลให้ผลการวิจัยคลาดเคลื่อนจากความเป็นจริง การประมาณค่าสูญหายจึงเป็นประเด็นที่หน่วยงานทางการศึกษาควรให้ความสำคัญเพื่อให้ได้สารสนเทศที่มีความถูกต้องน่าเชื่อถือและมีความเหมาะสมสอดคล้องกับสถานการณ์จริง (สังวร งัดกระโทก,2545; ศิริชัย กาญจนวาสี, 2548; สิวะโชติ ศรีสุทธิยากร, 2562)

เมื่อลักษณะของข้อมูลมีความสัมพันธ์กันหรือมีโครงสร้างหลายระดับ วิธีการทางสถิติจำนวนมากได้รับการพัฒนาขึ้นเพื่อจัดการกับข้อมูลที่สูญหายเนื่องจากวิธีในอดีตไม่สามารถวิเคราะห์ข้อมูลขนาดใหญ่ที่มีความซับซ้อนได้และไม่สามารถจัดการปฏิสัมพันธ์ระหว่างหน่วยข้อมูลที่มีความซับซ้อนได้จากการศึกษาวิธีการประมาณค่าข้อมูลสูญหายจะเห็นว่าวิธีการที่เหมาะสมในการประมาณค่าสูญหายสำหรับข้อมูลที่มีลักษณะพหุระดับ (Multilevel data) คือ วิธีแทนค่าหลายข้อมูล (Multiple Imputation: MI) เนื่องจากการวิเคราะห์พหุระดับจะวิเคราะห์หลายโมเดลโดยในแต่ละโมเดลจะมีจำนวนตัวแปรที่แตกต่างกัน การวิเคราะห์ด้วยด้วยการคำนวณแบบดั้งเดิม (Single Imputation) และการคำนวณความเป็นไปได้สูงสุดโดยตรง (Direct Maximum Likelihood; ML) ที่ได้กล่าวมาข้างต้นจึงไม่เหมาะสมเนื่องจากไม่สามารถนำโมเดลมาเปรียบเทียบกันได้

โดยการวิเคราะห์การประมาณค่าสูญหายด้วยวิธี Multiple Imputation คือ การประมาณค่าหรือการแทนที่ค่าสูญหายรูปแบบหนึ่งซึ่งมีวิธีการคือแทนค่าสูญหายด้วยค่าที่ทำนายได้บวกกับความผิดพลาดในการทำนายเพื่อไม่ให้ผลการวิเคราะห์ได้รับอิทธิพลจากความผิดพลาดในการทำนายที่สุ่มออกมาเพียงค่าเดียววิธีการนี้จึงทำการแทนค่าออกมาหลายครั้งหรือเรียกว่าการวนซ้ำ เพื่อให้ได้ข้อมูลหลายชุดและนำข้อมูลแต่ละชุดมาวิเคราะห์สถิติที่ต้องการทำการเลือกค่าที่ดีที่สุดในการวิเคราะห์ข้อมูลแบบพหุระดับ ใช้ MICE package ในโปรแกรม R จัดการกับข้อมูลสูญหาย หากข้อมูลมีการสูญหายไม่เกินร้อยละ 5 สามารถตัดข้อมูลทิ้งได้หรือใช้วิธีดั้งเดิมในการจัดการค่าสูญหายไม่ได้มีผลกระทบมากนัก แต่ถ้าใช้ Multiple Imputation จะผลการวิจัยที่เหมาะสมมากกว่า (สันทัด พรประเสริฐ, 2562; Audigier, 2018; Grund & el al, 2019)

จากการศึกษางานวิจัยของ Grund Simon (2016) Quartagno Matteo (2019) และสันทัด พรประเสริฐ (2019) ศึกษาวิธีวิธีการประมาณค่าสูญหายด้วยวิธี Multiple Imputation ให้ผลการวิจัยสอดคล้องกันว่า Multiple Imputation เป็นเครื่องมือสำหรับการประมาณค่าสูญหายและการอนุมานข้อมูลสูญหายได้มีประสิทธิภาพ แต่อย่างไรก็ตามวิธีการประมาณค่าสูญหายด้วย Multiple Imputation มีหลายวิธี ในการศึกษางานวิจัยครั้งนี้ผู้วิจัยเลือกวิธีการกำหนดค่าแบบเงื่อนไขสมบูรณ์ (Fully Conditional Specification; FCS) ซึ่งข้อดีคือผู้วิจัยสามารถกำหนดจำนวนชุดข้อมูลและจำนวนรอบในการวนซ้ำได้โดยการวนซ้ำแต่ละครั้งได้ตามลักษณะของข้อมูล ซึ่งสามารถขยายไปสู่ข้อมูลที่มีโครงสร้างหลายระดับ (Multilevel data)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

นอกจากนี้จากการศึกษาการประมาณค่าสูญหายด้วยวิธี Multiple Imputation Random Forest (MI-RF) พบว่าจากการศึกษางานวิจัยของ Jia and Wu (2019) Doov et al. (2014) และ Nissen, Donatello et al. (2019) ให้ผลการวิจัยสอดคล้องกัน วิธีการแทนที่ข้อมูลสูญหายด้วย Random Forest (RF) มีประสิทธิภาพสูงแม้ข้อมูลมีความซับซ้อน สามารถประมาณค่าสูญหายได้ดีที่สุดในทุกกรณี ขณะที่ผลการวิจัยของ Nissen, Donatello et al. (2019) แสดงให้เห็นว่าวิธี 1/2 MIN มีประสิทธิภาพต่ำในทุกกรณียกเว้นเมื่อข้อมูลมีการสูญหายแบบ MNAR วิธี 1/2 MIN สามารถประมาณค่าที่มีประสิทธิภาพมากกว่าวิธี Random Forest (RF) แต่อย่างไรก็ตามวิธี Random Forest (RF) ยังไม่ได้นำมาเปรียบเทียบในบริบทของข้อมูลโครงสร้างพหุระดับ (multilevel model) และจากการศึกษางานวิจัยของ Bertsimas, Pawlowski et al. (2018) ศึกษาการประมาณค่าสูญหายด้วยเทคนิค opt.impute วิธีการนี้ให้ความสำคัญกับการเพิ่มความหลากหลายในการใช้งาน

เพื่อลดข้อจำกัดที่ผ่านมา ปรับปรุงวิธีการให้สอดคล้องกับการใช้งานจริงซึ่งรวมกรอบแนวคิดการประมาณค่าสูญหายไว้ 3 วิธีคือ K- nearest neighbors, support vector machines, และ decision tree based method นำมาใช้สำหรับเทคนิค multiple Imputation เพื่อเพิ่มความหยืดหยุ่นและสะดวกในทางปฏิบัติมากขึ้นสามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง จากงานวิจัยของ Bertsimas, Pawlowski et al. (2018) แสดงให้เห็นว่าวิธี Opt.impute มีประสิทธิภาพอย่างมีนัยสำคัญทางสถิติ

จากการทบทวนเอกสารและงานวิจัยที่เกี่ยวข้องข้างต้น แสดงให้เห็นว่าวิธีการประมาณค่าสูญหายยังคงเป็นพื้นฐานที่มีความสำคัญสำหรับงานวิจัย แม้ในปัจจุบันจะมีวิธีการประมาณค่าข้อมูลสูญหายที่ได้รับความนิยมจำนวนมากแต่วิธีที่นิยมใช้ในปัจจุบันก็มีข้อจำกัดหลายประการยกตัวอย่างเช่นวิธีการสร้างแบบจำลองร่วม (Joint Modelling) ไม่มีประสิทธิภาพในการประมาณค่าเมื่อข้อมูลมีการละเมิดข้อตกลงเบื้องต้นเกี่ยวกับการแจกแจงปกติ และในทางปฏิบัติมักมีความผิดพลาดระหว่างการคำนวณเมทริกซ์ความแปรปรวนร่วม (Bertsimas, Pawlowski et al. 2018)

นอกจากนี้ Buuren (2012) สรุปข้อจำกัดของวิธี PMM กล่าวคือ “วิธี PMM ไม่มีประสิทธิภาพในการประมาณข้อมูลสูญหายเมื่อวิเคราะห์ในตัวอย่างขนาดเล็กและสามารถประมาณค่าข้อมูลสูญหายภายใต้ขอบเขตของข้อมูลที่สังเกตได้เท่านั้น” (น. 74) Kleinke (2018) นอกจากนี้ Bertsimas, Pawlowski et al. (2018) กล่าวว่า “วิธี pmm มีความน่าเชื่อถือในทางปฏิบัติแต่ไม่มีพื้นฐานทางทฤษฎีและไม่มีสูตรในการคำนวณที่ชัดเจนว่าสามารถปัญหาการปรับค่าสูญหายได้อย่างเหมาะสม”

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ในงานวิจัยนี้ผู้วิจัยศึกษาการประมาณค่าสูญหายเพื่อเพิ่มประสิทธิภาพการประมาณค่าสูญหาย ให้มีประสิทธิภาพมากยิ่งขึ้นโดยศึกษาเทคนิคการประมาณค่าสูญหายด้วยวิธี Opt.impute ซึ่งเป็นวิธีในการประมาณค่าสูญหายที่มีคุณภาพสูงจากงานวิจัยของ Bertsimas, Pawlowski et al. (2018) และเปรียบเทียบกับวิธีที่นิยมใช้กันในปัจจุบันได้แก่ Multiple imputation โดยการกำหนดค่าเงื่อนไขแบบสมบูรณ์ (Fully Conditional Specification: FCS), Random Forest (RF) ใช้เกณฑ์ค่าเฉลี่ยของค่า NRMSE, ค่าเฉลี่ยของค่า RB และค่าเฉลี่ยของค่า MCSE เพื่อเลือกวิธีที่ประมาณค่าของข้อมูลสูญหายที่มีประสิทธิภาพมากที่สุด (Black, Harel et al. 2011, Yin and Shi 2019)

ตารางที่ 2. 1

แสดงการเปรียบเทียบวิธีการประมาณค่าสูญหาย

วิธีทดแทนค่าข้อมูลสูญหาย	ข้อดี	ข้อเสีย
วิธีการทดแทนค่าข้อมูลสูญหายด้วยวิธีแบบดั้งเดิม		
Single Imputation (Mean/Mode substitution)	เป็นวิธีที่ง่าย สะดวกและรวดเร็ว จำนวนตัวอย่างไม่ลดลง	ข้อมูลไม่มีความหลากหลาย เนื่องจากแทนค่าด้วยค่าเดียวกันทั้งหมด ทำให้ผลการวิเคราะห์มีความเอนเอียงสูงอย่างรุนแรง (Severely biased) และไม่เหมาะสมที่จะนำไปวิเคราะห์หาความสัมพันธ์ เนื่องจากการแทนที่ค่าสูญหายไม่ได้คำนึกถึงความสัมพันธ์ระหว่างตัวแปร
Mean substitution	เป็นวิธีที่ง่าย สะดวกและรวดเร็ว จำนวนตัวอย่างไม่ลดลง	ลดความแปรปรวนระหว่างตัวแปร ส่งผลให้ผลการวิเคราะห์มีความเอนเอียง
Listwise deletion	เป็นวิธีที่ง่าย สะดวก และรวดเร็ว ในกรณีกลุ่มตัวอย่างใช้แบบสอบถามชุดเดียวกัน	ผลมีความเอนเอียงสูงอย่างรุนแรง อำนาจในการวิเคราะห์จะลดลงเนื่องจากจำนวนตัวอย่างลดลง
Pairwise deletion	เป็นวิธีที่ง่าย การวิเคราะห์ข้อมูลใช้ทุกหน่วยตัวอย่างในการวิเคราะห์ข้อมูล จำนวนตัวอย่างไม่ถูกตัดออก	ไม่สามารถวิเคราะห์ในเชิงเปรียบเทียบได้ เนื่องจากกลุ่มตัวอย่างมีจำนวนไม่เท่ากัน และให้ข้อมูลที่แตกต่างกัน
Hot-deck imputation	เป็นวิธีที่ง่ายและรวดเร็ว เหมาะกับข้อมูลที่มีขนาดใหญ่	ไม่คำนึงถึงความสัมพันธ์ระหว่างตัวแปร ทำให้ผลลัพธ์มีความเอนเอียง (biased)
K-nearest neighbor imputation	เป็นวิธีที่ง่าย ใช้กับข้อมูลที่มีลักษณะซับซ้อน	ใช้เวลานานกว่าวิธี Hot-deck imputation
Regression imputation	เป็นวิธีที่ง่าย สามารถใช้ข้อมูลที่เก็บรวบรวมได้ทุกหน่วยข้อมูล	ค่าที่ได้จากการพยากรณ์ไม่มีความแม่นยำ หากข้อมูลไม่อยู่ภายใต้สถานการณ์การสุ่มอย่างสมบูรณ์
Iterative robust model-based imputation	เป็นวิธีที่ง่าย สามารถใช้ข้อมูลที่เก็บรวบรวมได้ทุกหน่วยข้อมูล	ค่าที่ได้จากการพยากรณ์ไม่มีความแม่นยำ หากข้อมูลไม่อยู่ภายใต้สถานการณ์การสุ่มอย่างสมบูรณ์ (MCAR)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วิธีทดแทนค่าข้อมูลสูญหาย	ข้อดี	ข้อเสีย
วิธีการทดแทนค่าข้อมูลสูญหายที่นิยมใช้และมีความซับซ้อนมากขึ้น		
Maximum likelihood estimation	การวิเคราะห์ข้อมูลใช้ทุกหน่วยตัวอย่างในการวิเคราะห์ข้อมูล	ข้อจำกัดจำเป็นต้องใช้วิธีการเชิงตัวเลขเพื่อค้นหาฟังกชันโอกาสสูงสุด
Full information maximum likelihood	มีข้อได้เปรียบวิธี EM คือ มีขั้นตอนในการดำเนินการในขั้นตอนเดียว ได้ผลลัพธ์ที่ถูกต้อง มีประสิทธิภาพในการประมาณค่า ไม่เอนเอียง (unbiased)	การสร้างแบบจำลองสมการโครงสร้างมีความยุ่งยาก และการพิจารณาปรับสมการเชิงโครงสร้างอาจทำให้เกิดความคลาดเคลื่อน
Expectation Maximization algorithm	มีประสิทธิภาพในการประมาณค่า (efficient) ไม่เอนเอียง (unbiased) และมีประโยชน์อย่างมากในการวิเคราะห์เชิงสำรวจและการคำนวณความสอดคล้องภายใน ซึ่งไม่จำเป็นต้องทดสอบสมมติฐาน	ไม่แสดงค่าความคลาดเคลื่อนมาตรฐาน และช่วงความเชื่อมั่น ทำให้จำเป็นต้องมีขั้นตอนการทดสอบเพิ่มเติม ไม่สามารถวิเคราะห์สถิติเชิงอนุมานได้
K-Nearest Neighbor Imputation (KNN)	มีประสิทธิภาพเมื่อใช้กับข้อมูลประเภทนามบัญญัติ สามารถนำไปสร้างโมเดลที่มีประสิทธิภาพได้	อาจเกิดข้อผิดพลาดในการคำนวณค่าหากข้อมูลมีจำนวนมาก และวิธีนี้เหมาะกับเฉพาะข้อมูลประเภทนามบัญญัติ (Nominal) เท่านั้น เช่น ข้อมูลเพศชาย หญิง อาชีพ
Multiple Imputation Multivariate normal)MI-MVN(	มีความถูกต้องสูงเนื่องจากใช้การคำนวณหลายครั้ง เพื่อให้ได้ค่าที่ดีที่สุด และยังคำนึงถึงความสัมพันธ์ระหว่างกลุ่ม	กำหนดฟังก์ชันความหนาแน่นมีการแจกแจงปกติหลายตัวแปร (Multivariate normal)ทำให้ขาดความยืดหยุ่นสำหรับการประมวลผลในการวิเคราะห์ทางโปรแกรม
Fully Conditional Specification (FCS)	มีความหยืดหยุ่น สามารถวิเคราะห์ได้ทั้งตัวแปรต่อเนื่องและตัวแปรไม่ต่อเนื่อง วิธีการคือหาสมการทำนายค่าสูญหายของตัวแปรแรกจากตัวแปรที่สังเกตได้ในข้อมูล มีความถูกต้องสูงเนื่องจากใช้การคำนวณหลายครั้งเพื่อให้ได้ค่าที่ดีที่สุด	ไม่มีทฤษฎีทางสถิติอย่างสมบูรณ์มาสนับสนุน แต่สามารถประมาณค่าได้มีประสิทธิภาพ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วิธีทดแทนค่าข้อมูลสูญหาย	ข้อดี	ข้อเสีย
Random forest (RF)	มีความหยืดหยุ่นในการใช้งาน เนื่องจากไม่จำเป็นต้องทราบการแจกแจงของประชากร โดยวิธีนี้สามารถใช้ได้ทั้งข้อมูลต่อเนื่อง และข้อมูลไม่ต่อเนื่อง ผลลัพธ์ที่ได้จากการวิเคราะห์จะมีความถูกต้องแม่นยำและมีประสิทธิภาพที่สูง	การกำหนดจำนวน decision tree ไม่เพียงพอจะทำให้เกิด bias (จำนวน Tree มาก ค่า variance ก็ยิ่งลดลง)
Optimal Imputation (Opt. impute)	ให้ผลการประมาณค่าสูญหายที่มีประสิทธิภาพสูงแม้อัตราการสูญหายมาก มีความหยืดหยุ่น สามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง ลดข้อจำกัดในอดีตที่ผ่านมา	มีความซับซ้อนในการคำนวณและยาก

ในการวิจัยครั้งนี้ผู้วิจัยเปรียบเทียบวิธีทดแทนค่าสูญหายสำหรับการประมาณพารามิเตอร์ของ β_1เมื่อ β_1 แทนค่าเฉลี่ยของอิทธิพลภูมิหลังทางครอบครัวของนักเรียนในแต่ละสังกัด โดยเปรียบเทียบจำนวน 3 วิธี ได้แก่ Fully Conditional Specification (FCS), Random forest Imputation (RF) และ Optimal Imputation (Opt. impute) ประกอบด้วยวิธีย่อย Opt.knn, Opt.svm, Opt.tree และ Opt.cv ซึ่งมีรายละเอียดดังต่อไปนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

หลักการของวิธี Multiple Imputation (MI)

วิธี Multiple Imputation (MI) มีวิธีการประมาณค่าที่คล้ายกับวิธีใน Single Imputation แต่มีการดำเนินการมากกว่า 1 ครั้ง มีหลักการ คือ การทดแทนที่ข้อมูลสูญหายด้วยชุดข้อมูลของค่าที่เป็นไปได้มากกว่า 1 ค่า (D>1) ซึ่งวิธีนี้จะมีประสิทธิภาพมากขึ้นเมื่อมีการสร้างชุดข้อมูลที่สมบูรณ์ตั้งแต่ 5 ถึง 10 ชุด การแทนที่ค่าสูญหายมีความถูกต้องสูงแม้ว่าจะมีการสูญหายมากกว่าร้อยละ 50 เนื่องจากใช้การคำนวณหลายครั้ง เพื่อให้ได้ค่าที่ดีที่สุด (Rubin 1978) มีวิธีการดำเนินการ ดังนี้ กำหนดให้ ตัวแปรอิสระ X แบ่งออกเป็น 2 ส่วน ได้แก่ ตัวแปรอิสระที่ทราบค่า (X_{obs}) และตัวแปรอิสระที่ไม่ทราบค่า (X_{mis}) เมื่อสมมติให้ X มีการแจกแจงแบบมัลติโนเมียล (Multinomial distribution) ซึ่งพารามิเตอร์หาได้จาก Predictive Distribution โดยการจำลองค่าพารามิเตอร์จากค่าสังเกตที่ทราบค่า แล้วประมาณค่าสังเกตที่สูญหายทั้งหมด D ชุด ทำการวิเคราะห์ข้อมูล

ชุดที่สมบูรณ์และนำค่าสัมประสิทธิ์ถดถอยทั้ง D ชุด ที่ได้ไปหาค่าเฉลี่ยแทนที่ค่าสูญหายดังกล่าว ซึ่งมีขั้นตอนในการแทนที่ค่าสูญหาย ดังนี้

สมมติให้ X มีการแจกแจง $p(X \mid \theta)$ เมื่อ θ เป็นพารามิเตอร์ตัวแบบทั้งหมดของตัวแบบ Predictive Distribution หาได้จาก

$$p(X_{mis} \mid X_{obs}) = \int p(X_{mis} \mid X_{obs})\, d\theta$$
$$= \int p(X_{mis} \mid X_{obs}, \theta)\; p(\theta \mid X_{obs}) d\theta$$

จะเห็นว่าเราสามารถประมาณค่าสูญหาย จากการจำลองค่าพารามิเตอร์จากค่าสังเกตที่มีอยู่ ซึ่งมีการแจกแจงจากการเรียนรู้ $p(\theta \mid X_{obs})$แล้วจำลองค่าสูญหายจากการแจกแจงโดยการเรียนรู้ที่มีเงื่อนไข $p(X_{mis} \mid X_{obs}, \theta)$ ดังนั้นการประมาณค่าสูญหายด้วยวิธี MI ในขั้นตอนของการสร้างข้อมูล D ชุดนั้น ทำได้โดย การกำหนดค่าเริ่มต้นสำหรับการประมาณค่าสูญหายด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood: ML) ซึ่งแบ่งออกเป็น 2 ขั้นตอน ดังนี้

1. Imputation Step: $X_{mis}^{(r)} \sim p(X_{mis} \mid X_{obs}, \theta^{(r-1)})$

2. Posterior Step: $\theta^{(r)} \sim p(\theta \mid X_{obs}, X_{mis}^{(r)})$

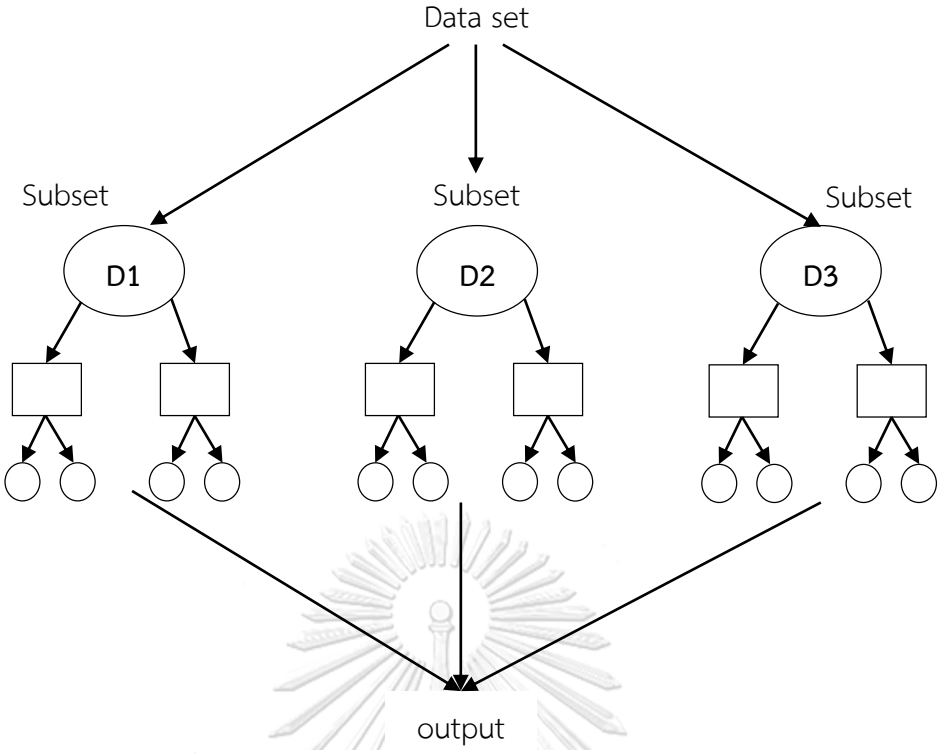
โดยที่ $(X_{obs}, X_{mis}^{(r)})$; r = 1,2,3,..... ทำวนซ้ำจนค่าลู่เข้าการแจกแจง $p(X_{mis}, \theta \mid X_{obs})$ จะได้ค่าประมาณค่าสังเกตที่สูญหายค่าใหม่ จากนั้นทำการวนซ้ำตามขั้นตอนดังกล่าวข้างต้นไปจนครบ D ครั้ง ก็จะได้ค่าประมาณสูญหายทั้งหมด D ชุด จากนั้นนำไปหาค่าเฉลี่ยของค่าประมาณที่ได้เพื่อนำค่าที่ได้ไปใช้ในการประมาณค่าสูญหาย การประมาณค่าสูญหายด้วยวิธี Multiple Imputation (MI) ข้างต้นมีหลายวิธีด้วยกัน เช่น สร้างการประมาณค่าโดยใช้วิธี predictive mean matching , Bayesian linear regression, logistic

การวิจัยครั้งนี้ ใช้การกำหนดค่าเงื่อนไขแบบสมบูรณ์ (Fully Conditional Specification: FCS) มีความหยืดหยุ่นโดยสามารถวิเคราะห์ได้ทั้งตัวแปรต่อเนื่องและตัวแปรไม่ต่อเนื่องวิธีการคือหาสมการทำนายค่าสูญหายของตัวแปรแรกจากตัวแปรที่สังเกตได้ในข้อมูล แล้ววนกลับไปทำนายตัวแปรตัวที่สองโดยอาศัยตัวแปรตัวแรกที่แทนค่าสูญหายแล้ว หลังจากนั้นทำซ้ำในตัวแปรตัวถัดไปจนตัวแปรสุดท้ายแล้ววนกลับตัวแปรตัวที่หนึ่งไปเรื่อย ๆ โดยผู้วิเคราะห์สามารถกำหนดจำนวนชุดข้อมูลและจำนวนรอบในการวนซ้ำได้โดยการวนซ้ำแต่ละครั้ง แต่มีข้อเสียคือวิธีการนี้ยังไม่มีทฤษฎีทางสถิติอย่างสมบูรณ์มาสนับสนุน (Bertsimas, Pawlowski et al. 2018, สันทัด พรประเสริฐมานิต 2562)

หลักการของการประมาณค่าสูญหายด้วยวิธี Random Forest Imputation

วิธี Random forest ถูกพัฒนาโดย Leo Breiman ในปัจจุบันการประมาณค่าข้อมูลสูญหายด้วยวิธี Random forest Imputation (RF) เป็นวิธีที่ได้รับความนิยมและเป็นแนวทางที่น่าสนใจสำหรับการประมาณค่าสูญหายของข้อมูล เป็นการประมาณค่าในสถิติที่ไม่อิงพารามิเตอร์ (nonparametric) ข้อมูลในประมาณค่ามีลักษณะแบบสุ่ม (random) โดยวิธีนี้สามารถใช้ได้ทั้งข้อมูลต่อเนื่อง (continuous data) และข้อมูลไม่ต่อเนื่อง (categorical data) มีความหยืดหยุ่นในการใช้งานเนื่องจากไม่จำเป็นต้องทราบการแจกแจงของประชากร ข้อมูลมีการกระจายแบบใดก็ได้ สามารถวิเคราะห์ข้อมูลที่มีการกระจายแบบไม่ปกติและความสัมพันธ์ที่ไม่เป็นเชิงเส้นระหว่างตัวแปรที่ไม่สามารถกำหนดพารามิเตอร์ได้ ซึ่งในปี ค.ศ. 2001 Breiman อธิบายแนวคิดของวิธี Random forest คือ การรวมกันของทฤษฎี Bagging การจำแนกประเภท (classification) การวิเคราะห์ถดถอย (regression) แผนภาพต้นไม้ตัดสินใจ (decision tree) และการสุ่มพื้นที่ย่อย (random forest algorithm) โดยมีหลักการคือการสร้างโมเดลจากการใช้แผนภาพต้นไม้ตัดสินใจ (decision tree) หลายโมเดลในการวิเคราะห์แต่ละครั้ง แสดงดังรูปภาพที่ 2.1.1 ซึ่งการวิเคราะห์แผนภาพต้นไม้ตัดสินใจ (decision tree) เป็นโมเดลทางคณิตศาสตร์และเป็นวิธีการในการทำเหมืองข้อมูลด้วยการสร้างโมเดลในลักษณะแผนภาพเป็นลำดับชั้นงานวิจัยส่วนมากจะใช้แผนภาพต้นไม้เพื่อสร้างเป็นโมเดลที่ช่วยทำนายค่าข้อมูลสูญหาย (Shichao, 2005) มีกระบวนการโดยในแต่ละโมเดลจะได้รับชุดข้อมูล (data set) ที่แตกต่างกันซึ่งเป็นซับเซต (subset) ของชุดข้อมูล (data set) ทั้งหมด และขณะการทำนายค่าข้อมูลสูญหาย (prediction) จะกำหนดให้แต่ละแผนภาพต้นไม้ตัดสินใจ (decision tree) ทำนายค่าข้อมูลสูญหาย (prediction) ของแต่ละโมเด]โดยแต่ละโมเดลเป็นอิสระต่อกัน หลังจากนั้นคำนวณผลการทำนายค่าข้อมูลสูญหาย (prediction) ด้วยการ vote output ซึ่งผลลัพธ์ที่ถูกเลือกโดยแผนภาพต้นไม้ตัดสินใจ (decision tree) มากที่สุด แสดงดังรูปที่ 2.1.1 ดังนั้นผลลัพธ์ที่ได้จากการวิเคราะห์จะมีความถูกต้อง แม่นยำ และมีประสิทธิภาพที่สูง (Doov et al., 2014; Witchapong Daroontham., 2018; Deng Wei and et, 2019)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปที่ 2.1 1 หลักการของวิธี Random forest Imputation

หลักการของการประมาณค่าสูญหายด้วยวิธี Optimal Imputation (Opt. impute)

จากการศึกษางานวิจัยของ Bertsimas, Pawlowski et al. (2018) แสดงให้เห็นว่า วิธี Opt.impute (Optimal Imputation) มีประสิทธิภาพสูงในการประมาณค่าสูญหาย ช่วยเพิ่มความหลากหลายในการ ใช้งาน สามารถลดข้อจำกัดในอดีตที่ผ่านมา มีความยืดหยุ่นและสะดวกในทางปฏิบัติมากขึ้น สามารถใช้ได้กับตัวแปรประเภทต่อเนื่องและไม่ต่อเนื่อง ซึ่งวิธีนี้ใช้ Julia โดยใช้ glmnet Package และ iai Package ในโปรแกรม R สำหรับการวิเคราะห์การประมาณค่าสูญหาย มีหลักการคือการรวมแนวคิดการประมาณค่าสูญหายไว้ 3 วิธีคือ K- nearest neighbors (knn), support vector machines (svm) และ decision tree (tree) โดยสร้างการแทนที่ค่าสูญหายหลายครั้งเช่นเดียวกันกับวิธี Multiple Imputation ซึ่งจะช่วยประมาณค่าได้ดีกว่าการใส่แบบจำลองกับชุดข้อมูลเพียงครั้งเดียว โดยมีขั้นตอนการกำหนดค่าเริ่มต้นสำหรับการประมาณค่าสูญหายด้วยวิธี Opt.impute แบ่งออกเป็น 2 ขั้นตอน ดังนี้ ขั้นตอนแรกเริ่มจากวิเคราะห์ Opt.knn ต่อมาวิเคราะห์ Opt.svm และ Opt.tree ดำเนินการวิเคราะห์ชุดข้อมูลแต่ละชุดแยกกัน ขั้นตอนที่สองเพื่อเพิ่มประสิทธิภาพที่สูงขึ้นจะรวมผลลัพธ์เข้าด้วยกัน และส่งผลลัพธ์ (output) สุดท้ายเพื่อให้ชุดข้อมูลที่สมบูรณ์ที่สุด

