การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม ในตัวแบบการถดถอยลอจิสติก

Comparison of Missing Data Imputation Methods for Dependent Variable in Logistic Regression Model

ธิดารัตน์ ธรรมโชโต

Tidarat Thammachoto

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied Statistics
Prince of Songkla University

2566

การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม ในตัวแบบการถดถอยลอจิสติก

Comparison of Missing Data Imputation Methods for Dependent Variable in Logistic Regression Model

ธิดารัตน์ ธรรมโชโต

Tidarat Thammachoto

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Applied Statistics
Prince of Songkla University

2566

ชื่อวิทยานิพนธ์ การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติก

ผู้เขียน นางสาวธิดารัตน์ ธรรมโชโต

สาขาวิชา สถิติประยุกต์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....

(รองศาสตราจารย์ ดร.ใกล้รุ่ง สามารถ)

คณะกรรมการสอบ

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.วราฤทธิ์ พานิชกิจโกศลกุล)

.....กรรมการ

(รองศาสตราจารย์ ดร.ใกล้รุ่ง สามารถ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.พรฑิตา ทิวทัศน์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์

.....

(ผู้ช่วยศาสตราจารย์ ดร.กวินพัฒน์ สิริกานติโสภณ)

รักษาการแทนคณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ .....
(รองศาสตราจารย์ ดร.ใกล้รุ่ง สามารถ)
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ .....
(นางสาวธิดารัตน์ ธรรมโชโต)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ .....

(นางสาวธิดารัตน์ ธรรมโชโต)

นักศึกษา

ชื่อวิทยานิพนธ์ การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติก
ผู้เขียน นางสาวธิดารัตน์ ธรรมโชโต
สาขาวิชา สถิติประยุกต์
ปีการศึกษา 2566

บทคัดย่อ

ข้อมูลสูญหายถือเป็นปัญหาที่สำคัญที่มีผลต่อการวิเคราะห์ข้อมูล ซึ่งจะนำไปสู่การสรุปผลที่มีความผิดพลาด การศึกษานี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค เมื่อเกิดการสูญหายบนตัวแปรตาม 7 วิธี ได้แก่ Mode imputation (Mode), Hot deck imputation (HD), Multiple imputation (MI), K-nearest neighbor imputation (KNN), Random forest imputation (RF), Logistic regression imputation (LR) และ Modified logistic regression imputation (MLR) ซึ่งเป็นวิธีประมาณค่าสูญหายที่พัฒนามาจากวิธี LR โดยการเปลี่ยนจากจุดตัดที่เท่ากับ 0.5 เป็นจุดตัดที่เหมาะสมสำหรับชุดข้อมูลนั้น ในการศึกษานี้จำลองให้มีการสูญหายแบบ Missing completely at random (MCAR), Missing at random (MAR), Missing not at random (MNAR) โดยกำหนดขนาดตัวอย่าง 20, 50, 100, 150, 200, 500 และ 1,000 มีเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (Estimated mean square error: EMSE) ผลการวิจัยพบว่า เมื่อข้อมูลมีขนาดเล็ก วิธี MLR ที่ได้พัฒนาขึ้นมาจะมีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่วิธี MI จะมีประสิทธิภาพดีที่สุด นอกจากนี้ยังพบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่า EMSE ลดลง และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มจะทำให้ค่า EMSE เพิ่มขึ้น

Thesis Title Comparison of Missing Data Imputation Methods for Dependent Variable in Logistic Regression Model
Author Miss Tidarat Thammachoto
Major Program Applied Statistics
Academic Year 2023

ABSTRACT

Missing data is an important issue affecting data analysis. It can lead to erroneous conclusions. The objective of this study is to compare and develop the performances of missing data imputation methods applied to binary logistic regression analysis. Seven imputation methods were applied: mode imputation (Mode), hot deck imputation (HD), multiple imputation (MI), k-nearest neighbor imputation (KNN), random forest imputation (RF), logistic regression imputation (LR), and modified logistic regression imputation (MLR), a method developed from the LR method by modifying the cutoff point from 0.5 to an optimal cutoff point for that dataset. In this study, missing data were simulated under three types of mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The simulation was run using sample sizes of 20, 50, 100, 150, 200, 500, and 1,000 and missing percentages of 10%, 20%, 30%, and 40%. The estimated mean square error (EMSE) was used to compare performances. The results revealed that the developed MLR method had the best performance with small sample sizes but the MI method had the best performance with large sample sizes. The performances of the imputation methods decreased when the percentage of missing data increased. However, when the sample size increased, performances increased.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องด้วยความกรุณาช่วยเหลือเป็นอย่างดีจาก รองศาสตราจารย์ ดร.ใกล้รุ่ง สามารถ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้สละเวลาให้คำปรึกษา ชี้แนะแนวทาง ข้อคิดเห็น และตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ตลอดระยะเวลาในการทำวิทยานิพนธ์นี้ ด้วยความเอาใจใส่อย่างดียิ่ง จนกระทั่งวิทยานิพนธ์นี้เสร็จสมบูรณ์ ผู้วิจัยรู้สึกซาบซึ้งและขอกราบขอบพระคุณอย่างสูงมา ณ โอกาสนี้

ผู้วิจัยขอกราบขอบพระคุณท่านคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ รองศาสตราจารย์ ดร. วราฤทธิ์ พานิชกิจโกศลกุล ประธานกรรมการสอบวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร.พรฑิตา ทิวทัศน์ กรรมการสอบวิทยานิพนธ์ ที่ได้สละเวลาและกรุณาให้คำแนะนำเพิ่มเติมในการแก้ไข รวมทั้งแนวคิดที่เป็นประโยชน์ต่อการปรับปรุงวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ และเจ้าหน้าที่ประจำสาขาวิทยาศาสตร์การคำนวณ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ทุก ๆ ท่าน ที่ได้ให้ความช่วยเหลือในการทำวิทยานิพนธ์ให้สำเร็จลุล่วงมาได้

นอกจากนี้ ผู้วิจัยขอกราบขอบพระคุณ “ทุนสนับสนุนบัณฑิตศึกษาจากกองทุนวิจัย คณะวิทยาศาสตร์ ประเภททุนตรี-โท ประจำปีการศึกษา 2565” ที่กรุณามอบโอกาสให้เข้าศึกษาในระดับบัณฑิตศึกษา พร้อมทั้งมอบเงินทุนการศึกษาให้แก่ข้าพเจ้า

สุดท้ายนี้ ขอกราบขอบพระคุณบิดา มารดา และสมาชิกในครอบครัวทุก ๆ คนที่ได้ให้ความช่วยเหลือในทุก ๆ ด้าน คอยสนับสนุนและให้กำลังใจอย่างดีที่สุดมาโดยตลอด จนประสบความสำเร็จ ทั้งนี้ขอขอบคุณเพื่อน ๆ และผู้ที่เกี่ยวข้องของผู้วิจัยทุกท่านที่มิได้ระบุนามที่ให้ความช่วยเหลือ และช่วยเป็นกำลังใจในการทำวิทยานิพนธ์จนสำเร็จลุล่วงไปได้ด้วยดี

ธิดารัตน์ ธรรมโชโต

สารบัญ

	หน้า
บทคัดย่อ	(5)
ABSTRACT	(6)
กิตติกรรมประกาศ	(7)
สารบัญ	(8)
สารบัญ (ต่อ)	(9)
สารบัญตาราง	(10)
สารบัญรูปภาพ	(11)
คำอธิบายสัญลักษณ์และคำย่อ	(12)
บทที่ 1 บทนำ	1
1.1 ที่มาและความสำคัญ	1
1.2 วัตถุประสงค์ของการศึกษา	3
1.3 ประโยชน์ที่คาดว่าจะได้รับ	3
1.4 ขอบเขตการวิจัย	3
1.5 นิยามศัพท์เฉพาะ	3
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง	5
2.1 การวิเคราะห์การถดถอยลอจิสติกทวิภาค	5
2.2 รูปแบบการสูญหายของข้อมูล	6
2.3 วิธีการประมาณค่าสูญหาย	7
2.3.1 วิธี Mode Imputation (Mode)	7
2.3.2 วิธี Hot Deck Imputation (HD)	7
2.3.3 วิธี Multiple Imputation (MI)	8
2.3.4 วิธี K-nearest Neighbor Imputation (KNN)	9
2.3.5 วิธี Random Forest Imputation (RF)	12
2.3.6 วิธี Logistic Regression Imputation (LR)	13
2.3.7 วิธี Modified Logistic Regression Imputation (MLR)	14

สารบัญ (ต่อ)

	หน้า
2.4 งานวิจัยที่เกี่ยวข้อง	16
2.5 กรอบแนวคิดการวิจัย	18
บทที่ 3 ระเบียบวิธีวิจัย	19
3.1 การศึกษาด้วยข้อมูลจำลอง	19
3.2 การศึกษาด้วยข้อมูลจริง	20
3.3 แผนผังขั้นตอนการดำเนินงาน	21
บทที่ 4 ผลการศึกษา	24
4.1 ผลการศึกษาจากข้อมูลจำลอง	24
4.2 ผลการศึกษาจากข้อมูลจริง	29
บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ	33
5.1 สรุปผลการศึกษา	33
5.2 อภิปรายผลการศึกษา	38
5.3 ข้อเสนอแนะ	39
บรรณานุกรม	40
ภาคผนวก	44
ประวัติผู้เขียน	51

สารบัญตาราง

หน้า

ตารางที่ 2.1 ตัวอย่างข้อมูลจำลองในการคำนวณ 10
ตารางที่ 2.2 ระยะห่างระหว่างจุด 3 ลำดับแรก 11
ตารางที่ 2.3 การประมาณค่าสูญหายด้วยวิธี KNN 11
ตารางที่ 2.4 การเปรียบเทียบระหว่างค่าพยากรณ์และค่าสังเกต 14
ตารางที่ 4.1 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจำลอง 24
ตารางที่ 4.2 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจำลอง 26
ตารางที่ 4.3 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจำลอง 27
ตารางที่ 4.4 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจริง 30
ตารางที่ 4.5 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจริง 31
ตารางที่ 4.6 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจริง 32
ตารางที่ 5.1 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ตัวแปรตามเกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจำลอง 35
ตารางที่ 5.2 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ตัวแปรตามเกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจริง 37

สารบัญรูปภาพ

หน้า

รูปที่ 2.1 ตัวอย่างการสูญหายของข้อมูลในรูปแปบการสูญหายแบบ MCAR, MAR และ MNAR 6
รูปที่ 2.2 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี MODE 7
รูปที่ 2.3 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี HD 8
รูปที่ 2.4 กระบวนการทำงานของ MICE ใน 1 รอบ 9
รูปที่ 2.5 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี RF (ขั้นตอนที่ 1) 12
รูปที่ 2.6 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี RF (ขั้นตอนที่ 2) 12
รูปที่ 2.7 ลักษณะทั่วไปของกราฟเส้นโค้ง ROC 15
รูปที่ 2.8 กรอบแนวคิดการวิจัย 18
รูปที่ 3.1 ขั้นตอนการดำเนินงาน 21
รูปที่ 3.2 ขั้นตอนการดำเนินงาน (ต่อ) 22
รูปที่ 3.3 ขั้นตอนการดำเนินงาน (ต่อ) 23
รูปที่ 4.1 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจริง 30
รูปที่ 4.2 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจริง 31
รูปที่ 4.3 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจริง 32

คำอธิบายสัญลักษณ์และคำย่อ

n	ขนาดตัวอย่าง
NA	ข้อมูลสูญหาย
$\%mis$	เปอร์เซ็นต์การสูญหายของข้อมูล
π	ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ
DD	Direct Deletion
EM	Expectation Maximization Algorithm
EMSE	Estimated Mean Square Error
HD	Hot Deck Imputation
KNN	K-nearest Neighbor Imputation
LD	Listwise Deletion
LR	Logistic Regression Imputation
MAR	Missing at Random
MCAR	Missing Completely at Random
MEAN	Mean Imputation
MED	Median Imputation
MI	Multiple Imputation
MICE	Multiple Imputation by Chained Equations
MLR	Modified Logistic Regression Imputation
MNAR	Missing Not at Random
Mode	Mode Imputation
RF	Random Forest Imputation
RI	Regression Imputation
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SRI	Stochastic Regression Imputation

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

การวิเคราะห์การถดถอยลอจิสติก (Logistic regression analysis) เป็นเทคนิคการวิเคราะห์ข้อมูลที่มีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตาม (Dependent variable) และตัวแปรอิสระ (Independent variable) โดยที่ตัวแปรอิสระอาจจะเป็นตัวแปรเชิงปริมาณเพียงอย่างเดียว หรืออาจมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพก็ได้ ในขณะที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มีค่าเป็นไปได้ 2 ค่า หรือมากกว่านั้น ซึ่งถ้ามี 2 ค่า เรียกว่า การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary logistic regression) ซึ่งในงานวิจัยส่วนใหญ่ได้นำเอาการวิเคราะห์การถดถอยลอจิสติกทวิภาค มาเป็นเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูล เพื่อใช้ในการพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจศึกษา สามารถแบ่งตัวแปรตามได้เป็น 2 กลุ่ม คือ กลุ่มที่เกิดเหตุการณ์ที่สนใจ และกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ ตัวอย่างเช่น ทำนายการเป็นโรค (เป็น, ไม่เป็น) หรือทำนายผลการสอบ (ผ่าน, ไม่ผ่าน) เป็นต้น ซึ่งการวิเคราะห์การถดถอยลอจิสติกทวิภาคนี้ ได้ถูกนำไปใช้อย่างแพร่หลายในด้านทางการแพทย์ ด้านเศรษฐศาสตร์ และอื่น ๆ อีกมากมาย

ปัญหาที่พบได้บ่อยในการปฏิบัติการดำเนินการเก็บรวบรวมข้อมูลจากการทดลอง การสำรวจ การสัมภาษณ์ และการสังเกต นั่นคือ การเกิดข้อมูลสูญหาย (Missing data) ซึ่งถือเป็นปัญหาที่มีความสำคัญสำหรับการวิเคราะห์ข้อมูล เนื่องจากหากนำข้อมูลที่ไม่มีความสมบูรณ์ไปใช้ในการวิเคราะห์การถดถอยย่อมทำให้ประสิทธิภาพในการวิเคราะห์ข้อมูลลดลง ผลลัพธ์ที่ได้อาจเกิดความเอนเอียง ซึ่ง Little & Rubin (2002) ได้กล่าวถึงการจำแนกประเภทของข้อมูลสูญหายไว้ 3 ประเภท (1) การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random: MCAR) เป็นลักษณะการสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด ไม่ขึ้นกับตัวแปรตัวอื่นหรือการสูญหายของตัวเอง (2) การสูญหายแบบสุ่ม (Missing at random: MAR) เป็นลักษณะการสูญหายอย่างสุ่มที่ขึ้นอยู่กับตัวแปรอื่น และ (3) การสูญหายแบบไม่สุ่ม (Missing not at random: MNAR) เป็นลักษณะการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มแต่ขึ้นอยู่กับตัวแปรเดียวกันรวมถึงตัวแปรตัวอื่น ๆ ด้วย

ในบางครั้งข้อมูลที่ได้จากการเก็บรวบรวมอาจเกิดการสูญหายบนตัวแปรตามสำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค ซึ่งเป็นตัวแปรเชิงคุณภาพตัวอย่างเช่น ในการวิเคราะห์การทำนายโรคของผู้ป่วย ซึ่งข้อมูลประกอบด้วยตัวแปรที่เป็นปัจจัยเสี่ยงที่มีผลต่อการเป็นโรคและตัวแปรที่บ่งบอกว่าผู้ป่วยคนนี้เป็นโรคหรือไม่ มีความเป็นไปได้ที่จะมีผู้ลืมตอบว่าเป็นโรคนี้หรือไม่ จะทำให้ข้อมูลนี้เกิดการสูญหายบนตัวแปรตาม จึงต้องมีการจัดการกับข้อมูลสูญหาย ซึ่งมีวิธีประมาณค่าสูญ

หายที่ง่ายที่สุดคือ การตัดข้อมูลที่สูญหายทิ้งและนำข้อมูลที่สมบูรณ์เท่านั้นมาวิเคราะห์ โดยวิธีนี้จะทำให้ขนาดของข้อมูลลดน้อยลง ส่งผลต่อการสูญเสียระดับความเชื่อมั่นและนำไปสู่การสรุปผลที่ผิดพลาด จึงมีนักวิจัยหลายท่านคิดค้นและพัฒนาวิธีการประมาณค่าสูญหายให้มีประสิทธิภาพที่ดีขึ้น

ในการวิจัยครั้งนี้ผู้วิจัยได้ทำการศึกษางานวิจัยที่เกี่ยวข้อง ซึ่งพบว่า งานวิจัยที่ศึกษาเกี่ยวกับวิธีประมาณค่าสูญหายในการวิเคราะห์การถดถอยลอจิสติกมีน้อย โดยเฉพาะเมื่อเกิดการสูญหายบนตัวแปรตาม ซึ่งมีงานวิจัยที่ศึกษาน้อยมาก ผู้วิจัยจึงได้ทำการศึกษาวิธีประมาณค่าสูญหายของตัวแปรเชิงคุณภาพในการวิเคราะห์การถดถอยลอจิสติกเพิ่มเติม เพื่อสนับสนุนงานวิจัยที่เกี่ยวข้องให้มากขึ้น