เกณฑ์ที่ในการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย

การวิจัยครั้งนี้มีวัตถุประสงค์หลักเพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ และนำวิธีทดแทนค่าสูญหายที่มีประสิทธิภาพไปประยุกต์ใช้กับข้อมูลจริงเพื่อศึกษาปัจจัยที่ส่งผลต่อความเหลื่อมล้ำทางการศึกษาโดยใช้โมเดลพหุระดับ โดยในการวิจัยครั้งนี้ศึกษาความเหลื่อมล้ำทางการศึกษาโดยวิเคราะห์สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน โดยเกณฑ์ที่ใช้ในการเปรียบเทียบความคลาดเคลื่อนของวิธีการประมาณค่าสูญหายว่าการประมาณค่าสูญหายด้วยวิธีการใดดีกว่า มีความแม่นยำมากกว่า ในการวิจัยครั้งนี้ผู้วิจัยใช้ 3 เกณฑ์ ได้แก่ (1) ค่าเฉลี่ยของรากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (NRMSE) (2) ค่าเฉลี่ยของความเอนเอียงสัมพัทธ์ (RB) และ (3) ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (MCSE) มีรายละเอียด ดังนี้

การหาค่าเฉลี่ยของ NRMSE, RB, และ MCSE

การวิจัยครั้งนี้หาค่าเฉลี่ยของ NRMSE, RB, และ MCSE ของแต่ละ วิธีทดแทนค่าสูญหายที่มีการทำซ้ำ 100 รอบ เป็นเกณฑ์ในการเปรียบเทียบประสิทธิของวิธีทดการทดแทนค่าสูญหายเพื่อประมาณค่าพารามิเตอร์ β_1 ในแต่ละสถานการณ์ที่ได้จากการจำลองข้อมูล มีรายละเอียดดังต่อไปนี้

เมื่อกำหนดให้

ℓ	แทน	กรณีที่ ℓ	โดยที่ $\ell = 1,2,3$
i	แทน	โรงเรียนที่ i	โดยที่ $i = 1,2,3,\ldots,N_{1\ell}$
j	แทน	สังกัดที่ j	โดยที่ $j = 1,2,3,\ldots,N_{2\ell}$
N	แทน	การจำลองรอบที่ N	โดยที่ $N = 1,2,3,\ldots,100$
β_1	แทน	ค่าสัมประสิทธ์ความถดถอยแบบสุ่มของสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาที่มีผลต่อผลสัมฤทธิ์ทางการเรียนของนักเรียน ระดับโรงเรียน	
θ_j	แทน	ค่าประมาณพารามิเตอร์ β_1 จากข้อมูลจริงของสังกัดที่ j	

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$\hat{\theta}_{ij}$ แทน ค่าประมาณพารามิเตอร์ β_1 ของโรงเรียนที่ i ในสังกัดที่ j

$\overline{\hat{\theta}_j}$ แทน ค่าเฉลี่ยของการประมาณพารามิเตอร์ β_1 จำนวน $N_{2\ell}$ สังกัด

$N_{1\ell}$ แทน ขนาดของตัวอย่างระดับที่หนึ่งหรือจำนวนโรงเรียนของกรณีศึกษาที่ ℓ โดยที่ $N_{1\ell} = 1{,}000, N_{1\ell} = 2{,}000, N_{1\ell} = 3{,}000$

$N_{2\ell}$ แทน ขนาดของตัวอย่างระดับที่สองหรือจำนวนสังกัดของกรณีศึกษาที่ ℓ โดยที่ $N_{2\ell} = 40, N_{2\ell} = 50, N_{2\ell} = 60$

$var\ (\theta_j)$ แทน ค่าความแปรปรวนของค่าประมาณพารามิเตอร์ β_1 จากข้อมูลจริงของสังกัดที่ j

$mean(\ \hat{\theta}j)$ แทน ค่าเฉลี่ยของค่าประมาณพารามิเตอร์ β_1 จำนวน $N_{2\ell}$ สังกัดที่ได้จากการจำลองจำนวน 100 รอบ

ค่าเฉลี่ยของรากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (NRMSE) ใช้เพื่อตรวจสอบความถูกต้องในการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย เมื่อ mean คือค่าเฉลี่ยตามการจำลอง 100 รอบ เมื่อค่าเฉลี่ยของ NRMSE ที่มีค่าน้อยจะแสดงถึงวิธีทดแทนค่าสูญหายที่มีความถูกต้องและมีประสิทธิภาพสูง สามารถคำนวณได้จากสูตรดังนี้

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

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ค่าเฉลี่ยของค่าความเอนเอียงสัมพัทธ์ (RB) ใช้เพื่อตรวจสอบความแม่นยำความเที่ยงตรง (preciseness) และความเอนเอียงของวิธีทดแทนค่าสูญหาย โดยค่าเฉลี่ยของ RB ที่มีค่าน้อยจะแสดงถึงวิธีทดแทนค่าสูญหายที่มีคุณสมบัติด้านความถูกต้องและแม่นยำ สามารถคำนวณได้จากสูตรดังนี้

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

ค่าเฉลี่ยของค่าคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (MCSE) ใช้เพื่อตรวจสอบความคงเส้นคงวา (consistency) ของวิธีทดแทนค่าสูญหายในการจำลองข้อมูล 100 รอบ โดยค่าเฉลี่ยของ MCSE ที่มีค่าน้อยจะแสดงถึงวิธีทดแทนค่าสูญหายที่มีคุณสมบัติความคงเส้นคงวาสูง สามารถคำนวณได้จากสูตรดังนี้ $MCSE\left(\overline{\hat{\theta}_j}\right) = \frac{\sqrt{\sum_{j=1}^{N_2} (\hat{\theta}_{(j)} - mean(\ \hat{\theta}_{j(j)}))2}}{100}$

ตอนที่ 2 ความเหลื่อมล้ำทางการศึกษา

ผู้วิจัยนำเสนอข้อมูลที่เกี่ยวข้องกับความเหลื่อมล้ำทางการศึกษาซึ่งแบ่งหัวข้อการนำเสนอดังนี้ ความหมายของความเหลื่อมล้ำทางการศึกษา ความสำคัญของความเหลื่อมล้ำทางการศึกษาในประเทศไทย และปัจจัยที่สะท้อนความเหลื่อมล้ำทางการศึกษา

ความหมายของความเหลื่อมล้ำทางการศึกษา

รัฐธรรมนูญแห่งราชอาณาจักรไทย พุทธศักราช 2560 มาตรา 54 กำหนดไว้ว่า "รัฐต้องดำเนินการให้เด็กทุกคนได้รับการศึกษาเป็นเวลาสิบสองปี ตั้งแต่ก่อนวัยเรียนจนจบการศึกษาภาคบังคับอย่างมีคุณภาพโดยไม่เก็บค่าใช้จ่าย โดยรัฐต้องดำเนินการให้ผู้ขาดแคลนทุนทรัพย์ได้รับการสนับสนุนค่าใช้จ่ายในการศึกษาตามความถนัดของตน โดยให้จัดตั้งกองทุนเพื่อใช้ในการช่วยเหลือผู้ขาดแคลนทุนทรัพย์เพื่อลดความเหลื่อมล้ำ" จากที่กล่าวมานั้นแสดงให้เห็นว่าการศึกษาเป็นปัจจัยขั้นพื้นฐานที่ประชาชนทุกคนมีสิทธิในการเข้าถึงและควรมีสิทธิที่จะได้รับการศึกษาอย่างเสมอภาคและเท่าเทียมกัน แต่อย่างไรก็ตาม จากการศึกษางานวิจัยตั้งแต่อดีตจนถึงปัจจุบันพบว่าระบบการศึกษาของประเทศไทยยังพบปัญหาที่เกี่ยวข้องกับความไม่เท่าเทียมทางการศึกษาการได้รับโอกาสทางการศึกษาที่ไม่เท่าเทียมกัน นักเรียนที่อยู่ในต่างจังหวัดได้รับโอกาสที่น้อยกว่านักเรียนที่อยู่ในกรุงเทพและปริมณฑลหรือนักเรียนที่ครอบครัวมีฐานะทางเศรษฐกิจต่ำย่อมได้รับโอกาสที่น้อยกว่านักเรียนที่ครอบครัวมีฐานะทางเศรฐกิจที่ดี (Jo Blanden and Stephen Machin 2004, Woessmann 2014, Weiss 2017, Alberto Alesina 2018, Marks and Pokropek 2019, กองทุนเพื่อความเสมอภาคทางการศึกษา 2563) เป็นต้น หากเราพิจารณาถึงพื้นฐานที่ก่อให้เกิดความแตกต่างสามารถแบ่งได้ 3 ปัจจัย ได้แก่ ปัจจัยส่วนบุคคล ปัจจัยในด้านครอบครัว และปัจจัยด้านความแตกต่างของโรงเรียนที่นักเรียนศึกษาอยู่ จะพบว่าปัจจัยในด้านต่าง ๆ เหล่านี้ต่างก็มีส่วนทำให้นักเรียนแต่ละคนเกิดความแตกต่างในด้านความรู้ความเข้าใจขั้นพื้นฐานที่ไม่เท่าเทียมกันสะท้อนได้จากผลสัมฤทธิ์ทางการเรียนหรือคะแนนสอบที่แตกต่างกัน (Woessmann 2014, สถาบันทดสอบทางการศึกษาแห่งชาติ 2560) ซึ่งความไม่เท่าเทียมกันหรือการได้รับโอกาสที่ไม่เท่ากันดังสถานการณ์ที่กล่าวมาข้างต้นเรียกว่า "ความเหลื่อมล้ำทางการศึกษา (educational inequality)"

ความสำคัญของความเหลื่อมล้ำทางการศึกษาในประเทศไทย

ประเทศไทยจำเป็นต้องให้ความสำคัญและส่งเสริมการศึกษาของเด็กและเยาวชนให้ได้รับโอกาสด้านการศึกษาที่เท่าเทียมกันเนื่องจากปัจจัยหนึ่งที่สำคัญและเป็นรากฐานต่อการพัฒนาประเทศคือปัจจัยด้านการศึกษา จะเห็นได้ว่าการศึกษามีความสำคัญต่อการพัฒนาทรัพยากรมนุษย์

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

และเป็นกลไกที่สำคัญในการพัฒนาเศรษฐกิจและสังคม ตั้งแต่อดีตจนถึงปัจจุบันการศึกษาของประเทศไทยได้รับการพัฒนาอย่างยาวนานซึ่งสะท้อนให้เห็นถึงวิวัฒนาการการเรียนการสอนในประเทศไทยและได้รับการปฏิรูปอยู่หลายครั้งเพื่อให้ประชาชนได้เข้าถึงการศึกษาให้มากที่สุด (สำนักงานเลขาธิการสภาการศึกษา กระทรวงศึกษาธิการ 2561) อย่างไรก็ตามแม้ว่าระบบการศึกษาของประเทศจะมีแนวทางในการปรับปรุงและพัฒนามาอย่างต่อเนื่องและมีกฎหมายที่ค่อนข้างชัดเจนไม่ว่าจะผ่านมากี่ยุคกี่สมัย แต่กลับพบว่าการแก้ไขปัญหาด้านการศึกษาของประเทศไทยนั้นยังไม่เป็นไปตามเจตนารมณ์ที่กำหนดไว้ในกฎหมายหลายอย่างและไม่ได้นำไปสู่การปฏิบัติได้ไม่ประสบความสำเร็จเท่าที่ควรโดยเฉพาะในประเด็นของความเหลื่อมล้ำทางการศึกษา เห็นได้จากอัตราการเข้าศึกษาต่อระดับอุดมศึกษาของเด็กนักเรียนไทยช่วงอายุ 19-24 ปี พบช่องว่างระหว่างกลุ่มนักเรียนจากครอบครัวที่มีฐานะร่ำรวยที่สุดกับกลุ่มนักเรียนที่ครอบครัวมีฐานนะยากจนที่สุดมีโอกาสเรียนต่อถึงขั้นอุดมศึกษากว้างมากขึ้นเรื่อย ๆ สังเกตได้จากข้อมูลในปี 2529 ทิ้งห่างกันร้อยละ 18.5 กลายเป็นร้อยละ 42.5 ในปี 2552 (สำนักงานส่งเสริมสังคมแห่งการเรียนรู้และคุณภาพเยาวชน 2552) และเมื่อพิจารณาอัตราการการเรียนต่อของนักเรียนปีการศึกษา 2561 พบว่ามีนักเรียนออกจากโรงเรียนกลางคันทั้งสิ้น 4,039 คน ด้วยเหตุผลหลักๆ คือ ปัญหาด้านครอบครัวย้ายตามผู้ปกครอง มีปัญหาการปรับตัว และต้องหาเลี้ยงครอบครัว เป็นนักเรียนชั้น ป.1 – ม.6 จำนวน 526 คน หรือคิดเป็นร้อยละ 13 ของนักเรียนที่ออกจากโรงเรียนกลางคันในปีการศึกษา 2561 ออกจากโรงเรียนเพราะต้องหาเลี้ยงครอบครัว ในขณะที่มีเด็กอีกประมาณ 2,900 คน เรียนไปด้วยและทำงานหาเลี้ยงครอบครัวไปด้วย (สำนักนโยบายและแผนงานการศึกษาขั้นพื้นฐาน 2562) แสดงให้เห็นว่าถึงแม้ว่าประเทศไทยจะมีนโยบายที่ชัดเจนในด้านการศึกษาแต่ก็มีเด็กจำนวนไม่น้อยที่ไม่ได้รับการศึกษาอย่างต่อเนื่องด้วยสาเหตุทางบ้านมีฐานะทางเศรษฐกิจสังคมอยู่ในระดับต่ำ

นอกจากนี้เมื่อพิจารณาผลสัมฤทธิ์ทางการเรียนของนักเรียนซึ่งถือว่าเป็นตัวชี้วัดความสำเร็จด้านการศึกษาที่สามารถประเมินคุณภาพของระบบการศึกษาในการเตรียมความพร้อมให้นักเรียนมีศักยภาพหรือความสามารถพื้นฐานที่จำเป็นต่อการดำรงชีวิต จากการพิจารณาคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 จำแนกตามสังกัดทั้ง 4 รายวิชา พบว่ากลุ่มโรงเรียนโฮมสคูลโรงเรียนสังกัดกระทรวงการอุดมศึกษาวิทยาศาสตร์วิจัยและนวัตกรรมและกลุ่มโรงเรียนจุฬาภรณ์ทำคะแนนได้สูงที่สุดในทุกรายวิชาและเมื่อพิจารณาลึกลงไปพบว่านักเรียนส่วนใหญ่มาจากครอบครัวที่มีฐานะเศรษฐกิจและสังคมที่ดีกว่าสถานศึกษาในสังกัดอื่น ๆ เมื่อพิจารณาคะแนนเฉลี่ยดังกล่าวจำแนกตามขนาดโรงเรียน ที่ตั้ง และภูมิภาคพบว่านักเรียนที่สามารถทำคะแนน

เฉลี่ยได้สูงที่สุดในทุกรายวิชาคือนักเรียนที่ศึกษาอยู่โรงเรียนขนาดใหญ่พิเศษ ที่ตั้งอยู่ในเขตเมือง และอยู่จังหวัดกรุงเทพมหานคร (สถาบันทดสอบทางการศึกษาแห่งชาติ 2562) จากสถานการณ์ที่กล่าวนั้นแสดงให้เห็นว่าประเทศไทยกำลังประสบปัญหาในด้านความแตกต่างความไม่เท่าเทียมกันของโอกาสทางการศึกษาอันเนื่องมาจากฐานะทางเศรษฐกิจและสังคมของนักเรียนและโรงเรียนที่แตกต่างกันของนักเรียนส่งผลต่อคะแนนผลสัมฤทธิ์ทางการศึกษา

ดังนั้นการสร้างความเสมอภาคทางการศึกษาเพื่อลดความเหลื่อมล้ำทางการศึกษาและลดช่องว่างด้านฐานะทางเศรษฐกิจและสังคมเป็นปัจจัยหนึ่งที่นับเป็นกุญแจสำคัญในการปฏิรูปการศึกษาเพื่อให้สอดคล้องและรองรับกับสถานการณ์ที่เกิดขึ้นในประเทศไทยในปัจจุบัน ซึ่งการแก้ปัญหาดังกล่าวนั้นต้องอาศัยความร่วมมือในการดำเนินงานจากหน่วยงานต่าง ๆ ไม่ว่าจะเป็นหน่วยงานภาครัฐ ท้องถิ่น และภาคส่วนที่เกี่ยวข้อง จึงควรให้ความสำคัญและร่วมกันอย่างมีแบบแผนเพื่อการบรรลุเป้าหมายดังกล่าว เนื่องจากการศึกษาของประเทศไทยเป็นพื้นฐานสำคัญในการพัฒนาทรัพยากรมนุษย์ที่มีคุณภาพในทุกด้านและนำไปสู่การพัฒนาประเทศให้มีความมั่นคงยั่งยืนและนำไปสู่มาตรฐานสากล (กลุ่มยุทธศาสตร์การศึกษา 2560, กองทุนเพื่อความเสมอภาคทางการศึกษา 2563)

ปัจจัยที่สะท้อนความเหลื่อมล้ำทางการศึกษา

ในปัจจุบันมีการศึกษาความเหลื่อมล้ำทางการศึกษาในหลากหลายมิติ (Gregory White 2016, Ma and Wu 2019, อัมมาร สยามวาลา และคณะ 2554, มัทยา บุตรงาม 2555, รัชวดี แสงมหะหมัด 2560, ณปภัช บรรณาการ 2562, สิวะโชติ ศรีสุทธิยากร 2562) ซึ่งผู้วิจัยได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับความเหลื่อมล้ำทางการศึกษาซึ่งในการศึกษาครั้งนี้ผู้วิจัยศึกษาความเหลื่อมล้ำทางการศึกษาสองด้าน ได้แก่ ปัจจัยด้านลักษณะของโรงเรียนและปัจจัยด้านภูมิหลังฐานะความเป็นอยู่ของครอบครัว

ปัจจัยด้านลักษณะของโรงเรียน

นักเรียนที่อยู่ในระบบการศึกษาไทยไม่ว่าจะอยู่ในโรงเรียนใดก็ตาม ควรมีสิทธิเท่าเทียมกัน กล่าวคือควรได้รับโอกาสทางการศึกษาที่เท่าเทียมกัน จากงานวิจัยของ สิวะโชติ ศรีสุทธิยากร (2562) เมื่อผู้วิจัยพิจารณาผลการวิเคราะห์คะแนนความเหลื่อมล้ำทางการศึกษาและคุณภาพทางการศึกษาระดับสังกัด พบว่า สังกัดของโรงเรียนจำนวน 364 สังกัด มีเพียงสังกัดเดียวเท่านั้นที่มีคะแนนคุณภาพทางการศึกษาอยู่ในระดับสูงมากคือสังกัดสำนักงานคณะกรรมการการอุดมศึกษาหรือโรงเรียนสาธิต นอกจากนี้ยังพบว่าสังกัดที่มีคะแนนคุณภาพต่ำที่สุดของประเทศเป็นสังกัดที่อยู่ในจังหวัดชายแดนภาพใต้ สะท้อนให้เห็นว่าแต่เมื่อนักเรียนอยู่ต่างโรงเรียนกัน มีสิทธิหรือโอกาส

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ที่ไม่เท่าเทียมกันทางการศึกษาจึงก่อให้เกิดความเหลื่อมล้ำทางการศึกษา แสดงให้เห็นว่าควรให้ความสำคัญเป็นอย่างยิ่งเนื่องจากสะท้อนให้เห็นว่าประเทศไทยมีการบริการจัดการทางการศึกษาที่ยังไม่ประสบความสำเร็จเท่าที่ควร

การบริหารการศึกษาของประเทศไทยมีการจัดสรรงบประมาณเงินอุดหนุนค่าใช้จ่ายรายหัวให้กับโรงเรียนกล่าวคือโรงเรียนขนาดใหญ่ที่มีนักเรียนจำนวนมากจะได้รับงบประมาณอุดหนุนที่มากกว่าโรงเรียนขนาดเล็กที่มีจำนวนนักเรียนที่น้อยกว่าจึงส่งผลให้โรงเรียนขนาดใหญ่มีการจัดการศึกษาที่คล่องตัวและยืดหยุ่นมากกว่าโรงเรียนขนาดเล็กขณะที่สภาพความเป็นจริงนั้นการจัดการศึกษาไม่เพียงแต่การบริหารนักเรียนภายในโรงเรียนเท่านั้นแต่ควรคำนึงถึงทรัพยากรที่ต้องใช้ภายในโรงเรียน เช่น สื่อการเรียนการสอน เทคโนโลยีต่าง ๆที่ทุกโรงเรียนจำเป็นต้องใช้ในการเรียนรู้ซึ่งปัจจัยดังกล่าวไม่ได้ขึ้นอยู่กับจำนวนนักเรียนเท่านั้น เห็นได้ชัดจากเมื่อพิจารณาถึงสถิติการออกจากระบบการศึกษาของนักเรียนพบว่านักเรียนที่เรียนอยู่โรงเรียนขนาดเล็กมีแนวโน้มออกจากระบบการศึกษาหรือออกจากโรงเรียนกลางคันมากกว่านักเรียนที่อยู่ในโรงเรียนขนาดใหญ่หรือเลือกทำงานและไม่ศึกษาต่อในระบบการศึกษาที่สูงขึ้นหลังจากจบการศึกษาภาคบังคับ เนื่องจากความไม่พร้อมในด้านต่าง ๆ ทำให้จำนวนนักเรียนในโรงเรียนลดลง ขนาดของโรงเรียนปรับเปลี่ยนเป็นโรงเรียนขนาดเล็ก ทำให้ต้องปรับค่าใช้จ่ายของนักเรียนต่อหัวตามไปด้วยทั้งที่ในความเป็นจริง นั้นแม้นักเรียนจะลดลงแต่ทรัพยากรที่ต้องใช้ภายในโรงเรียนไม่ได้ลดลงตามไปด้วยแต่กลับได้รับเงินสนับสนุนที่น้อยลง จะพบว่านักเรียนที่เรียนอยู่ในโรงเรียนขนาดเล็กสูญเสียโอกาสมากกว่านักเรียนที่เรียนอยู่ในโรงเรียนขนาดใหญ่ (Gauthier and Punyasavatsut 2019, สำนักงานเลขาธิการสภาการศึกษา กระทรวงศึกษาธิการ 2557-2558)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากสภาพปัญหาและการบริหารการศึกษาที่กล่าวข้างต้นทำให้เกิดความแตกต่างระหว่างโรงเรียนอย่างชัดเจนประกอบกับนโยบายของรัฐด้านการจัดสรรงบประมาณเงินอุดหนุนที่เอื้อต่อโรงเรียนขนาดใหญ่มากกว่าโรงเรียนขนาดเล็กจึงส่งผลกระทบให้เกิดความเหลื่อมล้ำทางการศึกษาอันเนื่องมาจากคุณลักษณะที่แตกต่างของโรงเรียนซึ่งสอดคล้องกับงานวิจัยของ อิทธิพล พลเหี้ยมหาญ (2562) สะท้อนให้เห็นถึงสภาพปัญหาของการบริหารโรงเรียนมัธยมศึกษาขนาดเล็กพบว่าโรงเรียนขนาดเล็กมีโอกาสที่น้อยกว่าโรงเรียนขนาดใหญ่โดยที่เด็กที่ศึกษาอยู่ในโรงเรียนขนาดเล็กกลายเป็นเด็กด้อยโอกาสทางการศึกษาทำให้เกิดความเหลื่อมล้ำทางการศึกษาอันเนื่องมาจากขนาดโรงเรียนและเมื่อพิจารณาคุณภาพของการศึกษาตัวชี้วัดหนึ่งซึ่งถือว่าเป็นตัวชี้วัดความสำเร็จของโรงเรียน คือผลการประเมินของนักเรียนหรือผลสัมฤทธิ์ทางการเรียนของนักเรียนจากผลการสอบโครงการ

ประเมินผลนักเรียนร่วมกับนานาชาติ PISA 2015 จำนวน 72 พบว่าประเทศไทยได้อันดับที่ 54 ในวิชาวิทยาศาสตร์และอันดับ 57 ในวิชาคณิตศาสตร์

นอกจากนั้นการประเมินระดับชาติสำหรับ 9 วิชาหลักของนักเรียนชั้นมัธยมศึกษาปีที่ 6 ประจำ พ.ศ. 2559 ปรากฏผลว่านักเรียนไทยสอบตกโดยเฉลี่ย 8 วิชาจากทั้งหมด 9 วิชาและเมื่อพิจารณาลึกลงไปพบว่ามีนักเรียนที่อยู่ในโรงเรียนขนาดใหญ่เท่านั้นที่มีคะแนนสูงกว่าค่าเฉลี่ย ซึ่งเป็นปัญหาที่สะท้อนให้เห็นถึงความเหลื่อมล้ำทางการศึกษาในด้านคุณภาพทางการศึกษาของโรงเรียนที่แตกต่างกัน โรงเรียนที่มีขนาดใหญ่มีแนวโน้มที่จะมีผลการสอบสูงกว่าโรงเรียนขนาดเล็กและโรงเรียนที่มีภูมิลำเนาอยู่ในเมืองมีการผลการสอบที่สูงกว่าโรงเรียนที่อยู่นอกเมือง (คณะกรรมการอิสระเพื่อการปฏิรูปการศึกษา, 2562) นอกจากนี้เมื่อพิจารณาคะแนนเฉลี่ยของผลการทดสอบ O-NET ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562 จำแนกตามสังกัดพบว่ากลุ่มโรงเรียนโฮมสคูล กลุ่มโรงเรียนสังกัดกระทรวงการอุดมศึกษาวิทยาศาสตร์วิจัยและนวัตกรรม และกลุ่มโรงเรียนจุฬาภรณ์ทำคะแนนได้สูงที่สุดในทุกรายวิชาดีกว่าโรงเรียนในสังกัดอื่น ๆ (สถาบันทดสอบทางการศึกษาแห่งชาติ 2562) จะเห็นว่านอกจากขนาดของโรงเรียนแล้วสังกัดของโรงเรียนก็มีผลทำให้เกิดความแตกต่างระหว่างโรงเรียนด้วยเช่นกัน

จากสถานการณ์ที่เกิดขึ้นนั้นทำให้มีแนวโน้มสูงที่ผู้ปกครองที่มีความพร้อมจะส่งลูกหลานไปเรียนโรงเรียนขนาดใหญ่และอยู่ในสังกัดที่สามารถทำคะแนนได้สูง นักเรียนที่ต้องอยู่ในโรงเรียนขนาดเล็กจึงมีโอกาสน้อยกว่าและบางส่วนต้องออกจากระบบการศึกษาในที่สุดเนื่องจากความไม่พร้อมในหลายด้าน ทำให้การบริหารการศึกษาไม่สามารถสะท้อนคุณภาพและประสิทธิภาพได้ดีเท่าที่ควร จากสถานการณ์ที่กล่าวนั้นแสดงให้เห็นว่าประเทศไทยกำลังประสบปัญหาในด้านความแตกต่างความไม่เท่าเทียมกันของโอกาสทางการศึกษาอันเนื่องมาจากคุณลักษณะโรงเรียนที่แตกต่างกันของนักเรียน ส่งผลต่อคะแนนผลสัมฤทธิ์ทางการศึกษา จากการศึกษาได้กล่าวมาข้างต้นนั้นสามารถสรุปได้ว่า ประเทศกำลังประสบปัญหาความเหลื่อมล้ำทางการศึกษาด้านคุณลักษณะของโรงเรียน คือ สังกัดของโรงเรียนและขนาดของโรงเรียนที่นักเรียนศึกษาอยู่

ปัจจัยด้านภูมิหลังฐานะความเป็นอยู่ของครอบครัว

นักเรียนที่อยู่ในระบบการศึกษาควรได้รับโอกาสทางการศึกษาที่เท่าเทียมกันไม่ขึ้นอยู่กับฐานะทางเศรษฐกิจและสังคมที่แตกต่างกันกล่าวคือไม่ว่าจะรวยหรือจนย่อมมีโอกาสทางการศึกษาที่เท่าเทียมกันขณะที่จากการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกลับพบว่าครอบครัวที่มีฐานะทางเศรษฐกิจและสังคมที่ดีจะสามารถให้การสนับสนุนลูกหลานให้ได้รับการศึกษาที่ดีได้มากกว่า

ครอบครัวที่มีฐานะทางเศรษฐกิจและสังคม ที่ต่ำกว่า สะท้อนให้เห็นถึงความเหลื่อมล้ำทางการศึกษาในด้านภูมิหลังฐานะทางเศรษฐกิจและสังคมของครอบครัวที่แตกต่างกันอย่างชัดเจน จากการศึกษาข้อมูลพบว่าเด็กนักเรียนไทยออกจากระบบการศึกษาหรือออกกลางคันจากสถานศึกษาซึ่งสร้างความเสียหายให้กับเศรษฐกิจและสังคมไทยสูงถึงปีละ 300,000 ล้านบาทหรือเท่ากับร้อยละ 3 ของผลิตภัณฑ์มวลรวมประชาชาติของไทยในแต่ละปี และพิจารณาสาเหตุของการออกกลางคันพบว่าตั้งท้องก่อนวัยอันควร ถูกดำเนินคดีโดยสถานพินิจ รวมไปถึงครอบครัวมีฐานะยากจนจึงทำให้ไม่สามารถศึกษาต่อได้ (สถาบันอนาคตไทยศึกษา 2562)

เมื่อพิจารณาปัญหาความยากจนพบว่าส่วนใหญ่เด็กไทยที่หลุดออกจากระบบการศึกษามีแนวโน้มขาดโอกาสทางการศึกษาที่มีคุณภาพนั้นเกิดจากครอบครัวของเด็กต้องแบกรับภาระค่าใช้จ่ายด้านการศึกษาสูงเมื่อเทียบกับรายได้ของครอบครัว ถึงแม้รัฐบาลจะมีงบประมาณในการสนับสนุนด้านการศึกษาแล้วก็ตาม (ข้อมูลจากบัญชีรายจ่ายด้านการศึกษาแห่งชาติ 2551-2559 , กองทุนเพื่อความเสมอภาคทางการศึกษา 2563) และเมื่อพิจารณาถึงการคงอยู่ในระบบการศึกษาจนมีโอกาสสำเร็จการศึกษาขั้นพื้นฐานและอุดมศึกษาของเด็กกลุ่มที่ครอบครัวมีฐานะที่ต่ำพบว่า มีความเหลื่อมล้ำด้านโอกาสการสำเร็จการศึกษามากถึง 7 เท่า โดยเด็กจากครัวเรือนที่มีรายได้ต่ำที่สุดร้อยละ 20 ของประเทศมีโอกาสศึกษาต่อระดับอุดมศึกษาโดยเฉลี่ยเพียงร้อยละ 5 ของประชากรในกลุ่มรายได้ต่ำเช่นเดียวกันเท่านั้น

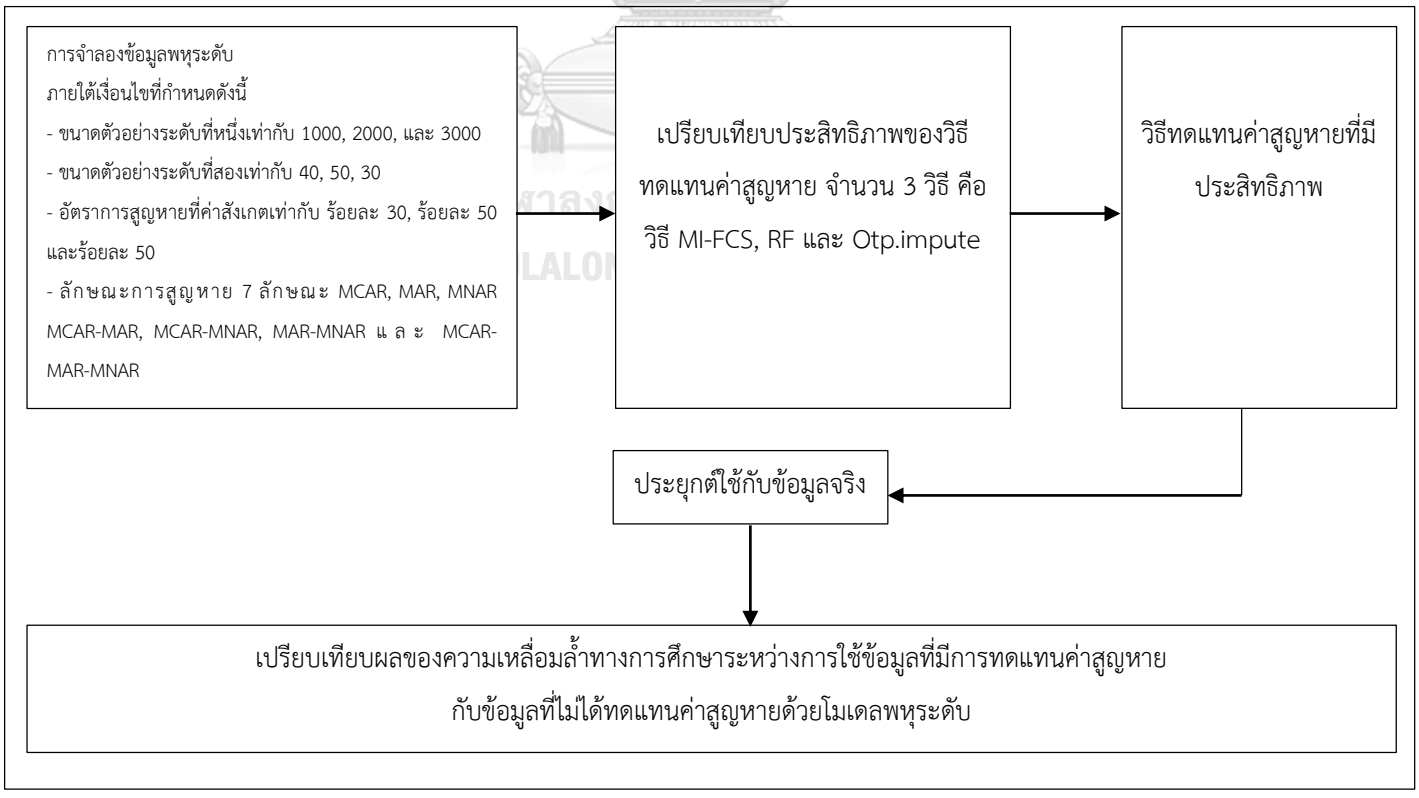
นอกจากนี้ความเหลื่อมล้ำทางการศึกษายังทำให้ปัจจุบันประเทศไทยมีกำลังแรงงานที่มีทักษะขั้นสูง (High-skilled Labor) ต่ำกว่าร้อยละ 20 ขณะที่การพัฒนาประเทศออกจากกับดักรายได้ปานกลาง (Middle Income Trap) ตามยุทธศาสตร์ชาติ 20 ปีของรัฐบาล จำเป็นต้องมีกำลังแรงงานที่มีทักษะหรือมีทักษะขั้นสูง ผู้สำเร็จการศึกษาขั้นพื้นฐานหรือการศึกษาระดับอุดมศึกษามากกว่าร้อยละ 50 ของกำลังแรงงานทั้งหมดในประเทศ (สำนักงบประมาณของรัฐสภา 2563, อ้างถึงใน ภาวรรณ ธนาเลิศสมบูรณ์ 2563)จากที่กล่าวมาข้างต้นพบว่าประชาชนไม่สามารถเข้าถึงการศึกษาได้อย่างทั่วถึงอันเนื่องมาจากความไม่พร้อมด้านภูมิหลังฐานะทางเศรษฐกิจและสังคมที่แตกต่างกันของครอบครัวนอกเหนือจากความไม่พร้อมด้านระบบการศึกษาของประไทยสะท้อนให้เห็นถึงความเหลื่อมล้ำทางการศึกษาในด้านภูมิหลังฐานะทางเศรษฐกิจและสังคมที่แตกต่างกัน

นอกจากนี้จากการศึกษาของ Jo Blanden and Stephen Machin (2004) Alberto Alesina (2018) และ Marks and Pokropek (2019) พบว่าผลสัมฤทธิ์ทางการเรียนของนักเรียน

ที่แตกต่างกันมีผลมาจากภูมิหลังทางฐานะเศรษฐกิจและสังคมของครอบครัวไม่ว่าจะเป็นด้านรายได้ของผู้ปกครอง สถานภาพความเป็นอยู่ของครอบครัว ลักษณะที่พักอาศัย ซึ่งปัจจัยดังกล่าวสะท้อนความเหลื่อมล้ำทางการศึกษาอันเนื่องมาจากภูมิหลังทางฐานะทางเศรษฐกิจและสังคมของครอบครัวที่แตกต่างกัน ซึ่งสอดคล้องกับการศึกษาของ ณปภัช บรรณาการ (2562) วิเคราะห์ความเหลื่อมล้ำทางการศึกษาของนักเรียนจาก รายได้ของผู้ปกครอง ลักษณะที่พักอาศัย สถานภาพครอบครัวและบุคคลที่พักอาศัยนอกจากนี้ มัทยา บุตรงาม (2555) ได้กล่าวถึงปัจจัยที่ส่งผลต่อความเหลื่อมล้ำทางการศึกษาเกิดจากความแตกต่างฐานะทางเศรษฐกิจ ได้แก่ จำนวนพี่น้องในครอบครัว ระดับการศึกษาของพ่อแม่ และความแตกต่างด้านที่อยู่อาศัย เป็นต้น

กรอบแนวคิดที่ใช้ในการวิจัย

การศึกษาวิจัยเรื่องการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับ: การประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ผู้วิจัยศึกษาเอกสารแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง สามารถสรุปกรอบแนวคิดในการวิจัยได้ในรูปที่ 2.1.2 ดังนี้



รูปที่ 2.1 2 กรอบแนวคิดที่ใช้ในการวิจัย

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง เพื่อต้องการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) โดยใช้การจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล (Monte Carlo Simulation Technique) และเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา ระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย โดยข้อมูลทุติยภูมิที่ใช้คือข้อมูลจากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ใช้การเขียนโปรแกรม R เวอร์ชั่น 4.2.1 และโปรแกรม Julia 1.7.2 ในการวิเคราะห์ข้อมูลและประมวลผล โดยวิธีทดแทนค่าสูญหาย ในโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มที่ศึกษาประกอบด้วย 3 วิธี ได้แก่

1. วิธี Fully Conditional Specification Multiple Imputation (MI-FCS)

2. วิธี Random Forest Imputation (RF)

3. วิธี Optimal imputation (Opt.impute) ซึ่งประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ K-Nearest Neighbors based (Opt.knn), Support Vector Machines based (Opt.svm), Decision Tree based (Opt.tree) และ Optimal Cross - validated (Opt.cv)

ผู้วิจัยทำการศึกษาประสิทธิภาพของวิธีทดแทนค่าสูญหาย ในข้างต้น โดยแบ่งการศึกษาออกเป็นสองระยะ ดังนี้ ระยะที่หนึ่ง คือ การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายโดยใช้การจำลองข้อมูล โดยพิจารณาประสิทธิภาพจาก 3 เกณฑ์ ได้แก่ ค่าเฉลี่ยของรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (NRMSE) ค่าเฉลี่ยของค่าความเอนเอียงสัมพัทธ์ (RB) และใช้ค่าเฉลี่ยของค่าคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (MCSE) และระยะที่สอง ผู้วิจัยจึงดำเนินการนำวิธีทดแทนข้อมูลสูญหายที่ได้จากผลระยะที่หนึ่งมาประยุกต์ใช้กับข้อมูลจริง เพื่อเปรียบเทียบการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหายในการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ทั้งนี้สามารถสรุปขั้นตอนและวิธีดำเนินการวิจัยออกเป็น 3 ตอน ได้แก่ ตอนที่ 1 ตัวแปรและโมเดลที่ใช้ในการศึกษา ตอนที่ 2 การศึกษาการจำลองข้อมูลและตอนที่ 3 การประยุกต์ใช้กับข้อมูลจริง โดยมีรายละเอียด ดังต่อนี้

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ตอนที่ 1 ตัวแปรและโมเดลที่ใช้ในการศึกษา

การวิจัยครั้งนี้เป็นการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ ภายใต้การวิเคราะห์แบบสองระดับโดยให้ระดับที่หนึ่ง คือ ระดับโรงเรียนและระดับที่สองคือระดับสังกัดโดยใช้การจำลองข้อมูลและการประยุกต์ใช้กับข้อมูลจริง เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา โดยตัวแปรที่ใช้ในการศึกษาครั้งนี้ มีรายละเอียดดังนี้

1. ตัวแปรที่ใช้ในการวิจัย

ตัวแปรที่ใช้ในการวิจัยครั้งนี้ ประกอบด้วย ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) และสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (x) มีรายละเอียดดังนี้

(1) ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) หมายถึง ความรู้ความเข้าใจขั้นพื้นฐานของนักเรียนระดับโรงเรียนซึ่งการวิจัยนี้พิจารณาจากผลคะแนนเฉลี่ยจากการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (O-NET) จำนวน 4 วิชา ได้แก่ วิชาภาษาไทย วิชาภาษาอังกฤษ วิชาคณิตศาสตร์ และวิชาวิทยาศาสตร์ (สถาบันทดสอบทางการศึกษาแห่งชาติ 2562) โดยผู้วิจัยประมาณค่าคะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนด้วยการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis : PCA)

(2) สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (x) หมายถึง จำนวนนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา ต่อจำนวนนักเรียนทั้งหมดในโรงเรียน ที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) โดยในการวิจัยครั้งนี้ได้พิจารณาจากสองตัวแปร ได้แก่ สัดส่วนของนักเรียนที่พักอาศัยอยู่กับบิดาและมารดา (x_1) และสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ (x_2) ซึ่งกำหนดให้แต่ละตัวแปรมีการให้คะแนนแบบสองค่า (dichotomous variables) รายละเอียดดังนี้

สัดส่วนของนักเรียนที่พักอาศัยอยู่กับบิดาและมารดา (x_1) พิจารณาจากค่าเฉลี่ยของนักเรียนในโรงเรียนที่พักอาศัยอยู่กับบิดาและมารดา ให้เท่ากับ 1 และอยู่กับบิดามารดาหรืออยู่กับมารดาอยู่กับญาติ หรืออื่น ๆ เท่ากับ 0

สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ (x_2) พิจารณาจากค่าเฉลี่ยของนักเรียนในโรงเรียนที่ผู้ปกครองมีรายได้ต่อปีไม่เกิน 300,000 บาท ให้เท่ากับ 1 และอื่น ๆ เท่ากับ 0

โดยอิงตามหลักเกณฑ์และคุณสมบัติของผู้กู้ยืมเงินกองทุนกู้ยืมเพื่อการศึกษา (กยศ) ได้นิยามคำว่า “ผู้ขาดแคลนทุนทรัพย์” หมายถึง ครอบครัวมีรายได้ต่อปีต่ำกว่า 300,000 ต่อปี (กองทุนกู้ยืมเพื่อการศึกษา, 2541)

ผู้วิจัยรวมคะแนนทั้งสองด้านเข้าด้วยกันด้วยค่าเฉลี่ยเรขาคณิต $x = \sqrt{x_1 x_2}$ เพื่อให้ได้คะแนนสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาแต่ละโรงเรียน (x)

2. โมเดลที่ใช้ในการศึกษา

จากการพิจารณาการศึกษาที่ผ่านมา จะพบว่าโดยส่วนใหญ่นิยมใช้ผลการทดสอบ PISA, O-NET เป็นข้อมูลในการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาของประเทศไทย โดยใช้ข้อมูลมาวิเคราะห์การถดถอย (Regression Analysis) เพื่อศึกษาปัจจัยที่ส่งผลกระทบต่อผลสัมฤทธิ์ทางการเรียน ซึ่งผลสัมฤทธิ์ทางการเรียนเป็นตัวบ่งชี้ที่สามารถสะท้อนสภาพของความเหลื่อมล้ำทางการศึกษาได้ขณะที่เมื่อพิจารณาสภาพการศึกษาของประเทศไทย พบว่ามีลักษณะการบริหารจัดการที่มีลำดับชั้น จากส่วนกลางลงไปยังเขตพื้นที่และโรงเรียน จึงทำให้ข้อมูลทางการศึกษาของประเทศมีโครงสร้างแบบข้อมูลพหุระดับ (multilevel data) ซึ่งก่อให้เกิดความสัมพันธ์ระหว่างหน่วยข้อมูล หากวิเคราะห์ข้อมูลที่มีโครงสร้างแบบพระดับโดยไม่ได้คำนึงโครงสร้างความสัมพันธ์ดังกล่าว อาจเป็นการละเมิดข้อตกลงเบื้องต้นของสถิติวิเคราะห์ ในด้านความเป็นอิสระของข้อมูลและทำให้ผลการประมาณความเหลื่อมล้ำที่ได้มีความน่าเชื่อถือต่ำ (สิวะโชติ, 2562)

ดังนั้นการศึกษาครั้งนี้ผู้วิจัยจึงใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มในการวิเคราะห์และประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) ของโรงเรียนในแต่ละสังกัด (Rothstein, 2019) เพื่อสะท้อนความเหลื่อมล้ำทางการศึกษา ซึ่งน่าจะทำให้ได้สารสนเทศของความเหลื่อมล้ำทางการศึกษาที่มีความน่าเชื่อถือ และสามารถให้สารสนเทศเชิงลึกได้ถึงระดับหน่วยย่อยของการศึกษา สามารถสะท้อนความเหลื่อมล้ำทางการศึกษาได้อย่างแม่นยำมากขึ้น

ผู้วิจัยจำลองข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ (Multilevel Model) ด้วยโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) ภายใต้การวิเคราะห์แบบสองระดับ เมื่อระดับที่หนึ่งคือระดับโรงเรียนและระดับที่สองคือระดับสังกัด กำหนดค่าพารามิเตอร์จากข้อมูลจริง เพื่อจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล นำค่าพารามิเตอร์

ที่สร้างขึ้นมาคำนวณหาผลสัมฤทธิ์ทางการเรียน (y_{ij}) ของนักเรียนในโรงเรียนที่ i และสังกัดที่ j จากสมการ ดังนี้

โมเดลระดับโรงเรียน $$y_{ij} = \alpha_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

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

เมื่อ y_{ij} แทน ผลสัมฤทธิ์ทางการเรียนของนักเรียนโดยเฉลี่ยในโรงเรียนที่ i สังกัดที่ j

x_{ij} แทน สัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา ระดับโรงเรียนที่ i และสังกัดที่ j

γ_{00} แทน สัมประสิทธิ์จุดตัดแกน

γ_{01} แทน อัตราการเปลี่ยนแปลงของสัดส่วนนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดาและมารดาที่มีผลต่อผลสัมฤทธิ์ทางการเรียน

$\varepsilon_{ij} \sim N(0,\sigma^2)$ แทน ความคลาดเคลื่อนสุ่มของโมเดลในระดับโรงเรียน

$U_{ij} \sim N(0,\Sigma_u)$ แทน ความคลาดเคลื่อนสุ่มของโมเดลในระดับระดับสังกัดโดยที่

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

$\Sigma_u = \begin{bmatrix} 0.588 & 0 \\ 0 & 1.272 \end{bmatrix}$ คือ เมทริกซ์ความแปรปรวนของเวกเตอร์ความคลาดเคลื่อนสุ่มในระดับสังกัด

ทั้งนี้ผู้วิจัยสรุปการกำหนดค่าพารามิเตอร์ตามข้อมูลจริงด้วยตารางการแจกแจงของตัวแปรและค่าพารามิเตอร์ที่ใช้ในการจำลองข้อมูล รายละเอียดปรากฏดังตารางที่ 3.1

ตารางที่ 3. 1 การแจกแจงของตัวแปรและค่าพารามิเตอร์ที่ใช้ในการจำลองข้อมูล

ตัวแปร/พารามิเตอร์	ลักษณะการแจกแจง/ค่าพารามิเตอร์
สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (x_{ij})	$x_{ij} \sim$ Beta (α,β) $\alpha = 0.0579$ $\beta = 15.278$
สัมประสิทธิ์จุดตัดแกน (γ_{00})	$-1.502 \text{ x } e^{-19}$
สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (γ_{01})	-1.02
ความคลาดเคลื่อนสุ่มของโมเดลในระดับโรงเรียน (ε_{ij})	$\text{N}(0,\sigma^2)$; $U_{0j} \sim \text{N}(0, 0.588)$
ความคลาดเคลื่อนสุ่มของโมเดลในระดับสังกัด (U_{ij})	$\text{N}(0,\sum u)$; $U_{ij} \sim \text{N}(0, 1.272)$

การสร้างแบบจำลองในครั้งนี้ ผู้วิจัยกำหนดให้สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน มีการแจกแจงแบบเบต้า (Beta distribution) โดยพิจารณาค่าเฉลี่ยแทนด้วย $\mathrm{E}[X] = 0.00377$ และค่าความแปรปรวน var [X] = 2.3005 x e^{-4} จากข้อมูลจริงดังกล่าวข้างต้น

สามารถนำค่าที่ได้มาคำนวณได้จากสมการ ดังนี้

$$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta} \tag{1}$$

$$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{2}$$

โดยกำหนดการแจกแจงของค่าพารามิเตอร์ตามลักษณะของข้อมูลจริงข้างต้น (ตารางที่ 3.1) สร้างโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มแล้วนำค่าพารามิเตอร์ที่สร้างขึ้นมาคำนวณหาค่าผลสัมฤทธิ์ทางการเรียน (y_{ij}) ของนักเรียนของโรงเรียนที่ i และสังกัดที่ j จะได้ข้อมูลการจำลองที่มีสมการ ดังนี้

$$y_{ij} = \alpha_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$

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

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตอนที่ 2 การศึกษาการจำลองข้อมูล

ผู้วิจัยศึกษาการจำลองข้อมูล (Simulation data) ภายใต้เงื่อนไขที่แตกต่างกัน 4 เงื่อนไข ประกอบด้วย (1) วิธีทดแทนค่าสูญหายจำนวน 3 วิธี ได้แก่ วิธี MI-FCS, วิธี RF และวิธี Opt.impute ซึ่งประกอบด้วยวิธีย่อย 4 วิธี ได้แก่วิธี Opt.knn, วิธี Opt.tree, วิธี Opt.svm และวิธี Opt.cv (2) ขนาดตัวอย่าง 3 ขนาด (3) ลักษณะการสูญหาย 7 ลักษณะ และ (4) อัตราการสูญหายของค่าสังเกตแต่ละลักษณะ เป็น 3 ระดับ โดยกำหนดให้ค่าสังเกตสูญหายในข้อมูลระดับโรงเรียน (ระดับที่หนึ่ง) ดังนั้นจากการพิจารณาเงื่อนไขที่ได้จากการจำลองข้อมูลพบว่ามีจำนวนทั้งสิ้น 378 เงื่อนไข (6x3x7x3) ทำซ้ำ 100 รอบ โดยผู้วิจัยสรุปรายละเอียดของการจำลองข้อมูล ดังต่อไปนี้

1. ข้อตกลงเบื้องต้นและเงื่อนไขที่ใช้ในการศึกษา

ขนาดตัวอย่าง

การวิจัยครั้งนี้ผู้วิจัยมุ่งศึกษาประสิทธิภาพของวิธีทดแทนค่าข้อมูลสูญหายสำหรับข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ (Multilevel data) และเพื่อให้สอดคล้องกับธรรมชาติของข้อมูลในการศึกษาครั้งนี้ ผู้วิจัยกำหนดขนาดตัวอย่างสองระดับ จากการพิจารณาลักษณะของข้อมูลทุติยภูมิของสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ปีการศึกษา 2562 - 2563 โดยที่ขนาดของตัวอย่างระดับที่หนึ่งเท่ากับ 1000, 2000, และ 3,000 หน่วย และขนาดตัวอย่างระดับที่สองเท่ากับ 40, 50, และ 60 หน่วย ตามลำดับ

อัตราการสูญหายและลักษณะการสูญหาย

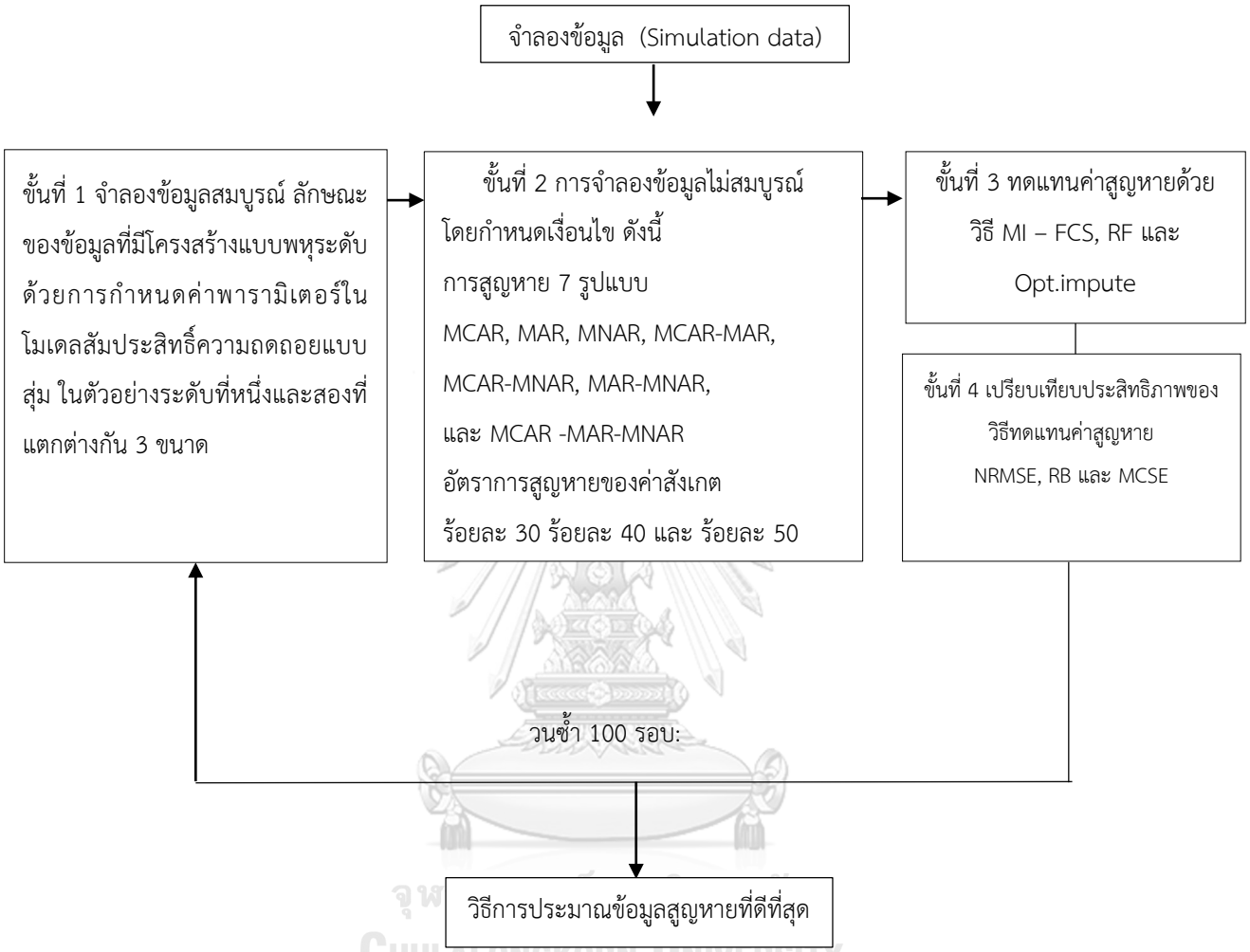
การศึกษาครั้งนี้ แบ่งลักษณะการสูญหายออกเป็น 7 ลักษณะ ได้แก่ การสูญหายแบบสุ่มสมบูรณ์ (MCAR), การสูญหายแบบสุ่ม (MAR), การสูญหายแบบไม่สุ่ม (MNAR) และการสูญหายแบบผสม 4 ลักษณะ ได้แก่ MCAR-MAR, MCAR-MNAR, MAR-MNAR และ MCAR-MAR-MNAR โดยผู้วิจัยกำหนดอัตราการสูญหายของค่าสังเกตในแต่ละลักษณะเป็น 3 ระดับ ได้แก่ ร้อยละ 30, ร้อยละ 40 และร้อยละ 50 สำหรับกรณีการสูญหายแบบผสมอัตราการสูญหายของค่าสังเกตแต่ลักษณะจะแบ่งสัดส่วนเท่า ๆ กัน กล่าวคือ สำหรับกรณีอัตราการสูญหายของค่าสังเกตแบบ MCAR - MAR ที่ระดับร้อยละ 30 จะแบ่งอัตราการสูญหายของค่าสังเกตแบบ MCAR ร้อยละ 15 และ MAR ร้อยละ 15

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2. ขั้นตอนในการจำลองข้อมูล (Simulation procedure)

ผู้วิจัยใช้โปรแกรม R เวอร์ชั่น 4.1.2 จำลองข้อมูลภายใต้สถานการณ์จำลองแบบมอนติคาร์โล (Monte Carlo Simulation) ที่มีโครงสร้างแบบพหุระดับภายใต้การวิเคราะห์แบบสองระดับ โดยให้ระดับที่หนึ่ง คือ ระดับโรงเรียน และระดับที่สอง คือ ระดับสังกัด ด้วยการกำหนดค่าพารามิเตอร์ในโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (random coefficients model) โดยใช้การจำลองข้อมูลมีรายละเอียดดังต่อไปนี้

แผนภาพสรุปขั้นตอนการดำเนินการวิจัยระยะที่หนึ่ง



รูปที่ 3.1. 1 แผนภาพขั้นตอนการจำลองข้อมูล

ขั้นตอนที่ 1 การจำลองข้อมูลสมบูรณ์ (complete data)

งานวิจัยนี้เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ ดังนั้นเพื่อให้ข้อมูลที่จำลองขึ้นมีความสอดคล้องกับธรรมชาติของข้อมูลจริง ผู้วิจัยจึงสร้างแบบจำลองจากสถานการณ์จริง โดยข้อมูลที่นำมาสร้างแบบจำลองในครั้งนี้ คือ ข้อมูลทุติยภูมิทางการศึกษาจากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ของนักเรียนระดับโรงเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563

กำหนดการแจกแจงของค่าพารามิเตอร์ตามลักษณะของข้อมูลจริงข้างต้น สร้างโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มแล้วนำค่าพารามิเตอร์ที่สร้างขึ้นมาคำนวณหาค่าผลสัมฤทธิ์ ทางการเรียน (y_{ij}) ของนักเรียนในโรงเรียนที่ i และสังกัดที่ j จะได้ข้อมูลการจำลองที่มีสมการ ดังนี้

$$y_{ij} = \alpha_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$$
$$\alpha_{0j} = \gamma_{00} + U_{0j},\ \beta_{1j} = \gamma_{01} + U_{1j}$$

ดังนั้น การวิจัยครั้งนี้ผู้วิจัยจึงจำลองข้อมูล โดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายและนำข้อมูลที่ทดแทนค่าสูญหายแล้วมาวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ทั้งนี้เมื่อพิจารณาบริบทของความเหลื่อมล้ำทางการศึกษาพบว่า ในปัจจุบันการวิเคราะห์และศึกษางานวิจัยด้านความเหลื่อมล้ำทางการศึกษามีหลากหลายมิติ เช่น ด้านคุณภาพการจัดการบริหารการศึกษาด้านการจัดการทรัพยากรครู เป็นต้น อย่างไรก็ตาม งานวิจัยนี้วิเคราะห์ความเหลื่อมล้ำทางการศึกษาโดยใช้ความแตกต่างสัมประสิทธิ์ความถดถอยแบบสุ่มของ สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (β_1) ที่ส่งผลให้ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนมีความแตกต่างกัน

ขั้นตอนที่ 2 จำลองข้อมูลไม่สมบูรณ์ (Missing data)

ผู้วิจัยจัดกระทำข้อมูลสมบูรณ์ที่สร้างขึ้นโดยให้เกิดการสูญหายภายใต้สถานการณ์และเงื่อนไขที่กำหนด โดยให้มีลักษณะการสูญหายที่แตกต่างกัน 7 ลักษณะ ทั้งนี้กำหนดให้มีการสูญหายที่ค่าสังเกตภายในตัวแปรระดับที่หนึ่ง จำลองข้อมูลโดยการทำวนซ้ำ 100 รอบ ผู้วิจัยสามารถแบ่งการจำลองข้อมูลที่ไม่สมบูรณ์ ได้ตามลักษณะการสูญหาย รายละเอียดดังนี้

เมื่อกำหนดให้ y_{ij} แทน ผลสัมฤทธิ์ทางการเรียนของนักเรียนในโรงเรียนที่ i และสังกัดที่ j

x_{ij} แทน สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน ในโรงเรียนที่ i และสังกัดที่ j

(1) การสูญหายแบบสุ่มสมบูรณ์ (MCAR) : การสูญหายที่ลักษณะของค่าสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด ค่าที่สูญหายเป็นอิสระจากตัวแปรต่าง ๆ กล่าวคือข้อมูลที่สูญหายไม่มีความสัมพันธ์กับตัวแปรใด ๆ หรือค่าอื่น ๆ ทั้งที่ทราบค่าและไม่ทราบค่าเพื่อให้สอดคล้องกับการสูญหายแบบสุ่มสมบูรณ์ในการศึกษาครั้งนี้ ผู้วิจัยจำลองข้อมูลไม่สมบูรณ์ โดยกำหนดให้ค่าสังเกตในตัวแปร x_{ij} เกิดการสูญหายอย่างสุ่มสมบูรณ์ โดยที่สุ่มให้ค่า x_{ij} เป็นค่าสูญหาย

ไม่มีความสัมพันธ์กับค่า x_{ij} และ y_{ij} ที่สังเกตได้ (ค่าที่เก็บได้) ซึ่งมีความน่าจะเป็นในการเกิดค่าสูญหายอย่างสุ่มสมบูรณ์เท่ากับ 0.3, 0.4 และ 0.5 ตามลำดับ ทั้งนี้เมื่อพิจารณาถึงสาเหตุของการเกิดค่าสูญหายแบบสุ่มสมบูรณ์ ในบริบทของข้อมูลความเหลื่อมล้ำทางการศึกษา เช่น เมื่อจำเป็นต้องเก็บข้อมูลทุติยภูมิอาจเกิดข้อผิดพลาดด้านการนำเข้าข้อมูล ผู้กรอกข้อมูลในหน่วยงานทางการศึกษาของรัฐไม่ใช่เจ้าของข้อมูลโดยตรงความหลายหลายของการจัดเก็บฐานข้อมูลที่แตกต่างกัน อาจส่งผลกระทบให้เกิดค่าสูญหายดังกล่าวได้

(2) การสูญหายแบบสุ่ม (MAR) หรืออาจเรียกว่าสูญหายแบบมีเงื่อนไข: การสูญหายแบบสุ่มคือความน่าจะเป็นที่เกิดค่าสูญหายขึ้นอยู่กับค่าตัวแปรสังเกตได้อื่นๆ ที่สนใจศึกษาเท่านั้น กล่าวคือค่าที่สูญหายมีความสัมพันธ์กับตัวแปรที่สังเกตได้ ในการศึกษาครั้งนี้ ผู้วิจัยจำลองข้อมูลโดยกำหนดให้ตัวแปร y_{ij} เป็นตัวแปรสังเกตได้เพื่อใช้เป็นตัวแปรทำนายค่าสูญหาย เมื่อมีการแจกแจงแบบปกติ $N(0,\sigma^2)$ โดยที่ σ^2 คือความคลาดเคลื่อนสุ่มของโมเดลในระดับโรงเรียน และแบ่งกลุ่มโรงเรียนตามระดับผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนออกเป็น 2 กลุ่ม คือกลุ่มโรงเรียน y_{ij} อยู่ในระดับสูง และกลุ่มโรงเรียน y_{ij} อยู่ในระดับต่ำ โดยกำหนดให้กลุ่มตัวอย่างที่มีระดับผลสัมฤทธิ์ทางการเรียนต่ำ มีความน่าจะเป็นในการไม่ให้ข้อมูลหรือไม่ตอบคำถามรายได้ผู้ปกครอง และสถานะการพักอาศัยอยู่กับบิดามารดา เป็นต้น โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขเท่ากับ 0.3, 0.4 และ 0.5 ตามลำดับ