Xu et al. (2020) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติก เมื่อข้อมูลที่สูญหายเป็นตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ โดยเสนอวิธีประมาณค่าสูญหาย 4 วิธี ได้แก่ Direct deletion (DD), Mode imputation (Mode), Hot deck imputation (HD) และ Multiple imputation (MI) ผลวิจัยพบว่า วิธี MI มีประสิทธิภาพดีที่สุด รองลงมา คือ วิธี HD, DD และ Mode ตามลำดับ ซึ่งผลการวิจัยนี้สอดคล้องกับงานวิจัยของ Tsiampalis & Panagiotakos (2020) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติกและตัวแบบปัวซง เมื่อข้อมูลที่สูญหายเป็นตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ ผลวิจัยพบว่า วิธี MI มีประสิทธิภาพดีที่สุด Waljee et al. (2013) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อข้อมูลเกิดการสูญหายแบบ MCAR บนตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพในตัวแบบการถดถอยลอจิสติก โดยเสนอวิธีการประมาณค่าสูญหาย 4 วิธี ได้แก่ Random forest imputation (RF), Mean imputation (Mean), K-nearest neighbor imputation (KNN) และ MI ผลการวิจัยพบว่า วิธี RF มีประสิทธิภาพดีที่สุด รองลงมาคือ วิธี MI, KNN และ Mean ในงานวิจัยของ Raghunathan et al. (2001) ได้กล่าวว่า Logistic Regression Imputation (LR) เหมาะแก่การประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติกทวิภาค สำหรับตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพ

ดังนั้นในงานวิจัยฉบับนี้ผู้วิจัยจึงมีความสนใจที่จะศึกษาวิธีการประมาณค่าสูญหายบนตัวแปรตาม ในการศึกษานี้ได้เสนอวิธี Modified logistic regression imputation (MLR) ซึ่งเป็นวิธีประมาณค่าสูญหายที่พัฒนามาจากวิธี LR โดยการเปลี่ยนจากจุดตัดที่เท่ากับ 0.5 เป็นจุดตัดที่เหมาะสมสำหรับชุดข้อมูลนั้นโดยเฉพาะ ซึ่งจุดตัดที่ดีที่สุดจะขึ้นอยู่กับเส้นโค้ง Receiver operating characteristic (ROC) เพื่อแบ่งผลลัพธ์ของการพยากรณ์ออกเป็น 2 กลุ่ม จากนั้นทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายบนตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค กับวิธีการประมาณค่าสูญหายที่ยอดนิยมอีก 6 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF และ วิธี LR โดยมีการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรตาม กำหนดขนาดตัวอย่าง

เท่ากับ 20, 50, 100, 150, 200, 500 และ 1,000 ที่มีเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพจะพิจารณาจากค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (Estimated mean square error: EMSE) โดยวิธีที่ให้ค่า EMSE ต่ำที่สุด คือ วิธีที่มีประสิทธิภาพดีที่สุด

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อพัฒนาวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1.3.1 สามารถใช้เป็นแนวทางในการเลือกใช้วิธีประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค ให้เหมาะสมกับลักษณะของข้อมูล

1.3.2 สามารถใช้เป็นแนวทางในการศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายในตัวแบบและสถานการณ์สูญหายอื่น ๆ

1.4 ขอบเขตการวิจัย

การศึกษาครั้งนี้จะมุ่งศึกษาถึงการพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 7 วิธี เมื่อการสูญหายเกิดขึ้นบนตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค มีการสูญหายแบบ MCAR, MAR และ MNAR โดยข้อมูลได้จากการจำลอง ประกอบด้วยตัวแปรอิสระ 2 ตัว ที่มีการแจกแจงปรกติมาตรฐาน และตัวแปรตาม 1 ตัว ที่เป็นตัวแปรไบนารี มีขนาดตัวอย่างเท่ากับ 20, 50, 100, 150, 200, 500 และ 1,000 กำหนดเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง นอกจากนี้มีการนำไปประยุกต์ใช้กับชุดข้อมูลในชีวิตจริงคือ ชุดข้อมูลการทำนายโรคหัวใจ และทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายด้วยการพิจารณาจากค่า EMSE

1.5 นิยามศัพท์เฉพาะ

1.5.1 ข้อมูลสูญหาย (Missing Data) หมายถึง ข้อมูลที่ไม่มีการระบุค่าหรือค่าว่าง ซึ่งอาจเกิดได้จากหลายสาเหตุ เช่น ผู้ให้ข้อมูลไม่ประสงค์จะตอบ ผู้ให้ข้อมูลข้ามคำถามบางข้อ ผู้ให้ข้อมูลไม่

สามารถตอบคำถามบางข้อได้และข้อมูลที่มีการเก็บรวบรวมไว้แล้วแต่เป็นข้อมูลที่ไม่สมบูรณ์ (ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และสุคนธ์ ประสิทธิ์วัฒนเสรี, 2559)

1.5.2 เปอร์เซ็นต์ของข้อมูลสูญหาย (Percentage of Missing Data) คือ สัดส่วนของข้อมูลที่สูญหายในรูปของเปอร์เซ็นต์เทียบกับจำนวนทั้งหมด

1.5.3 ขนาดตัวอย่าง (Sample size) หมายถึง จำนวนหน่วยตัวอย่างที่ต้องนำมาใช้ในการศึกษาวิจัย (ระพีพรรณ ฉลองสุข, 2550)

1.5.4 การสุ่มตัวอย่างอย่างง่าย (Simple Random Sampling) เป็นการสุ่มที่สมาชิกทุกหน่วยของประชากรที่มีจำนวนไม่มากนักแต่มีโอกาสอย่างเท่าเทียมกัน และเป็นอิสระจากกันที่จะได้เป็นตัวอย่าง เหมาะสมสำหรับใช้กับประชากรที่มีสภาพคล้ายคลึงกัน (สมชาย วรกิจเกษมสกุล, 2554)

1.5.5 วิธีการประมาณค่าข้อมูลสูญหาย (Imputation Methods) เป็นวิธีการประมาณค่าสูญหายโดยเอาหลักการทางคณิตศาสตร์ มาเติมเต็มค่าที่สูญหายไป ทำให้ผลลัพธ์สุดท้ายคล้ายกับว่าไม่เคยมีข้อมูลสูญหายเกิดขึ้นมาก่อน (จุฑารัตน์ จันชัยภูมิ, 2564)

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค จำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR การศึกษาในครั้งนี้ผู้วิจัยได้ทบทวนเอกสารและงานวิจัยที่เกี่ยวข้อง เพื่อมาสนับสนุนงานวิจัย ซึ่งสามารถอธิบายรายละเอียดได้ ดังนี้

2.1 การวิเคราะห์การถดถอยลอจิสติกทวิภาค

2.2 รูปแบบการสูญหายของข้อมูล

2.3 วิธีการประมาณค่าสูญหาย

2.4 งานวิจัยที่เกี่ยวข้อง

2.5 กรอบแนวคิดการวิจัย

2.1 การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression Analysis)

การวิเคราะห์การถดถอยลอจิสติกทวิภาค เป็นการวิเคราะห์หาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตาม เพื่อประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ โดยที่ตัวแปรอิสระ (X)เป็นตัวแปรเชิงปริมาณหรือเชิงคุณภาพก็ได้ แต่ตัวแปรตาม (Y) จะต้องเป็นตัวแปรเชิงคุณภาพ ซึ่งมีค่าได้ 2 ค่า คือ 0 และ 1 เมื่อ Y = 1 แทน การเกิดเหตุการณ์ที่สนใจ และ Y = 0 แทน ไม่เกิดเหตุการณ์ที่สนใจศึกษา นั่นคือ ตัวแปรตามมีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution) (ภัทฐิดา นิลภัทรฉัตร, 2559)

ตัวแปรตาม (Y) และตัวแปรอิสระ (X) มีความสัมพันธ์ภายใต้ตัวแบบการถดถอยลอจิสติกทวิภาค เขียนตัวแบบได้ดังนี้ (ปูเป้ สุดศิลา และคณะ, 2561)

$$\pi = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2}}$$

โดยที่ π แทน ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

$\beta_0, \beta_1, \beta_2$ แทน ค่าสัมประสิทธิ์การถดถอย

2.2 รูปแบบการสูญหายของข้อมูล

รูปแบบการสูญหายของข้อมูลเป็นปัจจัยหลักที่มีอิทธิพลต่อผลการวิเคราะห์มากกว่าขนาดการสูญหายของข้อมูล นักวิจัยจำเป็นต้องทราบรูปแบบการสูญหายของข้อมูลเพื่อใช้ในการเลือกวิธีการแก้ปัญหาที่เหมาะสม รูปแบบการสูญหายของข้อมูลนั้นสามารถแบ่งออกได้เป็น 3 ประเภทหลัก ดังนี้ (สิวะโชติ ศรีสุทธิยากร, 2557)

1. Missing completely at random (MCAR) เป็นรูปแบบการสูญหายที่เกิดขึ้นอย่างสุ่ม โดยข้อมูลที่สูญหายไม่มีความสัมพันธ์กับตัวแปรอื่น ๆ เช่น ผู้ให้ข้อมูลลืมตอบคำถาม

2. Missing at random (MAR) เป็นรูปแบบการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มโดยสมบูรณ์ทั้งชุดข้อมูลค่าสังเกต แต่เกิดขึ้นอย่างสุ่มเพียงบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือ ข้อมูลสูญหายมีความสัมพันธ์กับตัวแปรอื่น ๆ ที่ทราบค่า แต่ไม่มีความสัมพันธ์กับตัวแปรที่เกิดข้อมูลสูญหาย เช่น โอกาสในการตอบคำถามความพึงพอใจจะขึ้นอยู่กับคะแนน

3. Missing not at random (MNAR) เป็นรูปแบบการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายมีความสัมพันธ์กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรอื่น ๆ ด้วย เช่น โอกาสในการตอบคำถามความพึงพอใจจะขึ้นอยู่กับความพึงพอใจเอง

MCAR

ความพึงพอใจ	คะแนน
พอใจ	65
ไม่พอใจ	30
?	45
พอใจ	80
ไม่พอใจ	55
ไม่พอใจ	20
?	95

MAR

ความพึงพอใจ	คะแนน
พอใจ	65
?	30
ไม่พอใจ	45
พอใจ	80
ไม่พอใจ	55
?	20
พอใจ	95

MNAR

ความพึงพอใจ	คะแนน
พอใจ	65
ไม่พอใจ	30
?	45
พอใจ	80
?	55
ไม่พอใจ	20
พอใจ	95

รูปที่ 2.1 ตัวอย่างการสูญหายของข้อมูลในรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR

2.3 วิธีการประมาณค่าสูญหาย

2.3.1 วิธี Mode Imputation (Mode)

Xu et al. (2020) ได้กล่าวว่าวิธี Mode เป็นวิธีประมาณค่าสูญที่ง่ายที่สุด ซึ่งเหมาะสำหรับตัวแปรเชิงคุณภาพ ใช้หลักการแทนค่าสูญหายด้วยฐานนิยมของข้อมูลที่ไม่ได้สูญหาย โดยข้อมูลที่สูญหายในตัวแปรเดียวกันจะถูกแทนด้วยฐานนิยม ซึ่งฐานนิยม หมายถึง ข้อมูลที่เกิดซ้ำมากที่สุดหรือข้อมูลที่มีความถี่มากที่สุด

y	x_1	x_2
1	-2.987	-0.042
1	-0.312	-0.562
0	0.572	0.173
1	2.192	0.015
0	-0.998	-0.837
NA	0.168	-0.767
1	0.034	-0.704
0	-0.222	-1.498
1	-1.133	0.752
0	-0.215	0.455

y	x_1	x_2
1	-2.987	-0.042
1	-0.312	-0.562
0	0.572	0.173
1	2.192	0.015
0	-0.998	-0.837
1	0.168	-0.767
1	0.034	-0.704
0	-0.222	-1.498
1	-1.133	0.752
0	-0.215	0.455

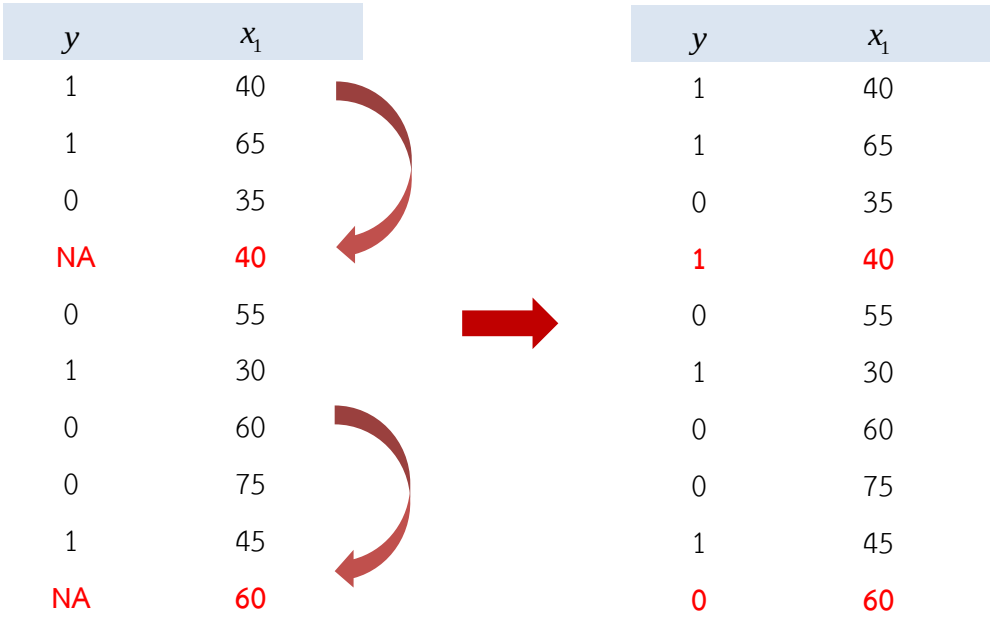
	ความถี่
1	5
0	4

Mode คือ 1

รูปที่ 2.2 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี Mode

2.3.2 วิธี Hot Deck Imputation (HD)

วิธี HD เป็นวิธีการประมาณค่าสูญหายที่พิจารณาโดยการเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกันมากที่สุดกับหน่วยตัวอย่างที่เกิดค่าสูญหาย จากนั้นแทนค่าที่สูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงนั้น (พัชนา สุวรรณแสน, 2562) หรือในบางกรณีค่าสังเกตที่นำมาประมาณค่าสูญหายอาจจะเกิดจากการสุ่มที่เรียกว่า “ random hot deck methods” หรือใช้วิธีการคำนวณหาระยะห่างระหว่างค่าสังเกตที่สมบูรณ์กับค่าสังเกตที่มีการสูญหาย (Andridge & Little, 2010) ซึ่งการเลือกค่าสังเกตที่นำมาประมาณค่าสูญหายจะพิจารณาจากระยะห่างระหว่างค่าสังเกตที่สมบูรณ์กับค่าสังเกตที่มีการสูญหายที่มีระยะห่างน้อยที่สุด โดยทั่วไปคำนวณระห่างระหว่างจุดด้วยวิธี Euclidian Distance (Peyre et al., 2010)



รูปที่ 2.3 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี HD

2.3.3 วิธี Multiple Imputation (MI)

วิธี MI ถือว่าเป็นวิธีการประมาณค่าสูญหายที่รับความนิยมนำมาใช้กันแพร่หลายอย่างมาก เนื่องจากวิธีนี้ ข้อมูลที่สูญหายจะถูกแทนที่ด้วยชุดข้อมูลของค่าที่เป็นไปได้มากกว่า 1 (Rubin, 1986)

สำหรับการประมาณค่าสูญหายด้วยวิธี MI ในโปรแกรม RStudio โดยใช้ package ที่มีชื่อว่า MICE ซึ่งจะมีการรันชุดของตัวแบบการถดถอย โดยที่ตัวแปรแต่ละตัวที่มีข้อมูลสูญหายจะถูกจำลองแบบมีเงื่อนไขตามตัวแปรอื่น ๆ ในข้อมูล นั่นคือตัวแปรแต่ละตัวสามารถสร้างตัวแบบตามลักษณะของข้อมูลได้ เช่น ตัวแปรเชิงคุณภาพจะใช้ตัวแบบการถดถอยลอจิสติก และตัวแปรเชิงคุณภาพจะใช้ตัวแบบการถดถอยเชิงเส้น เพื่อจะได้ข้อมูลที่มีความเหมาะสม ซึ่งวิธี MICE มีขั้นตอนในการดำเนินการดังนี้ (Azur et al., 2011)

ขั้นตอนที่ 1 สำหรับตัวแปรที่มีข้อมูลสูญหาย ค่าที่สูญหายจะถูกแทนที่ด้วยค่าเฉลี่ยสำหรับตัวแปรที่เป็นตัวแปรเชิงปริมาณ หรือฐานนิยมสำหรับตัวแปรที่เป็นตัวแปรเชิงคุณภาพ ซึ่งค่าที่ถูกแทนที่เข้าไปด้วยค่าเฉลี่ยและฐานนิยมนั้น เรียกว่า Place holder

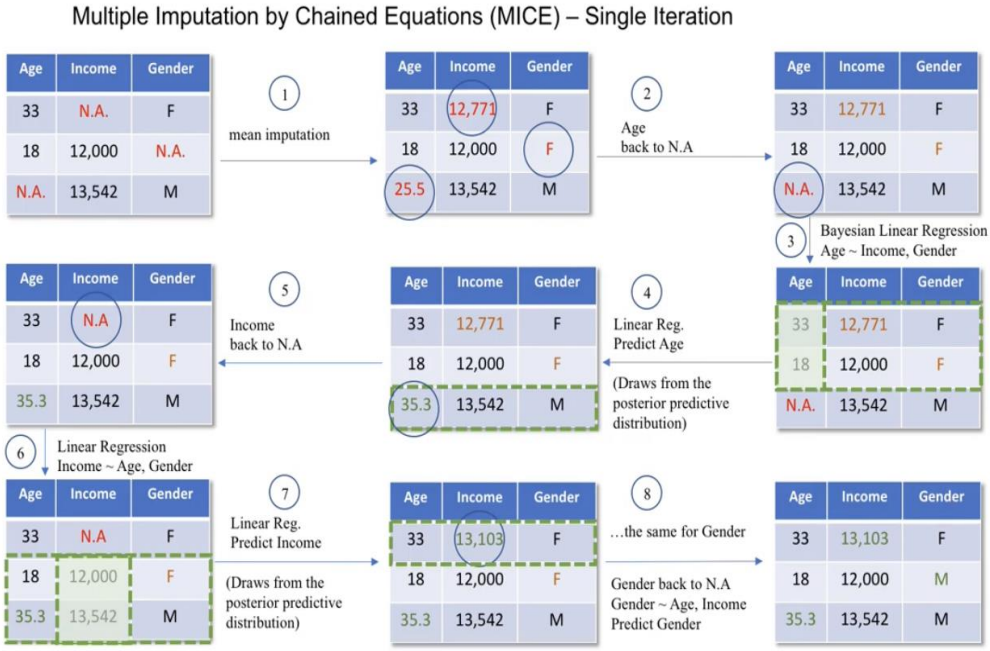
ขั้นตอนที่ 2 จากนั้น Place holder ของตัวแปรที่ต้องการทำการประมาณค่าสูญหาย จะถูกทำให้กลับมาเป็นค่าสูญหายอีกครั้ง ตัวแปรนี้เรียกว่า Var