(3) การสูญหายแบบไม่สุ่ม (MNAR): ลักษณะของข้อมูลสูญหายไม่ได้เกิดขึ้นอย่างสุ่มโดยค่าสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน จึงไม่สามารถอธิบายการสูญหายได้ด้วยตัวแปรสังเกตได้เพียงอย่างเดียว เนื่องจากการสูญหายในลักษณะนี้อาจมีความสัมพันธ์กับตัวแปรอื่นที่ไม่สามารถวัดได้จากค่าสังเกต โดยการศึกษานี้ผู้วิจัยสนใจการสูญหายที่ตัวแปรเดียวกัน จึงกำหนดให้ตัวแปร x_{ij} สูญหายมีความสัมพันธ์กับค่า x_{ij} ของตนเอง กล่าวคือ การที่นักเรียนไม่ตอบรายได้ของผู้ปกครองขึ้นอยู่รายได้ของผู้ปกครอง โดยกำหนดความน่าจะเป็นเท่ากับ 0.3, 0.4 และ 0.5 ตามลำดับ

(4) การสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม (MCAR-MAR): นำชุดข้อมูลสมบูรณ์จากการจำลองในขั้นตอนที่ 1 มาดำเนินการโดยกำหนดให้มีการสุ่มการสูญหายแบบสุ่มสมบูรณ์ (MCAR) ตามข้อที่ (1) โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขข้างต้นเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับ หลังจากนั้นนำชุดข้อมูลที่ได้มาสุ่มการสูญหายแบบสุ่ม

(MAR) ตามข้อที่ (2) โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขข้างต้นเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับ ดังนั้นจะได้ลักษณะการสูญหาย MCAR-MAR ด้วยอัตราการสูญหายของค่าสังเกตเท่ากับร้อยละ 30, ร้อยละ 30 และร้อยละ 50 ตามลำดับ

(5) การสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบไม่สุ่ม (MCAR-MNAR): นำชุดข้อมูลสมบูรณ์จากการจำลองในขั้นตอนที่ (1) มาดำเนินการกำหนดให้มีการสุ่มการสูญหายแบบสุ่มสมบูรณ์ (MCAR) โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขข้างต้นเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับ หลังจากนั้นนำชุดข้อมูลที่ได้สุ่มการสูญหายแบบไม่สุ่ม (MNAR) ตามข้อที่ (3) โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขข้างต้นเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับ ดังนั้นจะได้ลักษณะการสูญหาย MCAR-MNAR ด้วยอัตราการสูญหายของค่าสังเกตเท่ากับร้อยละ 30, ร้อยละ 40 และร้อยละ 50 ตามลำดับ

(6) การสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม (MAR-MNAR): นำชุดข้อมูลสมบูรณ์จากการจำลองในขั้นตอนที่ (2) มาดำเนินการโดยกำหนดให้มีการสุ่มการสูญหายแบบสุ่ม (MAR) โดยกำหนดความน่าจะเป็นของการไม่ตอบตามเงื่อนไขข้างต้นเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับ หลังจากนั้นนำชุดข้อมูลที่ได้สุ่มการสูญหายแบบไม่สุ่ม (MNAR) ตามข้อที่ (3) โดยกำหนดความน่าจะเป็นของการไม่ตอบเท่ากับ 0.15, 0.20 และ 0.25 ตามลำดับดังนั้นจะได้การสูญหาย MAR-MNAR ที่กำหนด

(7) การสูญหายแบบผสม (MCAR-MAR-MNAR): นำชุดข้อมูลสมบูรณ์มาดำเนินการตามการสูญหายในการจำลองตามขั้นตอนที่ (1) – (3) ตามลำดับ กำหนดความน่าจะเป็นของการไม่ตอบในเงื่อนไขแต่ละข้อ ให้มีอัตราการสูญหายของค่าสังเกตโดยรวมเฉลี่ยเท่ากับร้อยละ 30, ร้อยละ 40 และร้อยละ 50 ตามลำดับ จะได้การสูญหายแบบผสม MCAR-MAR-MNAR

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ขั้นตอนที่ 3 วิธีทดแทนค่าสูญหาย (Imputed data)

ประมาณค่าพารามิเตอร์ที่สูญหายด้วยวิธีทดแทนค่าสูญหาย 3 วิธี ได้แก่ วิธี MI-FCS, วิธี RF และวิธี Opt.impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่วิธี Opt.knn, วิธี Opt.tree, วิธี Opt.svm และวิธี Opt.cv

ผู้วิจัยนำข้อมูลที่ไม่สมบูรณ์ (Missing data) ที่ได้จากการจำลองข้อมูลภายใต้สถานการณ์ที่กำหนดตามขั้นตอนที่ (1) – (2) ข้างต้น มาทดแทนค่าสูญหายด้วยวิธีทดแทนค่าสูญหายจำนวน 3 วิธี ได้แก่ วิธี MI-FCS, วิธี RF และ วิธี Opt.impute ซึ่งประกอบด้วยวิธีย่อย 4 วิธี ได้แก่

วิธี Opt.knn, วิธี Opt.tree, วิธี Opt.svm และวิธี Opt.cv โดยใช้โปรแกรม R และโปรแกรม Julia ในการประมวลผลและคำนวณผลการประมาณค่าทดแทนค่าสูญหาย ทั้งนี้ในการทดแทนค่าสูญหายด้วยวิธี Opt.impute ได้รับการอนุญาตให้ใช้ Interface to 'Interpretable AI' Modules หรือ “iai” Package (AI software) สำหรับวิเคราะห์การทดแทนค่าสูญหาย Bertsimas et al. (2018) จากสถาบันเทคโนโลยีแมสซาชูเซตส์ (Massachusetts Institute of Technology: MIT) นอกจากนี้ในขั้นตอนของการทดแทนค่าสูญหายด้วยวิธี MI-FCS, วิธี RF ผู้วิจัยใช้ Package สำเร็จรูปในโปรแกรม R สำหรับการทดแทนค่าสูญหายแต่ละวิธี ประกอบด้วย “mice” Package และ “missForest” Package

ขั้นตอนที่ 4 เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย

เมื่อได้ผลจากวิธีทดแทนค่าสูญหายในแต่ละสถานการณ์เรียบร้อยแล้ว นำผลที่ได้ไปวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายจำนวน 3 วิธี ได้แก่ วิธี MI-FCS, วิธี MI-RF และวิธี Opt.impute โดยใช้เกณฑ์ 3 เกณฑ์ ได้แก่ ค่าเฉลี่ยของรากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมาตรฐาน (NRMSE) ใช้เพื่อตรวจสอบความถูกต้องในการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย ตามการจำลอง 100 รอบ เมื่อค่าเฉลี่ยของ NRMSE ที่มีค่าน้อยจะแสดงถึงวิธีทดแทนค่าสูญหาย ที่มีความถูกต้องและมีประสิทธิภาพสูง ค่าเฉลี่ยของค่าความเอนเอียงสัมพัทธ์ (RB) ใช้เพื่อตรวจสอบความถูกต้องแม่นยำ (preciseness) และความเอนเอียงของวิธีทดแทนค่าสูญหายโดยค่าเฉลี่ยของ RB ที่มีค่าน้อยจะแสดงถึงวิธีทดแทนค่าสูญหายที่มีคุณสมบัติด้านความแม่นยำและเที่ยงตรงสูง และค่าเฉลี่ยของค่าคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (MCSE) ใช้เพื่อตรวจสอบความคลาดเคลื่อนมาตรฐานและความคงเส้นคงวาของวิธีทดแทนค่าสูญหายในการจำลองข้อมูล 100 รอบ โดยค่าเฉลี่ยของ MCSE ที่มีค่าน้อยจะแสดงถึงมีความคลาดเคลื่อนมาตรฐาน จากการจำลองข้อมูล 100 รอบอยู่ในระดับต่ำ และมีคุณสมบัติด้านความคงเส้นคงวาสูง (consistency)

3. การตรวจสอบเงื่อนไขและความถูกต้องของโปรแกรมที่ใช้ในการจำลองข้อมูล

การวิจัยในครั้งนี้เป็นศึกษาประสิทธิภาพของวิธีทดแทนค่าสูญหายโดยใช้การจำลองข้อมูลภายใต้เงื่อนไขจำนวนทั้งสิ้น 378 เงื่อนไข ซึ่งสิ่งสำคัญสำหรับการศึกษาโดยแบบจำลองข้อมูล คือ การตรวจสอบความถูกต้องของเงื่อนไขที่ได้จำลองขึ้น ได้แก่ ลักษณะการสูญหาย 7 ลักษณะ คือ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MNAR, MAR-MNAR, และ MCAR -MAR-MNAR

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

อัตราการสูญหายของค่าสังเกตโดยกำหนดให้สูญหายที่ระดับที่หนึ่ง คือ ร้อยละ 30, ร้อยละ 40 และร้อยละ 50 ตามลำดับ เพื่อให้การดำเนินการจัดการข้อมูลสูญหายตามเงื่อนไขที่กำหนดสำหรับเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย 3 วิธี ได้แก่ วิธี MI-FCS, วิธี RF และวิธี Opt.impute ประกอบด้วยวิธีย่อย 4 วิธี ได้แก่ วิธี Opt.knn, วิธี Opt.tree, วิธี Opt.svm และวิธี Opt.cv ในข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ ให้มีความถูกต้องน่าเชื่อถือ โดยการตรวจสอบตรวจสอบเงื่อนไขและความถูกต้องของโปรแกรมที่ใช้ในการจำลองข้อมูลในครั้งนี้

ผู้วิจัยแบ่งการตรวจสอบเงื่อนไขตามหัวข้อ ได้แก่

3.1 การตรวจสอบลักษณะการสูญหาย

3.2 การตรวจสอบอัตราการสูญหายของค่าสังเกต รายละเอียดดังนี้

3.1 การตรวจสอบลักษณะการสูญหาย

เมื่อดำเนินการจำลองข้อมูลสูญหาย ในข้างต้นเรียบร้อยแล้ว เพื่อให้ข้อมูลสูญหายเป็นไปตามเงื่อนไขที่กำหนด ผู้วิจัยตรวจสอบความถูกต้องและตรวจสอบเงื่อนไขของแบบจำลองที่สร้างขึ้นโดยการใช้ตารางแสดงผลการตรวจสอบเงื่อนไขการสูญหายในการจำลองข้อมูล และใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation) เพื่อศึกษาความสัมพันธ์ของค่าสูญหายกับค่าสังเกตได้ภายใต้ประเภทของการสูญหายหลัก 3 รูปแบบ และใช้การสร้างแผนภาพด้วย package "naniar" ในโปรแกรม R เพื่อตรวจสอบระดับอัตราการสูญหายตามเงื่อนไขที่กำหนด มีรายละเอียดดังนี้

ตารางที่ 3. 2 แสดงผลการตรวจสอบเงื่อนไขการสูญหายในการจำลองข้อมูล

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

เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MCAR โดยใช้ Little's test of MCAR ในการทดสอบสมมติฐาน ดังนี้

H_0: การสูญหายในการจำลองข้อมูลเป็นแบบ MCAR

H_1: การสูญหายในการจำลองข้อมูลไม่เป็นแบบ MCAR

จากการพิจารณาผลการทดสอบในตารางที่ 3.2 พบว่า p-value = 0.838 จะเห็นว่าค่า p-value > 0.05 ดังนั้นจึงสามารถสรุปได้ว่าการสูญหายที่เกิดขึ้นในตัวแปร x_{ij} ในการจำลองข้อมูลในสถานการณ์นี้ เป็นไปตามเงื่อนไขที่กำหนด

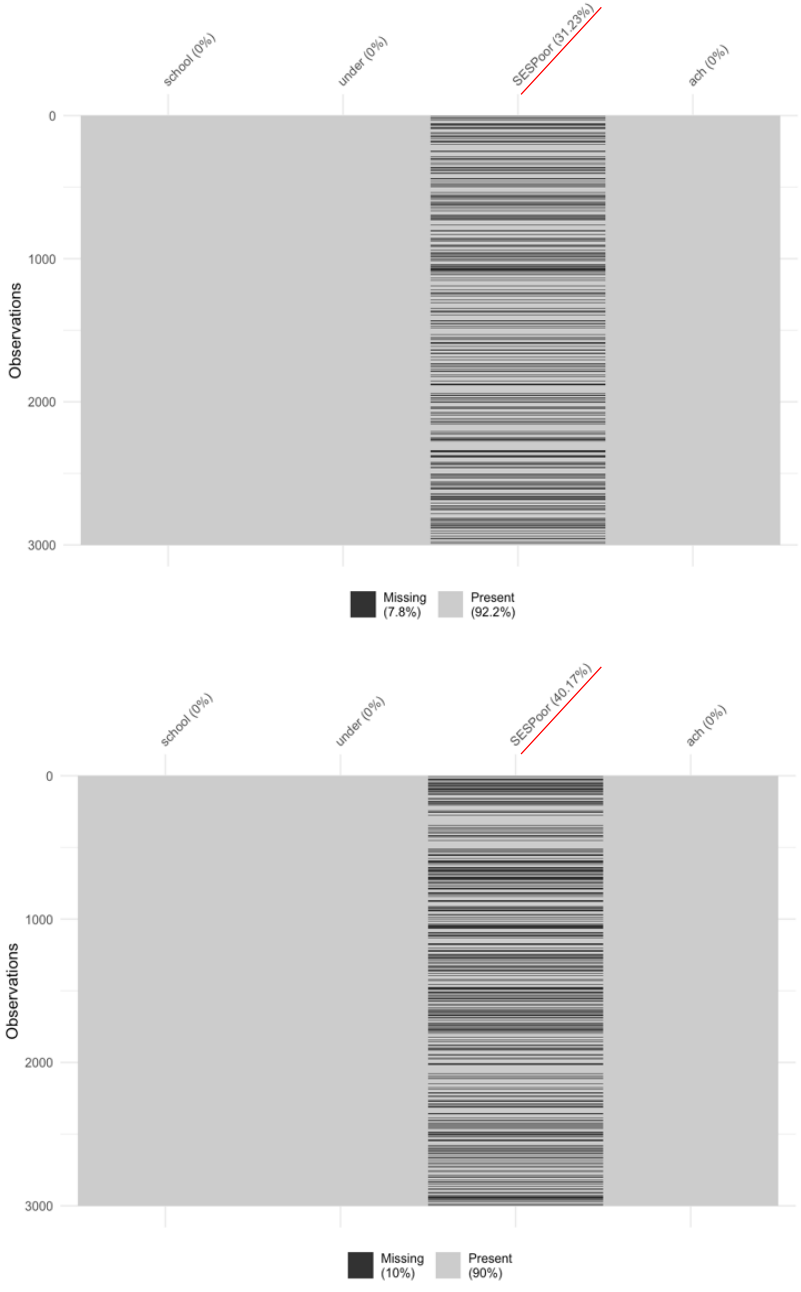
เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MAR ผู้วิจัยดำเนินการแปลงตัวแปร x_{ij} ที่สูญหายให้มีค่าเท่ากับ 1 และค่ารวบรวมได้ในตัวแปร x_{ij} ให้เท่ากับ 0 หลังจากนั้นทดสอบความสัมพันธ์ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสูญหายในตัวแปร x_{ij} กับค่ารวบรวมได้ในตัวแปร y_{ij} จากตารางที่ 3.2 พบว่ามีความสัมพันธ์ในทิศทางบวกเท่ากับ 0.3202 ดังนั้นจึงสามารถสรุปได้ว่าเกิดการสูญหายแบบมีเงื่อนไขหรือการสูญหายแบ MAR ดังนั้นการจำลองข้อมูลในสถานการณ์นี้เป็นไปตามเงื่อนไขที่กำหนด

เมื่อพิจารณาผลการตรวจสอบเงื่อนไขการจำลองข้อมูลสูญหายแบบ MNAR ด้วยการทดสอบสมมติฐาน โดยใช้ Little's test of MCAR จากตารางที่ 3.2 ผลทดสอบพบว่าปฏิเสธสมมติฐานกล่าวคือข้อมูลที่จำลองขึ้นมีลักษณะการสูญหายแบบ MCAR อย่างมีนัยสำคัญทางสถิติ (p-value = 0.000) กล่าวคือข้อมูลที่จำลองขึ้นไม่ได้เกิดขึ้นอย่างสุ่มสมบูรณ์ เมื่อพิสูจน์ได้แล้วว่าการจำลองข้อมูลสูญหายไม่ได้เป็น MCAR อย่างไรก็ตามการสูญหายแบบ MNAR เป็นการสูญหายที่มีสาเหตุมาจากตัวแปรที่เราไม่ทราบค่า หรือไม่มีข้อมูลให้นำมาวิเคราะห์ได้ ดังนั้นในทางปฏิบัติจึงทำได้เพียงตรวจสอบจากข้อมูลที่เก็บรวบรวมได้ว่ามีแนวโน้มที่จะมีการสูญหายแบบ MAR หรือไม่ หากผลการตรวจสอบอธิบายการสูญหายแบบ MAR ได้น้อย จะทำให้สามารถสรุปได้ว่าข้อมูลมีแนวโน้มที่จะเป็น MNAR มากขึ้น โดยพิจารณาความสัมพันธ์ของข้อมูลที่รวบรวมได้กับ ข้อมูลสูญหาย ในตัวแปร x_{ij} พบว่า ค่าสูญหายที่เกิดขึ้นมีความสัมพันธ์กับค่าที่รวบรวมได้ y_{ij} เพียง 0.002 ดังนั้นสรุปได้ว่ามีแนวโน้มที่ข้อมูลจำลองในกรณีนี้เป็นไปตามเงื่อนไขการสูญหายแบบ MNAR

ทั้งนี้ในส่วนของการตรวจสอบลักษณะการสูญหายแบบผสม 3 รูปแบบ ได้แก่ MCAR - MAR, MCAR - MNAR และ MAR - MNAR ผู้วิจัยมีหลักการตรวจสอบเช่นเดียวกับลักษณะการสูญหายหลัก 3 ลักษณะ ปรากฏตารางที่ 3.2

3.2 การตรวจสอบอัตราการสูญหายของค่าสังเกต

เมื่อพิจารณาการตรวจสอบลักษณะการสูญหายของข้อมูลในการทดสอบข้างต้นว่าถูกต้องแล้ว ผู้วิจัยจึงตรวจสอบความถูกต้องของอัตราการสูญหายในแต่ละระดับทั้งหมด 3 ระดับ ได้แก่ ร้อยละ 30 ร้อยละ 40 และร้อยละ 50 ตามลำดับ โดยในการจำลองข้อมูลในแต่ละกรณี ผู้วิจัยตรวจสอบอัตราการ สูญหายด้วยการสร้างแผนภาพด้วย package "naniar" ในโปรแกรม R เพื่อตรวจสอบอัตราการสูญหายที่เกิดขึ้นว่าถูกต้องโดยจะดำเนินการตรวจสอบก่อนทดแทนค่าสูญหายในแต่ละกรณี ทั้งนี้ขอแสดงตัวอย่างแผนภาพอัตราการสูญหายทั้งหมด 3 ระดับ แสดงดังรูปที่ 3.1.2

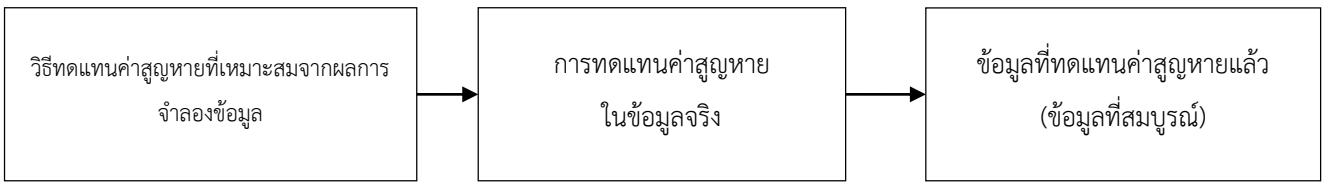


รูปที่ 3.1. 2 การตรวจสอบระดับอัตราการสูญหายโดยใช้การจำลองข้อมูล

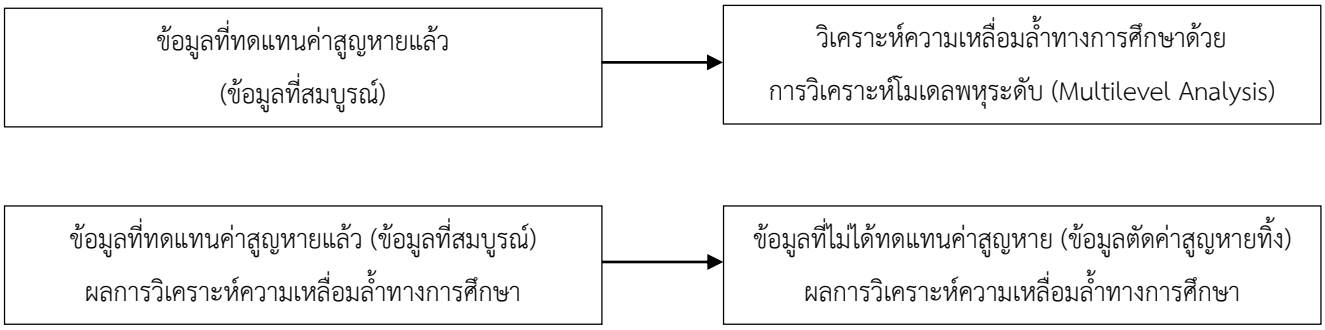
รูปที่ 3.1.2 การตรวจสอบระดับอัตราการสูญหายโดยใช้การจำลอง

ตอนที่ 3 การประยุกต์ใช้กับข้อมูลจริง

ผู้วิจัยดำเนินการนำวิธีทดแทนข้อมูลสูญหายที่ได้จากผลการวิจัยระยะที่หนึ่งมาประยุกต์ใช้กับข้อมูลจริง โดยข้อมูลจริงที่ใช้ในการวิจัยครั้งนี้ คือ ข้อมูลทุติยภูมิจากสถาบันทดสอบทางการศึกษาแห่งชาติ หรือ สทศ. ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ระดับโรงเรียน จำนวน 2,109 โรงเรียน ที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) จำนวน 42 สังกัด ปีการศึกษา 2563 ประกอบด้วย ข้อมูลผลการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน (O-NET) แทนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน และข้อมูลด้านสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน หลังจากนั้นนำข้อมูลที่ทดแทนค่าสูญหายแล้วมาวิเคราะห์ความเหลื่อมล้ำทางการศึกษาด้วยการวิเคราะห์โมเดลพหุระดับ ซึ่งสามารถสรุปขั้นตอนได้ดังรูปที่ 3.1.3 ถึง 3.1.4



รูปที่ 3.1. 3 การประยุกต์ใช้กับข้อมูลจริง



รูปที่ 3.1. 4 การวิเคราะห์ความเหลื่อมล้ำทางการศึกษา

การวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ใช้การวิเคราะห์โมเดลพหุระดับโดยกำหนดให้มีโครงสร้างของข้อมูล 2 ระดับ ได้แก่ ระดับโรงเรียน และระดับสังกัด นำมาวิเคราะห์อิทธิพลสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียนระดับโรงเรียน โดยพิจารณาจากค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) ที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน หากผลการวิเคราะห์พบอิทธิพลเท่ากับศูนย์หมายถึงเกิดความเสมอภาคหรือมีความเท่าเทียมกัน กล่าวคือไม่พบความเหลื่อมล้ำทางการศึกษา ขณะที่หากอิทธิพลไม่เท่ากับศูนย์จะหมายถึงเกิดความไม่เสมอภาคด้านการศึกษา สะท้อนความเหลื่อมล้ำทางการศึกษาอันเนื่องมาจากสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน หลังจากนั้นเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย รูปที่ 3.1.3

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย จำนวน 3 วิธี ได้แก่ วิธีทดแทนค่าสูญหาย MI-FCS, วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Otp.impute ประกอบด้วย วิธี Otp.knn, วิธี Otp.tree, วิธี Otp.svm และวิธี Otp.cv ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา โดยใช้การจำลองข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับ ในโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) โดยมีเงื่อนไขในการศึกษาดังนี้ ลักษณะการสูญหาย 7 ลักษณะ ได้แก่ MCAR, MAR, MNAR, MCAR-MAR, MCAR-MAR, MAR-MNAR และ MCAR-MAR-MNAR ขนาดตัวอย่างระดับที่หนึ่งเท่ากับ 1000, 2000, 3000 หน่วย และขนาดตัวอย่างระดับที่สองเท่ากับ 40, 50, 60 หน่วย มีอัตราการสูญหายในระดับที่หนึ่งเท่ากับร้อยละ 30, ร้อยละ 40 และร้อยละ 50 ตามลำดับ เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ β_1 โดยใช้เกณฑ์ค่าเฉลี่ยของรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย มาตรฐาน (NRMSE) ค่าเฉลี่ยของความเอนเอียงสัมพัทธ์ (RB) แทนค่าความถูกต้องของวิธีทดแทนค่าสูญหายในค่าประมาณค่าพารามิเตอร์ β_1 และค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐานแบบมอนติคาร์โลสัมพัทธ์ (MCSE) สำหรับการประมาณค่าความคลาดเคลื่อนมาตรฐาน ซึ่งอาศัยการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล (Monte Carlo Simulation) กำหนดการทำซ้ำ 100 รอบในแต่ละเงื่อนไข โดยสูตรที่ใช้ในการคำนวณเกณฑ์ทั้งสามได้กล่าวไว้ในบทที่ 2 และจะนำวิธีทดแทนค่าสูญหายที่มีประสิทธิภาพที่ได้จากผลการจำลองข้อมูลมาประยุกต์ใช้กับข้อมูลจริงทางการศึกษา เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษาและเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ดังนั้นเพื่อให้เกิดความเข้าใจได้อย่างสะดวกและรวดเร็ว ผู้วิจัยจึงนำเสนอผลการวิจัยโดยใช้สัญลักษณ์ต่าง ๆ ในตาราง รูปภาพ และสรุปผลการวิจัยโดยมีความหมาย ดังนี้

β_1	หมายถึง พารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่มในการประมาณค่าของความเหลื่อมล้ำทางการศึกษาของนักเรียนระดับโรงเรียน
MI – FCS	หมายถึง วิธีทดแทนค่าสูญหาย Multiple Imputation Fully Conditional Specification

RF	หมายถึง วิธีทดแทนค่าสูญหาย Random Forest
Opt.impute	หมายถึง วิธีทดแทนค่าสูญหาย Optimal Impute ได้แก่ วิธี Opt.knn, Opt.svm และ Opt.tree
Opt.knn	หมายถึง วิธีทดแทนค่าสูญหาย Optimal Impute K-Nearest Neighbors based
Opt.svm	หมายถึง วิธีทดแทนค่าสูญหาย Optimal Impute Support Vector Machines based
Opt.tree	หมายถึง วิธีทดแทนค่าสูญหาย Optimal Impute Decision Tree based
Opt.cv	หมายถึง วิธีทดแทนค่าสูญหาย Optimal Impute cross – validated
NRMSE	หมายถึง เกณฑ์ค่าเฉลี่ยของรากที่สองของความคลาดเคลื่อน กำลังสองเฉลี่ยมาตรฐาน
RB	หมายถึง เกณฑ์ค่าเฉลี่ยของความเอนเอียงสัมพัทธ์
MCSE	หมายถึง เกณฑ์ค่าเฉลี่ยของความคลาดเคลื่อนมาตรฐาน แบบมอนติคาร์โลสัมพัทธ์
MCAR	หมายถึง ลักษณะการสูญหายแบบสุ่มสมบูรณ์
MAR	หมายถึง ลักษณะการสูญหายแบบสุ่ม
MNAR	หมายถึง ลักษณะการสูญหายแบบไม่สุ่ม
MCAR - MAR	หมายถึง ลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม
MCAR - MNAR	หมายถึง ลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับ การสูญหายแบบไม่สุ่ม
MAR - MNAR	หมายถึง ลักษณะการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม
MCAR – MAR – MNAR	หมายถึง ลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม และการสูญหายแบบไม่สุ่ม
N_1	หมายถึง ขนาดตัวอย่างในระดับที่ 1
N_2	หมายถึง ขนาดตัวอย่างในระดับที่ 2
M	หมายถึง อัตราการสูญหายของตัวอย่างในระดับที่ 1

การนำเสนอผลการวิเคราะห์ข้อมูลในบทนี้ ผู้วิจัยแบ่งการนำเสนอออกเป็น 3 ตอน ได้แก่ ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายโดยใช้การจำลองข้อมูล ตอนที่ 2 ผลการวิเคราะห์สถิติพื้นฐานและการประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริงและตอนที่ 3 ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาด้วยโมเดลพหุระดับ ซึ่งมีรายละเอียดดังต่อไปนี้

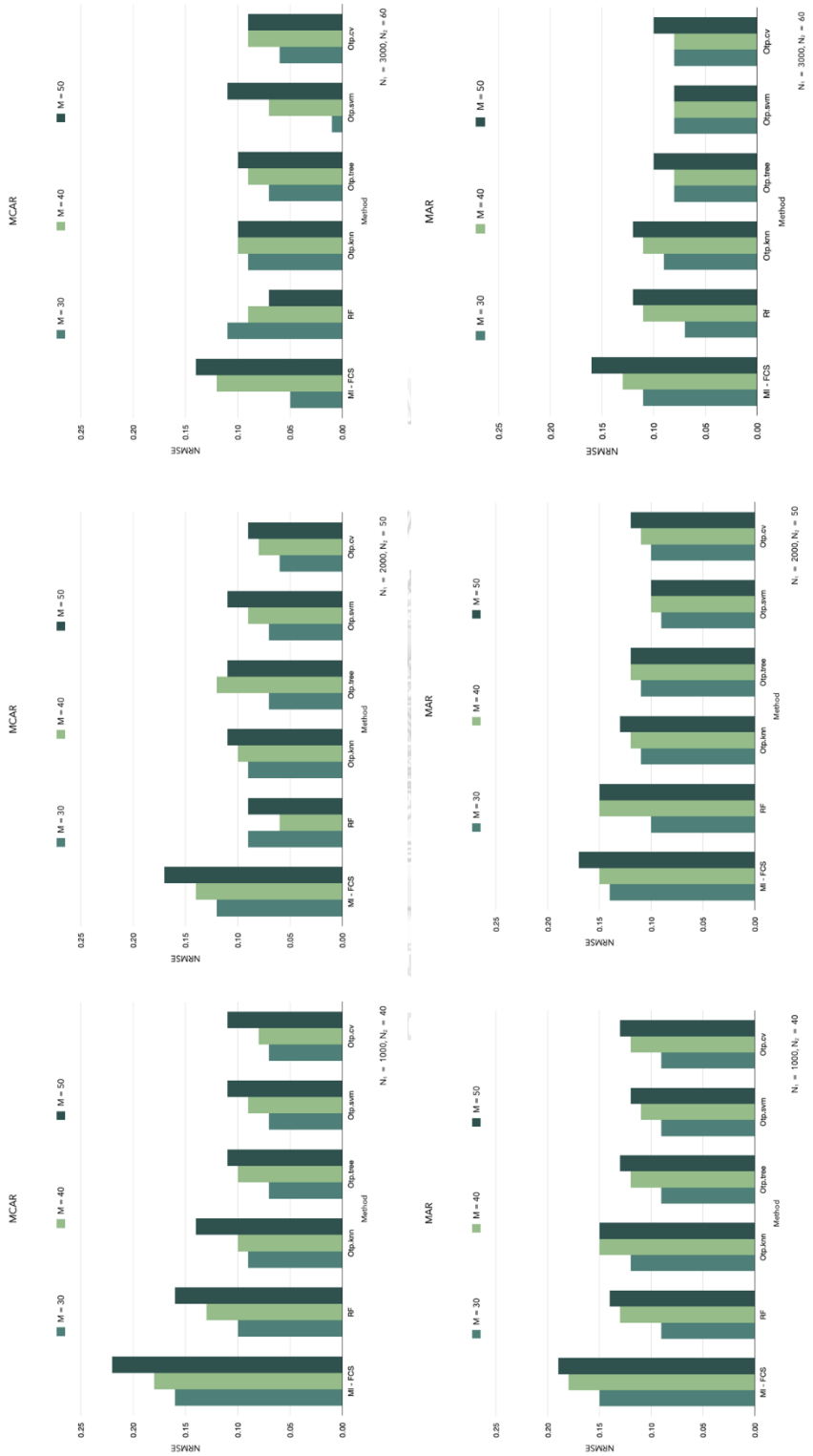
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 4.1 แสดงค่าเฉลี่ย NRMSE ของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1)

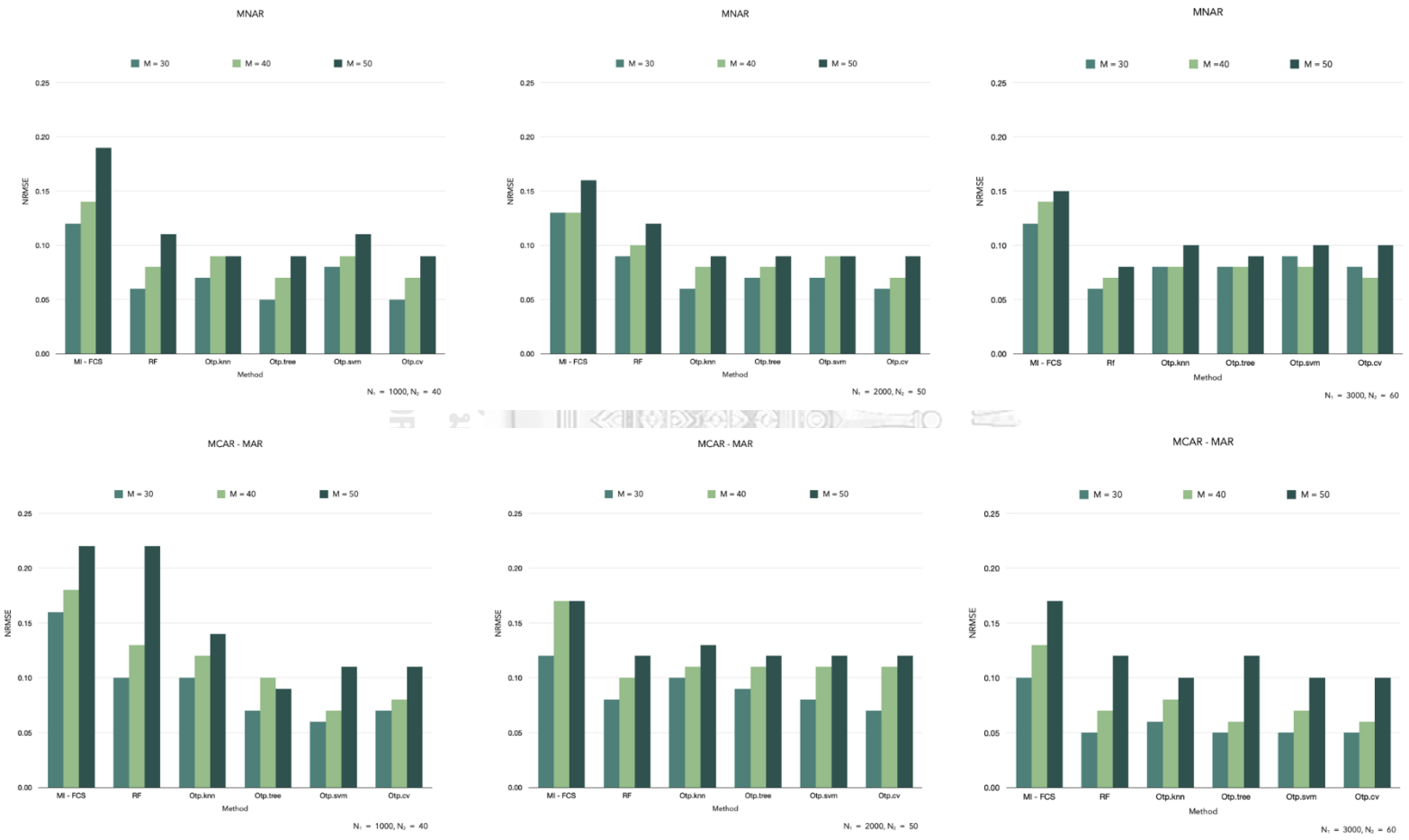
Methods	N_1	N_2	m	MCAR	MAR	MNAR	MCAR - MAR	MCAR - MNAR	MAR - MNAR	MCAR – MAR - MNAR
				NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)
MI - FCS	1000	40	30	0.16	0.15	0.12	0.16	0.14	0.14	0.14
			40	0.18	0.18	0.14	0.18	0.16	0.18	0.17
			50	0.22	0.19	0.19	0.22	0.20	0.19	0.19
	2000	50	30	0.12	0.14	0.13	0.12	0.15	0.18	0.11
			40	0.14	0.15	0.13	0.17	0.16	0.17	0.15
			50	0.17	0.17	0.16	0.17	0.17	0.16	0.19
	3000	60	30	0.05	0.11	0.12	0.10	0.12	0.17	0.11
			40	0.12	0.13	0.14	0.13	0.12	0.20	0.13
			50	0.14	0.16	0.15	0.17	0.17	0.22	0.14
RF	1000	40	30	0.10	0.09	0.06	0.10	0.09	0.10	0.08
			40	0.13	0.13	0.08	0.13	0.11	0.12	0.13
			50	0.16	0.14	0.11	0.22	0.13	0.14	0.15
	2000	50	30	0.09	0.10	0.09	0.08	0.08	0.14	0.06
			40	0.06	0.15	0.10	0.10	0.09	0.11	0.11
			50	0.09	0.15	0.12	0.12	0.09	0.15	0.10
	3000	60	30	0.11	0.07	0.06	0.05	0.10	0.08	0.07
			40	0.09	0.11	0.07	0.07	0.11	0.15	0.08
			50	0.07	0.12	0.08	0.12	0.11	0.19	0.08

Methods	N_1	N_2	m	MCAR	MAR	MNAR	MCAR - MAR	MCAR - MNAR	MAR - MNAR	MCAR – MAR - MNAR
				NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)
Opt.knn	1000	40	30	0.09	0.12	0.07	0.10	0.09	0.04	0.09
			40	0.10	0.15	0.09	0.12	0.10	0.06	0.13
			50	0.14	0.15	0.09	0.14	0.11	0.08	0.14
	2000	50	30	0.09	0.11	0.06	0.10	0.09	0.05	0.09
			40	0.10	0.12	0.08	0.11	0.11	0.05	0.12
			50	0.11	0.13	0.09	0.13	0.12	0.06	0.14
	3000	60	30	0.09	0.09	0.08	0.06	0.02	0.07	0.09
			40	0.10	0.11	0.08	0.08	0.07	0.08	0.10
			50	0.10	0.12	0.10	0.10	0.09	0.09	0.11
Opt.tree	1000	40	30	0.07	0.09	0.05	0.07	0.07	0.05	0.09
			40	0.10	0.12	0.07	0.10	0.09	0.10	0.13
			50	0.11	0.13	0.09	0.09	0.11	0.11	0.13
	2000	50	30	0.07	0.11	0.07	0.09	0.08	0.04	0.08
			40	0.12	0.12	0.08	0.11	0.12	0.05	0.12
			50	0.11	0.12	0.09	0.12	0.11	0.06	0.13
	3000	60	30	0.07	0.08	0.08	0.05	0.05	0.08	0.08
			40	0.09	0.08	0.08	0.06	0.07	0.11	0.11
			50	0.10	0.10	0.09	0.12	0.09	0.10	0.11

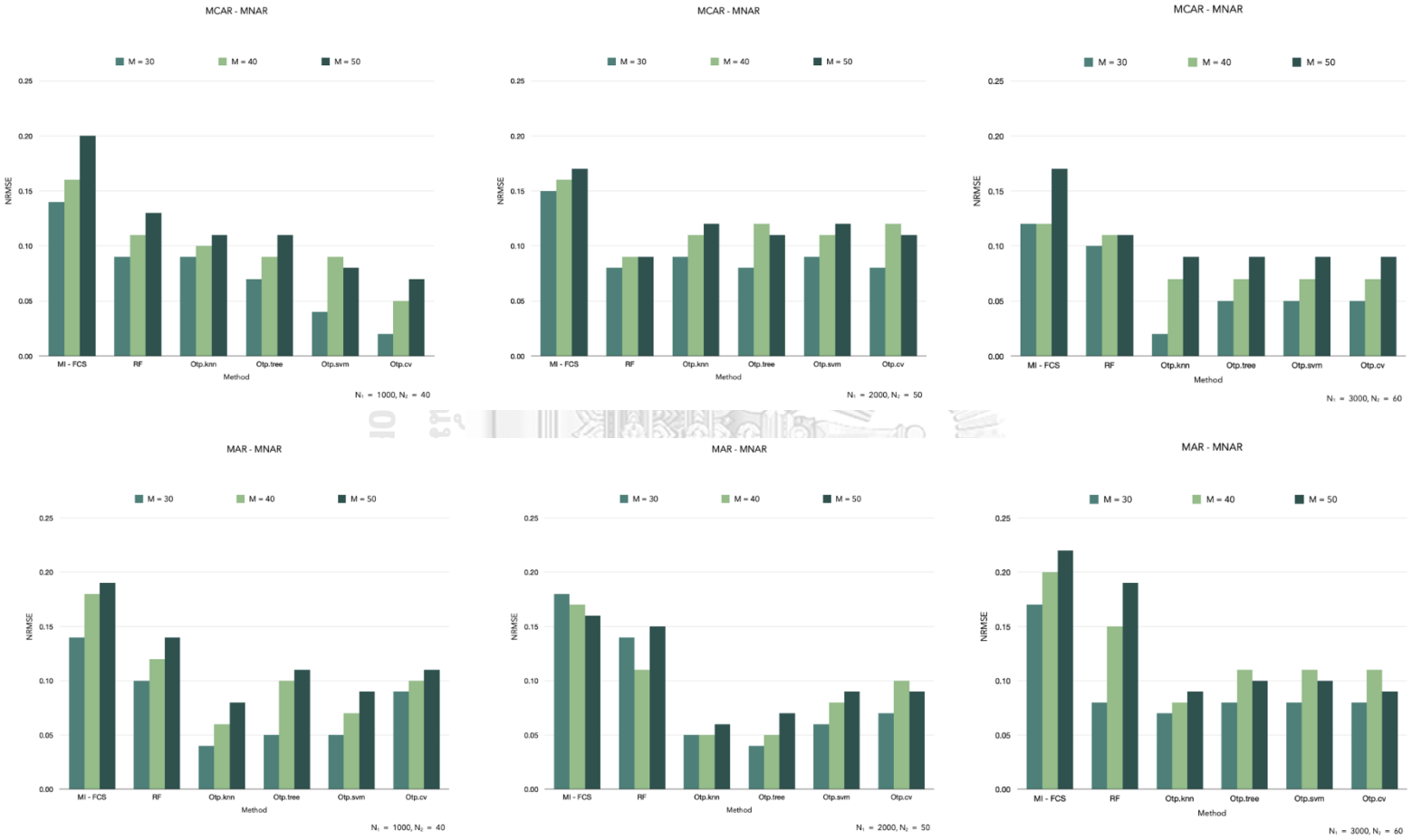
Methods	N_1	N_2	m	MCAR	MAR	MNAR	MCAR - MAR	MCAR - MNAR	MAR - MNAR	MCAR – MAR - MNAR
				NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)	NRMSE (β_1)
Opt.svm	1000	40	30	0.07	0.09	0.08	0.06	0.04	0.05	0.08
			40	0.09	0.11	0.09	0.07	0.09	0.07	0.12
			50	0.11	0.12	0.11	0.11	0.08	0.09	0.15
	2000	50	30	0.07	0.09	0.07	0.08	0.09	0.06	0.08
			40	0.09	0.10	0.09	0.11	0.11	0.08	0.11
			50	0.11	0.10	0.09	0.12	0.12	0.09	0.13
	3000	60	30	0.01	0.08	0.09	0.05	0.05	0.08	0.09
			40	0.07	0.08	0.08	0.07	0.07	0.11	0.11
			50	0.11	0.08	0.10	0.10	0.09	0.10	0.11
Opt.cv	1000	40	30	0.07	0.09	0.05	0.07	0.02	0.09	0.08
			40	0.08	0.12	0.07	0.08	0.05	0.10	0.09
			50	0.11	0.13	0.09	0.11	0.07	0.11	0.14
	2000	50	30	0.06	0.10	0.06	0.07	0.08	0.07	0.07
			40	0.08	0.11	0.07	0.11	0.12	0.10	0.11
			50	0.09	0.12	0.09	0.12	0.11	0.09	0.12
	3000	60	30	0.06	0.08	0.08	0.05	0.05	0.09	0.08
			40	0.09	0.08	0.07	0.06	0.07	0.11	0.09
			50	0.09	0.10	0.10	0.10	0.09	0.09	0.10



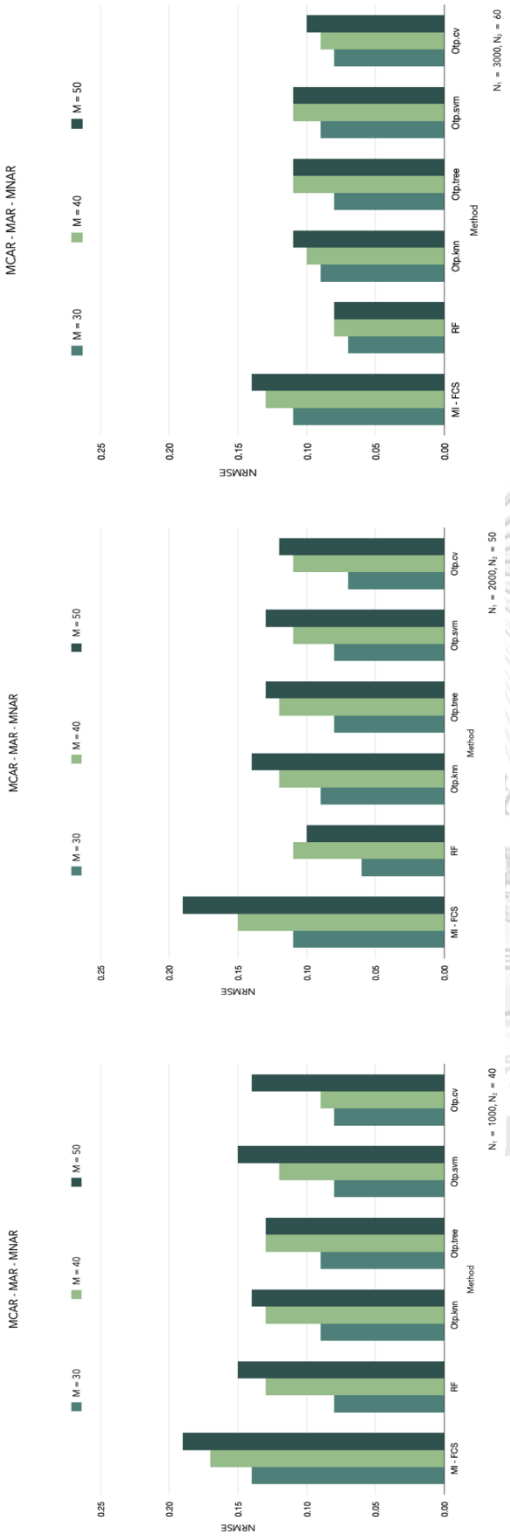
รูปที่ 4.1. 1 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม เมื่อข้อมูลมีการสูญหายแบบ MCAR, MAR จำแนกตามขนาดตัวอย่างและอัตราการสูญหาย



รูปที่ 4.1. 2 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม เมื่อข้อมูลมีการสูญหายแบบ MNAR, MCAR - MAR จำแนกตามขนาดตัวอย่างและอัตราการสูญหาย



เมื่อข้อมูลมีการสูญหายแบบ MCAR - MNAR, MAR - MNAR จำแนกตามขนาดตัวอย่างและอัตราการสูญหาย



รูปที่ 4.1. 4 แสดงการเปรียบเทียบค่า NRMSE ของวิธีทดแทนค่าสูญหาย ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบกลุ่ม
เมื่อข้อมูลมีการสูญหายแบบ MCAR - MAR - MNAR จำแนกตามขนาดตัวอย่างและอัตราการสูญหาย

ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล

1.1 การเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย NRMSE

จากตารางที่ 4.1 และแผนภูมิแท่งในรูปที่ 4.1.1 ถึง 4.1.4 สามารถสรุปผลวิเคราะห์ได้ ดังนี้

เมื่อพิจารณาภาพรวมของผลการวิเคราะห์ประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย NRMSE พบว่าวิธีทดแทนค่าสูญหาย MI – FCS ให้ค่าเฉลี่ย NRMSE สูงที่สุดเกือบทุกรณี รองมาคือ วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Otp.impute ตามลำดับ โดยจะสังเกตได้ว่า เมื่อขนาดตัวอย่างและอัตราการสูญหายเพิ่มสูงขึ้น มีแนวโน้มที่จะทำให้ค่าเฉลี่ย NRMSE เพิ่มสูงขึ้น ทั้งนี้หากพิจารณาประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ β_1 ในโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม จำแนกตามลักษณะการสูญหาย 7 ลักษณะ สามารถสรุปผลได้ดังนี้

1.1.1 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ (MCAR)

เมื่อพิจารณาการสูญหายแบบสุ่มสมบูรณ์ หรือ MCAR พบว่าค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Opt.impute ให้ค่าเฉลี่ย NRMSE น้อยที่สุด ซึ่งอยู่ในช่วง 0.01 ถึง 0.14 โดยจะเห็นว่าวิธีทดแทนค่าสูญหาย Opt.knn, Opt.tree, Opt.svm และ Opt.cv ให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกันมาก ขณะที่วิธีทดแทนค่าสูญหาย RF มีค่าเฉลี่ย NRMSE ระหว่างช่วง 0.06 ถึง 0.16 และวิธีทดแทนค่าสูญหาย MI – FCS มีค่าเฉลี่ย NRMSE อยู่ระหว่างช่วง 0.05 ถึง 0.22 จะสังเกตได้ว่า โดยรวมวิธีทดแทนค่าสูญหาย MI – FCS มีแนวโน้มให้ค่าเฉลี่ย NRMSE สูงที่สุด เมื่อเทียบกับวิธีทดแทนค่าสูญหาย Opt.impute และวิธี RF ตามลำดับ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.1.2 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่ม (MAR)

เมื่อพิจารณาการสูญหายแบบสุ่ม หรือ MAR จากแผนภูมิแท่งรูปที่ 4.1.1 พบว่าค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหายทั้งสามวิธีให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกันมาก แต่หากพิจารณาโดยให้ขนาดตัวอย่างคงที่ จะพบว่าค่าเฉลี่ย NRMSE มีแนวโน้มเพิ่มสูงขึ้นเมื่ออัตราการสูญหายเพิ่มขึ้น โดยจะสังเกตได้ว่าวิธีทดแทนค่าสูญหาย RF มีแนวโน้มให้ค่าเฉลี่ย NRMSE ต่ำที่สุด ซึ่งอยู่ระหว่าง 0.07 ถึง 0.15 รองลงมาคือวิธีทดแทนค่าสูญหาย Opt.impute ให้ค่าเฉลี่ย NRMSE อยู่ระหว่าง 0.08 ถึง 0.15 โดยที่วิธี Opt.cv Opt.svm, Opt.tree, และ Opt.knn ให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกัน ขณะที่วิธีทดแทนค่าสูญหาย MI – FCS ให้ค่าเฉลี่ย NRMSE สูงที่สุดในช่วง 0.11 ถึง 0.19

1.1.3 เมื่อข้อมูลมีลักษณะการสูญหายแบบไม่สุ่ม (MNAR)

เมื่อพิจารณาการสูญหายแบบไม่สุ่ม หรือ MNAR จะพบว่าค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Opt.impute อยู่ระหว่างช่วง 0.05 ถึง 0.11 และวิธี RF มีค่าเฉลี่ย NRMSE ระหว่าง 0.05 ถึง 0.09 จะสังเกตเห็นว่าให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกันมากนัก ขณะที่วิธีทดแทนค่าสูญหาย MI – FCS มีแนวโน้มให้ค่าเฉลี่ย NRMSE สูงที่สุด ซึ่งอยู่ระหว่างช่วง 0.13 ถึง 0.19 ตามลำดับ นอกจากนี้จะสังเกตได้ว่าจากรูปที่ 4.1.2 เมื่อขนาดตัวอย่างเพิ่มสูงขึ้น จะสังเกได้ว่ามีแนวโน้มให้ค่าเฉลี่ย NRMSE ลดลง

1.1.4 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม (MCAR - MAR)

จากการพิจารณาค่าเฉลี่ย NRMSE เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่ม โดยให้ขนาดตัวอย่างคงที่ จะสังเกตได้ชัดว่าค่าเฉลี่ย NRMSE มีแนวโน้มสูงขึ้นเมื่ออัตราการสูญหายเพิ่มสูงขึ้น และเมื่อพิจารณาค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Opt.impute จะพบว่าวิธี Opt.knn ให้ค่าเฉลี่ย NRMSE อยู่ในช่วง 0.06 ถึง 0.14 ขณะที่วิธี Opt.tree วิธี Opt.svm และวิธี Opt.cv ให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกันมาก ซึ่งอยู่ระหว่าง 0.05 ถึง 0.12 ตามลำดับ นอกจากนี้เมื่อพิจารณาค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย RF จะพบว่าให้ค่าเฉลี่ย NRMSE อยู่ในช่วง 0.05 ถึง 0.22 และวิธีทดแทนค่าสูญหาย MI – FCS ให้ค่าเฉลี่ย NRMSE อยู่ในช่วง 0.10 ถึง 0.22 จะเห็นว่าวิธีทดแทนค่าสูญหาย Opt.impute มีแนวโน้มให้ค่าเฉลี่ย NRMSEต่ำที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และวิธี MI – FCS ตามลำดับ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.1.5 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบไม่สุ่ม (MCAR - MNAR)

พิจารณาผลการวิเคราะห์จากตารางที่ 4.1 เมื่อการสูญหายเกิดขึ้นแบบผสม MCAR - MNAR พบว่าค่าเฉลี่ย NRMSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Opt.impute ในขนาดตัวอย่างระดับที่หนึ่งและระดับที่สองเท่ากับ 1000 และ 40 หน่วย ให้ค่าเฉลี่ย NRMSE อยู่ระหว่างช่วง 0.02 ถึง 0.12 โดยที่วิธี Opt.cv มีแนวโน้มให้ค่าเฉลี่ย NRMSE ต่ำที่สุด รองลงมาคือวิธี Opt.svm, Opt.tree วิธี Opt.knn และ วิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย NRMSE อยู่ในช่วง 0.09 ถึง 0.13 ขณะที่วิธีทดแทนค่าสูญหาย MI – FCS ให้ค่าเฉลี่ย NRMSE สูงที่สุดอยู่ในช่วง 0.14 ถึง 0.20 และมีแนวโน้มสูงขึ้นเมื่ออัตราการสูญหายและขนาดอย่างเพิ่มสูงขึ้น สังเกตได้จากรูปที่ 4.1.3

1.1.6 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม (MAR - MNAR)

เมื่อพิจารณาการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม หรือ MAR - MNAR จากตารางที่ 4.1 และแผนภูมิแท่งรูปที่ 4.1.3 โดยให้ขนาดตัวอย่างคงที่ พบว่าวิธีทดแทนค่าสูญหาย Opt.impute มีค่าเฉลี่ย NRMSE อยู่ระหว่าง 0.04 ถึง 0.11 วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI – FCS ให้ค่าเฉลี่ย NRMSE อยู่ในช่วง 0.08 ถึง 0.19 และ 0.14 ถึง 0.22 ตามลำดับนอกจากนี้เมื่อพิจารณาวิธีทดแทนค่าสูญหาย Opt.impute ในตัวอย่างระดับที่หนึ่งและระดับที่สองเท่ากับ 1000 และ 40 หน่วยจะสังเกตได้ว่าวิธี Opt.knn มีแนวโน้มให้ค่าเฉลี่ย NRMSE ต่ำที่สุด รองลงมาคือ Opt.svm, Opt.tree และ Opt.cv ทั้งนี้จะสังเกตได้ว่าค่าเฉลี่ย NRMSE มีแนวโน้มเพิ่มสูงขึ้นเมื่ออัตราการสูญหายเพิ่มสูงขึ้น

1.1.7 เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมกับการสูญหายแบบสุ่มและการสูญหายแบบไม่สุ่ม (MCAR – MAR – MNAR)

จากการพิจารณาค่าเฉลี่ย NRMSE เมื่อข้อมูลมีลักษณะการสูญหายแบบสุ่มสมบูรณ์ผสมสามลักษณะ จากแผนภูมิแท่งรูปที่ 4.1.3 เมื่อให้ขนาดตัวอย่างคงที่จะสังเกตได้ว่าค่าเฉลี่ย NRMSE มีแนวโน้มสูงขึ้นเมื่ออัตราการสูญหายเพิ่มสูงขึ้น โดยวิธีทดแทนค่าสูญหาย Opt.impute และวิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย NRMSE ไม่แตกต่างกัน อยู่ในช่วง 0.07 ถึง 0.15 และ 0.06 ถึง 0.15 ตามลำดับ ขณะที่วิธีทดแทนค่าสูญหาย MI – FCS มีแนวโน้มให้ค่าเฉลี่ย NRMSE สูงที่สุดอยู่ในช่วง 0.11 ถึง 0.19 ตามลำดับ จะเห็นว่าวิธีทดแทนค่าสูญหาย MI – FCS มีแนวโน้มให้ค่าเฉลี่ย NRMSE สูงที่สุด เป็นต้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จากการพิจารณาประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย NRMSE จำแนกตามลักษณะการสูญหาย 7 ลักษณะ ที่ได้กล่าวมาในข้างต้น ทำให้สามารถสรุปได้ว่าในภาพรวมวิธีทดแทนค่าสูญหาย Opt.impute มีแนวโน้มให้ค่าเฉลี่ย NRMSE ต่ำที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI ถึง FCS ตามลำดับ โดยที่จะสังเกตได้ว่าเมื่อขนาดตัวอย่างและอัตราการสูญหายเพิ่มสูงขึ้นจะทำให้ค่าเฉลี่ย NRMSE จะมีแนวโน้มเพิ่มขึ้นสูงขึ้น ตามลำดับ ผลการวิเคราะห์ปรากฏดังตารางที่ 4.1 และรูปที่ 4.1.1 - 4.1.3

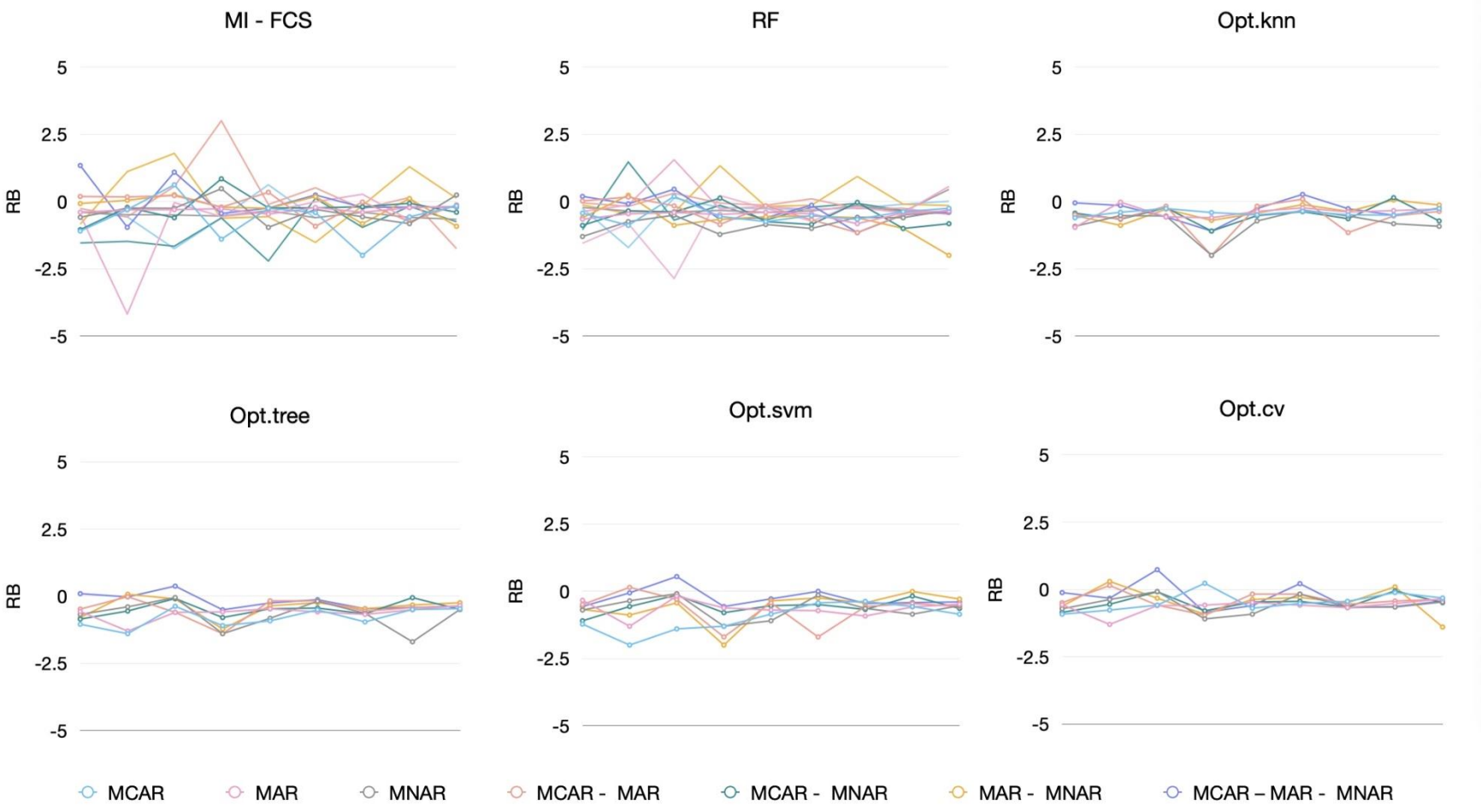
ตารางที่ 4. 2 แสดงค่าเฉลี่ย RB, MCSE ของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1)