ขั้นตอนที่ 3 ตัวแปร Var จะถูกใช้เป็นตัวแปรตาม และตัวแปรอื่น ๆ ทั้งหมดจะเป็นตัวแปรอิสระในการทำการถดถอย เพื่อหาตัวแบบการทำนายสำหรับตัวแปร Var

ขั้นตอนที่ 4 เมื่อได้ตัวแบบการทำนายมาแล้ว ค่า Place holder ของตัวแปร Var จะถูกแทนที่ด้วยค่าทำนายที่เกิดจากตัวแบบ และใช้ค่าทั้งหมดของตัวแปร Var เป็นตัวแปรอิสระ สำหรับการหาตัวแบบการทำนายของตัวแปร Var ตัวถัดไป

ขั้นตอนที่ 5 ทำซ้ำตั้งแต่ขั้นตอนที่ 2–4 เพื่อหาตัวแบบการทำนายของทุกตัวแปร Var ที่มีค่าสูญหายจนทุกตัวแปร Var ถูกเติมเต็มทั้งหมด

ขั้นตอนที่ 6 ทำซ้ำตั้งแต่ขั้นตอนที่ 2–4 เพื่อหาตัวแบบการทำนายในแต่ละรอบ เพราะในแต่ละรอบตัวแบบอาจมีค่าทำนายที่เปลี่ยนแปลงไป



รูปที่ 2.4 กระบวนการทำงานของ MICE ใน 1 รอบ

ที่มา: Philip9876, 2018

2.3.4 วิธี K-nearest Neighbor Imputation (KNN)

วิธี KNN เป็นวิธีการประมาณค่าสูญหายที่นำมาใช้กันอย่างแพร่หลาย เนื่องจากเป็นวิธีที่ใช้ง่ายและเป็นวิธีที่มีประสิทธิภาพ ซึ่งเป็นวิธีที่สามารถนำไปประยุกต์ใช้ในการจำแนกของข้อมูลได้

สำหรับขั้นตอนการทำงานของวิธี KNN เพื่อให้เกิดความเข้าใจในกระบวนการทำงานของการศึกษานี้ ผู้วิจัยจึงได้ยกตัวอย่างข้อมูลที่ได้จากการจำลอง โดยกำหนดให้ NA คือ ข้อมูลที่สูญหาย ซึ่งเกิดการสูญหายบนตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพ

ตารางที่ 2.1 ตัวอย่างข้อมูลจำลองในการคำนวณ

	y	x_1	x_2
R1	1	-2.987	-0.042
R2	1	-0.312	-0.562
R3	0	0.572	0.173
R4	0	2.192	0.015
R5	1	-0.998	-0.837
R6	NA	0.168	-0.767
R7	1	0.034	-0.704
R8	0	-0.222	-1.498
R9	1	-1.133	0.752
R10	0	-0.215	0.455

พัชนา สุวรรณแสน (2562) ได้อธิบายถึงขั้นตอนในการดำเนินงานของวิธี KNN ดังนี้

ขั้นตอนที่ 1 กำหนดค่า k ที่เหมาะสม โดย $k=\sqrt{c}$ เมื่อ c คือ จำนวนค่าสังเกตที่มีข้อมูลครบถ้วน จากตัวอย่างนี้คำนวณค่า $k=\sqrt{9}=3$

ขั้นตอนที่ 2 คำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidian Distance ระหว่างข้อมูลที่เกิดค่าสูญหายที่ต้องการพิจารณากับข้อมูลที่มีความสมบูรณ์ ดังสมการ

$$dist(R_i,R_j)=\sqrt{\sum_{p=1}^{c}(x_{i,p}-x_{j,p})^2}$$

โดยที่ $dist(R_i,R_j)$ แทน ระยะห่างระหว่างข้อมูลแถวที่ i และข้อมูลแถวที่ j

$x_{i,p}$ แทน ค่าข้อมูลที่เกิดการสูญหาย แถวที่ i คอลัมน์ที่ p

$x_{j,p}$ แทน ค่าข้อมูลที่มีความสมบูรณ์ แถวที่ j คอลัมน์ที่ p

วิธีการคำนวณหาระยะห่าง

$$dist(R_6, R_1) = \sqrt{(0.168-(-2.987))^2+(-0.767-(-0.042))^2} = 3.237$$
$$dist(R_6, R_2) = \sqrt{(0.168-(-0.312))^2+(-0.767-(-0.562))^2} = 0.522$$
$$dist(R_6, R_3) = \sqrt{(0.168-0.572)^2+(-0.767-0.173)^2} = 1.023$$
$$dist(R_6, R_4) = \sqrt{(0.168-2.192)^2+(-0.767-0.015)^2} = 2.170$$
$$dist(R_6, R_5) = \sqrt{(0.168-(-0.998))^2+(-0.767-(-0.837))^2} = 1.168$$
$$dist(R_6, R_7) = \sqrt{(0.168-0.034)^2+(-0.767-(-0.704))^2} = 0.148$$
$$dist(R_6, R_8) = \sqrt{(0.168-(-0.222))^2+(-0.767-(-1.498))^2} = 0.828$$
$$dist(R_6, R_9) = \sqrt{(0.168-(-1.133))^2+(-0.767-0.752)^2} = 1.999$$
$$dist(R_6, R_{10}) = \sqrt{(0.168-(-0.215))^2+(-0.767-0.455)^2} = 1.281$$

ขั้นตอนที่ 3 เรียงลำดับระยะห่างระหว่างจุดโดยพิจารณาจากข้อมูลที่ใกล้ที่สุดตามจำนวน $k=3$

ตารางที่ 2.2 ระยะห่างระหว่างจุด 3 ลำดับแรก

	y	x_1	x_2	dist	sort
R7	1	0.034	-0.704	0.148	1
R2	1	-0.312	-0.562	0.522	2
R8	0	-0.222	-1.498	0.828	3

ขั้นตอนที่ 4 ประมาณค่าข้อมูลสูญหายจากฐานนิยมของข้อมูลที่อยู่ใกล้ที่สุด 3 ตัวจากตัวแปรเดียวกันกับตำแหน่งที่มีการสูญหาย ซึ่งจะเห็นได้ว่าตัวแปร y ของข้อมูลที่อยู่ใกล้ที่สุด 3 ตัว ฐานนิยม คือ 1

ตารางที่ 2.3 การประมาณค่าสูญหายด้วยวิธี KNN

	y	x_1	x_2
R1	1	-2.987	-0.042
R2	1	-0.312	-0.562
R3	0	0.572	0.173
R4	0	2.192	0.015
R5	1	-0.998	-0.837
R6	1	0.168	-0.767
R7	1	0.034	-0.704
R8	0	-0.222	-1.498
R9	1	-1.133	0.752
R10	0	-0.215	0.455

2.3.5 วิธี Random Forest Imputation (RF)

Hong & Lynn (2020) กล่าวว่า วิธี RF สำหรับโปรแกรม Rstudio โดยใช้ package ที่มีชื่อว่า missForest ซึ่งวิธี RF มีขั้นตอนดำเนินงานดังนี้

ขั้นตอนที่ 1 แทนค่าที่เกิดการสูญหายด้วยฐานนิยมสำหรับตัวแปรเชิงคุณภาพ

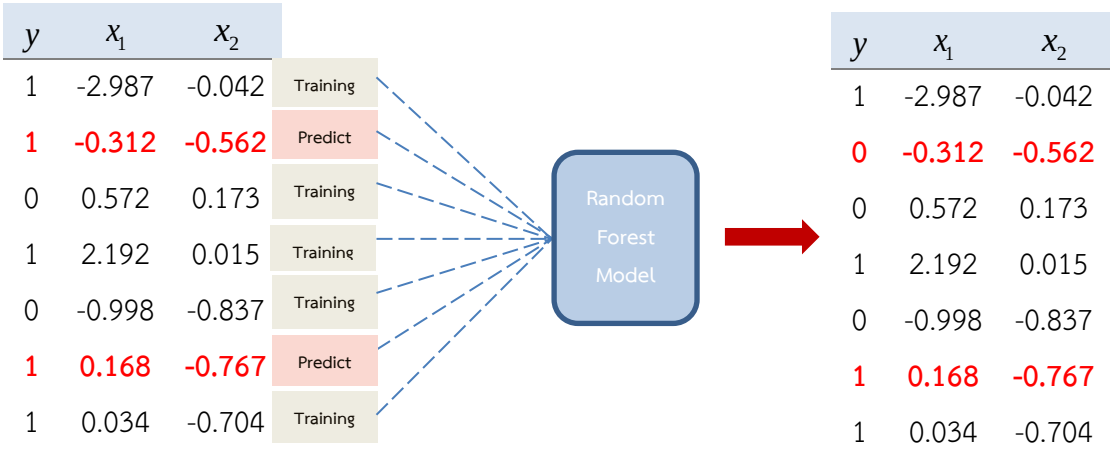
y	x_1	x_2
1	-2.987	-0.042
NA	-0.312	-0.562
0	0.572	0.173
1	2.192	0.015
0	-0.998	-0.837
NA	0.168	-0.767
1	0.034	-0.704

Impute missing values using mode

y	x_1	x_2
1	-2.987	-0.042
1	-0.312	-0.562
0	0.572	0.173
1	2.192	0.015
0	-0.998	-0.837
1	0.168	-0.767
1	0.034	-0.704

รูปที่ 2.5 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี RF (ขั้นตอนที่ 1)

ขั้นตอนที่ 2 กระบวนการประมาณค่าจะทำตามลำดับของตัวแปรที่มีข้อมูลสูญหาย โดยเรียงจากมากไปน้อย จากนั้นแบ่งข้อมูลเป็น 2 ชุด โดยที่ชุดข้อมูลที่สมบูรณ์เป็น training และข้อมูลชุดที่มีค่าสูญหายเป็น predict นำข้อมูลมาเข้าในตัวแบบ random forest โดยนำ training มาหารูปแบบของการทำนาย ใช้วิธีการสุ่มซ้ำหลายๆครั้ง เพื่อที่จะหาว่าควรแทนค่ารูปแบบใด แล้วนำมาใช้กับตัว predict เพื่อแทนค่าสูญหาย จากนั้นค่าที่สูญหายจะถูกเติมด้วยค่าทำนายที่ได้จากตัวแบบ random forest



รูปที่ 2.6 ตัวอย่างการประมาณค่าสูญหายด้วยวิธี RF (ขั้นตอนที่ 2)

ขั้นตอนที่ 3 คำนวณค่า Proportion of Falsely Classified Entries (PFC) สำหรับตัวแปรเชิงคุณภาพ ดังสมการนี้ (Guo et al., 2021)

$$PFC = \frac{\sum count(y_{true} \neq \hat{y}_{imp})}{\sum count(y_{true})}$$

โดยที่ y_{true} แทน ค่า y ที่สูญหายและได้รับการแทนค่าสูญหายด้วยฐานนิยม

$\hat{y}_{imp}$ แทน ค่าประมาณของ y ที่ได้จากตัวแบบ random forest

จากนั้นทำกระบวนการวนซ้ำจนกว่าค่า PFC_t มีค่ามากกว่าค่า PFC_{t-1} (โดยที่ t คือ ชุดข้อมูลที่ได้จากตัวแบบ random forest) โดยทั่วไปจะทำซ้ำ 4-5 ครั้ง อย่างไรก็ตามจำนวนการทำซ้ำอาจจะขึ้นอยู่กับขนาดของข้อมูลและจำนวนข้อมูลสูญหาย (Hong & Lynn, 2020)

2.3.6 วิธี Logistic Regression Imputation (LR)

วิธี LR เป็นวิธีที่ได้นำเทคนิคการวิเคราะห์การถดถอยลอจิสติกทวิภาค มาใช้ในการประมาณค่าสูญหายสำหรับตัวแปรเชิงคุณภาพ ซึ่งมีขั้นตอน ดังนี้ (ยุทธ ไกยวรรณ์, 2555)

ขั้นตอนที่ 1 นำชุดข้อมูลที่ไม่เกิดการสูญหาย มาประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation: MLE)

ขั้นตอนที่ 2 นำค่าประมาณสัมประสิทธิ์การถดถอยที่ได้จากขั้นตอนที่ 1 มาทำนายค่าสูญหายในตัวแปรตาม เขียนตัวแบบได้ ดังนี้

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}$$

เมื่อ $p(y = 1) = \hat{\pi}$

$\hat{\pi}$ แทน ความน่าจะเป็นประมาณของการเกิดเหตุการณ์ที่สนใจ

$\hat{\beta}_0$ แทน ค่าประมาณคงที่ (เมื่อไม่มีอิทธิพลจากตัวแปรอิสระใด)

$\hat{\beta}_1, \hat{\beta}_2$ แทน ค่าประมาณสัมประสิทธิ์การถดถอยของตัวแปรอิสระ x_1 และ x_2 ตามลำดับ

ขั้นตอนที่ 3 เมื่อทราบค่าความน่าจะเป็นประมาณแล้วนำมาพิจารณาด้วยจุดตัด (Cut off point) โดยทั่วไปนิยม ใช้ค่า 0.5 เป็นจุดตัด เพื่อจำแนกผลลัพธ์ออกเป็น 2 กลุ่ม ได้แก่

$\hat{\pi} \geq 0.5$ นั่นคือ ผลลัพธ์อยู่ในกลุ่มที่เกิดเหตุการณ์ที่สนใจ $(y = 1)$

$\hat{\pi} < 0.5$ นั่นคือ ผลลัพธ์อยู่ในกลุ่มที่เกิดเหตุการณ์ที่ไม่สนใจ $(y = 0)$

ขั้นตอนที่ 4 นำผลลัพธ์ที่ได้จากการพิจารณาด้วยจุดตัดมาแทนในตำแหน่งที่เกิดการสูญหายของข้อมูล

2.3.7 วิธี Modified Logistic Regression Imputation (MLR)

วิธี MLR เป็นวิธีที่ผู้วิจัยได้พัฒนามาจากการศึกษาวิธี LR เพื่อให้เกิดประสิทธิภาพให้มากยิ่งขึ้น โดยวิธี MLR มีขั้นตอนที่เพิ่มขึ้นมาจากวิธี LR คือการหาจุดตัดที่เหมาะสม (Optimal cut off point) เพื่อให้การจำแนกผลลัพธ์ที่แบ่งออกเป็น 2 กลุ่มมีความถูกต้องและแม่นยำยิ่งขึ้น จากนั้นจะมีขั้นตอนการดำเนินงานเช่นเดียวกับวิธี LR

จุดตัดที่เหมาะสมสามารถพิจารณาได้จากการวิเคราะห์ด้วยกราฟเส้นโค้ง ROC เป็นกราฟที่แสดงให้เห็นถึงประสิทธิภาพของการจำแนกประเภทแบบไบนารี ซึ่งตัวแปรตามเป็นตัวแปรเชิงคุณภาพแบ่งออกเป็น 2 กรณี คือ $y=1$ เมื่อเกิดเหตุการณ์ที่สนใจหรือผลการทดสอบเป็นบวก และ $y=0$ เมื่อเกิดเหตุการณ์ที่ไม่ได้สนใจหรือผลการทดสอบเป็นลบ เบญจพร เอี่ยมประโคน และณัตติ ฤดี เจริญรักษ์ (2560) กล่าวว่า จุดตัด หมายถึง จุดที่ใช้จำแนกเหตุการณ์ออกเป็นเหตุการณ์ที่สนใจกับเหตุการณ์ที่ไม่ได้สนใจพบว่า สามารถแบ่งกรณีการเปรียบเทียบระหว่างค่าพยากรณ์และค่าสังเกต ซึ่งแบ่งออกเป็น 4 กรณี ดังนี้

ตารางที่ 2.4 การเปรียบเทียบระหว่างค่าพยากรณ์และค่าสังเกต

		ค่าสังเกต	
		Positive	Negative
ค่าพยากรณ์	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

โดยที่

True positive (TP) คือ จำนวนเหตุการณ์ที่สนใจมีผลเป็นบวก และมีผลจากการพยากรณ์เป็นบวก

False positive (FP) คือ จำนวนเหตุการณ์ที่สนใจมีผลเป็นลบ แต่มีผลจากการพยากรณ์เป็นบวก

False negative (FN) คือ จำนวนเหตุการณ์ที่สนใจมีผลเป็นบวก และมีผลจากการพยากรณ์เป็นลบ

True negative (TN) คือ จำนวนเหตุการณ์ที่สนใจมีผลเป็นลบ และมีผลจากการพยากรณ์เป็นลบ

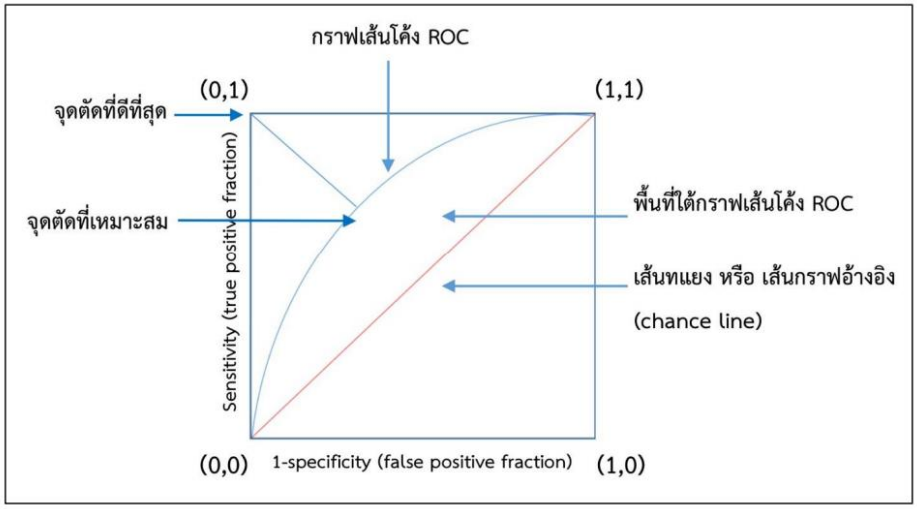
ตัวสถิติที่ใช้วัดความถูกต้องของการพยากรณ์ซึ่งผลที่เกิดจากการพยากรณ์มีเพียง 2 ค่า คือ Sensitivity or True positive rate (TPR) คือ อัตราส่วนของจำนวนค่าพยากรณ์ที่มีผลเป็นบวกที่ทำนายถูกต่อจำนวนเหตุการณ์ที่สนใจที่มีผลเป็นบวก จะได้

$$sensitivity = \frac{TP}{TP + FN}$$

Specificity or True negative rate (TNR) คือ อัตราส่วนของจำนวนค่าพยากรณ์ที่มีผลเป็นลบที่ทำนายถูกต่อ จำนวนเหตุการณ์ที่สนใจที่มีผลเป็นลบ จะได้

$$specificity = \frac{TN}{FP + TN}$$

กราฟเส้นโค้ง ROC จะแสดงความสัมพันธ์ระหว่าง แกน X ซึ่งแทน “1-specificity” กับแกน Y ซึ่งแทน “sensitivity” โดยจุดบนกราฟเส้นโค้ง ROC ที่เกิดขึ้น คือ จุดตัดที่เป็นไปได้ระหว่างความสัมพันธ์ของทั้งสองแกนดังกล่าว (พงษ์เดช สารการ และภัทรนันท์ หมั่นพลศรี, 2564)



รูปที่ 2.7 ลักษณะทั่วไปของกราฟเส้นโค้ง ROC
ที่มา: พงษ์เดช สารการ และภัทรนันท์ หมั่นพลศรี, 2564

พงษ์เดช สารการ และภัทรนันท์ หมั่นพลศรี (2564) กล่าวว่า การพิจารณาจุดตัดที่เหมาะสมด้วยวิธี Youden’s index แทนด้วยสัญลักษณ์ "J" เป็นวิธีการที่ง่ายและตรงไปตรงมา อีกทั้งยังถูกยอมรับและนำมาใช้ค่อนข้างแพร่หลาย ซึ่งเป็นวิธีการกำหนดจุดตัดแบบค่าเดียว โดยเป็นการพิจารณาระยะห่างระหว่างจุด (1-specificity, Sensitivity) ที่เป็นไปได้บนกราฟเส้นโค้ง ROC กับเส้นทแยงหรือ chance line ในแนวดิ่ง ซึ่งมีสูตรในการพิจารณา ดังนี้ (Youden, 1950)

$$J = sensitivity + specificity - 1$$
$$J = sensitivity - (1 - specificity)$$

เกณฑ์การพิจารณาจุดตัดที่เหมาะสมของวิธี Youden’s index ได้แก่ จุดตัดที่มีค่าระยะห่างระหว่างจุด (1-specificity, Sensitivity) ที่เป็นไปได้บนกราฟเส้นโค้ง ROC กับเส้นทแยงที่มีค่ามากที่สุด

2.4 งานวิจัยที่เกี่ยวข้อง

Peng & Zhu (2007) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายวิธี MI และวิธี EM สำหรับตัวแปรร่วมเชิงคุณภาพในตัวแบบการถดถอยลอจิสติก เมื่อข้อมูลเกิดการสูญหายแบบ MCAR และ MAR โดยขนาดตัวอย่างเท่ากับ 200 และ 400 เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการประมาณค่าสูญหาย ได้แก่ bias, efficiency, coverage และ rejection rate ผลการวิจัยพบว่า วิธี MI มีประสิทธิภาพดีกว่า EM

Waljee et al. (2013) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อข้อมูลเกิดการสูญหายแบบ MCAR บนตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ ในตัวแบบการถดถอยลอจิสติกและตัวแบบ Random Forest โดยวิธีการประมาณค่าสูญหายที่ใช้มี 4 วิธี ได้แก่ วิธี RF, วิธี MEAN, วิธี KNN และ วิธี MI เปอร์เซ็นต์ของการสูญหายเท่ากับ 10%, 20% และ 30% เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการประมาณค่าสูญหายคือ average relative error สำหรับตัวแปรเชิงปริมาณ และ misclassification error สำหรับตัวแปรเชิงคุณภาพ ผลการวิจัยพบว่า วิธี RF มีประสิทธิภาพดีที่สุด รองลงมาคือ วิธี MI, วิธี KNN และ วิธี MEAN ตามลำดับ ทั้งในตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ

Meeyai (2016) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อข้อมูลมีการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรอิสระ มีวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ ได้แก่ วิธี LD, วิธี MEAN, วิธี RI, วิธี SRI และวิธี MI โดยการจำลองแบบ Monte Carlo จะถูกนำมาใช้เพื่อจำลองข้อมูลแล้วสามารถนำไปประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยลอจิสติก กำหนดขนาดตัวอย่างเท่ากับ 10, 20, 30, 40, 50, 100, 250, 500, 1000, 2500 และ 5000 ที่เปอร์เซ็นต์ของการสูญหายเท่ากับ 10%, 20% 30%, 40%, 50%, 60%, 70% และ 80% ผลการวิจัยพบว่า วิธี MI ให้ประสิทธิภาพได้ดีสำหรับข้อมูลที่เกิดการสูญหายแบบ MCAR และ MAR แต่ไม่มีวิธีการใดที่ให้ประสิทธิภาพได้ดีที่สุดในกรณีที่ใช้การจัดการข้อมูลสูญหายแบบ MNAR

ภัทฐิดา นิลภัทรฉัตร (2559) ได้ศึกษาการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระที่มีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยลอจิสติกทวิภาค เมื่อมีตัวแปรอิสระ 3 ตัว และเกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง โดยวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ ได้แก่ วิธี MEAN, วิธี MED, วิธี KNN และ วิธี MI ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองข้อมูล โดยกำหนดขนาดตัวอย่าง 70, 100 และ 200 ตัวแปรอิสระที่เกิดการสูญหายมีเปอร์เซ็นต์ของการสูญหายโดยเฉลี่ย 3 ระดับ คือ 10%, 20% และ 30% มีระดับการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับ คือ การสูญหายแบบอิกนอร์เรเบิล การสูญหายแบบนอนอิกนอร์เรเบิลระดับปานกลาง และการสูญหายแบบนอนอิกนอร์เรเบิลระดับสูง กำหนดค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระทั้ง 3 ตัว คือ 0.5, 1 และ 1.5 ตามลำดับ ทำการจำลองในแต่ละสถานการณ์เป็น

จำนวน 5,000 รอบ และเกณฑ์ที่ใช้ในการเปรียบเทียบแต่ละวิธี คือ ค่า EMSE พบว่า วิธี MI จะมีประสิทธิภาพ เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายมีค่าต่ำ และขนาดตัวอย่างมีขนาดเล็ก วิธี MEAN และวิธี MED จะมีประสิทธิภาพที่ดี เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายที่มีเปอร์เซ็นต์ของการสูญหายและขนาดตัวอย่างมีขนาดใหญ่ นอกจากนี้พบว่า ค่า EMSE มีแนวโน้มเพิ่มขึ้น เมื่อเปอร์เซ็นต์การสูญหายและสัดส่วนการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น และค่า EMSE มีแนวโน้มเพิ่มขึ้น เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายมีค่าสูง

ปูเป้ สุดศิลา และคณะ (2561) ได้ศึกษาการเปรียบเทียบวิธีการประมาณค่าสูญหาย 4 วิธี ได้แก่ วิธี MEAN, วิธี MI, วิธี KNN และ วิธี Weight Locally Linear Reconstruction (WLLR) สำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค กำหนดตัวแปรอิสระเชิงปริมาณจำนวนสองตัว (x_1, x_2) เมื่อมีการสูญหายแบบ MCAR ในตัวแปรอิสระ x_1 เพียงตัวเดียวเท่านั้น ศึกษาด้วยข้อมูลจำลองและข้อมูลจริง โดยจำลองข้อมูลด้วยวิธีมอนติคาร์โลทำซ้ำ 1000 ครั้ง กำหนดขนาดตัวอย่างเท่ากับ 60, 100, 300 และ 500 เปอร์เซ็นต์การสูญหายที่ระดับ 5%, 10%, 15%, 20%, 30%, 40% และ 50% เกณฑ์ในการเปรียบเทียบคือ ค่า EMSE ผลการวิจัยพบว่า ที่ทุกเปอร์เซ็นต์การสูญหายที่กำหนด เมื่อขนาดตัวอย่างเท่ากับ 60 100 และ 300 วิธี MEAN มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่างเท่ากับ 500 วิธี MI มีประสิทธิภาพที่ดีที่สุด นอกจากนี้พบว่า ค่า EMSE เพิ่มขึ้นเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น และลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

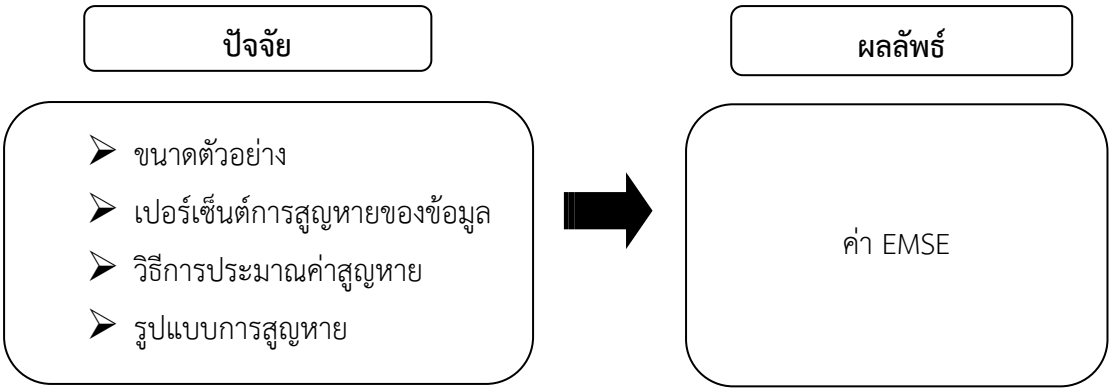
Abonazel & Ibrahim (2018) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายด้วยวิธี RI และ วิธี EM เมื่อข้อมูลเกิดการสูญหายแบบ MCAR บนตัวแปรตามและตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกทวิภาค ผลการวิจัยพบว่า วิธี RI มีประสิทธิภาพที่ดีและมีความเหมาะสมกับตัวแบบการถดถอยนี้มากกว่าวิธี EM

Stavseth et al. (2019) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อข้อมูลเกิดการสูญหายแบบ MAR บนตัวแปรเชิงคุณภาพ โดยวิธีการประมาณค่าสูญหายที่ใช้ในการศึกษามี 6 วิธี ได้แก่ วิธี Expectation–Maximization with Bootstrapping, วิธี Multiple Correspondence Analysis, วิธี Latent Class Analysis, วิธี MI, วิธี HD และ RF โดยใช้ข้อมูลจริงจากแบบสอบถามในโครงการบำรุงรักษาฝีน เมื่อขนาดตัวอย่างเท่ากับ 200 และ 1,000 เปอร์เซ็นต์การสูญหายที่ระดับ 5%, 10%, 20%, และ 40% ผลการวิจัยพบว่า ทุกวิธีมีประสิทธิภาพค่อนข้างดีเมื่อตัวอย่างมีขนาดใหญ่ แต่เมื่อตัวอย่างมีขนาดเล็ก วิธี Multiple Correspondence Analysis มีประสิทธิภาพดีที่สุดในการประมาณค่าสูญหาย

Xu et al. (2020) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติก เมื่อข้อมูลมีการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรอิสระที่เป็นตัวแปรชิงคุณภาพและตัวแปรเชิงปริมาณ โดยมีวิธีประมาณค่าสูญหาย 4 วิธี ได้แก่ วิธี DD, วิธี Mode, วิธี HD และวิธี MI กำหนดเปอร์เซ็นต์การสูญหายที่ระดับ 5%, 10%, 15% และ 20% โดยใช้ข้อมูลจริงจากแบบสอบถามทางจิตวิทยา ใช้ RMSE ในการเปรียบเทียบวิธีการประมาณค่าสูญหาย ผลวิจัยพบว่า วิธี MI มีประสิทธิภาพดีที่สุด รองลงมา คือ วิธี HD, วิธี DD และวิธี Mode ตามลำดับ

Tsiampalis & Panagiotakos (2020) ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติกและตัวแบบการถดถอยปัวซง เมื่อข้อมูลมีการสูญหายแบบ MCAR, MAR และ MNAR สำหรับตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ ข้อมูลที่ใช้ในการศึกษาได้เป็นข้อมูลจริงที่ได้จากการศึกษาทางระบาดวิทยาของ ATTICA มีขนาดตัวอย่างเท่ากับ 2,194 โดยวิธีการประมาณค่าสูญหายที่ใช้มี 7 วิธี ได้แก่ Complete case analysis (CCA), วิธี Proration, วิธี Score mean imputation (SMI), วิธี Item mean imputation (IMI), วิธี Person mean imputation (PMI), วิธี SRI และวิธี MI ผลวิจัยพบว่า วิธี MI มีประสิทธิภาพดีที่สุด

2.5 กรอบแนวคิดการวิจัย



รูปที่ 2.8 กรอบแนวคิดการวิจัย

บทที่ 3
ระเบียบวิธีวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค จำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR ซึ่งมีรายละเอียดในการดำเนินการศึกษา ดังนี้

3.1 การศึกษาด้วยข้อมูลจำลอง

3.2 การศึกษาด้วยข้อมูลจริง

3.3 แผนผังขั้นตอนการดำเนินงาน

3.1 การศึกษาด้วยข้อมูลจำลอง

จำลองข้อมูลตามสถานการณ์ต่างๆ โดยทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ด้วยโปรแกรม R Studio มีวิธีการดังนี้

3.1.1 สร้างตัวแปรอิสระสองตัว 2 ตัว ที่มีการแจกแจงปรกติมาตรฐาน $X_1 \sim N(0,1)$ และ $X_2 \sim N(0,1)$ โดยที่ตัวแปรอิสระทั้งสองตัวไม่มีความสัมพันธ์กัน ซึ่งมีจำนวน 1,000 records

3.1.2 สร้างชุดข้อมูลที่มีความสัมพันธ์กับตัวแปรอิสระภายใต้ตัวแบบการถดถอยลอจิสติกทวิภาค โดยกำหนดให้ค่าสัมประสิทธิ์การถดถอย $\beta_0, \beta_1, \beta_2 = 1$ จะเขียนตัวแบบได้ดังนี้

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$$

โดยที่ π คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

$\beta_0, \beta_1, \beta_2$ คือ ค่าสัมประสิทธิ์การถดถอย

3.1.3 กำหนดค่า (0, 1) ให้กับตัวแปรตามจากการแจกแจงแบร์นูลลี $Y \sim Ber(\pi)$ ด้วยความน่าจะเป็น (π) ที่คำนวณได้จากตัวแบบการถดถอยลอจิสติกทวิภาค ในข้อ 3.1.2

3.1.4 เลือกตัวอย่างจากการสุ่มตัวอย่างแบบง่าย (simple random sampling) โดยกำหนดขนาดตัวอย่าง (n) เท่ากับ 20, 50, 100, 150, 200, 500 และ 1,000

3.1.5 หาตัวแบบการถดถอยลอจิสติกประมาณจากตัวอย่าง

3.1.6 กำหนดให้มีรูปแบบการสูญหายของข้อมูล 3 แบบ คือ MCAR, MAR และ MNAR บนตัวแปรตามเพียงตัวเดียว โดยที่มีเปอร์เซ็นต์การสูญหายทั้งสิ้น 4 ระดับ คือ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง

3.1.7 นำข้อมูลที่จำลองการสูญหายของข้อมูลแล้วมาทำการประมาณค่าข้อมูลและแทนที่ข้อมูลที่สูญหายของตัวแปรตาม ด้วยวิธีการสูญหายทั้ง 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR

3.1.8 เมื่อแทนค่าข้อมูลสูญหายแล้วทำการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกใหม่ด้วยวิธีภาวะน่าจะเป็นสูงสุด (MLE) จะได้ค่าสัมประสิทธิ์การถดถอยที่แตกต่างกันในแต่ละสถานการณ์

3.1.9 คำนวณค่า EMSE จากวิธีการประมาณค่าสูญหายทั้ง 7 วิธี โดยการทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณสูญหาย ดังสมการนี้ (Ozkale & Arican, 2015)

$$EMSE = \frac{1}{1000}\sum_{t=1}^{1000}\sum_{b=0}^{2}(\hat{\beta}_{(b,t)} - \hat{\beta}^{*}_{(b,t)})^2$$

โดยที่

$\hat{\beta}_{(b,t)}$ คือ สัมประสิทธิ์การถดถอยประมาณของข้อมูลตัวอย่างที่ไม่มีค่าสูญหายในรอบที่ t