Methods	N_1	N_2	m	MCAR		MAR		MNAR		MCAR - MAR		MCAR - MNAR		MAR - MNAR		MCAR – MAR - MNAR	
				RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)
MI - FCS	1000	40	30	-0.08	3.37	-0.40	3.24	-0.26	3.90	-1.54	3.90	-0.09	3.80	-0.84	3.83	21.82	3.26
			40	-0.50	3.38	-0.49	3.24	-0.57	3.89	-1.48	3.55	-0.61	3.58	1.11	3.89	-0.11	3.54
			50	-4.18	3.74	-0.50	3.24	0.56	3.82	-1.66	3.74	-0.25	3.94	1.79	3.92	-12.29	3.40
	2000	50	30	-0.04	3.53	-0.54	3.70	3.00	3.53	-0.61	3.53	6.96	3.50	-0.62	3.77	-0.36	3.67
			40	-0.56	3.59	-0.34	3.79	-0.11	3.68	-2.21	3.60	-0.58	3.56	-0.56	3.76	-0.57	3.53
			50	-0.39	3.54	-0.60	3.60	0.51	3.51	0.12	3.41	-0.58	3.52	-1.52	3.50	-1.75	3.45
	3000	60	30	-0.01	3.50	-0.40	3.69	-0.27	3.77	-0.96	3.66	0.64	3.56	-0.15	3.60	-0.61	3.60
			40	0.27	3.80	-0.59	3.60	0.15	3.62	-0.18	3.57	-0.86	3.57	1.29	3.59	0.62	3.67
			50	-0.69	3.62	-0.66	3.58	-1.75	3.61	-0.75	3.57	-0.40	3.77	0.12	3.58	-0.62	3.57
RF	1000	40	30	-0.37	0.05	-0.22	0.04	-0.05	0.05	-1.05	0.04	-0.13	0.04	-0.25	0.04	-1.55	0.04
			40	-0.13	0.04	-0.35	0.04	-0.18	0.05	1.47	0.05	-0.45	0.04	-1.71	0.03	-0.82	0.04
			50	1.55	0.04	-0.39	0.03	0.31	0.04	-0.69	0.04	-0.41	0.03	0.13	0.03	-2.86	0.04
	2000	50	30	-0.28	0.06	-0.33	0.02	-0.06	0.03	-0.14	0.03	1.33	0.03	-0.22	0.03	0.22	0.03
			40	-0.18	0.03	-0.40	0.02	-0.14	0.03	-0.69	0.02	-0.18	0.03	-0.27	0.02	-0.25	0.02
			50	-0.30	0.03	-0.35	0.02	0.10	0.03	-0.21	0.02	-0.14	0.03	-0.29	0.02	-0.73	0.02
	3000	60	30	-0.22	3.62	-0.15	0.02	-0.26	0.02	-0.07	0.02	0.93	0.02	-0.25	0.02	-0.07	0.02
			40	-0.42	0.02	-0.38	0.02	-0.25	0.02	-0.34	0.02	-0.10	0.02	-0.10	0.02	-0.32	0.02
			50	0.56	0.02	0.45	0.02	-0.48	0.02	-0.44	0.02	-0.15	0.02	0.01	0.02	-0.40	0.02

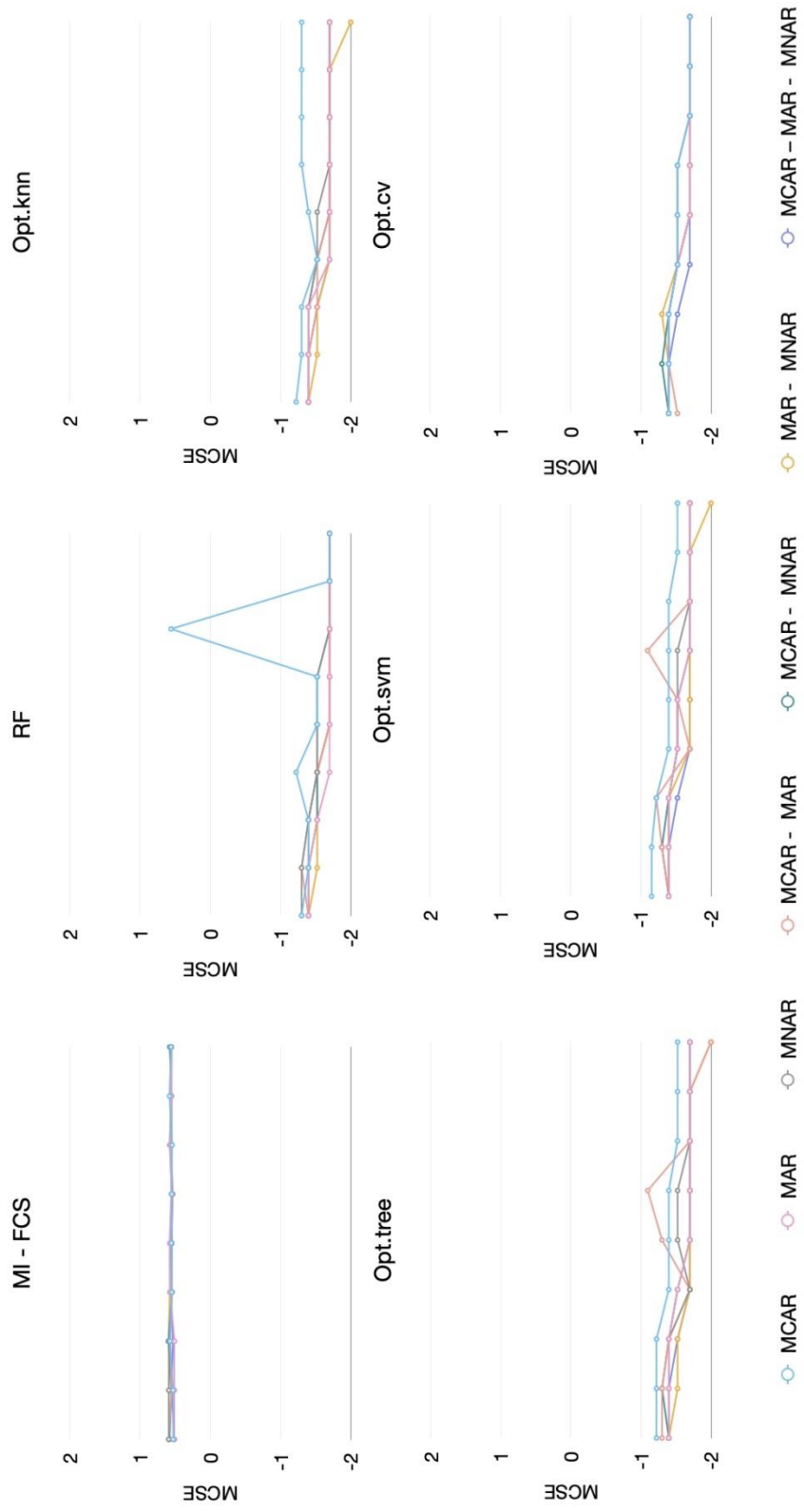
CHULALONGKORN UNIVERSITY

Methods	N_1	N_2	m	MCAR		MAR		MNAR		MCAR - MAR		MCAR - MNAR		MAR - MNAR		MCAR – MAR - MNAR	
				RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)
Opt.knn	1000	40	30	-0.25	0.06	-0.11	0.04	-0.12	0.04	-0.32	0.04	-0.37	0.04	-0.30	0.04	0.89	0.04
			40	-0.40	0.05	0.95	0.04	-0.30	0.04	-0.23	0.04	-0.22	0.04	-0.13	0.03	0.71	0.04
			50	-0.55	0.05	-0.27	0.04	-0.29	0.04	-0.67	0.03	-0.59	0.04	-0.52	0.03	-0.27	0.03
	2000	50	30	-0.39	0.03	-0.25	0.02	-0.01	0.03	0.01	0.03	0.08	0.03	-0.20	0.02	0.08	0.02
			40	-0.32	0.04	-0.43	0.02	-0.19	0.03	-0.67	0.02	-0.30	0.02	-0.37	0.02	-0.55	0.02
			50	-0.44	0.05	-0.58	0.02	-0.48	0.02	-1.23	0.02	-0.42	0.02	-0.77	0.02	-1.84	0.02
	3000	60	30	-0.32	0.05	0.41	0.02	-0.29	0.02	-0.07	0.02	-0.23	0.02	-0.43	0.02	-0.54	0.02
			40	-0.31	0.05	-0.46	0.02	-0.15	0.02	-0.29	0.02	1.39	0.02	-1.14	0.02	-0.31	0.02
			50	-0.55	0.05	-0.53	0.02	-0.12	0.02	-0.43	0.02	-0.19	0.02	-0.74	0.01	-0.57	0.02
Opt.tree	1000	40	30	-0.09	0.06	-0.27	0.04	-0.21	0.04	-0.33	0.05	-0.14	0.04	-0.16	0.04	1.24	0.04
			40	0.04	0.06	-0.05	0.04	-0.40	0.04	0.95	0.05	-0.28	0.05	-1.18	0.03	0.94	0.04
			50	-0.42	0.06	-0.25	0.04	-0.89	0.04	-0.25	0.04	-0.82	0.04	-0.79	0.03	-2.37	0.03
	2000	50	30	-0.08	0.04	-0.26	0.03	-0.04	0.02	0.04	0.02	-0.16	0.03	-0.06	0.02	-0.31	0.02
			40	-0.12	0.04	-0.34	0.02	-0.15	0.03	0.67	0.05	-0.34	0.02	-0.44	0.02	-0.56	0.02
			50	0.31	0.04	-0.26	0.02	-0.65	0.03	-0.67	0.08	-0.36	0.02	-0.57	0.02	-0.74	0.02
	3000	60	30	-0.11	0.03	-0.21	0.02	-0.23	0.02	-0.28	0.02	-0.22	0.02	-0.34	0.02	-0.34	0.02
			40	-0.32	0.03	-0.31	0.02	-0.02	0.02	-0.37	0.02	0.88	0.02	-0.47	0.02	-0.38	0.02
			50	-0.34	0.03	-0.45	0.02	-0.32	0.02	-0.42	0.01	-0.32	0.02	-0.57	0.01	-0.42	0.02

Methods	N_1	N_2	m	MCAR		MAR		MNAR		MCAR - MAR		MCAR - MNAR		MAR - MNAR		MCAR – MAR - MNAR	
				RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	RB (β_1)	MCSE (β_1)	R (β_1)	MCSE (β_1)
Opt.svm	1000	40	30	-0.06	0.07	-0.46	0.04	-0.20	0.04	-0.32	0.04	-0.08	0.04	-0.21	0.04	0.27	0.04
			40	-0.01	0.07	-0.05	0.04	-0.44	0.04	1.39	0.05	-0.27	0.05	-0.13	0.04	0.87	0.04
			50	-0.04	0.06	-0.66	0.04	-0.81	0.04	-0.50	0.06	-0.72	0.04	-0.37	0.04	-3.50	0.03
	2000	50	30	-0.05	0.04	-0.24	0.03	-0.05	0.03	0.02	0.02	-0.16	0.03	-0.01	0.02	-0.27	0.02
			40	-0.14	0.04	-0.20	0.03	-0.08	0.03	-0.37	0.03	-0.29	0.03	-0.44	0.02	-0.52	0.02
			50	-0.37	0.04	-0.19	0.02	-0.72	0.03	0.02	0.08	-0.32	0.02	-0.56	0.02	-0.99	0.02
	3000	60	30	-0.42	0.04	-0.12	0.02	-0.23	0.02	-0.27	0.02	-0.21	0.02	-0.37	0.02	-0.35	0.02
			40	-0.27	0.03	-0.26	0.02	-0.14	0.02	-0.35	0.02	0.66	0.02	-0.98	0.02	-0.38	0.02
			50	-0.14	0.03	-0.33	0.02	-0.29	0.02	-0.27	0.02	-0.23	0.02	-0.51	0.01	-0.40	0.02
Opt.cv	1000	40	30	-0.12	0.04	-0.27	0.04	-0.19	0.04	-0.32	0.03	0.14	0.04	-0.24	0.04	0.76	0.04
			40	-0.17	0.04	-0.05	0.04	-0.42	0.04	1.39	0.04	-0.28	0.05	-1.98	0.04	-0.47	0.04
			50	-0.26	0.04	-0.25	0.04	-0.83	0.04	-0.26	0.04	-0.82	0.04	0.47	0.05	-5.40	0.03
	2000	50	30	-1.69	0.03	-0.26	0.03	-0.08	0.03	-0.11	0.03	-0.16	0.03	-0.12	0.03	-0.14	0.02
			40	-0.20	0.03	-0.33	0.02	-0.12	0.03	-0.67	0.02	-0.34	0.02	-0.42	0.02	-0.25	0.02
			50	0.30	0.03	-0.26	0.02	-0.67	0.03	-0.67	0.02	-0.36	0.02	-0.50	0.02	-1.63	0.02
	3000	60	30	-0.36	0.02	-0.21	0.02	-0.22	0.02	-0.28	0.02	-0.22	0.02	-0.32	0.02	-0.21	0.02
			40	-0.76	0.02	-0.30	0.02	-0.22	0.02	-0.37	0.02	0.88	0.02	-1.24	0.02	-0.23	0.02
			50	-0.48	0.02	-0.44	0.02	-0.34	0.02	-0.43	0.02	-0.32	0.02	-0.04	0.02	-0.37	0.02



รูปที่ 4.1.5 แสดงการเปรียบเทียบลอการิทึมของค่าเฉลี่ย RB ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่มด้วยวิธีทดแทนค่าสูญหาย



รูปที่ 4.1. 6 แสดงการเปรียบเทียบลอการิทึมของค่าเฉลี่ย MCSE ในการประมาณค่าพารามิเตอร์ที่มีประสิทธิ์ความเอนเอียงแบบสุ่มด้วยวิธีทดแทนค่าสูญหาย

1.2 ผลการวิเคราะห์เปรียบเทียบประสิทธิภาพของวิธีการทดแทนค่าสูญหายด้วยค่าเฉลี่ย RB และค่าเฉลี่ย MCSE

จากการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย ในภาพรวมโดยใช่ค่าเฉลี่ย NRMSE ในหัวข้อที่ 1.1 พบว่าวิธีทดแทนค่าสูญหาย Otp.impute มีประสิทธิภาพสูงที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI – FCS ตามลำดับ อย่างไรก็ตามผู้วิจัยใช้ค่าเฉลี่ยความเอนเอียงสัมพันธ์ (RB) และค่าเฉลี่ยความคลาดเคลื่อนแบบมอนติคาร์โลสัมพันธ์ (MCSE) เพื่อตรวจสอบความถูกต้องแม่นยำ และความคลาดเคลื่อนมาตรฐานและคุณสมบัติความคงเส้นคงวาของการประมาณค่าพารามิเตอร์ β_1 ด้วยวิธีทดแทนค่าสูญหายที่แตกต่างกัน จากการจำลองข้อมูล 100 รอบ โดยสามารถสรุปผลการวิเคราะห์ได้ดังนี้

1.2.1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย RB

เมื่อพิจารณาค่าเฉลี่ย RB ในตารางที่ 4.2 จะพบว่าค่าเฉลี่ย RB ที่ประมาณค่าสัมประสิทธิ์ความถดถอยแบบสุ่ม โดยวิธีทดแทนค่าสูญหาย MI - FCS ให้ค่าเฉลี่ย RB อยู่ในช่วง 12.29 ถึง 21.82 วิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย RB อยู่ในช่วง -2.86 ถึง 1.47 และวิธีทดแทนค่าสูญหาย Otp.impute ให้ค่าเฉลี่ย RB ต่ำที่สุด อยู่ระหว่าง -5.40 ถึง 0.89 ตามลำดับ จะเห็นได้ว่าการประมาณค่าสัมประสิทธิ์ความถดถอยแบบสุ่มโดยวิธีทดแทนค่าสูญหาย MI - FCS ให้ช่วงของค่าเฉลี่ย RB ที่สูงที่สุดซึ่งพบค่าสูงสุดเท่ากับ 21.82 ขณะที่การประมาณค่าโดยวิธีทดแทนค่าสูญหาย RF และ Otp.impute พบค่าสูงสุดเท่ากับ 1.47 สังเกตได้ว่าช่วงของค่าเฉลี่ย RB มีค่าแตกต่างกันอยู่มาก

ดังนั้น ผู้วิจัยจึงได้ปรับสเกลค่าเฉลี่ย RB ที่ได้จากการจำลองข้อมูลให้อยู่ในรูปมาตราส่วนลอการิทึม (logarithm scale) เพื่อให้มีขนาดที่เหมาะสม สามารถนำเสนอแผนภาพเพื่อแสดงการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย โดยใช้ค่าเฉลี่ย RB ได้ง่ายมากยิ่งขึ้น จากรูปที่ 4.1.5 เมื่อกำหนดให้ขนาดตัวอย่างและอัตราการสูญหายคงที่ จะสังเกตได้ว่าวิธีทดแทนค่าสูญหาย MI - FCS มีแนวโน้มให้ค่าเฉลี่ย RB ค่อนข้างสูงที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และ วิธีทดแทนค่าสูญหาย Otp.impute ตามลำดับ ขณะที่เมื่อพิจารณาตารางที่ 4.2 ซึ่งจะพบว่าค่าเฉลี่ย RB มีแนวโน้มเข้าใกล้ศูนย์เมื่อขนาดตัวอย่างเพิ่มสูงขึ้นและอัตราการสูญหายลดลง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.2.2 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยค่าเฉลี่ย MCSE

ในทำนองเดียวกันกับหัวข้อ 1.2.1 เมื่อพิจารณาค่าเฉลี่ยค่าเฉลี่ย MCSE ในตารางที่ 4.2 จะพบว่าช่วงของค่าเฉลี่ย MCSE ที่ประมาณโดยวิธีทดแทนค่าสูญหายทั้ง 3 วิธี มีค่าแตกต่างกันอย่างมาก โดยเมื่อพิจารณาค่าเฉลี่ย MCSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย MI - FCS ให้ค่าเฉลี่ย MCSE อยู่ในช่วง 3.24 ถึง 3.94 วิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย MCSE อยู่ในช่วง 0.62 ถึง 3.62 และวิธีทดแทนค่าสูญหาย Otp.impute ให้ค่าเฉลี่ย MCSE อยู่ระหว่าง 0.01 ถึง 0.08 ตามลำดับ

ดังนั้น ผู้วิจัยจึงได้ปรับสเกลค่าเฉลี่ย MCSE ที่ได้จากการจำลองข้อมูลให้อยู่ในรูปมาตราส่วนลอการิทึม (logarithm scale) เพื่อให้มีขนาดที่เหมาะสม สามารถนำเสนอแผนภาพเพื่อแสดงการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย โดยใช้ค่าเฉลี่ย MCSE ได้ง่ายมากยิ่งขึ้น จากรูปที่ 4.1.6 เมื่อพิจารณาค่าเฉลี่ย MCSE โดยกำหนดให้ขนาดตัวอย่างและอัตราการสูญหายคงที่ จะสังเกตได้ชัดว่าวิธีทดแทนค่าสูญหาย MI - FCS ให้มีแนวโน้มให้ค่าเฉลี่ย MCSE ค่อนข้างสูงที่สุด เมื่อเทียบกับการประมาณโดยวิธีทดแทนค่าสูญหาย RF ขณะที่จะเห็นว่าวิธีทดแทนค่าสูญหาย Otp.impute ให้ค่าเฉลี่ย MCSE ค่อนข้างต่ำที่สุด ซึ่งหมายความว่าการประมาณค่าพารามิเตอร์ β_1 โดยใช้วิธีทดแทนค่าสูญหาย Otp.impute จะให้ค่าความคลาดเคลื่อนมาตรฐานอยู่ในระดับต่ำ นอกนี้จากการสังเกตการเปลี่ยนของค่าเฉลี่ย MCSE ในรูปที่ 4.1.6 จะพบว่าค่าเฉลี่ย MCSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Otp.impute มีค่าลู่เข้ามากที่สุด แสดงให้เห็นว่ามีความคงเส้นคงวาในการประมาณค่าพารามิเตอร์ β_1

จากการพิจารณาผลในหัวข้อ 1.1 ถึง 1.2 ในข้างต้น ทำให้สามารถสรุปได้ว่าปัจจัยสำคัญที่นักวิจัยไม่ควรมองข้ามในการประมาณค่าเพื่อทดแทนค่าสูญหาย คือ ลักษณะของการสูญหาย (type of missing) ขนาดตัวอย่าง (simple size) และร้อยละของอัตราค่าสูญหาย (Percentage of missing rate) เป็นต้น

ตอนที่ 2 ผลการวิเคราะห์สถิติพื้นฐานและการประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง

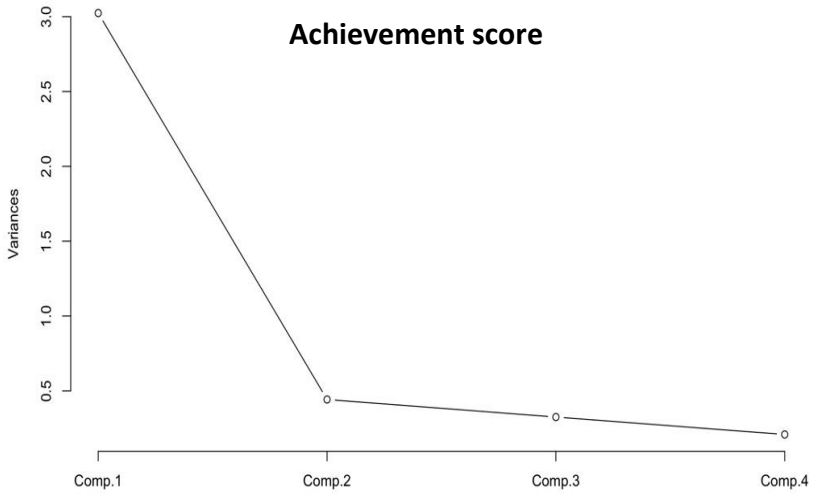
จากผลการวิเคราะห์ประสิทธิภาพของวิธีทดแทนค่าสูญหายทั้ง 3 วิธี โดยใช้การจำลองข้อมูลในตอนที่ 1 ข้างต้น จะพบว่าวิธีทดแทนค่าสูญหาย Otp.impute มีประสิทธิภาพสูงที่สุด เมื่อเทียบกับวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI - FCS ตามลำดับ ดังนั้นการวิจัยในตอนนี้ผู้วิจัยจึงนำวิธีทดแทนค่าสูญหายที่ได้จากการจำลองข้อมูลในตอนที่ 1 มาประยุกต์ใช้กับข้อมูลจริง โดยใช้ข้อมูลทุติยภูมิจากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวน 2,109 โรงเรียน ที่อยู่ในสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) ทั้งหมด 42 แห่ง

ผู้วิจัยนำข้อมูลในข้างต้นมาวิเคราะห์โดยใช้สถิติพื้นฐานเพื่อศึกษาสภาพและลักษณะของข้อมูลจริงทางการศึกษาที่มีโครงสร้างแบบพหุระดับ โดยแบ่งการนำเสนอผลการวิจัยออกเป็นสองส่วน ได้แก่ 2.1 ผลการวิเคราะห์สถิติพื้นฐานในข้อมูลจริงระดับโรงเรียน 2.2 การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง โดยรายละเอียดดังนี้

2.1 ผลการวิเคราะห์สถิติพื้นฐานในข้อมูลจริงระดับโรงเรียน

ผู้วิจัยใช้การวิเคราะห์องค์ประกอบหลัก (principle component analysis: PCA) และใช้สถิติบรรยาย (descriptive statistics) เพื่อประมาณคะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน และเพื่อวิเคราะห์ข้อมูลของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ และไม่ได้อาศัยอยู่กับบิดาและมารดา ระดับโรงเรียน ซึ่งมีรายละเอียด ดังนี้

2.1.1 ผลการวิเคราะห์ผลสัมฤทธิ์ทางการเรียนของนักเรียน ระดับโรงเรียน



รูปที่ 4.1.7 แสดงผลการวิเคราะห์องค์ประกอบหลักของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน

ผู้วิจัยใช้การวิเคราะห์องค์ประกอบหลัก (PCA) ประมาณคะแนนผลสัมฤทธิ์ทางการเรียน โดยจากการพิจารณาค่าสัมประสิทธิสหสัมพันธ์ของเพียร์สันระหว่างคะแนนเฉลี่ย O – NET ระดับโรงเรียนทั้ง 4 รายวิชาพบว่าค่าสัมประสิทธิ์สหสัมพันธ์มีค่าอยู่ในช่วง 0.58 – 0.78 อย่างมีนัยสำคัญทางสถิติ ทั้งนี้เมื่อพิจารณาผลการวิเคราะห์องค์ประกอบหลัก พบว่าสามารถสกัดองค์ประกอบของผลสัมฤทธิ์ทางการเรียนได้เพียง 1 องค์ประกอบ ผลปรากฏดังรูปที่ 4.1.7 โดยองค์ประกอบดังกล่าวสามารถอธิบายความแปรปรวนในคะแนนสอบ O-NET คิดเป็นค่าไอเกนเท่ากับ 3.02 ซึ่งคิดเป็นร้อยละ 75.60 ของความแปรปรวนในตัวแปรเดิม มีค่าน้ำหนักองค์ประกอบของคะแนนสอบทั้ง 4 รายวิชาอยู่ในช่วง 0.473 ถึง 0.516 โดยวิชาภาษาอังกฤษมีน้ำหนักองค์ประกอบสูงที่สุดเท่ากับ 0.516 รองลงมาคือ วิชาคณิตศาสตร์เท่ากับ 0.508 วิชาวิทยาศาสตร์เท่ากับ 0.502 และวิชาภาษาไทยมีน้ำหนักองค์ประกอบน้อยที่สุดเท่ากับ 0.473 ตามลำดับ

ตารางที่ 4. 3 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน

O - NET	M	SD	Min	Max
ภาษาไทย	54.74	7.97	30.62	83.75
ภาษาอังกฤษ	32.83	7.37	17.50	76.92
คณิตศาสตร์	24.54	7.86	4.00	86.92
วิทยาศาสตร์	29.67	4.16	16.50	63.42

เมื่อพิจารณาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของผลสัมฤทธิ์ทางการเรียนของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ระดับโรงเรียน ปีการศึกษา 2563 จำนวน 2,109 โรงเรียน ที่อยู่ภายใต้สังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา 42 เขตพื้นที่ จากตารางที่ 4.3 จะสังเกตได้ว่าผลสัมฤทธิ์ทางการเรียนของนักเรียนโดยเฉลี่ยอยู่ในระดับต่ำ โดยที่วิชาภาษาไทยมีคะแนนเฉลี่ยสูงที่สุด (M = 54.74 คะแนน) รองลงมาคือวิชาภาษาอังกฤษ (M = 32.83 คะแนน) วิชาวิทยาศาสตร์ (M = 24.54 คะแนน) และวิชาคณิตศาสตร์ (M = 29.67 คะแนน) ตามลำดับ นอกจากนี้จะเห็นว่าวิชาที่มีคะแนนสูงที่สุด (Max = 86.92) และมีคะแนนต่ำที่สุด (Min = 4.00) คือวิชาคณิตศาสตร์

2.1.2 ผลการวิเคราะห์ข้อมูลของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดาและมารดา

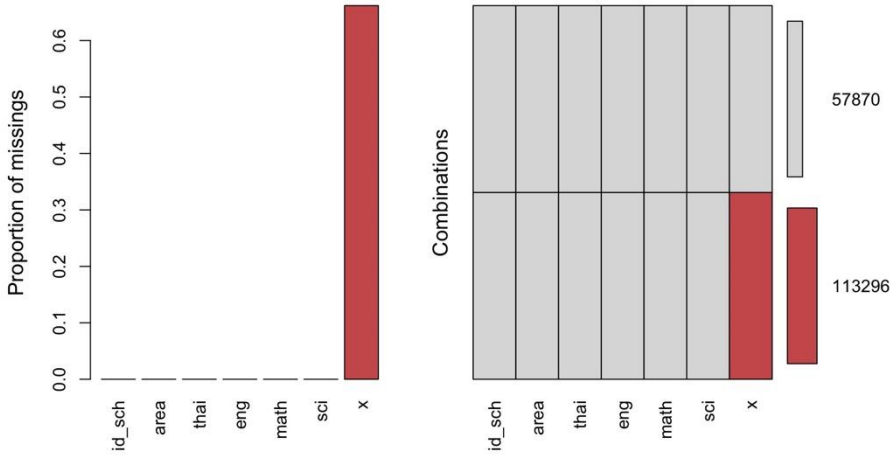
เมื่อพิจารณาจำนวนนักเรียนที่ไม่ได้พักอาศัยอยู่กับบิดาและมารดาและจำนวนนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ จากตารางที่ 4.4 จะพบว่ามีจำนวนนักเรียนที่ไม่ได้พักอาศัยอยู่กับบิดาและมารดาจำนวน 44,359 คน คิดเป็นเพียงร้อยละ 25.05 และนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์จำนวน 5,287 คน คิดเป็นร้อยละ 2.99

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

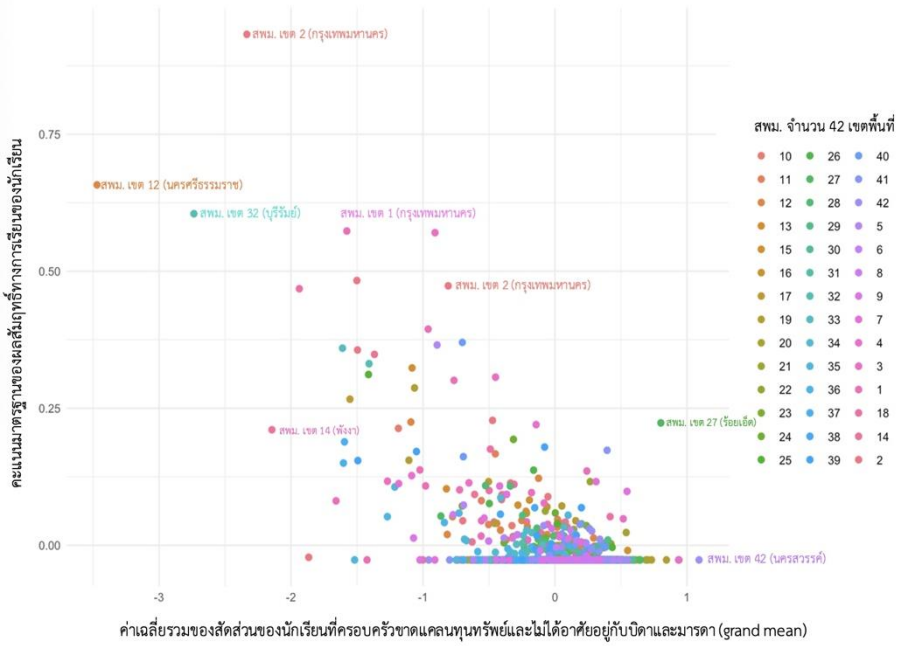
ตารางที่ 4. 4 แสดงจำนวนและร้อยละของข้อมูลนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563

ข้อมูลของนักเรียนชั้นมัธยมศึกษาปีที่ 3	**จำนวน (ร้อยละ)**	**จำนวนค่าสูญหาย (ร้อยละ)**
นักเรียนที่ไม่ได้พักอาศัยอยู่กับบิดาและมารดา	44,359 คน (25.05)	111,224 คน (62.81)
อื่นๆ	21,514 คน (12.14)	
นักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์	5,287 คน (2.99)	111,126 คน (62.74)
อื่นๆ	60,684 คน (34.27)	
สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และ	57,870 คน (33.81)	113,296 คน (66.19)
ไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (x)	969 โรงเรียน (45.95)	1,140 โรงเรียน (54.05)

ทั้งนี้ผู้วิจัยรวมคะแนนทั้งสองด้านเข้าด้วยกันด้วยค่าเฉลี่ยเรขาคณิตเพื่อให้ได้คะแนนสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน $(\boldsymbol{x})$ โดยที่พบว่าค่าเฉลี่ยของ $\boldsymbol{x}$ มีค่าเท่ากับ 0.06 นอกจากนี้เมื่อพิจารณารูปที่ 4.1.8 จะสังเกตได้ว่าเมื่อรวมคะแนนทั้งสองด้านเข้าด้วยกันแล้วจะพบค่าสูญหายอยู่ในระดับสูงมาก เท่ากับร้อยละ 66.19 ขณะที่เมื่อพิจารณาค่าสูญหายในระดับโรงเรียน จะพบว่ามีค่าสูญหายจำนวน 1,140 โรงเรียน คิดเป็นร้อยละ 54.05 ของโรงเรียนทั้งหมด สังเกตได้จากตารางที่ 4.4