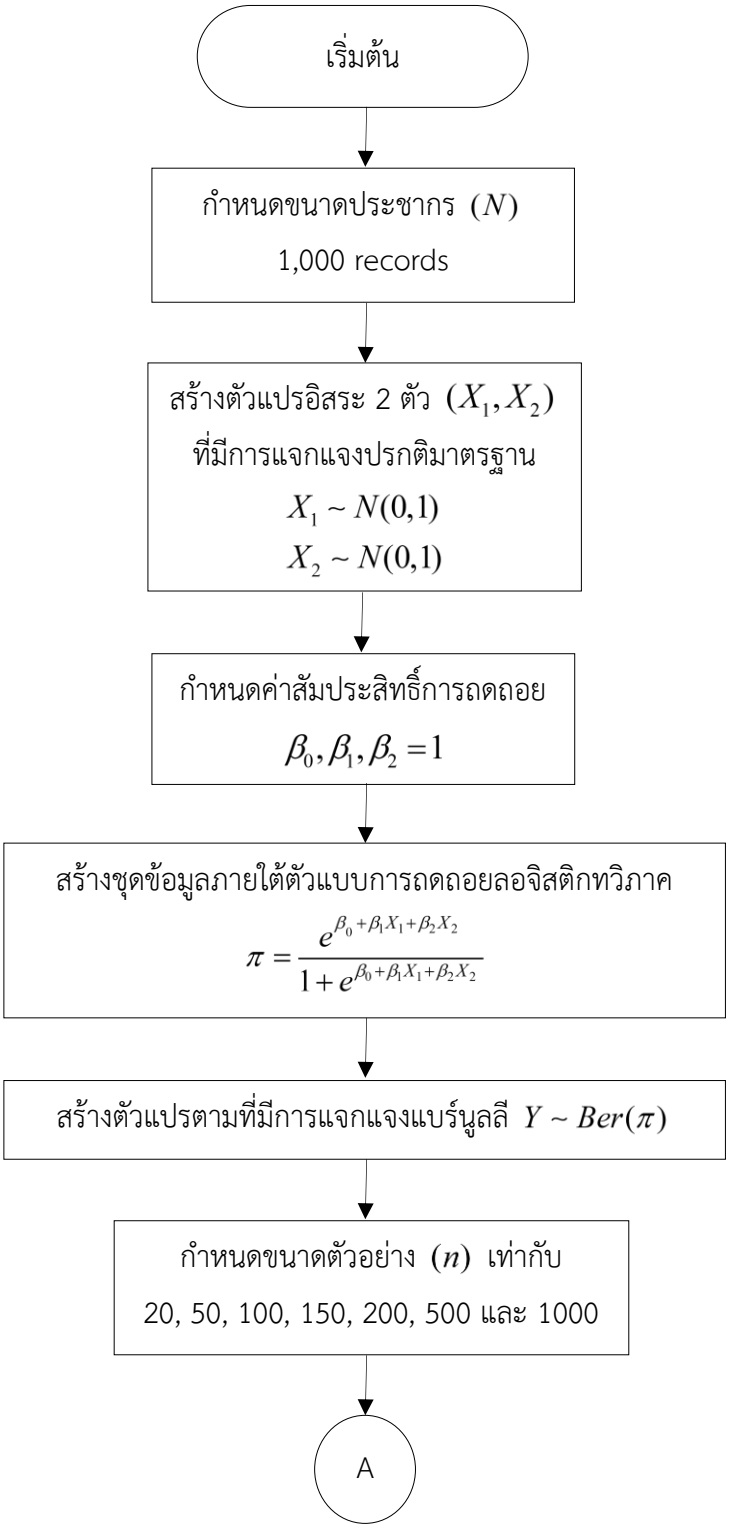
$\hat{\beta}^{*}_{(b,t)}$ คือ สัมประสิทธิ์การถดถอยประมาณของข้อมูลตัวอย่างที่มีการแทนค่าสูญหายแล้วในรอบที่ t

โดยพิจารณาจากค่า EMSE วิธีการประมาณค่าสูญหายวิธีใดมีค่า EMSE ต่ำกว่า แสดงว่าวิธีการประมาณค่าสูญหายนั้นมีประสิทธิภาพที่ดีกว่า (ปูเป้ สุดศิลา และคณะ, 2561)

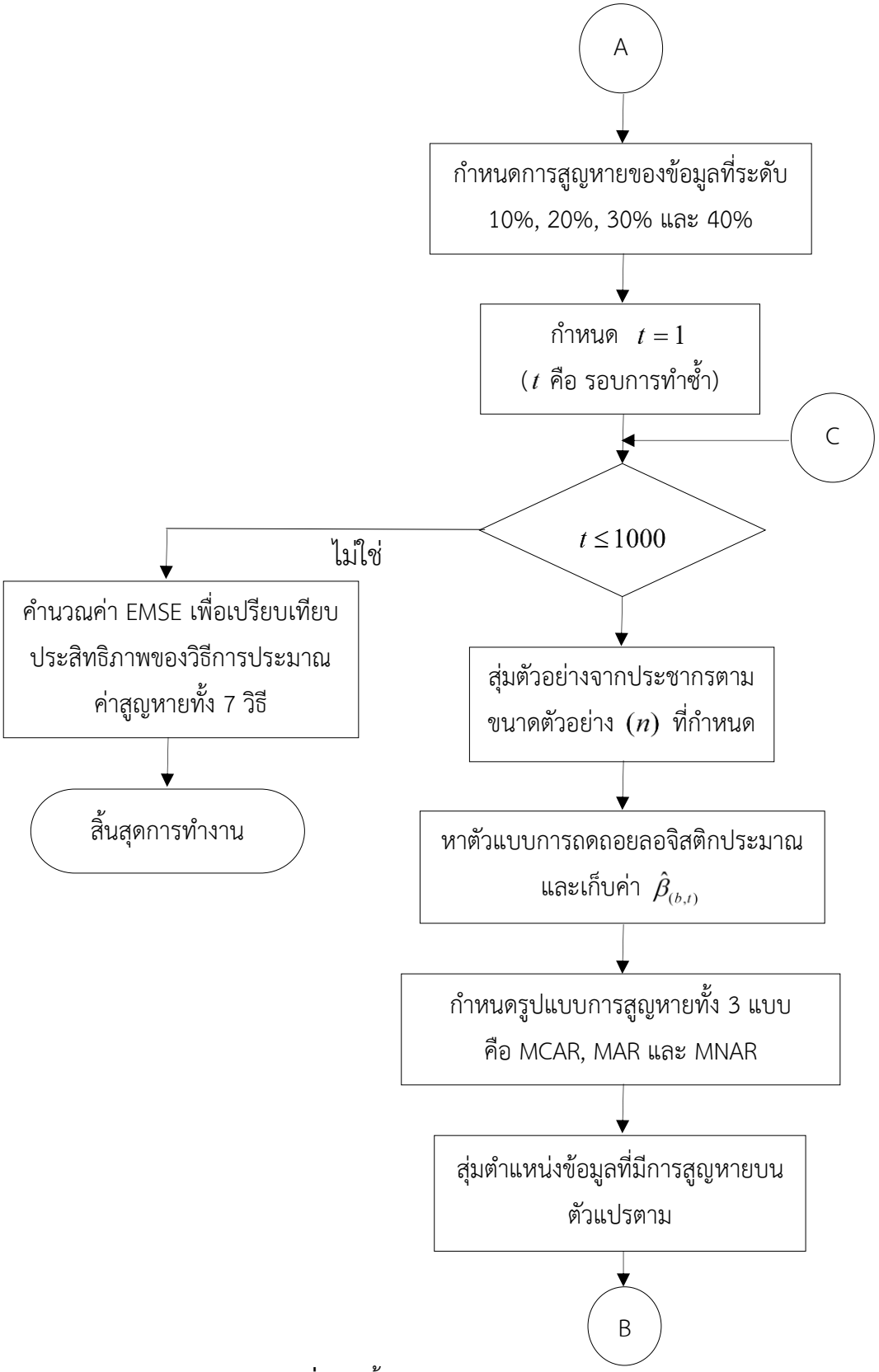
3.2 การศึกษาด้วยข้อมูลจริง

ข้อมูลจริงที่นำมาใช้ในการศึกษานี้คือ ชุดข้อมูลการทำนายโรคหัวใจ จากฐานข้อมูลออนไลน์ www.Kaggle.com (Geb, 2023) ชุดข้อมูลนี้มีตัวอย่างจำนวน 1,025 คน ซึ่งมีจำนวนตัวอย่างที่เป็นโรคหัวใจ 526 คน และจำนวนตัวอย่างที่ไม่เป็นโรคหัวใจ 499 คน สำหรับการศึกษานี้ได้ทำการสุ่มขนาดตัวอย่างมา 50 และ 500 คน เพื่อเป็นตัวแทนของข้อมูลที่มีขนาดเล็กและใหญ่ ผู้วิจัยนำตัวแปรอิสระที่มีอิทธิพลต่อการเป็นโรคหัวใจมาจำนวน 2 ตัว ได้แก่ ความดันโลหิต (x_1) และ ระดับคอเลสเตอรอล (x_2) โดยที่ตัวแปรตามคือ โอกาสการเป็นโรคหัวใจ กำหนดให้ 1 แทน เป็นโรคหัวใจ และ 0 แทน ไม่เป็นโรคหัวใจ กำหนดให้มีรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรตาม โดยที่เปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง จากนั้นประมาณค่าสูญหายจากวิธีการประมาณค่าสูญหายทั้ง 7 วิธี แล้วนำไปประมาณค่าพารามิเตอร์ด้วยวิธี MLE และเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายด้วย EMSE จากการทำซ้ำ 1,000 รอบ โดยวิธีที่ให้ค่า EMSE ต่ำที่สุดคือวิธีที่มีประสิทธิภาพดีที่สุดเช่นเดียวกับชุดข้อมูลจำลอง

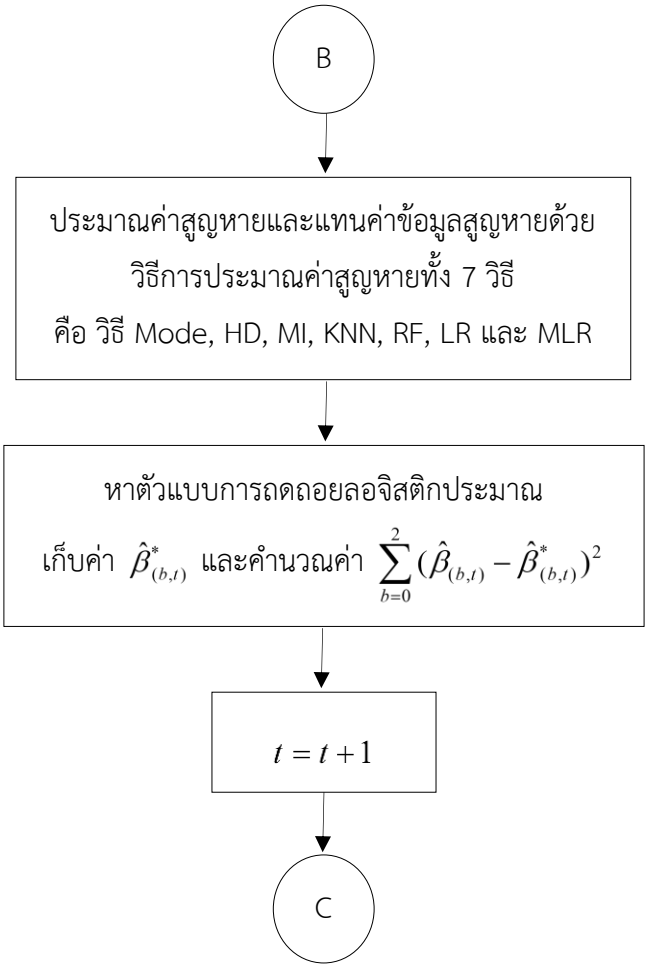
3.3 แผนผังขั้นตอนการดำเนินงาน



รูปที่ 3.1 ขั้นตอนการดำเนินงาน



รูปที่ 3.2 ขั้นตอนการดำเนินงาน (ต่อ)



รูปที่ 3.3 ขั้นตอนการดำเนินงาน (ต่อ)

บทที่ 4
ผลการศึกษา

ในการศึกษาครั้งนี้ เพื่อให้ได้ผลสรุปในการเปรียบเทียบประสิทธิภาพและพัฒนาของวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค จำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR โดยกำหนดขนาดตัวอย่าง 20, 50, 100, 150, 200, 500 และ 1,000 ที่มีเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง ใช้เกณฑ์ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายพิจารณาจากค่า EMSE ผู้วิจัยได้ทำการศึกษาจากชุดข้อมูล 2 ชุด คือ ชุดข้อมูลที่ได้จากการจำลองและชุดข้อมูลจริง ซึ่งมีผลการศึกษาดังนี้

4.1 ผลการศึกษาจากข้อมูลจำลอง

ตารางที่ 4.1 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจำลอง

n	*%mis*	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
20	10	0.5911	0.6101	0.5024	0.457	0.4949	0.4275	**0.3665**
	20	1.0897	1.2101	1.3069	1.1856	1.4505	0.9588	**0.6981**
	30	1.8378	1.901	1.9077	2.341	2.6995	1.8096	**1.7392**
	40	2.4658	2.1801	2.5308	2.7798	4.2039	2.5269	**2.1017**
50	10	0.2041	0.2986	0.2493	0.2317	0.2361	0.2213	**0.2021**
	20	0.4278	0.607	0.4574	0.4545	0.4804	0.4716	**0.3808**
	30	0.7703	0.946	**0.7068**	0.8208	0.88	0.86	0.7595
	40	1.2003	1.2932	**1.1851**	1.4074	1.8442	1.8719	1.7713
100	10	0.0907	0.1154	**0.0534**	0.0594	0.0547	0.0593	0.0568
	20	0.2483	0.2543	**0.1255**	0.1842	0.1668	0.2022	0.1891
	30	0.4832	0.4494	**0.2225**	0.379	0.3364	0.4659	0.4423
	40	0.8189	0.6261	**0.3415**	0.7002	0.5931	0.9147	0.8598
150	10	0.0708	0.0872	**0.0376**	0.0458	0.0412	0.049	0.0505
	20	0.2137	0.2244	**0.0768**	0.1429	0.1097	0.1618	0.1542
	30	0.4218	0.3904	**0.129**	0.2796	0.2172	0.3725	0.3636
	40	0.7516	0.5799	**0.2141**	0.5904	0.4305	0.7976	0.7421

ตารางที่ 4.1 (ต่อ)

n	%*mis*	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
200	10	0.0632	0.076	**0.0253**	0.0374	0.0311	0.0412	0.0407
	20	0.1969	0.1952	**0.0579**	0.1304	0.0892	0.1524	0.1471
	30	0.4004	0.3544	**0.101**	0.2718	0.1884	0.3483	0.3373
	40	0.715	0.5325	**0.1659**	0.5556	0.3631	0.756	0.7202
500	10	0.0479	0.0532	**0.0098**	0.0255	0.0151	0.0295	0.0289
	20	0.1701	0.1661	**0.0223**	0.101	0.0532	0.1249	0.1193
	30	0.3732	0.3031	**0.0403**	0.2534	0.1203	0.3233	0.3061
	40	0.6673	0.4653	**0.0627**	0.4937	0.2142	0.6731	0.6278
1,000	10	0.0437	0.0468	**0.005**	0.0237	0.0105	0.0265	0.0255
	20	0.1613	0.1541	**0.0108**	0.0996	0.0389	0.116	0.1104
	30	0.3548	0.2896	**0.0195**	0.2421	0.0863	0.2962	0.2799
	40	0.6478	0.4565	**0.0311**	0.5014	0.1706	0.6485	0.6059

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์

จากตารางที่ 4.1 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR ในขนาดตัวอย่างเท่ากับ 20 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายระดับ 10% และ 20% การประมาณค่าสูญหายด้วยวิธี MLR ให้ค่า EMSE ต่ำที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายระดับ 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 100, 150, 200, 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า EMSE ต่ำที่สุด

ตารางที่ 4.2 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจำลอง

n	*%mis*	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
20	10	0.4301	0.7124	0.3918	0.3585	0.3737	0.2969	**0.256**
	20	1.0542	1.4064	1.2863	1.3938	1.4202	1.0551	**0.9712**
	30	2.1896	2.506	1.5899	1.7556	1.918	1.4243	**1.1774**
	40	3.4126	2.8294	2.4708	3.0721	3.699	1.863	**1.6621**
50	10	0.1336	0.4619	0.1247	0.0989	0.1076	0.095	**0.0831**
	20	0.4248	0.9198	0.2876	0.2926	0.2845	0.2934	**0.246**
	30	0.9642	1.2982	0.6243	0.6465	0.6413	0.6841	**0.5283**
	40	1.9623	1.6311	1.0512	1.1536	1.1559	1.3444	**0.971**
100	10	0.0528	0.1951	0.0538	0.0513	0.0512	0.0533	**0.0481**
	20	0.1812	0.4574	**0.1366**	0.1684	0.1479	0.1871	0.154
	30	0.4921	0.7131	**0.2352**	0.3376	0.3075	0.4314	0.3332
	40	0.9779	1.0158	**0.3791**	0.6607	0.5612	0.8869	0.6544
150	10	0.0363	0.2425	0.0346	0.0367	0.033	0.038	**0.0329**
	20	0.1365	0.3938	**0.0838**	0.1336	0.1001	0.1483	0.1195
	30	0.3239	0.666	**0.1564**	0.293	0.2232	0.3889	0.2954
	40	0.7651	0.9347	**0.2519**	0.6051	0.4419	0.7818	0.5855
200	10	0.0344	0.1492	**0.0282**	0.0345	0.0288	0.0369	0.0319
	20	0.1162	0.3933	**0.0629**	0.1173	0.0815	0.1324	0.1059
	30	0.2951	0.6621	**0.1127**	0.2656	0.1826	0.342	0.2586
	40	0.6816	0.8894	**0.1948**	0.6001	0.371	0.772	0.5608
500	10	0.0239	0.1237	**0.0098**	0.0248	0.0148	0.0265	0.0319
	20	0.0995	0.3331	**0.0242**	0.1011	0.0516	0.1159	0.0925
	30	0.256	0.594	**0.0437**	0.2758	0.1246	0.316	0.2434
	40	0.5261	0.8513	**0.072**	0.5594	0.2425	0.6861	0.4989
1,000	10	0.0209	0.1168	**0.0047**	0.0225	0.0104	0.0235	0.0193
	20	0.0939	0.3254	**0.0114**	0.1036	0.0394	0.1113	0.0868
	30	0.2418	0.5695	**0.0204**	0.268	0.0956	0.3008	0.2237
	40	0.5093	0.8245	**0.035**	0.5631	0.2029	0.6716	0.4729