รูปที่ 4.1. 8 แสดงค่าสูญหายในสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน



รูปที่ 4.1. 9 แสดงความสัมพันธ์ระหว่างผลสัมฤทธิ์ทางการเรียนของนักเรียนกับสัดส่วนนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดา

เมื่อพิจารณาความสัมพันธ์ของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนกับสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน ที่เก็บรวบรวมได้จำนวน 969 โรงเรียน จากรูปที่ 4.1.9 จะสังเกตได้ว่าส่วนใหญ่โรงเรียนที่อยู่ในสังกัด สพม. มีคะแนนผลสัมฤทธิ์ทางการเรียนระดับโรงเรียนอยู่ในระดับต่ำ โดยที่เกือบทุกสังกัดมีคะแนนเฉลี่ยสัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดา อยู่ในระดับปานกลางถึงต่ำ ขณะที่มีเพียงบางสังกัดเท่านั้นที่มีคะแนนผลสัมฤทธิ์ทางการเรียนอยู่ในระดับสูง ได้แก่ สพม. เขต 1 และ สพม. เขต 2 จังหวัดกรุงเทพมหานคร สพม. เขต 12 จังหวัดนครศรีธรรมราช สพม. เขต 32 จังหวัดบุรีรัมย์ และ สพม. เขต 14 จังหวัดพังงา เป็นต้น โดยจะสังเกตได้ว่าโรงเรียนที่มีผลสัมฤทธิ์ทางการเรียนอยู่ในระดับสูง มีแนวโน้มที่จะมีสัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดา อยู่ในระดับต่ำ

นอกจากนี้ ผู้วิจัยจึงดำเนินการแปลงคะแนนสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดา ระดับโรงเรียน ที่สูญหาย โดยกำหนดให้มี 2 ค่า (dummy variable) เพื่อศึกษาลักษณะความสัมพันธ์ของ “ค่าสูญหาย” กับ “ค่าที่สังเกตได้” โดยที่ให้เท่ากับ “1” เมื่อพบค่าสูญหาย และเท่ากับ “0” เมื่อพบค่าสังเกตได้ จากการพิจารณาความสัมพันธ์ของค่าสังเกตได้ ในตัวแปรผลสัมฤทธิ์ทางการเรียนกับค่าสูญหายในสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน จะพบว่าค่าสูญหายมีความสัมพันธ์กับค่าสังเกตซึ่งพบค่าสัมประสิทธิ์สหสัมพันธ์อยู่ในช่วง 0.04 – 0.10 อย่างมีนัยสำคัญทางสถิติ

2.2 การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง

การนำเสนอผลการวิจัยในหัวข้อนี้ ผู้วิจัยประยุกต์ใช้วิธีทดแทนค่าสูญหายที่ได้การจำลองข้อมูลกับข้อมูลจริงทางการศึกษาของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) โดยมีรายละเอียด ดังนี้

2.2.1 การพิจารณาตัวแปรและค่าสูญหายในข้อมูลที่สนใจศึกษา

เมื่อพิจารณาข้อมูลจริง ที่นำมาประยุกต์ใช้ในการวิจัยครั้งนี้ จะพบว่าข้อมูลดังกล่าวมีโครงสร้างแบบพหุระดับสองระดับ โดยระดับที่หนึ่งคือระดับโรงเรียน มีค่าเท่ากับ 2,109 โรงเรียน และระดับที่สองคือระดับสังกัด มีค่าเท่ากับ 42 สังกัด โดยที่ข้อมูลชุดนี้ประกอบด้วยสองส่วน ส่วนแรกคือข้อมูลผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน และส่วนที่สองคือข้อมูลสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดาและมารดา ระดับโรงเรียน ตามลำดับ ขณะที่เมื่อพิจารณาค่าสูญหายของข้อมูล พบว่าเกิดค่าสูญหายในตัวแปรสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน ซึ่งเป็นตัวแปรระดับที่หนึ่ง จำนวน 1,140 โรงเรียน คิดเป็นร้อยละ 54.05 ของโรงเรียนทั้งหมด แสดงให้เห็นว่า

ในการเก็บรวบรวมข้อมูลทุติยภูมิจากหน่วยงานของรัฐหรือหน่วยงานทางการศึกษา มีโอกาสพบค่าสูญหายมากกว่าร้อยละ 50 โดยเฉพาะการรวบรวมข้อมูลส่วนบุคคลที่มีความอ่อนไหว (sensitive data) โดยจะสังเกตได้จากข้อมูลชุดนี้ พบค่าสูญหายในตัวแปรสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน ขณะที่มีเพียง 969 โรงเรียนที่มีข้อมูลครบถ้วนสมบูรณ์ การวิจัยครั้งนี้ผู้วิจัยเปรียบเทียบ ผลที่ได้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่ไม่ได้ทดแทนค่าสูญหาย เพื่อให้เกิดประโยชน์แก่ผู้ที่สนใจ สามารถนำวิธีทดแทนค่าสูญหายไปปรับใช้กับข้อมูลทางการศึกษาที่มีโครงสร้างแบบพหุระดับได้อย่างถูกต้องและเหมาะสมมากที่สุด

2.2.2 การพิจารณาลักษณะของข้อมูลสูญหาย (Type of missing data)

เมื่อพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ของค่าสูญหายในตัวแปรสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียนในหัวข้อ 2.1 จะพบว่าค่าสูญหายมีความสัมพันธ์กับค่าสังเกตได้ในตัวแปรผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน อย่างมีนัยสำคัญทางสถิติ จึงทำให้ผู้วิจัยสามารถสรุปได้ว่าข้อมูลชุดนี้มีลักษณะการสูญหายแบบสุ่มหรือ MAR

นอกจากนี้หากผู้วิจัยพิจารณาค่าสูญหายด้วยปัจจัยอื่นนอกเหนือจากการสูญหายในรูปแบบ MAR ผู้วิจัยต้องทำการสำรวจนักเรียนที่ไม่ตอบคำถามด้านครอบครัวรวมไปถึงการขาดแคลนทุนทรัพย์ทางการศึกษาและการพักอาศัยอยู่กับบิดามารดาเพิ่มเติม ยกตัวอย่างเช่น การโทรศัพท์ไปยังกลุ่มของนักเรียนที่ไม่ตอบคำถาม โดยสอบถามถึงสาเหตุของการไม่ตอบ เป็นต้น ซึ่งสามารถนำคำตอบมาพิจารณาลักษณะการสูญหายได้หากคำตอบเหล่านั้นมีความแตกต่างกันมาก แสดงว่าข้อมูลนั้นมีการสูญหายในรูปแบบ MNAR แต่อย่างไรก็ตามการวิจัยในครั้งนี้เป็นการเก็บข้อมูลทุติยภูมิทำให้ไม่สามารถกระทำเช่นนั้นได้ ผู้วิจัยจึงไม่สามารถสรุปได้อย่างแน่ชัดว่าข้อมูลชุดนี้มีลักษณะการสูญหายในรูปแบบ MNAR ผสมอยู่ด้วยหรือไม่

อย่างไรก็ตาม ถ้าหากพิจารณาถึงสาเหตุของการเกิดค่าสูญหายในตัวแปรดังกล่าว จะพบว่าค่าสูญหายที่เกิดขึ้น เกิดขึ้นที่ตัวแปรที่มีการถามข้อมูลส่วนบุคคลที่มีความอ่อนไหว (sensitive data) เพราะฉะนั้นจึงมีความเป็นได้ว่าสาเหตุของการสูญหายในข้อมูลชุดนี้อาจมาจากปัจจัยของตนเอง เช่น นักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์ หรือนักเรียนที่ไม่ได้อยู่กับบิดาและมารดา จะไม่อยากบอกข้อมูลส่วนบุคคลของตนเอง ให้ผู้อื่นทราบ เป็นต้น ดังนั้นจากการพิจารณาสาเหตุของการเกิดค่าสูญหายที่ได้กล่าวมานี้ จึงทำให้ผู้วิจัยทำให้สามารถสรุปได้ว่าข้อมูลชุดนี้มีการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม หรือ MAR – MNAR

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2.3 การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง

เมื่อพิจารณาผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย จากผลการวิจัยในตอนที่ 1 ภาพรวมพบว่าส่วนใหญ่วิธีทดแทนค่าสูญหาย Opt.impute มีประสิทธิภาพสูงที่สุด กล่าวคือ มีแนวโน้มให้ค่าเฉลี่ย NRMSE, ค่าเฉลี่ย RB และค่าเฉลี่ย MCSE ต่ำที่สุด รองลงมาคือวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI – FCS สังเกตได้จากรูปที่ 4.1.3 ถึง 4.1.5 ตามลำดับ และจากการพิจารณาสาเหตุของการเกิดค่าสูญหายในหัวข้อ 2.2.3 ในข้างต้น พบว่าข้อมูลชุดนี้ มีการสูญหายแบบสุ่มผสมกับการสูญหายแบบไม่สุ่ม ผู้วิจัยจึงพิจารณาประสิทธิภาพของวิธีทดแทนค่าสูญหาย Opt.impute เมื่อข้อมูลมีลักษณะการสูญหายแบบผสม MAR - MNAR พบว่าวิธี Opt.knn มีแนวโน้มให้ประสิทธิภาพสูงที่สุดแม้ว่าขนาดตัวอย่างและอัตราการสูญหายเพิ่มสูงขึ้น

ดังนั้น การวิจัยครั้งนี้ผู้วิจัยเลือกใช้วิธีทดแทนค่าสูญหาย Opt.knn มาประยุกต์ใช้กับข้อมูลจริงในการศึกษาครั้งนี้โดยนำข้อมูลที่ทดแทนค่าสูญหายมาวิเคราะห์ความเหลื่อมล้ำทางการศึกษาโดยใช้โมเดลพหุระดับ เปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย โดยจะแสดงรายละเอียดของผลการวิจัยในตอนที่ 3

ตอนที่ 3 การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) เพื่อนำไปประยุกต์ใช้กับการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา เพื่อสร้างองค์ความรู้ในทางทฤษฎีและทางปฏิบัติ สามารถนำวิธีทดแทนค่าสูญหายไปประยุกต์กับสถานการณ์จริงได้อย่างถูกต้องและเหมาะสม สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงความจริงมากยิ่งขึ้น และเกิดความผิดพลาดน้อยที่สุด โดยการนำเสนอผลการวิจัยในตอนนี้ แสดงการเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย ซึ่งมีรายละเอียด ดังนี้

เมื่อพิจารณาการเปรียบเทียบผลของการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา ระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย จากตารางที่ 4.5 จะเห็นได้ชัดว่า ข้อมูลทั้งสองชุดมีขนาดตัวอย่างระดับที่หนึ่งที่แตกต่างกัน โดยข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย กล่าวคือข้อมูลที่ตัดค่าสูญหายทิ้ง มีขนาดตัวอย่างในระดับที่หนึ่งเท่ากับ 969 โรงเรียน ขณะที่ข้อมูลที่มีการทดแทนค่าสูญหายมีขนาดตัวอย่างในระดับที่หนึ่งเท่ากับ 2,109 ซึ่งแตกต่างกันมากถึง 1,140 โรงเรียน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 4. 5

แสดงผลการเปรียบเทียบระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย

	ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย (Listwise Deletion)			ข้อมูลที่ทดแทนค่าสูญหาย (Opt.knn)		
ขนาดตัวอย่างในระดับที่ 1	969 โรงเรียน			2,109 โรงเรียน		
ขนาดตัวอย่างในระดับที่ 2	42 สังกัด			42 สังกัด		
	ผลการวิเคราะห์ความแปรปรวนของผลสัมฤทธิ์ทางการเรียนระดับโรงเรียน					
ระดับโรงเรียน (Level 1)	**Variance (SD)**	**ICC**		**Variance (SD)**		**ICC**
	0.0017 (0.0411)	0.8034		0.0016 (0.0173)		0.8421
ระดับสังกัด (Level 2)	**Variance (SD)**	**ICC**	**LRT (Df)**	**Variance (SD)**	**LRT (Df)**	**ICC**
	0.0004 (0.0204)	0.1966	112.79 (1)	0.0003 (0.0400)	278.08 (1)	0.1579
อิทธิพลคงที่ (Fixed effect): สัมประสิทธิ์จุดตัดแกน	- 0.1752 (SE = 0.00346)	p-value < 0.000		- 0.1508 (SE = 0.00282)	p-value < 0.000	
p-value	< 0.000			< 0.000		
	ผลการวิเคราะห์โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่มของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน					
อิทธิพลคงที่ (Fixed effect)	**Estimate**			**Estimate**		
สัมประสิทธิ์จุดตัดแกน (γ_{00})	- 0.1759 (SE = 0.00314) (p-value < 0.000)			- 0.1524 (SE = 0.0022) (p-value < 0.000)		
	ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย (Listwise Deletion)			**ข้อมูลที่ทดแทนค่าสูญหาย (Opt.knn)**		
อิทธิพลคงที่ (Fixed effect)	**Estimate**			**Estimate**		
สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (γ_{10})	- 0.5776 (SE = 0.0953) (p-value < 0.000)			-1.7024 (SE = 0.1895) (p-value < 0.000)		

อิทธิพลสุ่ม (Random effect)	Variance	Variance
สัมประสิทธิ์จุดตัดแกน (U_{0j})	0.000332 (SD = 0.01822)	0.00017 (SD = 0.01313)
สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (U_{1j})	0.186539 (SD = 0.4319)	1.20922 (SD = 1.09965)
เศษเหลือ (ε_{ij})	0.001462 (SD = 0.03823)	0.0012016 (SD = 0.03466)
สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_{1j}) ของสัดส่วนนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาของโรงเรียนในแต่ละสังกัด สพม. **(ความเหลื่อมล้ำทางการศึกษา)**	**อยู่ในช่วง - 0.0018 ถึง - 1.1966**	**อยู่ในช่วง 0.5803 ถึง - 4.7667**

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เมื่อพิจารณาค่าความคลาดเคลื่อนมาตรฐาน (Standard Error) ของอิทธิพลคงที่ (Fixed effect) จากการประมาณค่าความแปรปรวนของผลสัมฤทธิ์ทางการเรียนระดับโรงเรียนในตารางที่ 4.5 จะพบว่าข้อมูลที่ไม่ได้ทดแทนค่าสูญหายให้ค่าความคลาดเคลื่อนมาตรฐานเท่ากับ 0.00346 (SE = 0.00346) ขณะที่ข้อมูลที่มีการทดแทนค่าสูญหายให้ค่าความคลาดเคลื่อนมาตรฐานเท่ากับ 0.00282 (SE = 0.00282) แสดงให้เห็นว่าเมื่อเกิดปัญหาข้อมูลสูญหาย หากเลือกใช้วิธีการแก้ปัญหาด้วยการตัดค่าสูญหายทิ้ง อาจจะส่งกระทบผลให้ค่าคลาดเคลื่อนมาตรฐาน (standard error) ของค่าสถิติมีค่าสูงขึ้น อำนาจในการทดสอบสถิติลดลง เมื่อเปรียบเทียบกับการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย

นอกจากนี้ จะสังเกตได้ว่าหากพิจารณาความแตกต่างของผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา โดยใช้ความแตกต่างของสัมประสิทธิ์ความถดถอยแบบสุ่มของสัดส่วนนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดา จะพบว่าข้อมูลทั้งสองชุดให้ผลการวิเคราะห์ที่แตกต่างกัน รายละเอียดแสดงดังตารางที่ 4.5

ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา โดยพิจารณาจากความแตกต่างของค่า β_{1j} ในแต่ละสังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) จำนวน 42 สังกัด ระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย จากตารางที่ 4.5 ในข้างต้น จะพบว่าข้อมูลทั้งสองชุดดังกล่าว ให้ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาที่แตกต่างกัน มีรายละเอียดดังนี้

3.1 ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา เมื่อใช้ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย

เมื่อพิจารณาอิทธิพลสุ่ม (Random effect) จะสังเกตได้ว่าข้อมูลที่ไม่ได้ทดแทนค่าสูญหายให้ค่า β_{1j} ในแต่ละสังกัดอยู่ในช่วง - 0.0018 ถึง - 1.1966 จะเห็นว่าทุกสังกัดมีค่าสัมประสิทธิ์ความถดถอยในทิศทางลบ กล่าวคือหากโรงเรียนในสังกัด สพม. มีสัดส่วนของนักเรียนที่ขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับพ่อแม่เพิ่มสูงขึ้น จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนลดลง อย่างมีนัยสำคัญทางสถิติ เมื่อพิจารณาอิทธิพลคงที่ (Fixed effect) จากการประมาณค่าสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (γ_{10}) พบว่า สัมประสิทธิความถดถอยมีค่าเท่ากับ - 0.5776 กล่าวคือ เมื่อมีสัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อยู่กับพ่อแม่ เพิ่มสูงขึ้น 1 หน่วย จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนลดลง 0.5776 คะแนน อย่างมีนัยสำคัญทางสถิติ

3.2 ผลการวิเคราะห์ความเหลื่อมล้ำทางการศึกษา เมื่อใช้ข้อมูลที่มีการทดแทนค่าสูญหาย

เมื่อพิจารณาอิทธิพลสุ่ม (Random effect) จะพบค่า β_{1j} อยู่ในช่วง 0.5803 ถึง - 4.7667 ซึ่งจะเห็นว่ามีสังกัด สพม. บางแห่งให้ค่า β_{1j} ในทิศทางบวก กล่าวคือหากโรงเรียนในสังกัด สพม. มีสัดส่วนของนักเรียนที่ขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับพ่อแม่ เพิ่มสูงขึ้น 1 หน่วย จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน เพิ่มขึ้นหรือลดลง อยู่ในช่วง 0.5803 ถึง - 4.7667 อย่างมีนัยสำคัญทางสถิติ เมื่อพิจารณาอิทธิพลคงที่ (Fixed effect) จากการประมาณค่าสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (γ_{10}) พบว่าสัมประสิทธิความถดถอย มีค่าเท่ากับ -1.7024 กล่าวคือ เมื่อมีสัดส่วนของนักเรียนในโรงเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อยู่กับพ่อแม่เพิ่มสูงขึ้น 1 หน่วย จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนลดลง 1.7024 อย่างมีนัยสำคัญทางสถิติ

การวิจัยในตอนนี้เปรียบเทียบการใช้ข้อมูลที่มีการทดแทนค่าสูญหายด้วย Opt.knn กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหายหรือตัดค่าสูญหายทิ้ง Listwise Deletion ขณะที่เมื่อพิจารณาลักษณะการสูญหายที่เกิดขึ้นในข้อมูลชุดนี้ จากการพิจารณาผลในตอนที่ 2 หัวข้อที่ 2.2.3 จะพบว่าข้อมูลมีลักษณะการสูญหายแบบผสม MAR - MNAR กล่าวคือค่าสูญหายมีความสัมพันธ์ขึ้นอยู่กับค่าสังเกตได้ หรือค่าสูญหายในข้อมูลชุดนี้ เกิดการสูญหายจากปัจจัยของตนเอง จะเห็นว่าการนำข้อมูลส่วนที่เหลือมาวิเคราะห์ (Listwise Deletion) ถึงแม้ว่าจะสามารถกระทำได้ง่าย สะดวก แต่มีข้อควรระวังคือหากลักษณะการสูญหายไม่ได้อยู่ในแบบสุ่มสมบูรณ์ MCAR กล่าวคือค่าสูญหายไม่มีความสัมพันธ์หรือขึ้นอยู่กับค่าสังเกตได้ การใช้วิธีการตัดข้อมูลแบบ Listwise Deletion อาจจะไม่ใช่วิธีทดแทนค่าสูญหายที่เหมาะสมที่สุด เนื่องจากหากประมาณค่าโดยการตัดค่าสูญหายทิ้ง อาจจะส่งผลกระทบต่อโครงสร้างความสัมพันธ์ของข้อมูล โดยอาจจะทำให้การประมาณค่าพารามิเตอร์มีความลำเอียง (biased) หรืออาจก่อให้เกิดผลกระทบอย่างรุนแรงต่อผลการศึกษา นอกจากนี้เมื่อพิจารณาระดับของอัตราการสูญหาย จะพบว่าข้อมูลดังกล่าวมีอัตราการสูญหายอยู่ในระดับสูงมากกว่าถึงร้อยละ 50 จะเห็นว่า หากนำข้อมูลส่วนที่เหลือมาวิเคราะห์ผลการวิจัยอาจทำให้สูญเสียรายละเอียดที่สำคัญ บางอย่าง ทำให้ผลการวิจัยที่ได้ อาจไม่ครอบคลุมประชากรที่ต้องการศึกษา

จากการเปรียบเทียบผลของการวิเคราะห์ความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย จากผลการวิจัยในข้างต้น แสดงให้เห็นว่าหากใช้ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย อาจจะส่งผลกระทบต่อการวิจัยที่สนใจศึกษาทั้งในส่วนของการวิเคราะห์ การประมาณค่าพารามิเตอร์ รวมไปถึงการสรุปผลตีความผลของการวิจัย

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

โดยที่ระดับความรุนแรงของผลกระทบ อาจขึ้นอยู่กับหลายปัจจัย ได้แก่ ขนาดการสูญหายของข้อมูล (size of missing data) วิธีการแก้ไขปัญหาการสูญหายของข้อมูล (method for treating missing data) และที่สำคัญคือการพิจารณาลักษณะการสูญหายที่เกิดขึ้นหรือสาเหตุของการเกิดค่าสูญหาย (pattern of missing data or type of missing data) เป็นต้น ดังนั้นนักวิจัยจึงมีความจำเป็นต้องทราบแนวทางและกระบวนการจัดค่าสูญหาย เพื่อสามารถเลือกใช้วิธีทดแทนค่าสูญหายได้อย่างเหมาะส เพื่อให้เกิดประสิทธิภาพสูงที่สุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่ม (β_1) ในโมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม (Random Coefficients Model) เพื่อวิเคราะห์ความเหลื่อมล้ำทางการศึกษา และเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษาระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย โดยการวิจัยครั้งนี้ศึกษาภายใต้สถานการณ์ต่าง ๆ ที่กำหนดขึ้น มีเงื่อนไขในการศึกษาดังนี้

1. ลักษณะการสูญหาย 7 รูปแบบ คือ MCAR, MAR, MNAR, MCAR – MAR, MCAR – MNAR, MAR – MNAR และ MCAR – MAR - MNAR

2. ระดับอัตราการสูญหายของค่าสังเกตในตัวอย่างระดับที่หนึ่งเท่ากับร้อยละ 30, ร้อยละ 40 และร้อยละ 50

3. ขนาดตัวอย่างระดับที่หนึ่งเท่ากับ 1000, 2000, 3000 หน่วย ขนาดตัวอย่างระดับที่สองเท่ากับ 40, 50, 60 หน่วย วนซ้ำ 100 รอบ ตามลำดับ

4. การประยุกต์ใช้กับข้อมูลทุติยภูมิจากสถาบันทดสอบทางการศึกษาแห่งชาติ (สทศ.) ซึ่งแบ่งข้อมูลออกเป็นสองส่วน ได้แก่ ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) และข้อมูลสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (x)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ดังนั้นในบทนี้ ผู้วิจัยสรุปผลการวิจัย อภิปรายผลและข้อเสนอแนะ ออกเป็น 3 ตอน ดังนี้

ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล

ตอนที่ 2 ผลการวิเคราะห์การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง

ตอนที่ 3 การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา ระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย

เมื่อพิจารณาผลการวิเคราะห์มีประเด็นที่น่าสนใจและนำไปสู่การสรุปและอภิปรายผล โดยมีรายละเอียดดังนี้

สรุปผลการวิจัย

ตอนที่ 1 ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายด้วยการจำลองข้อมูล

ผลการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าข้อมูลสูญหายระหว่างวิธี MI-FCS, วิธี RF และ วิธี Opt.impute ซึ่งประกอบด้วย 4 วิธีย่อยได้แก่ Opt.knn, Opt.tree, Opt.svm และ Opt.cv ในการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยแบบสุ่มของสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้พักอาศัยอยู่กับบิดามารดาระดับโรงเรียน (β_1) ในข้อมูลพหุระดับ การนำเสนอสรุปผลการวิจัยในบทนี้ ผู้วิจัยนำเสนอตามเกณฑ์ของค่าเฉลี่ย NRMSE, ค่าเฉลี่ย RB และค่าเฉลี่ย MCSE ในการประมาณค่าพารามิเตอร์ (β_1) โดยมีข้อสรุปรายละเอียด ดังนี้

1.1 ประสิทธิภาพของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ (β_1) ในภาพรวม

เมื่อพิจารณาการเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในภาพรวมด้วยค่าเฉลี่ยของ NRMSE ในการศึกษาในครั้งนี้ พบว่าโดยภาพรวมวิธีทดแทนค่าสูญหาย Opt.impute จะมีแนวโน้มให้ค่าเฉลี่ย NRMSE ต่ำที่สุด แม้ว่าขนาดตัวอย่างและอัตราการสูญหายจะเพิ่มสูงขึ้น โดยมีค่าอยู่ในช่วง 0.01 - 0.15 ขณะที่วิธีทดแทนค่าสูญหาย RF มีค่าเฉลี่ย NRMSE อยู่ระหว่าง 0.06 - 0.16 และวิธีทดแทนค่าสูญหาย MI - FCS มีค่าเฉลี่ย NRMSE อยู่ในช่วง 0.05 - 0.22 จะเห็นว่าวิธีทดแทนค่าสูญหาย Opt.impute มีแนวโน้มให้ค่าประมาณพารามิเตอร์ที่มีประสิทธิภาพมากที่สุด รองลงมาคือ วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI - FCS ตามลำดับ

1.2 ความถูกต้องและความแม่นยำในการประมาณค่าพารามิเตอร์ (β_1)

ผู้วิจัยใช้ค่าเฉลี่ยความเอนเอียงสัมพันธ์ (RB) เพื่อตรวจสอบความถูกต้องและความแม่นยำของวิธีทดแทนค่าสูญหายในการประมาณค่าพารามิเตอร์ β_1 จะสังเกตเห็นว่าค่าเฉลี่ย RB มีค่าใกล้เคียงกันเมื่อพิจารณาภายในเงื่อนไขเดียวกัน ขณะที่หากพิจารณารายเงื่อนไขพบว่าค่าเฉลี่ย RB จะเข้าใกล้ศูนย์เมื่ออัตราการสูญหายอยู่ในระดับต่ำ โดยที่วิธีทดแทนค่าสูญหาย Opt.impute และวิธีทดแทนค่าสูญหาย RF จะเข้าใกล้ศูนย์มากกว่าวิธีทดแทนค่าสูญหาย MI-FCS แสดงให้เห็นว่ามีค่าใกล้เคียงกับค่าพารามิเตอร์ที่แท้จริงมากกว่า ดังนั้นจึงสามารถสรุปได้ว่าวิธีทดแทนค่าสูญหาย Otp.impute มีแนวโน้มให้ค่าประมาณพารามิเตอร์ที่ถูกต้องและแม่นยำมากที่สุด รองลงมาคือ วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI - FCS ตามลำดับ

1.3 ความผิดพลาดในการประมาณค่าคลาดเคลื่อนมาตรฐานและความคงเส้นคงวา (consistency) ของการประมาณค่าพารามิเตอร์ (β_1)

เมื่อพิจารณาค่าเฉลี่ย MCSE โดยกำหนดให้ขนาดตัวอย่างและอัตราการสูญหายคงที่ จะเห็นว่า ค่าเฉลี่ย MCSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย MI - FCS ให้ค่าเฉลี่ย MCSE อยู่ในช่วง 3.24 ถึง 3.94 วิธีทดแทนค่าสูญหาย RF ให้ค่าเฉลี่ย MCSE อยู่ในช่วง 0.62 ถึง 3.62 ขณะที่วิธีทดแทนค่าสูญหาย Otp.impute มีค่าเฉลี่ย MCSE อยู่ในช่วง 0.01 ถึง 0.07 ซึ่งจะสังเกตได้ชัดว่า ค่าเฉลี่ย MCSE ที่ประมาณโดยวิธีทดแทนค่าสูญหาย Otp.impute มีแนวโน้มการลู่เข้ามากที่สุดเมื่อเปรียบเทียบกับวิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย MI - FCS ตามลำดับ แสดงว่าวิธี Otp.impute มีความผิดพลาดในการประมาณค่าพารามิเตอร์ (β_1) โดยให้ค่าคลาดเคลื่อนมาตรฐานน้อยที่สุด การจำลองวิธีทดแทนค่าสูญหายด้วยวิธีนี้มีความคงเส้นคงวา (consistency) ในการประมาณค่าพารามิเตอร์ (β_1) มากที่สุด

ตอนที่ 2 ผลการวิเคราะห์การประยุกต์ใช้วิธีทดแทนค่าสูญหายกับข้อมูลจริง

ผลการวิเคราะห์ข้อมูลพื้นฐานของนักเรียนระดับโรงเรียนชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวน 2,109 โรงเรียน ที่อยู่ภายใต้สังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา (สพม.) ทั้งหมด 42 สังกัด ซึ่งประกอบด้วยข้อมูลสองส่วน ได้แก่ ข้อมูลผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน และข้อมูลของสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน ดังนี้

เมื่อพิจารณาผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) ระดับโรงเรียน พบว่าภาพรวมคะแนนเฉลี่ยของผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนอยู่ในระดับต่ำ โดยวิชาภาษาไทยมีคะแนนเฉลี่ยสูงที่สุด (M = 54.74คะแนน) รองลงมาคือวิชาภาษาอังกฤษ (M = 32.83 คะแนน) วิชาวิทยาศาสตร์ (M = 29.67 คะแนน) และวิชาคณิตศาสตร์ (M = 24.67 คะแนน) ตามลำดับ เมื่อพิจารณาค่าสูญหายในข้อมูลสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน จะสังเกตได้ชัดว่าค่าสูญหายมีลักษณะการสูญหายเป็นรายโรงเรียน โดยมีอัตราการสูญหายสูงมากกว่าร้อยละ 50 ซึ่งจะพบว่าอัตราการสูญหายอยู่ในระดับที่สูงมากเมื่อเทียบกับข้อมูลที่เก็บรวบรวมได้

จากผลการพิจารณาความสัมพันธ์ของค่าสูญหาย ในตัวแปรสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน (x) จะพบว่าค่าที่สูญหายมีความสัมพันธ์ กับผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) ที่เก็บรวบรวมได้

โดยโรงเรียนที่มีคะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน (y) อยู่ในระดับต่ำในบางโรงเรียน จะมีโอกาสพบอัตราการสูญหายที่สูงกว่าโรงเรียนที่มีคะแนนผลสัมฤทธิ์ทางการเรียน (y) อยู่ในระดับสูงอย่างสุ่ม จึงสรุปได้ว่าการสูญหายมีลักษณะแบบสุ่ม และการวิจัยครั้งนี้ใช้ข้อมูลทุติยภูมิจึงไม่สามารถเก็บข้อมูลได้เพิ่มเติม จากการพิจารณาค่าสูญหายจากข้อมูลที่เก็บรวบรวมได้ไม่สามารถสรุปได้ว่ามีการสูญหายแบบ MCAR หรือ MNAR ด้วยหรือไม่ แต่อย่างไรก็ตามหากพิจารณาจากตัวแปรที่เกิดค่าสูญหายในสถานการณ์จริง จะพบว่ามีโอกาสที่อาจเกิดการสูญหายแบบ MNAR รวมอยู่ด้วย ผู้วิจัยจึงเลือกใช้วิธีทดแทนค่าสูญหาย Opt.knn ในการประยุกต์ใช้กับข้อมูลจริง