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์

จากตารางที่ 4.2 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MAR ในขนาดตัวอย่างเท่ากับ 20 และ 50 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 100 และ 150 ที่เปอร์เซ็นต์การสูญหายระดับ 10% การประมาณค่าสูญหายด้วยวิธี MLR ให้ค่า EMSE ต่ำที่สุด แต่ขนาดตัวอย่างเท่ากับ 100 และ 150 ที่เปอร์เซ็นต์การสูญหายระดับ 20%, 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 200, 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า EMSE ต่ำที่สุด

ตารางที่ 4.3 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจำลอง

n	$\%mis$	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
20	10	0.3994	0.7139	0.4798	0.5166	0.4328	**0.2206**	0.3362
	20	1.0047	1.4151	1.1079	1.31	1.2588	**0.6539**	0.8567
	30	2.0273	2.396	2.2275	2.513	2.8411	**1.4421**	1.7712
	40	3.9258	3.2697	3.2232	3.5661	4.2104	**1.9326**	2.9982
50	10	0.0841	0.3098	0.126	0.0908	0.1161	**0.0801**	0.1068
	20	0.3767	0.7109	0.3467	0.3754	0.348	**0.2049**	0.3192
	30	1.1525	1.3832	0.7652	0.9411	0.7569	**0.5301**	1.0625
	40	2.8354	2.1069	1.4433	1.7065	1.6658	**1.0871**	1.6995
100	10	**0.0206**	0.1971	0.0648	0.0498	0.0632	0.0446	0.0682
	20	**0.1306**	0.5444	0.184	0.1616	0.1873	0.139	0.221
	30	0.7673	1.09	0.4099	0.5434	0.451	**0.3007**	0.5377
	40	2.7483	1.7785	0.845	1.3566	0.9989	**0.6277**	1.19
150	10	**0.0125**	0.1635	0.0459	0.0416	0.0435	0.0322	0.0516
	20	**0.0802**	0.5321	0.1602	0.1531	0.1644	0.1185	0.2124
	30	0.6978	1.039	0.3707	0.5019	0.3963	**0.2654**	0.5072
	40	2.5076	1.7683	0.6728	0.917	0.7969	**0.5266**	1.0475

ตารางที่ 4.3 (ต่อ)

n	*%mis*	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
200	10	**0.008**	0.155	0.0353	0.0288	0.0359	0.0253	0.0455
	20	**0.0493**	0.499	0.134	0.125	0.1386	0.0961	0.1896
	30	0.5171	1.0149	0.313	0.4474	0.3618	**0.2557**	0.5057
	40	2.6185	1.7021	0.6168	0.8798	0.7281	**0.4933**	1.0227
500	10	**0.0033**	0.0854	0.0243	0.0136	0.0198	0.014	0.0273
	20	**0.0124**	0.339	0.0997	0.0848	0.0887	0.0585	0.1193
	30	**0.0776**	0.7386	0.2563	0.2367	0.2395	0.1449	0.3082
	40	2.1623	1.2611	0.5197	0.6274	0.5036	**0.2955**	0.6412
1000	10	**0.0025**	0.0819	0.02	0.0126	0.0164	0.0114	0.0247
	20	**0.0105**	0.3216	0.0934	0.0756	0.0784	0.0533	0.1175
	30	**0.0305**	0.7123	0.2422	0.2536	0.22	0.1364	0.307
	40	2.0713	1.2355	0.4879	0.6703	0.4579	**0.2753**	0.6318

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์

จากตารางที่ 4.3 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR ผลที่ได้แตกต่างจากแบบ MCAR และ MAR ในขนาดตัวอย่างเท่ากับ 20 และ 50 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี LR ให้ค่า EMSE ต่ำที่สุด ส่วนขนาดตัวอย่างเท่ากับ 100, 150 และ 200 ที่เปอร์เซ็นต์การสูญหายระดับ 10% และ 20% รวมถึงขนาดตัวอย่างเท่ากับ 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% การประมาณค่าสูญหายด้วยวิธี Mode ให้ค่า EMSE ต่ำที่สุด แต่ที่ขนาดตัวอย่างเท่ากับ 100, 150 และ 200 ที่เปอร์เซ็นต์การสูญหายระดับ 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายระดับ 40% การประมาณค่าสูญหายด้วยวิธี LR ให้ค่า EMSE ต่ำที่สุด

4.2 ผลการศึกษาจากข้อมูลจริง

การนำเสนอผลการศึกษาจะแสดงในรูปแบบตารางและกราฟ เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายในแต่ละสถานการณ์ โดยมีขนาดตัวอย่างเท่ากับ 50 และ 500

ผลการศึกษาจากตารางที่ 4.4 และรูปที่ 4.1 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR ในขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MLR ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี LR วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี Mode และ KNN ซึ่งให้ค่า EMSE ที่ใกล้เคียงกัน สำหรับขนาดตัวอย่างเท่ากับ 500 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี RF วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี LR นอกจากนี้ยังพบว่า ในวิธีการประมาณค่าสูญหายทั้ง 7 วิธี เมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่า EMSE ลดลง และเมื่อเปอร์เซ็นต์การสูญหายเพิ่ม จะทำให้ค่า EMSE เพิ่มขึ้น

ผลการศึกษาจากตารางที่ 4.5 และรูปที่ 4.2 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MAR ในขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MLR ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี LR วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี KNN สำหรับขนาดตัวอย่างเท่ากับ 500 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี HD วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี LR นอกจากนี้ยังพบว่า ในวิธีการประมาณค่าสูญหายทั้ง 7 วิธี เมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่า EMSE ลดลง และเมื่อเปอร์เซ็นต์การสูญหายเพิ่ม จะทำให้ค่า EMSE เพิ่มขึ้น

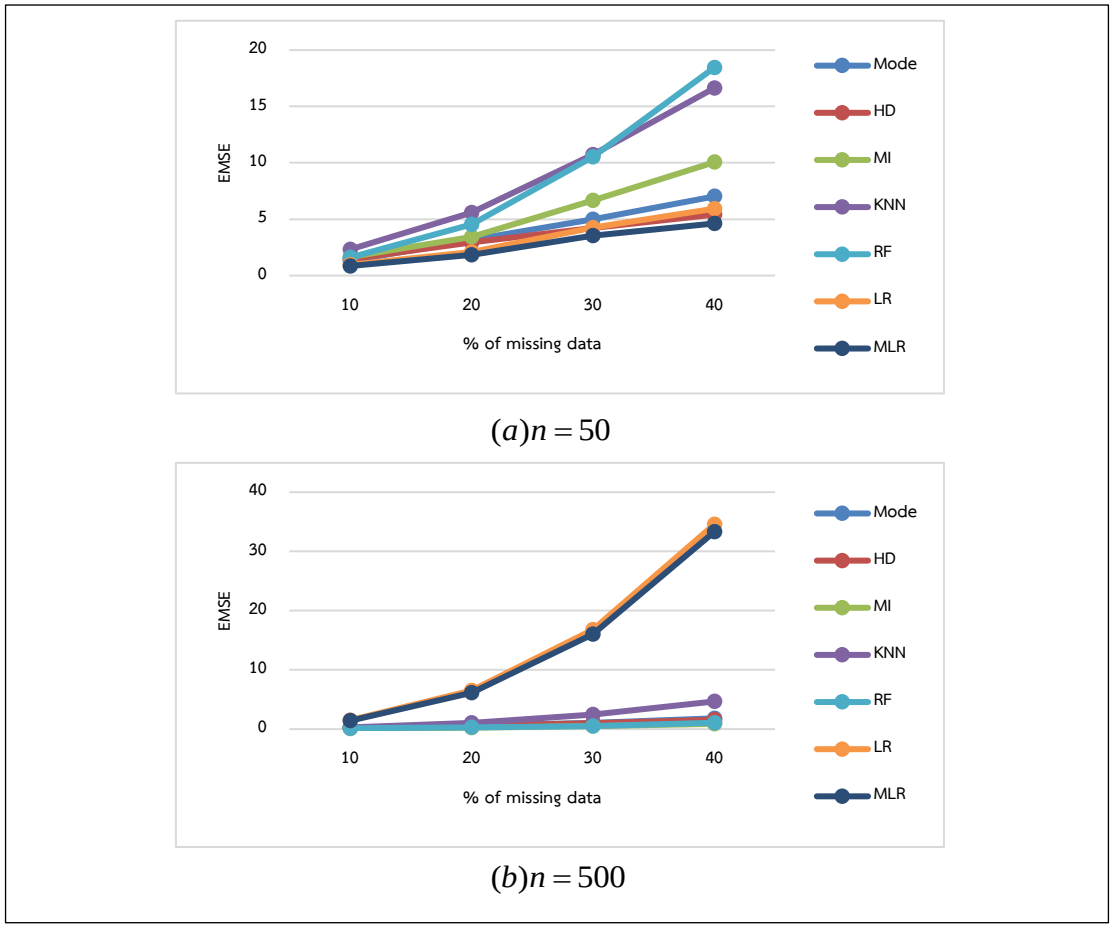
ผลการศึกษาจากตารางที่ 4.6 และรูปที่ 4.3 พบว่า กรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR ในขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี MLR ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี LR วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี KNN สำหรับขนาดตัวอย่างเท่ากับ 500 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 10%, 20%, 30% และ 40% การประมาณค่าสูญหายด้วยวิธี Mode ให้ค่า EMSE ต่ำที่สุด รองลงมาคือวิธี HD วิธีที่ให้ค่า EMSE สูงที่สุดคือวิธี MLR นอกจากนี้ยังพบว่า ในวิธีการประมาณค่าสูญหายทั้ง 7 วิธี เมื่อขนาดตัวอย่างเพิ่มขึ้นจะทำให้ค่า EMSE ลดลง และเมื่อเปอร์เซ็นต์การสูญหายเพิ่ม จะทำให้ค่า EMSE เพิ่มขึ้น

ผลการศึกษาการเปรียบเทียบค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจริงแสดงในรูปแบบตารางและกราฟ ดังนี้

ตารางที่ 4.4 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจริง

n	$\%mis$	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
50	10	1.2985	1.4545	1.6443	2.3165	1.6019	0.9565	**0.8521**
	20	3.1659	2.9561	3.4293	5.5947	4.5412	2.1009	**1.8231**
	30	4.9746	4.2238	6.6518	10.735	10.5311	4.2487	**3.5256**
	40	7.0184	5.4066	10.0569	16.6316	18.4505	5.9324	**4.6212**
500	10	0.2402	0.2092	**0.086**	0.2917	0.1475	1.501	1.4183
	20	0.6162	0.5436	**0.2529**	1.0377	0.3059	6.488	6.1542
	30	1.0724	0.9855	**0.4875**	2.4703	0.492	16.7636	16.0368
	40	1.8198	1.5662	**0.8787**	4.6676	1.0745	34.563	33.3094

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์

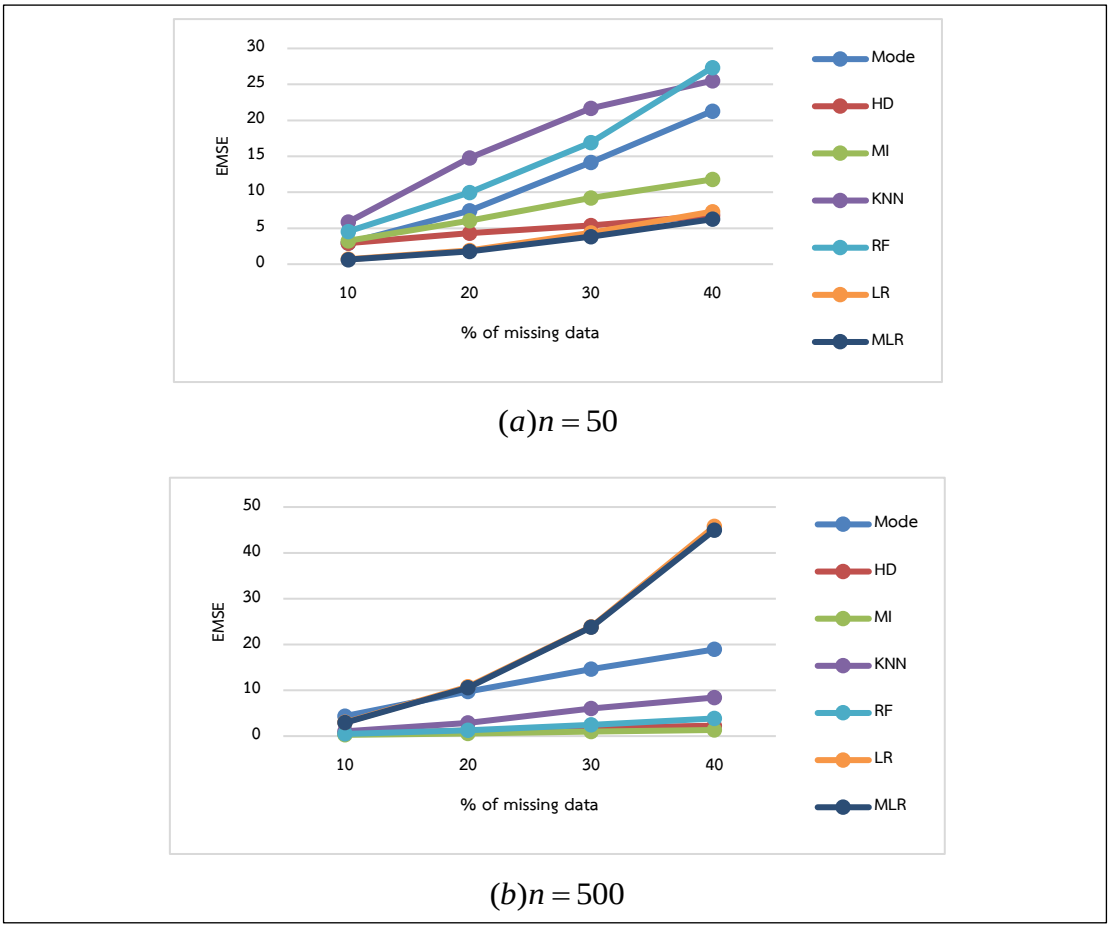


รูปที่ 4.1 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR สำหรับข้อมูลจริง

ตารางที่ 4.5 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจริง

n	$\%mis$	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
50	10	2.9291	2.9227	3.2495	5.8502	4.55	0.6868	**0.6363**
	20	7.4511	4.3272	6.0424	14.7784	9.9751	1.929	**1.7608**
	30	14.1573	5.4033	9.2139	21.6973	16.9056	4.3768	**3.8069**
	40	21.2927	6.7895	11.7816	25.5103	27.3465	7.3129	**6.2516**
500	10	4.3728	0.5855	**0.275**	1.0421	0.4702	2.9816	2.9114
	20	9.7376	1.2011	**0.5511**	2.9072	1.2992	10.7696	10.5844
	30	14.5897	1.8053	**0.985**	6.0274	2.4289	23.8251	23.7744
	40	18.9135	2.2727	**1.3366**	8.4251	3.8492	45.7501	44.9565

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์

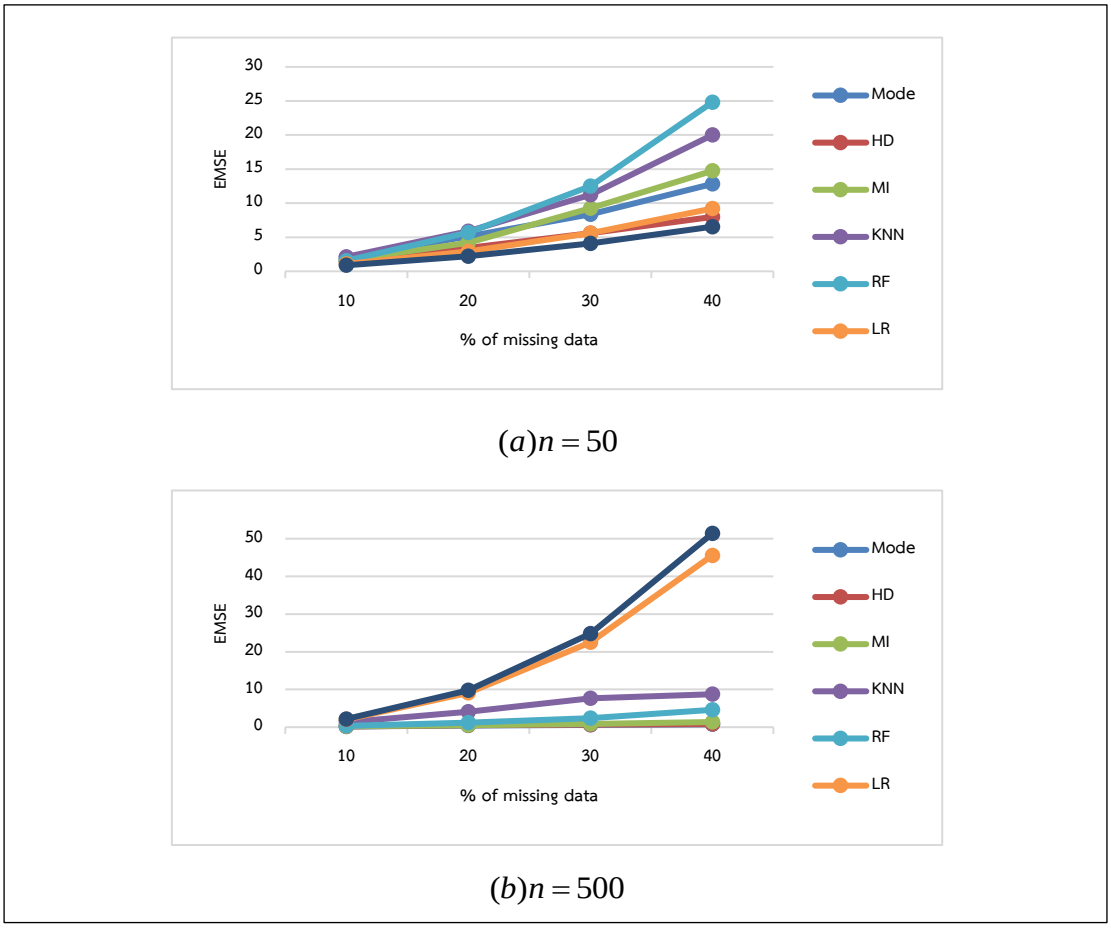


รูปที่ 4.2 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MAR สำหรับข้อมูลจริง