ตอนที่ 3 การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา

การเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา ระหว่างการใช้ข้อมูลที่มีการทดแทนค่าสูญหาย กับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย โดยใช้โมเดลพหุระดับที่ลักษณะของข้อมูลมี 2 ระดับ ระดับที่หนึ่ง คือ ระดับโรงเรียน และระดับที่สอง คือ ระดับสังกัด โดยที่ตัวแปรอิสระ คือ สัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน และตัวแปรตาม คือ ผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน ชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2563 จำนวนโรงเรียน 2,109 โรงเรียน พบว่ามีจำนวนโรงเรียน 1,140 โรงเรียน คิดเป็นร้อยละ 54.05 ไม่ตอบคำถามด้านรายได้และการพักอาศัยอยู่กับบิดามารดา ทำให้สามารถสรุปการเปรียบเทียบผลของความเหลื่อมล้ำทางการศึกษา ได้ ดังนี้

จากการเปรียบเทียบ ผลของการใช้ข้อมูลที่มีการทดแทนค่าสูญหายกับข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย พบว่าเมื่อใช้ข้อมูลที่มีการทดแทนค่าสูญหาย จะสังเกตได้ว่าหากโรงเรียนในสังกัด สพม. มีสัดส่วนของนักเรียนที่ขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับพ่อแม่เพิ่มสูงขึ้น 1 หน่วย จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียน เพิ่มขึ้นหรือลดลง อยู่ในช่วง 0.5803 ถึง - 4.7667 อย่างมีนัยสำคัญทางสถิติ ขณะที่เมื่อใช้ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย พบว่าหากโรงเรียนในสังกัด สพม. มีสัดส่วนของนักเรียนที่ขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับพ่อแม่เพิ่มสูงขึ้น จะส่งผลให้คะแนนผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนลดลง อยู่ในช่วง - 0.0018 ถึง - 1.1966 อย่างมีนัยสำคัญทางสถิติ

จะเห็นว่า หากใช้ข้อมูลที่ไม่ได้ทดแทนค่าสูญหาย อาจจะส่งผลกระทบต่อการวิจัยที่สนใจศึกษา ทั้งในส่วนของการวิเคราะห์ การประมาณค่าพารามิเตอร์ รวมไปถึงการสรุปผลตีความผลของการวิจัย โดยที่ระดับความรุนแรงของผลกระทบ อาจขึ้นอยู่กับหลายปัจจัย ได้แก่

ขนาดการสูญหายของข้อมูล (size of missing data) วิธีการแก้ไขปัญหาการสูญหายของข้อมูล (method for treating missing data) และที่สำคัญคือการพิจารณาลักษณะการสูญหายที่เกิดขึ้นหรือสาเหตุของการเกิดค่าสูญหาย (pattern of missing data or type of missing data) เป็นต้น ดังนั้นนักวิจัยจึงมีความจำเป็นต้องทราบแนวทางและกระบวนการจัดค่าสูญหาย เพื่อสามารถเลือกใช้วิธีทดแทนค่าสูญหายได้อย่างเหมาะสม เพื่อให้เกิดประสิทธิภาพสูงที่สุด นอกจากนี้ผลความเหลื่อมล้ำทางการศึกษาในครั้งนี้สะท้อนให้เห็นว่าถึงแม้ว่าประเทศไทยจะมีนโยบายการศึกษาภาคบังคับเพื่อให้นักเรียนได้รับโอกาสอย่างเท่าเทียมกัน แต่ปัญหาความไม่เท่าเทียมด้านการศึกษาก็ยังไม่หมดไป โดย จะเห็นได้จากความเหลื่อมล้ำทางการศึกษาที่เกิดขึ้นในข้อมูลชุดนี้ แม้จะอยู่ในระดับที่ไม่ได้สูงมากนัก แต่ก็นับเป็นปัญหาที่สำคัญที่ควรได้รับการแก้ไขอย่างจริงจัง เพื่อให้เกิดประชากรที่มีคุณภาพมากขึ้นประเทศก็จะพัฒนาเร็วขึ้นอย่างเห็นได้ชัด

อภิปรายผล

จากการเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหาย จำนวน 3 วิธี ได้แก่ วิธีทดแทนค่าสูญหาย FCS, วิธีทดแทนค่าสูญหาย RF และวิธีทดแทนค่าสูญหาย Opt.impute ประกอบด้วยวิธี Opt.knn, Opt.tree, Opt.svm และ Opt.cv พบว่าวิธีทดแทนค่าสูญหาย Opt.impute มีประสิทธิภาพสูงที่สุดแม้ในสถานการณ์ที่อัตราการสูญหายสูงมากถึงร้อยละ 50 เมื่อเทียบกับวิธีทดแทนค่าสูญหาย RF และวิธี FCS นอกจากนี้จะสังเกตได้ว่า เมื่อขนาดของตัวอย่างเพิ่มขึ้นส่งผลให้ค่าเฉลี่ยของ NRMSE และ RB มีแนวโน้มลดลง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เมื่อพิจารณาผลการวิจัยมีประเด็นที่น่าสนใจและนำไปสู่การอภิปรายผล โดยมีรายละเอียดดังนี้

1. ผลการเปรียบเทียบประสิทธิภาพวิธีการทดแทนค่าสูญหายด้วยค่าเฉลี่ย NRMSE และค่าเฉลี่ย RB ได้แก่ Opt.knn, Opt.tree, Opt.svm และ Opt.cv จากการจำลองข้อมูลจะเห็นได้ชัดว่าส่วนใหญ่ วิธีทดแทนค่าสูญหาย Opt.impute มีประสิทธิภาพสูงที่สุด เมื่อเทียบกับวิธีทดแทนค่าสูญหาย MI - FCS และวิธี RF แม้ในสถานการณ์ที่อัตราการสูญหายสูงมากถึงร้อยละ 50 ซึ่งให้ผลการวิจัยไม่แตกต่างกับงานวิจัยของ Bertsimas, Pawlowski, and Zhuo ปี ค.ศ. 2018 ที่พบว่า Opt.impute สามารถทดแทนค่าสูญหายได้อย่างมีประสิทธิภาพสูงแม้อัตราการสูญหายมาก เพิ่มประสิทธิภาพให้กับวิธีการทดแทนค่าสูญหายได้อย่างถูกต้องและเหมาะสมมากขึ้น นอกจากนี้เป็นที่น่าสังเกตว่าในภาพรวมเมื่อขนาดของตัวอย่างระดับที่หนึ่งและตัวอย่างขนาดที่สองมีขนาดมากขึ้นส่งผลให้ค่าเฉลี่ย NRMSE และค่าเฉลี่ย RB มีค่าเฉลี่ยลดลง ไม่แตกต่างจากงานวิจัยของ

ชนินันท์ พฤกษ์ประมูลและคณะ ที่พบว่าขนาดตัวอย่างเป็นปัจจัยที่มีอิทธิพลโดยขนาดของกลุ่มตัวอย่างที่มากขึ้นส่งผลให้การวิเคราะห์ข้อมูลมีความคลาดเคลื่อนลดลง

2. จากผลการวิเคราะห์ชี้ให้เห็นว่าสามารถนำวิธีทดแทนค่าสูญหาย Opt.impute มาประยุกต์ใช้กับข้อมูลที่มีโครงสร้างแบบพหุระดับได้อย่างสมเหตุสมผล กล่าวคือ จากการศึกษางานวิจัยของ Bertsimas, Pawlowski, and Zhuo ปี ค.ศ. 2018 พบว่า วิธีทดแทนค่าสูญหาย Opt.impute ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อวิเคราะห์และเรียนรู้รูปแบบการสูญหายโดยใช้การฝึก (train) เมื่อขนาดตัวอย่างเพิ่มสูงขึ้น ทำให้กระบวนการเรียนรู้ (Algorithm) รูปแบบการสูญหาย มีมากขึ้น จากหลักการดังกล่าว จึงทำให้การทดแทนค่าสูญหายด้วยวิธี Opt.impute มีความแม่นยำและ มีความน่าเชื่อถือมากขึ้นเมื่อขนาดตัวอย่างเพิ่มสูงขึ้นส่งผลให้การประมาณค่าพารามิเตอร์ในโมเดลมีความตรงและเข้าใกล้พารามิเตอร์ค่าจริงมากขึ้น อย่างไรก็ตามหากไม่นำปัจจัยด้านการสูญหายมาพิจารณาจะพบว่าเมื่อขนาดตัวอย่างเพิ่มสูงขึ้นยังคงช่วยให้การประมาณค่าพารามิเตอร์ดีขึ้นเช่นกัน (J. Lorah and A. Womac, 2019)

3. ถึงแม้ว่าจากผลการวิจัยในภาพรวม จะพบว่าวิธีทดแทนค่าสูญหาย 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 แต่ในทางปฏิบัติ ผู้วิจัยมีความเห็นว่าสามารถเลือกใช้วิธี RF ทดแทนได้ เนื่องจากเมื่อพิจารณาหลักการของวิธี Opt.cv พบว่าวิธี Opt.cv รวมการคำนวณ Opt.knn, Opt.svm และ Opt.tree เพื่อเลือกค่าที่ดีที่สุด จึงทำให้ใช้ระยะเวลาในการคำนวณที่ค่อนข้างนานเมื่อเทียบกับวิธีทดแทนค่าสูญหาย RF ซึ่งอาจทำให้ผู้วิจัยเสียเวลาวิเคราะห์ผลการวิจัยทั้งที่วิธี Opt.cv และวิธี RF ให้ประสิทธิภาพ ไม่แตกต่างกันมาก ดังนั้นเมื่อเกิดการสูญหายแบบ MCAR ควรเลือกใช้วิธี RF แทนวิธี Opt.cv จึงจะมีความเหมาะสมมากกว่า

3. ด้วยข้อจำกัดด้านการใช้ข้อมูลทุติยภูมิ การวิจัยครั้งนี้จึงกำหนดให้ตัวอย่างระดับที่หนึ่ง คือ ระดับโรงเรียน ระดับที่สอง คือ ระดับสังกัด แต่ถึงแม้ว่าการวิจัยครั้งนี้จะกำหนดให้ระดับที่หนึ่ง คือ โรงเรียน แต่นักวิจัยสามารถนำวิธีทดแทนค่าสูญหายจากการวิจัยครั้งนี้ภายใต้เงื่อนไขที่กำหนดไปประยุกต์ใช้กับงานวิจัยที่ข้อมูลในระดับที่หนึ่งคือระดับนักเรียนได้ เนื่องจากการจำลองข้อมูลระดับโรงเรียนในครั้งนี้ ผู้วิจัยจำลองขนาดของตัวอย่างระดับที่หนึ่งเท่ากับ 1000, 2000, และ 3000 หน่วย ขนาดของตัวอย่างระดับที่สองเท่ากับ 40, 50, และ 60 หน่วย จะสังเกตเห็นว่าขนาดของตัวอย่าง

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ระดับที่หนึ่งมีขนาดที่สอดคล้องกับจำนวนนักเรียนในสถานการณ์จริง ดังนั้นจึงสามารถนำไปประยุกต์กับนักเรียนในโรงเรียนเป็นตัวอย่างระดับที่หนึ่งได้ตามความเหมาะสม

ข้อเสนอแนะ

1. ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1.1 ผลการวิจัยระยะที่หนึ่งสะท้อนให้เห็นว่าการทดแทนค่าสูญหายก่อนนำข้อมูลไปวิเคราะห์ผลการวิจัยนั้นเป็นขั้นตอนสำคัญที่ผู้วิจัยไม่ควรละเลย โดยผู้วิจัยสามารถนำวิธีทดแทนค่าสูญหายจากการวิจัยนี้ไปประยุกต์ใช้ได้อย่างครอบคลุมเพื่อให้ผลการวิจัยอนุมานไปสู่ค่าพารามิเตอร์ได้ใกล้เคียงมากที่สุด ยกตัวอย่างเช่น จากผลการวิจัยพบว่าการทดแทนค่าสูญหายด้วยวิธี Opt.cv, วิธี Opt.knn, วิธี Opt.tree เป็นวิธีที่ทดแทนค่าสูญหายได้อย่างมีประสิทธิภาพสามารถจัดการค่าสูญหายได้อย่างถูกต้องและแม่นยำใกล้เคียงกัน สามารถเลือกใช้วิธีใดวิธีหนึ่งตามความเหมาะสม

1.2 จากผลการวิเคราะห์ข้อมูลทุติยภูมิของนักเรียนชั้นมัธยมศึกษาปีที่ 3 ระดับโรงเรียน พบว่าผลสัมฤทธิ์ทางการเรียนของนักเรียนระดับโรงเรียนมีลักษณะเบ้ขวาเล็กน้อยซึ่งสะท้อนให้เห็นว่าโดยส่วนใหญ่โรงเรียนมีค่าเฉลี่ยอยู่ในระดับต่ำกว่าค่าเฉลี่ยในขณะที่บางโรงเรียนมีนักเรียนที่สอบได้คะแนนเกือบเต็มดังนั้นควรจัดการระบบการศึกษาอย่างจริงจังโดยให้มีการพัฒนาหลักสูตรไปในมีทิศทางเดียวกันในทุกโรงเรียน เพิ่มจำนวนครูที่มีความรู้ความสามารถเฉพาะด้าน และควรส่งเสริมให้ความช่วยเหลือผู้เรียนที่ครอบครัวมีสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน อยู่ในระดับต่ำให้ได้รับโอกาสมากยิ่งขึ้น โดยหากสนับสนุนให้กับนักเรียนกลุ่มนี้อย่างเต็มที่จะสามารถเพิ่มผลสัมฤทธิ์ทางการเรียนของนักเรียนได้ประมาณ 6 เท่าหรือประมาณ 400,000 คน (ณปภัช, 2562)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.3 การแนะนำด้านการใช้โปรแกรมสำหรับวิเคราะห์ผลการวิจัย เนื่องจากในงานวิจัยครั้งนี้ใช้โปรแกรม R และโปรแกรม Julia ในการวิเคราะห์ผลการวิจัยควรระมัดระวังระหว่างการใช้โปรแกรมคือการเลือกใช้ package ให้ตรงกับ version ของโปรแกรมและถึงแม้ว่าในปัจจุบันจะมี package สำเร็จรูปมากมายที่สามารถนำมาใช้สำหรับการวิเคราะห์ได้ แต่บางครั้งการจะนำ package มาใช้วิเคราะห์ผลอาจจะต้องขออนุญาตผู้พัฒนาก่อนเพื่อไม่ให้เกิดการละเมิดลิขสิทธิ์โดยไม่รู้ตัว และควรศึกษาคำสั่งในการเขียนโปรแกรมของ package (Source code) ที่ทำการวิเคราะห์ด้วยเพราะบางครั้งชื่อและวิธีการเดียวกันแต่กระบวนการภายในของโปรแกรมที่ต้องการศึกษาแตกต่างกับทฤษฎีที่จะนำมาใช้ในการวิจัย เป็นต้น

2. ข้อเสนอแนะในการนำการวิจัยครั้งต่อไป

2.1 ผลการวิจัยในครั้งนี้เปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหาย ลักษณะการสูญหาย 7 ลักษณะ ในค่าสังเกตของข้อมูล การวิจัยในครั้งต่อไปอาจศึกษาภายใต้ลักษณะการสูญหาย 7 ลักษณะในตัวแปรเดียวกล่าวคือค่าสังเกต หนึ่งค่าอาจเกิดลักษณะการสูญหายสองลักษณะ เป็นต้น เพื่อทดสอบประสิทธิภาพของวิธีทดแทนค่าสูญหายซึ่งผลการวิจัยที่ได้จะทำให้ได้สารสนเทศที่เป็นประโยชน์ต่อการวิเคราะห์ผลการวิจัยให้มีประสิทธิภาพมากยิ่งขึ้น

2.2 การวิจัยในครั้งนี้ใช้ข้อมูลทุติยภูมิในการประยุกต์ใช้กับข้อมูลจริงทำให้มีข้อจำกัดด้านตัวแปรสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน สองประเด็นดังนี้ ประแรกคือเนื่องจากเป็นตัวแปรที่ผู้เรียนไม่สามารถเปลี่ยนแปลงได้หรือเปลี่ยนแปลงได้ยาก เช่น รายได้และลักษณะความเป็นอยู่ ประเด็นที่สองคือผู้วิจัยไม่สามารถเก็บข้อมูลได้เพิ่มเติมจึงทำให้ไม่สามารถพิสูจน์ลักษณะการสูญหายของข้อมูลได้อย่างแน่ชัด ดังนั้นในการวิจัยครั้งต่อไปอาจจะเก็บรวมรวมข้อมูลปฐมภูมิเพื่อศึกษาและเปรียบเทียบผลการใช้ข้อมูลสูญหาย เป็นต้น

2.3 ในการวิจัยครั้งนี้ศึกษาอิทธิพลของสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน เพื่อสะท้อนความเหลื่อมล้ำทางการศึกษาด้วยการวิเคราะห์พหุระดับโดยใช้โมเดลสัมประสิทธิ์ความถดถอยแบบสุ่ม ดังนั้นการวิจัยในครั้งต่อไปอาจจะพิจารณาปัจจัยอื่นเพิ่มเติมเพื่ออธิบายเศษเหลือของผลสัมฤทธิ์ทางการเรียนของนักเรียนในระดับโรงเรียน เช่น การศึกษาความสามารถของผู้ปกครอง ลักษณะการเลี้ยงดูหรือระดับการดูแลเอาใจใส่ของผู้ปกครอง และการศึกษาของผู้ปกครอง เป็นต้น

2.4 การจำลองข้อมูลสูญหายในการวิจัยครั้งนี้ ผู้วิจัยกำหนดค่าสูญหายในตัวแปรระดับที่หนึ่ง ดังนั้นผลวิธีทดแทนค่าสูญหายในการวิจัยครั้งนี้สามารถนำไปประยุกต์กับการสูญหายในระดับที่หนึ่งได้ การวิจัยครั้งต่อไปควรศึกษาเพิ่มเติมเมื่อเกิดการสูญหายในตัวแปรระดับที่สอง ผลการวิจัยที่ได้จะครอบคลุม และเป็นประโยชน์ต่อการวิเคราะห์ผลการวิจัยให้มีประสิทธิภาพมากยิ่งขึ้น

2.5 การจำลองข้อมูลในการวิจัยครั้งนี้เพื่อเปรียบเทียบประสิทธิภาพของวิธีทดแทนค่าสูญหายในข้อมูลพหุระดับสำหรับการประมาณค่าพารามิเตอร์อิทธิพลของสัดส่วนของนักเรียนที่ครอบครัวขาดแคลนทุนทรัพย์และไม่ได้อาศัยอยู่กับบิดามารดาระดับโรงเรียน ผู้วิจัยประมาณค่าพารามิเตอร์ β_{1j} เพียงตัวเดียว ในบริบทของความเหลื่อมล้ำทางกาศึกษา ดังนั้นการวิจัยครั้งต่อไปควรเปรียบเทียบผลการประมาณค่าพารามิเตอร์มากกว่าหนึ่งตัวแปร เช่น ด้านบริหารจัดการของโรงเรียน เป็นต้น

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บรรณานุกรม

Alberto Alesina, S. H., Stelios Michalopoulos, Elias Papaioannou. (2018). Educational Inequality and Intergenerational Mobility in Africa.

Angela M Wood, I. R. W., Simon G Thompson. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals.

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2018). From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research* 18 (2018) 1-39.

Black, A. C., Harel, O., & Betsy McCoach, D. (2011). Missing data techniques for multilevel data: implications of model misspecification. *Journal of Applied Statistics,* 38(9), 1845-1865. doi:10.1080/02664763.2010.529882

Buuren, S. v. (2012). *Flexible Imputation of Missing Data*.

Carol M. Musil, C. B. W., Piyanee Klainin Yobas, Susan L. Jones. (2002). A Comparison of Imputation Techniques for Handling Missing Data. *Western Journal of Nursing,* 24(7), 815-829. doi:10.1177/019394502237390

Chen, G., & Åstebro, T. (2016). How to Deal with Missing Categorical Data: Test of a Simple Bayesian Method. *Organizational Research Methods,* 6(3), 309-327. doi:10.1177/1094428103254672

Cro, S., Morris, T. P., Kenward, M. G., & Carpenter, J. R. (2020). Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Stat Med*. doi:10.1002/sim.8569

Daniel P. Loucks , E. v. B. (2017). *Water Resource Systems Planning and Management*.

Gauthier, B., & Punyasavatsut, C. (2019). Inequalities in Presence of a School Funding Formula: The 15-year Free Education Program in Thailand. *International Journal of Educational Development,* 70. doi:10.1016/j.ijedudev.2019.102100

Gregory White, M. R., Joan Kahn. (2016). Educational Inequality in India: An Analysis of Gender Differences in Reading and Mathematics.

Harshad Hegde, N. S., Aloksagar Panny, Ingrid Glurich, Pamela Christie, Amit Acharya. (2019). MICE vs PPCA_ Missing data imputation in healthcare.

doi:10.1016//.imu.2019.1002/5

Howie, M. (2020). Educational inequality: closing the gap. *Educational Research and Evaluation*, 1-3. doi:10.1080/13803611.2020.1730085

Hughes, R. A., Heron, J., Sterne, J. A. C., & Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol,* 48(4), 1294-1304. doi:10.1093/ije/dyz032

Jia, F., & Wu, W. (2019). Evaluating methods for handling missing ordinal data in structural equation modeling. *Behav Res Methods,* 51(5), 2337-2355. doi:10.3758/s13428-018-1187-4

Jo Blanden and Stephen Machin. (2004). EDUCATIONAL INEQUALITY AND THE EXPANSION OF UK HIGHER EDUCATION. *Scottish Journal of Political Economy, Vol.* 51, *No.* 2, *May* 2004.

Kleinke, K. (2018). Multiple Imputation by Predictive Mean Matching When Sample Size Is Small. *Methodology,* 14(1), 3-15. doi:10.1027/1614-2241/a000141

Ma, G., & Wu, Q. (2019). Social capital and educational inequality of migrant children in contemporary China: A multilevel mediation analysis. *Children and Youth Services Review,* 99, 165-171. doi:10.1016/j.childyouth.2019.02.002

Marks, G. N., & Pokropek, A. (2019). Family income effects on mathematics achievement: their relative magnitude and causal pathways. *Oxford Review of Education,* 45(6), 769-785. doi:10.1080/03054985.2019.1620717

National Research Counci. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*.

Neidhöfer, G., Serrano, J., & Gasparini, L. (2018). Educational inequality and intergenerational mobility in Latin America: A new database. *Journal of Development Economics,* 134, 329-349. doi:10.1016/j.jdeveco.2018.05.016

Nissen, J., Donatello, R., & Van Dusen, B. (2019). Missing data and bias in physics education research: A case for using multiple imputation. *Physical Review Physics Education Research,* 15(2). doi:10.1103/PhysRevPhysEducRes.15.020106

Nugroho, H., & Surendro, K. (2019). *Missing Data Problem in Predictive Analytics*. Paper presented at the Proceedings of the 2019 8th International Conference on Software and Computer Applications - ICSCA '19.

CHULALONGKORN UNIVERSITY

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *J Couns Psychol,* 57(1), 1-10. doi:10.1037/a0018082

Singh, U. (2016). Applying Multiple Imputation in the Analysis of Low Birth Weight in Nepal.

Van Buuren. (2007). Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. doi:10.1177/0962280206074463.

Van Buuren. (2011). mice: Multivariate Imputation by Chained Equations in R. *Statistical Software,* 45(3).

Vidotto, D., Vermunt, J. K., & van Deun, K. (2015). Multiple Imputation of Missing Categorical Data using Latent Class Models: State of the Art. 57, 542-546.

Weiss, E. G. a. E. (2017). Education inequalities at the school starting gate. *Economic Policy Institute*.

Woessmann, L. (2014). The Economic Case for Education. *EENEE Analytical Report No.* 20.

Yin, P., & Shi, J. Q. (2019). Simulation-based sensitivity analysis for non-ignorably missing data. *Stat Methods Med Res,* 28(1), 289-308. doi:10.1177/0962280217722382

Yuan, Y. (2014). Sensitivity Analysis in Multiple Imputation for Missing Data.

Zaninotto, P., & Sacker, A. (2017). Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques. *Journal of Modern Applied Statistical Methods,* 16(2), 378-402. doi:10.22237/jmasm/1509495600

ภาษาไทย

กลุ่มยุทธศาสตร์การศึกษา. (2560). รายงานผลการทดสอบการศึกษาระดับชาติ ประจำปี 2560.

กองทุนเพื่อความเสมอภาคทางการศึกษา. (2563). กองทุนเพื่อความเสมอภาคทางการศึกษา. Retrieved from https://www.eef.or.th/.

ข้อมูลจากบัญชีรายจ่ายด้านการศึกษาแห่งชาติ 2551-2559. ข้อมูลจากบัญชีรายจ่ายด้านการศึกษาแห่งชาติ 2551-2559.

ณปภัช บรรณาการ. (2562). ความเหลื่อมล้ำทางการศึกษาของนักเรียนในโรงเรียน สังกัดสำนักงานเขตพื้นที่การศึกษามัธยมศึกษา กรุงเทพมหานคร. วารสารอิเล็กทรอนิกส์ทางการศึกษา.

ปิยะเมธ ดิลกธรสกุล. (2559). แบบจำลองการตัดสินใจ: แบบจำลองมาร์คอฟและการจัดการความไม่แน่นอน. Retrieved from

ภัทฐิดา นิลภัทรฉัตร. (2559). การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์ถดถอยโลจิสติก. (ปริญญาสถิติศาสตร์มหาบัณฑิต). จุฬาลงกรณ์ มหาวิทยาลัย

มัทยา บุตรงาม. (2555). ความสัมพันธ์ระหว่างความเหลื่อมล้ำทางการศึกษาและความเหลื่อมล้ำทางการรายได้ในประเทศไทย.

รัชวดี แสงมหะหมัด. (2560). ความเหลื่อมล้ำทางการศึกษา: คุณภาพสังคมที่คนไทยมองไม่เห็น. วารสารรัฐศาสตร์และรัฐประศาสนศาสตร์ ปีที่ 8 ฉบับที่ 1 (มกราคม-มิถุนายน 2560), 33-66.

รัฐธรรมนูญแห่งราชอาณาจักรไทย. (2560). รัฐธรรมนูญแห่งราชอาณาจักรไทย. Retrieved from http://www.ratchakitcha.soc.go.th/DATA/PDF/2560/A/040/1.PDF.

วราพร ลิ่มชูเชื้อ. (2556). การเปรียบเทียบวิธีการใส่ค่าสูญหาย ในตัวแบบการถดถอยเชิงเส้นพหุ เมื่อร้อยละของตัวแปรตามและตัวแปรอิสระต่างกัน สำหรับการสูญหายแบบนอนอิกนอร์เรเบิล. (ปริญญาสถิติศาสตร์มหาบัณฑิต). จุฬาลงกรณ์ มหาวิทยาลัย,

สถาบันทดสอบทางการศึกษาแห่งชาติ. (2560). จำนวนของนักเรียนตามช่วงคะแนนผลการทดสอบ O-NET ม.6 ปีการศึกษา 2557-2560 จำแนกตามรายวิชา. Retrieved from https://www.niets.or.th/th/

สถาบันทดสอบทางการศึกษาแห่งชาติ. (2562). สรุปผลการทดสอบทางการศึกษาแห่งชาติขั้นพื้นฐาน ระดับชั้นมัธยมศึกษาปีที่ 3 ปีการศึกษา 2562. Retrieved from http://www.newonetresult.niets.or.th/AnnouncementWeb/PDF/SummaryONETM3_2562.pdf.

สถาบันอนาคตไทยศึกษา. (2562). ข้อมูลสถาบันอนาคตไทยศึกษา: Future Trend. Retrieved from https://www.thailandff.org/. https://www.thailandff.org/

สังวรณ์ งัดกระโทก. (2545). หลักการวิเคราะห์อิทธิพลเชิงสาเหตุแบบพหุระดับ การวิเคราะห์องค์ประกอบแบบพหุระดับและการวิเคราะห์โมเดลโค้งพัฒนาการแบบพหุระดับ: วิธีการของ Muthen. วารวิธีวิทยาการวิจัย ปีที่ 15 ฉบับที่ 1.

สำนักงบประมาณของรัฐสภา. (2563). รายงานการศึกษากองทุนเพื่อความเสมอภาคทางการศึกษา ปีงบประมาณ พ.ศ. 2562. Retrieved from https://www.parliament.go.th/ewtadmin/ewt/parbudget/ewt_dl_link.php?nid=783&filename=index.

สำนักงานเลขาธิการสภาการศึกษา กระทรวงศึกษาธิการ. (2557-2558). สถิติการศึกษาของประเทศไทย.

สำนักงานเลขาธิการสภาการศึกษา กระทรวงศึกษาธิการ. (2561). รายงานการศึกษาไทย พ.ศ. 2561.

สำนักงานส่งเสริมสังคมแห่งการเรียนรู้และคุณภาพเยาวชน. (2552). ความไม่เท่าเทียมทางการศึกษา

ภัยบั่นทอนปัญญาของชาติ.

สำนักทดสอบทางการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ. (2557). ผลการประเมินคุณภาพผู้เรียนระดับชาติ ปีการศึกษา 2557. Retrieved from http://bet.obec.go.th/index/wp-content/uploads/2016/08/O-NET57.pdf.

สำนักนโยบายและแผนงานการศึกษาขั้นพื้นฐาน. (2562). สำนักนโยบายและแผนงานการศึกษาขั้นพื้นฐาน, . Retrieved from http://plan.bopp-obec.info/Home/. http://plan.bopp-obec.info/Home/

สิวะโชติ ศรีสุทธิยากร. (2562). ความเหลื่อมล้ำทางการศึกษาและปัจจัยที่ส่งผล: การวิเคราะห์พหุระดับร่วมกับดัชนีเอ็นโทรปีนัยทั่วไปที่อิงมัธยฐาน. วารสาริธีวิทยาการวิจัย ปีที่ 32 ฉบับที่ 3 (กันยายน-ธันวาคม 2562), 356-386.

สุพร ชัยเดชสุริยะ และ พัชรี วรจรัสรังสี. (2555). การเปรียบเทียบผลสัมฤทธิ์ทางการเรียนของนักเรียนโรงเรียนสาธิตจุฬาลงกรณ์มหาวิทยาลัย ฝ่ายประถม ที่มีภูมิหลัง และการสนับสนุนทางด้านการเรียนของผู้ปกครองแตกต่างกัน. วารสารครุศาสตร์.

อัมมาร สยามวาลา และคณะ. (2554). การปฏิรูปการศึกษารอบใหม่: การศึกษาอย่างมีคุณภาพทั่วถึง. Retrieved from http://tdri.or.th/wp-content/uploads/2012/10/s1_ammar.pdf.

อิทธิพล พลเหี้ยมหาญ. (2562). โรงเรียนมัธยมศึกษาขนาดเล็ก ปัญหาใหญ่ของการศึกษาไทย. *Phimoldhamma Research Institute Journal, Vol.* 6 *No.* 2 (*July - December* 2019), 93-103.

อุษณีย์ วงษ์อำมาตย์. (2555). การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์ถดถอยเชิงเส้นพหุ. (ปริญญาสถิติศาสตร์มหาบัณฑิต). จุฬาลงกรณ์ มหาวิทยาลัย,

Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวนวลรัตน์ ฉิมสุด
วัน เดือน ปี เกิด	14 มิถุนายน 2536
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	ปริญญาตรี: วิทยาศาสตรบัณฑิต (สถิติ) มหาวิทยาลัยธรรมศาสตร์
ที่อยู่ปัจจุบัน	267 ซ.รังสิต - นครนายก 39 ต.ประชาธิปัตย์ อ.ธัญบุรี จ.ปทุมธานี 12130
ผลงานตีพิมพ์	ผลงานทางวิชาการที่เคยนำเสนอในที่ประชุมวิชาการ

Roumporn Sittimongkol, Sirichan Vesarcahasart, Nuanrat Chimsud & et. (2017). An Application of Cluster Analysis on the Applying Major Selection in the Department of Mathematics and Statistics, Thammasat University, Thailand. In Hiroshima Japan, International

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University