ตารางที่ 4.6 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจริง

n	$\%mis$	Imputation methods						
		Mode	HD	MI	KNN	RF	LR	MLR
50	10	2.003	1.5238	1.6362	2.1273	1.5676	1.073	**0.8829**
	20	5.1217	3.426	4.1891	5.8669	5.6942	2.9301	**2.1978**
	30	8.3482	5.6179	9.2561	11.2109	12.4961	5.5865	**4.0799**
	40	12.8224	7.9778	14.7367	19.9994	24.8089	9.2117	**6.5458**
500	10	**0.172**	0.1844	0.2519	1.3387	0.3395	2.0271	2.1636
	20	**0.4177**	0.4796	0.5116	4.0507	1.1619	9.0693	9.7956
	30	**0.5698**	0.6684	0.8527	7.6202	2.4013	22.5309	24.8214
	40	**0.7405**	0.8216	1.3661	8.7293	4.597	45.55	51.3829

หมายเหตุ : ตัวหนา หมายถึง ค่า EMSE ต่ำที่สุดในแต่ละสถานการณ์



รูปที่ 4.3 ค่า EMSE ของวิธีการประมาณสูญหาย 7 วิธี กรณีที่เกิดการสูญหายแบบ MNAR สำหรับข้อมูลจริง

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบการถดถอยลอจิสติกทวิภาค จำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR ชุดข้อมูลที่ใช้ในการศึกษามี 2 ชุด คือ ชุดข้อมูลที่ได้จากการจำลองและชุดข้อมูลจริงจากฐานข้อมูลออนไลน์ www.Kaggle.com ซึ่งเป็นข้อมูลการทำนายโรคหัวใจ โดยกำหนดขนาดตัวอย่างของข้อมูลจำลองเท่ากับ 20, 50, 100, 150, 200, 500 และ 1,000 ส่วนขนาดตัวอย่างของข้อมูลจริงเท่ากับ 50 และ 500 กำหนดเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง เกณฑ์ที่ใช้ในการเปรียบเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายคือ ค่า EMSE ซึ่งหัวข้อที่จะกล่าวในบทนี้ มีดังนี้

5.1 สรุปผลการศึกษา

5.2 อภิปลายผลการศึกษา

5.3 ข้อเสนอแนะ

5.1 สรุปผลการศึกษา

5.1.1 สรุปผลการศึกษาข้อมูลจำลอง

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 20 ในทุกระดับเปอร์เซ็นต์การสูญหาย รวมถึงขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายระดับ 10% และ 20% วิธี MLR จะให้ประสิทธิภาพดีที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 50 ที่เปอร์เซ็นต์การสูญหายระดับ 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 100, 150, 200, 500 และ 1,000 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MI จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n \leq 50)$ วิธี MLR มีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่ $(n \geq 100)$ วิธี MI มีประสิทธิภาพดีที่สุด

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 20 และ 50 ในทุกระดับเปอร์เซ็นต์การสูญหาย รวมถึงขนาดตัวอย่างเท่ากับ 100 และ 150 ที่เปอร์เซ็นต์การสูญหายระดับ 10% วิธี ระดับ 20%, 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 200, 500 และ 1,000 ในทุกระดับ

เปอร์เซ็นต์การสูญหาย วิธี MI จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n \leq 50)$ วิธี MLR มีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่ $(n \geq 100)$ วิธี MI มีประสิทธิภาพดีที่สุด ซึ่งจะเห็นได้ว่ากรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR และ MAR ให้ผลสรุปที่คล้ายกัน

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 20 และ 50 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี LR จะให้ประสิทธิภาพดีที่สุด ส่วนขนาดตัวอย่างเท่ากับ 100, 150 และ 200 ที่เปอร์เซ็นต์การสูญหายระดับ 10% และ 20% รวมถึงขนาดตัวอย่างเท่ากับ 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% วิธี Mode จะให้ประสิทธิภาพดีที่สุด แต่ที่ขนาดตัวอย่างเท่ากับ 100, 150 และ 200 ที่เปอร์เซ็นต์การสูญหายระดับ 30% และ 40% รวมถึงขนาดตัวอย่างเท่ากับ 500 และ 1,000 ที่เปอร์เซ็นต์การสูญหายระดับ 40% วิธี LR จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n \leq 50)$ วิธี LR มีประสิทธิภาพดีที่สุด เมื่อข้อมูลมีขนาดใหญ่ $(n \geq 100)$ ที่เปอร์เซ็นต์การสูญหายระดับต่ำ วิธี Mode มีประสิทธิภาพดีที่สุด แต่เมื่อเมื่อข้อมูลมีขนาดใหญ่ $(n \geq 100)$ ที่เปอร์เซ็นต์การสูญหายระดับสูง วิธี LR มีประสิทธิภาพดีที่สุด ซึ่งจะเห็นได้ว่ากรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR ให้ผลสรุปที่แตกต่างกับ MCAR และ MAR

นอกจากนี้จากผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรตาม จะสังเกตเห็นได้ว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพของประมาณค่าสูญหายทั้ง 7 วิธี ได้แก่ ขนาดตัวอย่าง และเปอร์เซ็นต์การสูญหายของข้อมูล เมื่อขนาดตัวอย่างเพิ่มขึ้นจะส่งผลให้ค่า EMSE มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้ และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้นจะส่งผลให้ค่า EMSE มีแนวโน้มเพิ่มขึ้นด้วย เนื่องจากข้อมูลที่มีเปอร์เซ็นต์การสูญหายเพิ่มขึ้นจะทำให้การประมาณค่าสูญหายและการประมาณค่าพารามิเตอร์เกิดความคลาดเคลื่อน จึงสรุปได้ว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นจะส่งผลให้ประสิทธิภาพของวิธีการประมาณค่าสูญหายดีขึ้น และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มจะส่งผลให้ประสิทธิภาพของวิธีการประมาณค่าสูญหายลดลง

ผลการศึกษาการเปรียบเทียบค่า EMSE ของวิธีการประมาณสูญหายทั้ง 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจำลอง ในแต่ละสถานการณ์สามารถสรุปได้ดังตาราง 5.1 ดังนี้

ตารางที่ 5.1 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ตัวแปรตามเกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจำลอง

n	*%mis*	**รูปแบบการสูญหาย**		
		MCAR	MAR	MNAR
20	10	MLR	MLR	LR
	20	MLR	MLR	LR
	30	MLR	MLR	LR
	40	MLR	MLR	LR
50	10	MLR	MLR	LR
	20	MLR	MLR	LR
	30	MI	MLR	LR
	40	MI	MLR	LR
100	10	MI	MLR	Mode
	20	MI	MI	Mode
	30	MI	MI	LR
	40	MI	MI	LR
150	10	MI	MLR	Mode
	20	MI	MI	Mode
	30	MI	MI	LR
	40	MI	MI	LR
200	10	MI	MI	Mode
	20	MI	MI	Mode
	30	MI	MI	LR
	40	MI	MI	LR
500	10	MI	MI	Mode
	20	MI	MI	Mode
	30	MI	MI	Mode
	40	MI	MI	LR
1,000	10	MI	MI	Mode
	20	MI	MI	Mode
	30	MI	MI	Mode
	40	MI	MI	LR

5.1.2 สรุปผลการศึกษาข้อมูลจริง

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 50 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MLR จะให้ประสิทธิภาพดีที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 500 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MI จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n=50)$ วิธี MLR มีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่ $(n=500)$ วิธี MI มีประสิทธิภาพดีที่สุด ซึ่งจะเห็นได้ว่าผลการศึกษาจากข้อมูลจริงมีความสอดคล้องกับผลการศึกษาที่ได้จากข้อมูลจำลอง

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 50 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MLR จะให้ประสิทธิภาพดีที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 500 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MI จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n=50)$ วิธี MLR มีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่ $(n=500)$ วิธี MI มีประสิทธิภาพดีที่สุด ซึ่งจะเห็นได้ว่ากรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR และ MAR ให้ผลสรุปที่คล้ายกัน และพบว่าผลการศึกษาจากข้อมูลจริงมีความสอดคล้องกับผลการศึกษาที่ได้จากข้อมูลจำลอง

ผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR บนตัวแปรตาม พบว่า ที่ขนาดตัวอย่างเท่ากับ 50 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี MLR จะให้ประสิทธิภาพดีที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 500 ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี Mode จะให้ประสิทธิภาพดีที่สุด จึงสรุปได้ว่า ในข้อมูลที่มีขนาดเล็ก $(n=50)$ วิธี MLR มีประสิทธิภาพดีที่สุด แต่เมื่อข้อมูลมีขนาดใหญ่ $(n=500)$ วิธี Mode มีประสิทธิภาพดีที่สุด ซึ่งจะเห็นได้ว่ากรณีที่ข้อมูลเกิดการสูญหายแบบ MNAR ให้ผลสรุปที่แตกต่างกับ MCAR และ MAR ในส่วนของข้อมูลที่มีขนาดใหญ่ $(n=500)$ และพบว่าผลการศึกษาจากข้อมูลจริงไม่สอดคล้องกับผลการศึกษาที่ได้จากข้อมูลจำลอง

นอกจากนี้จากผลการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ข้อมูลเกิดการสูญหายแบบ MCAR, MAR และ MNAR บนตัวแปรตาม จะสังเกตเห็นได้ว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นจะส่งผลให้ค่า EMSE มีแนวโน้มลดลง และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้นจะส่งผลให้ค่า EMSE มีแนวโน้มเพิ่มขึ้นด้วย ซึ่งผลการศึกษาจากข้อมูลจริงมีความสอดคล้องกับผลการศึกษาที่ได้จากข้อมูลจำลอง

ผลการศึกษาการเปรียบเทียบค่า EMSE ของวิธีการประมาณสูญหายทั้ง 7 วิธี กรณีที่เกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจริง ในแต่ละสถานการณ์สามารถสรุปได้ดังตาราง 5.2 ดังนี้

ตารางที่ 5.2 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 7 วิธี กรณีที่ตัวแปรตามเกิดการสูญหายแบบ MCAR, MAR และ MNAR สำหรับข้อมูลจริง

n	$\%mis$	**รูปแบบการสูญหาย**		
		MCAR	MAR	MNAR
50	10	MLR	MLR	MLR
	20	MLR	MLR	MLR
	30	MLR	MLR	MLR
	40	MLR	MLR	MLR
500	10	MI	MI	Mode
	20	MI	MI	Mode
	30	MI	MI	Mode
	40	MI	MI	Mode

จุดประสงค์หลักของการศึกษานี้คือการพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในการถดถอยลอจิสติกทวิภาค จำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR ในการศึกษานี้ได้มีการพัฒนาวิธีการประมาณค่าสูญหายมา 1 วิธี คือ วิธี MLR จากผลการศึกษาสรุปได้ว่า เมื่อข้อมูลมีขนาดเล็ก วิธี MLR ที่ได้พัฒนาขึ้นมาจะมีประสิทธิภาพดีที่สุด และเมื่อขนาดตัวอย่างใหญ่ วิธี MI จะมีประสิทธิภาพดีที่สุด นอกจากนี้เมื่อขนาดตัวอย่างเพิ่มขึ้นจะส่งผลให้ประสิทธิภาพของวิธีการประมาณค่าสูญหายดีขึ้น และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มจะส่งผลให้ประสิทธิภาพของวิธีการประมาณค่าสูญหายลดลง ซึ่งผลการศึกษาจากข้อมูลจำลองและข้อมูลจริงมีความสอดคล้องกัน ในข้อมูลที่เกิดการสูญหายแบบ MCAR และ MAR

จากผลการศึกษาจะเห็นได้ว่า วิธี MLR ซึ่งเป็นวิธีประมาณค่าสูญหายที่ผู้วิจัยได้เสนอขึ้นมาใหม่ โดยพัฒนามาจากวิธี LR เมื่อนำทั้ง 2 วิธี มาเปรียบเทียบกัน พบว่า ในสถานการณ์ส่วนใหญ่วิธี MLR มีประสิทธิภาพดีกว่าวิธี LR และเมื่อเปรียบวิธีประมาณค่าสูญหายทั้ง 7 วิธี ในกรณีข้อมูลมีขนาดเล็ก นอกเหนือจากที่วิธี MLR มีประสิทธิภาพดีกว่าวิธี LR แล้ว ก็ยังมีประสิทธิภาพดีกว่าวิธีอื่น ๆ ด้วย

5.2 อภิปรายผลการศึกษา

การศึกษาครั้งนี้เป็นการศึกษาเพื่อพัฒนาและเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตามในการถดถอยลอจิสติกทวิภาค เมื่อตัวแปรตามมีการสูญหายแบบ MCAR, MAR และ MNAR วิธีการประมาณค่าสูญหายที่สนใจศึกษามีจำนวน 7 วิธี ได้แก่ วิธี Mode, วิธี HD, วิธี MI, วิธี KNN, วิธี RF, วิธี LR และ วิธี MLR ซึ่งวิธี MLR เป็นวิธีที่ผู้วิจัยได้พัฒนาขึ้นมา ข้อมูลที่ใช้ในการศึกษามี 2 ชุด คือ ชุดข้อมูลที่ได้จากการจำลองและชุดข้อมูลจริงเป็นข้อมูลการทำนายโรคหัวใจ สำหรับชุดข้อมูลจำลองกำหนดขนาดตัวอย่างเท่ากับ 20, 50, 100, 150, 200, 500, และ 1,000 ส่วนข้อมูลจริงกำหนดขนาดตัวอย่างเท่ากับ 50 และ 500 ซึ่งเป็นตัวแทนของข้อมูลขนาดเล็กและข้อมูลขนาดใหญ่ตามลำดับ โดยกำหนดเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% ของขนาดตัวอย่าง เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพจะพิจารณาจากค่า EMSE โดยวิธีที่ให้ค่า EMSE ต่ำที่สุด คือ วิธีที่มีประสิทธิภาพดีที่สุด

จากผลการศึกษาจะเห็นได้ว่า การประมาณค่าสูญหายด้วยวิธี MI มีประสิทธิภาพดีที่สุด ในกรณีที่ข้อมูลมีขนาดใหญ่ ซึ่งเป็นไปตามผลการศึกษาของ Xu et al. (2020) และ Tsiampalis & Panagiotakos (2020) ที่ได้ศึกษาการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรเชิงคุณภาพในตัวแบบการถดถอยลอจิสติก เมื่อมีการสูญหายแบบ MCAR, MAR และ MNAR พบว่า วิธี MI มีประสิทธิภาพดีที่สุดเช่นเดียวกัน เนื่องจากวิธี MI มีการแทนค่าหลายครั้ง แล้วนำค่าที่ได้จากการประมาณค่ามาสรุปเพื่อให้ได้ค่าที่ดีที่สุด จะทำให้ได้ข้อมูลที่มีความถูกต้องและแม่นยำสูง ดังนั้นวิธี MI จึงมีประสิทธิภาพที่ดีกว่าวิธีอื่น ๆ ในกรณีที่ข้อมูลมีขนาดใหญ่

ในกรณีที่ข้อมูลมีขนาดเล็ก การประมาณค่าสูญหายด้วยวิธี MLR มีประสิทธิภาพดีทีสุด เนื่องจากข้อมูลที่สูญหายเป็นตัวแปรเชิงคุณภาพ 2 กลุ่ม ซึ่งเหมาะแก่การนำมาวิเคราะห์การถดถอยลอจิสติก โดยผู้วิจัยได้นำเทคนิคการวิเคราะห์การถดถอยลอจิสติกทวิภาค มาใช้ในการประมาณค่าสูญหายด้วยวิธี MLR แต่เนื่องจากเมื่อข้อมูลขนาดใหญ่มีความซับซ้อนมากกว่าข้อมูลขนาดเล็ก วิธี MI ซึ่งเป็นวิธีที่มีความยุ่งยากซับซ้อน จึงมีความเหมาะสมในกรณีที่ข้อมูลมีขนาดใหญ่ และจะเห็นได้ว่าในสถาณการณ์ส่วนใหญ่วิธี MLR มีประสิทธิภาพดีกว่าวิธี LR เนื่องจาก วิธี MLR มีการพิจารณาจุดตัดที่เหมาะสมกับข้อมูลชุดนั้น ๆ ในขณะที่วิธี LR ใช้จุดตัดเท่ากับ 0.5 ที่ใช้โดยทั่วไป นั่นคือวิธี MLR จะให้ประสิทธิภาพที่ดีกว่าวิธี LR และวิธีอื่น ๆ ในกรณีที่ข้อมูลมีขนาดเล็ก

นอกจากนี้ยังพบว่า ผลการศึกษาจากข้อมูลจำลองและข้อมูลจริงมีความสอดคล้องกัน ในรูปแบบการสูญหายแบบ MAR และ MCAR ส่วน MNR จะให้ผลที่ต่างออกไป เนื่องจากการสูญหายแบบ MNAR ทำให้การประมาณค่าพารามิเตอร์มีความคลาดเคลื่อนมากกว่า MCAR และ MAR ซึ่งเป็นผลกระทบที่รุนแรงในการวิเคราะห์ (Little & Rubin, 2002) และเมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า EMSE มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นช่วยให้ค่าประมาณที่ได้จะใกล้เคียงค่า

จริงมากขึ้น (สุชาดา กีระนันทน์, 2545) นั่นคือช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้ และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น ส่งผลให้ค่า EMSE มีแนวโน้มเพิ่มขึ้นด้วย เนื่องจากเมื่อข้อมูลมีเปอร์เซ็นต์การสูญหายเพิ่มขึ้นทำให้ขนาดตัวอย่างลดลง จะส่งผลให้เกิดความคลาดเคลื่อนในการประมาณค่าพารามิเตอร์ (สุกัญญา ศิริโยธา, 2558)

5.3 ข้อเสนอแนะ

แนวทางในการเลือกวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติก เมื่อข้อมูลสูญหายเกิดขึ้นแบบ MCAR, MAR และ MNAR บนตัวแปรตาม สามารถแนะแนวทางได้เป็น 2 ด้าน ดังนี้

5.3.1 ด้านการนำไปใช้ประโยชน์

ผลการศึกษาการเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติกทวิภาค เมื่อเกิดปัญหาข้อมูลสูญหายบนตัวแปรตามที่เป็นข้อมูลเชิงคุณภาพ ในการทดลองได้จำลองสถานการณ์ที่แตกต่างกัน เพื่อวิเคราะห์ว่าวิธีการประมาณค่าสูญหายใดเหมาะสมกับสถานการณ์ใด สามารถสรุปได้ว่า กรณีที่ข้อมูลมีขนาดเล็ก ควรเลือกใช้การประมาณค่าสูญหายด้วยวิธี MLR และกรณีที่ข้อมูลมีขนาดใหญ่ ควรเลือกใช้การประมาณค่าสูญหายด้วยวิธี MI

สำหรับการวิเคราะห์ด้วยข้อมูลจริง ก่อนที่จะนำข้อมูลมาวิเคราะห์ในตัวแบบการถดถอยลอจิสติกทวิภาค ตามแนวทางการเลือกใช้วิธีการประมาณค่าสูญหายที่เหมาะสม ควรทำการตรวจสอบข้อตกลงเบื้องต้นของข้อมูล และตรวจสอบว่าข้อมูลมีความสัมพันธ์กันหรือไม่ด้วย

5.3.2 ด้านการศึกษาวิจัย

เพื่อเป็นแนวทางในการศึกษาการเปรียบเทียบวิธีการประมาณค่าสูญหายใจขอบเขตการวิจัยที่มีความหลากหลายขึ้นจากการศึกษาในครั้งนี้ ซึ่งครั้งต่อไปอาจทำการศึกษากรณีอื่น ๆ ดังนี้

1. ทำการศึกษาและพัฒนาวิธีการประมาณค่าสูญหายใหม่ ๆ และมีความหลากหลายมาเปรียบเทียบประสิทธิภาพ
2. ทำการศึกษาเพิ่มเติมในกรณีที่มีข้อมูลสูญหายเกิดขึ้นในตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพและตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพ
3. ทำการศึกษาเพิ่มเติมในกรณีที่มีตัวแปรอิสระมากกว่า 3 ตัว และมีตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพอย่างน้อย 1 ตัว
4. ทำการศึกษาเพิ่มเติมในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันในระดับต่าง ๆ
5. เพิ่มเกณฑ์ที่นำมาใช้ในการเปรียบเทียบวิธีการประมาณค่าสูญหายที่นอกเหนือจากค่า EMSE เพื่อให้การวิเคราะห์มีความน่าเชื่อถือมากยิ่งขึ้น

บรรณานุกรม

จุฑารัตน์ จันชัยภูมิ. (2564, 27 กรกฎาคม). *วิธีการจัดการข้อมูลสูญหาย.* https://roots.tech/th/blog/products-services-1/how-to-manage-missing-data-25

เบญจพร เอี่ยมประโคน, และณัตติฤดี เจริญรักษ์. (2561). วิธีการเปรียบเทียบพื้นที่ใต้เส้นโค้ง ROC สำหรับข้อมูลชุดเดียวกัน: กรณีศึกษาแบบจำลองคะแนนเครดิต. ใน *การประชุมวิชาการและนำเสนอผลงานวิชาการระดับชาติ UTCC Academic Day ครั้งที่ 2* (หน้า 1756-1769). มหาวิทยาลัยหอการค้าไทย.

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี, และสุคนธ์ ประสิทธิ์วัฒนเสรี. (2559). *Missing data and manage ment.* http://dmbj.ejnal.com/ejournal/showdetail/ ?show_detail=T&art_id=123

ปูเป้ สุดศิลา, อำไพ ทองธีรภาพ, และบุญอ้อม โฉมที. (2561). การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระ ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม. ใน *การประชุมวิชาการและนำเสนอผลงานวิชาการระดับชาติ UTCC Academic Day ครั้งที่ 2* (หน้า 1713-1727). มหาวิทยาลัยหอการค้าไทย.

พงษ์เดช สารการ, และภัทรนันท์ หมั่นพลศรี. (2564). จุดตัดที่เหมาะสมสำหรับการวิเคราะห์เส้นโค้ง Receiver Operating Characteristic (ROC) ในการพัฒนาเครื่องมือนวัตกรรมทางสุขภาพ: กรณีตัวอย่างโดยใช้โปรแกรม STATA. *Thai Bulletin of Pharmaceutical Sciences*, *16*(1), 93-108.

พัชนา สุวรรณแสน. (2562). การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์. *วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา, 4*(1), 1-9.

ภัทฐิดา นิลภัทรฉัตร. (2559). *การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์การถดถอยโลจิสติก* [วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต]. จุฬาลงกรณ์มหาวิทยาลัย.

ยุทธ ไกยวรรณ์. (2555). หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย. *วารสารวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย*, *4*(1), 1-12.

ระพีพรรณ ฉลองสุข. (2550). การกำหนดขนาดตัวอยาง. *วารสารไทยไภษัชยนิพนธ์*, *4*(1), 1-15. https://doi.org/10.14456/tbps.2009.1

สมชาย วรกิจเกษมสกุล. (2554). *การสุ่มตัวอย่าง*. https://www.gotoknow.org/posts/537047

สิวะโชติ ศรีสุทธิยากร. (2557). การวิเคราะห์ข้อมูลสูญหาย. *วารสารครุศาสตร์จุฬาลงกรณ์มหาวิทยาลัย*, *42*(1), 217-223.

สุกัญญา ศิริโยธา. (2558). *การประมาณค่าพารามิเตอร์ในตัวแบบถดถอยโลจิสติกเมื่อตัวแปรอิสระมีค่าสูญหาย* [วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต]. มหาวิทยาลัยนเรศวร.

สุชาดา กีระนันทน์. (2545). *การอนุมานเชิงสถิติ: ทฤษฎีชั้นต้น* (พิมพ์ครั้งที่ 3). สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

Abonazel, M. R., & Ibrahim, M. G. (2018). On estimation methods for binary logistic regression model with missing values. *International Journal of Mathematics and Computational Science*, *4*(3), 79-85.

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40-64.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work?. *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

Geb, D. (2023). *Heart disease predictions.* https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/input

Guo, C. Y., Yang, Y. C., & Chen, Y. H. (2021). The optimal machine learning-based missing data imputation for the cox proportional hazard model. *Frontiers in Public Health*, *9*, 1-8. https://doi.org/10.3389/fpubh.2021.680054

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC Medical Research Methodology*, *20*(1), 199. https://doi.org/10.1186/s12874-020-01080-1

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119013563

Meeyai, S. (2016). Logistic regression with missing data: A comparison of handling methods, and effects of percent missing values. *Journal of Traffic and Logistics Engineering*, *4*(2), 128-134.

Ozkale, M. R., & Arican, E. (2015). First-order r – d Class Estimator in Binary Logistic Regression Model. *Statistics & Probability Letters*, *106*, 19-29. https://doi.org/10.1016/j.spl.2015.06.021

Peng, C. Y. J., & Zhu, J. (2007). Comparison of two approaches for handling missing covariates in logistic regression. *Educational and Psychological Measurement, 68*(1), 58-77. https://doi.org/10.1177/0013164407305582

Peyre, H., Leplège, A., & Coste, J. (2010). Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Quality of Life Research*, *20*(2), 287-300.

Philip9876. (2018, June 12). *You tube: Multiple imputation by chained equations.* https://www.youtube.com/watch?v=zX-pacwVyvU

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., & Solenberger P. (2001). A multi variate technique for multiply imputing missing values using a sequence of Regression models. *Survey Methodology*, *27*(1), 85-95.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87-94. https://doi.org/10.2307/1391390

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, *7*, 1-12. https://doi.org/10.1177/2050312118822912

Tsiampalis, T., & Panagiotakos, D. B. (2020). Missing-data analysis: Socio- demographic, clinical and lifestyle determinants of low response rate on self- reported psychological and nutrition related multiitem instruments in the context of the ATTICA epidemiological study. *BMC Medical Research Methodology*, *20*(1), 148. https://doi.org/10.1186/s12874-020-01038-3

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), 1-7. http://dx.doi.org/10.1136/bmjopen-2013-002847

Xu, X., Xia, L., Zhang, Q., Wu, S., Wu, M., & Liu, H. (2020). The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology*, *20*(1), 42. https://doi.org/10.1186/s12874-020-00932-0

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

ภาคผนวก

คำสั่งโปรแกรม R-Studio

ชุดคำสั่งที่ใช้ในการวิเคราะห์การเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหาย 7 วิธี ทั้งชุดข้อมูลจำลองและข้อมูลจริง โดยทำการจำลองข้อมูลตามสถานการณ์ที่กำหนด ซึ่งมีคำสั่ง ดังต่อไปนี้

```
# ข้อมูลจำลอง #
##### Simulate data #####
set.seed(123)
x1 <- rnorm(1000)
x2 <- rnorm(1000)
z <- 1+x1+x2
pr <- 1/(1+exp(-z))
y <- rbinom(1000,1,pr)
data <- data.frame(y,x1,x2)

# ข้อมูลจริง #
##### Input heart disease predictions data #####
attach(heart)
data <- heart

#- - - - - - - - - - - - - - - - - - Install and Library Packages - - - - - - - - - - - - - - - - - - -#

#install.packages("logistf")
#install.packages("missMethods")
#install.packages("VIM")
#install.packages("mice")
#install.packages("missForest")
#install.packages("MKmisc")
#install.packages("caret")
```

```
#library(logistf)
#library(missMethods)
#library(VIM)
#library(mice)
#library(missForest)
#library(MKmisc)
#library(caret)

SampleSize <- 20        # 20 50 100 150 200 500 1000#
PercenMissing <- 0.1    # 0.1 0.2 0.3 0.4 #
Repeat <- 1000

Sum_Cal_Mode <- 0
Sum_Cal_HD <-  0
Sum_Cal_MI <- 0
Sum_Cal_KNN <- 0
Sum_Cal_RF <- 0
Sum_Cal_LR <- 0
Sum_Cal_MLR <- 0

i <- 1
while (i <= Repeat) {
#- - - - - - - - - - - - - - - - - - - - - - Simulate missing data - - - - - - - - - - - - -  - - - - - - - #
  Sample <- data[sample(nrow(data),SampleSize),];   #Sample
  full <- logistf(y~x1+x2,data = Sample)
  summary(full)
  beta0full <- full$coefficients[1]
  beta1full <- full$coefficients[2]
  beta2full <- full$coefficients[3]
```

```
  Missing <- ampute(Sample,prop = PercenMissing,patterns = c(0,1,1),mech = "MCAR")
  MissingData <- Missing$amp; #Missing data patterns are MCAR, MAR and MNAR
  MissingData

### - - - - - - - - - - - - - - - - - - - - - -METHODS - - - - - - - - - - - - - - - - - - - - - - -###

#- - - - - - - - - - - - - - - - - - - - - - Mode Imputation - - - - - - - - - - - - - - - - - - - - - -#
  Mode <- impute_mode(MissingData, type = "columnwise", convert_tibble = TRUE)
  fit_Mode <- logistf(y~x1+x2,data = Mode)
  beta0fit_Mode <- fit_Mode$coefficients[1]
  beta1fit_Mode <- fit_Mode$coefficients[2]
  beta2fit_Mode <- fit_Mode$coefficients[3]
  Cal_Mode <- ((beta0full-beta0fit_Mode)^2)+((beta1full-
          beta1fit_Mode)^2)+((beta2full-beta2fit_Mode)^2)
  Sum_Cal_Mode <- Sum_Cal_Mode + Cal_Mode

#- - - - - - - - - - - - - - - - - - - - - - Hot Deck Imputation - - - - - - - - - - - - - - - - - - - - - - #
  HD <- hotdeck(MissingData)
  fit_HD <- logistf(y~x1+x2,data = HD)
  beta0fit_HD <- fit_HD$coefficients[1]
  beta1fit_HD <- fit_HD$coefficients[2]
  beta2fit_HD <- fit_HD$coefficients[3]
  Cal_HD <- ((beta0full-beta0fit_HD)^2)+((beta1full-beta1fit_HD)^2)+((beta2full-
          beta2fit_HD)^2)
  Sum_Cal_HD <- Sum_Cal_HD + Cal_HD

#- - - - - - - - - - - - - - - - - - - - - - Multiple Imputation - - - - - - - - - - - - - - - - - - - - - -#
  imp_MI <- mice(MissingData)
  MI <- complete(imp_MI)
  fit_MI <- logistf(y~x1+x2,data = MI)
  beta0fit_MI <- fit_MI$coefficients[1]
```

```
  beta1fit_MI <- fit_MI$coefficients[2]
  beta2fit_MI <- fit_MI$coefficients[3]
  Cal_MI <- ((beta0full-beta0fit_MI)^2)+((beta1full-beta1fit_MI)^2)+((beta2full-
        beta2fit_MI)^2)
  Sum_Cal_MI <- Sum_Cal_MI + Cal_MI

#- - - - - - - - - - - - - - - - - - K-nearest Neighbor Imputation - - - - - - - - - - - - - - - - - #
  Sqrt_M <- round(sqrt(SampleSize-(PercenMissing*SampleSize)))
  KNN <- kNN(MissingData, k=Sqrt_M)
  fit_KNN <- logistf(y~x1+x2,data = KNN)
  beta0fit_KNN <- fit_KNN$coefficients[1]
  beta1fit_KNN <- fit_KNN$coefficients[2]
  beta2fit_KNN <- fit_KNN$coefficients[3]
  Cal_KNN <- ((beta0full-beta0fit_KNN)^2)+((beta1full-beta1fit_KNN)^2)+((beta2full-
        beta2fit_KNN)^2)
  Sum_Cal_KNN <- Sum_Cal_KNN + Cal_KNN

#- - - - - - - - - - - - - - - - - - Random Forest Imputation - - - - - - - - - - - - - - - - - -#
  Y <- factor(MissingData$y)
  miss <- data.frame(Y,MissingData$x1,MissingData$x2)
  MissForest <- missForest(miss)
  RF <-  MissForest$ximp
  fit_RF <- logistf(RF[,1]~RF[,2]+RF[,3],data = RF)
  beta0fit_RF <- fit_RF$coefficients[1]
  beta1fit_RF <- fit_RF$coefficients[2]
  beta2fit_RF <- fit_RF$coefficients[3]
  Cal_RF <- ((beta0full-beta0fit_RF)^2)+((beta1full-beta1fit_RF)^2)+((beta2full-
        beta2fit_RF)^2)
  Sum_Cal_RF <- Sum_Cal_RF + Cal_RF
```

```
#- - - - - - - - - - - - - - - - - - Logistic Regression Imputation - - - - - - - - - - - - - - - - - -#
  MissingData1<-MissingData
  predLR <- ifelse(predfull >= 0.5, 1, 0)
  MissingData1$y[which(is.na(MissingData1$y))] = predLR[is.na(MissingData1$y)]
  LR <- MissingData1
  fit_LR <- logistf(y~x1+x2,data = LR)
  beta0fit_LR <- fit_LR$coefficients[1]
  beta1fit_LR <- fit_LR$coefficients[2]
  beta2fit_LR <- fit_LR$coefficients[3]
  Cal_LR <- ((beta0full-beta0fit_LR)^2)+((beta1full-beta1fit_LR)^2)+((beta2full-
          beta2fit_LR)^2)
  Sum_Cal_LR <- Sum_Cal_LR + Cal_LR

#- - - - - - - - - - - - - - - - Modified Logistic Regression Imputation - - - - - - - - - - - - - -#
  # Optimalcutoff #
  predfull <- predict(full, type = "response")
  opt <- optCutoff(predfull, truth = Sample$y, namePos = 1)
  optcutoff <- as.numeric(opt[1])

  MissingData2<-MissingData
  predMLR <- ifelse(predfull >= optcutoff, 1, 0)
  MissingData2$y[which(is.na(MissingData2$y))] = predMLR[is.na(MissingData2$y)]
  MLR <- MissingData2
  fit_MLR  <- logistf(y~x1+x2,data = MLR)
  beta0fit_MLR  <- fit_MLR$coefficients[1]
  beta1fit_MLR  <- fit_MLR$coefficients[2]
  beta2fit_MLR  <- fit_MLR$coefficients[3]
  Cal_MLR <- ((beta0full-beta0fit_MLR)^2)+((beta1full-beta1fit_MLR)^2)+((beta2full-
          beta2fit_MLR)^2)
  Sum_Cal_MLR <- Sum_Cal_MLR + Cal_MLR
```

```
### - - - - - - - - - - - - - - - - - - - - - - - - Repeat - - - - - - - - - - - - - - - - - - - - - - - -###

  i = i+1
}

# EMSE #

EMSE_Mode <- Sum_Cal_Mode/Repeat;round(EMSE_Mode,4)
EMSE_HD <- Sum_Cal_HD/Repeat;round(EMSE_HD,4)
EMSE_MI <- Sum_Cal_MI/Repeat;round(EMSE_MI,4)
EMSE_KNN <- Sum_Cal_KNN/Repeat;round(EMSE_KNN,4)
EMSE_RF <- Sum_Cal_RF/Repeat;round(EMSE_RF,4)
EMSE_LR <- Sum_Cal_LR/Repeat;round(EMSE_LR,4)
EMSE_MLR <- Sum_Cal_MLR/Repeat;round(EMSE_MLR,4)

##### - - - - - - - - - - - - - - - - - - - - - - - - END - - - - - - - - - - - - - - - - - - - - - - - -#####
```

ประวัติผู้เขียน

ชื่อ สกุล นางสาวธิดารัตน์ ธรรมโชโต

รหัสประจำตัวนักศึกษา 6510220054

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วิทยาศาสตรบัณฑิต (สถิติ)	มหาวิทยาลัยสงขลานครินทร์	2564

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ทุนสนับสนุนบัณฑิตศึกษาจากกองทุนวิจัย คณะวิทยาศาสตร์ ประเภททุนตรี-โท สัญญาเลขที่ 1-2565-02-006