การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ

Comparison of Missing Data Imputation Methods for Independent Variables in Logistic Regression Model with Multicollinearity Data

ธัญพิชชา ฤทธิ์เทวา

Thanpitcha Ritthewa

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

การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ

Comparison of Missing Data Imputation Methods for Independent Variables in Logistic Regression Model with Multicollinearity Data

ธัญพิชชา ฤทธิ์เทวา

Thanpitcha Ritthewa

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 การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ

ผู้เขียน นางสาวธัญพิชชา ฤทธิ์เทวา

สาขาวิชา สถิติประยุกต์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....

(รองศาสตราจารย์ ดร. ใกล้รุ่ง สามารถ)

คณะกรรมการสอบ

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร. วราฤทธิ์ พานิชกิจโกศลกุล)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. พรฑิตา ทิวทัศน์)

.....กรรมการ

(รองศาสตราจารย์ ดร. ใกล้รุ่ง สามารถ)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาสถิติประยุกต์

.....

(ผู้ช่วยศาสตราจารย์ ดร.กวินพัฒน์ สิริกานติโสภณ)

รักษาการแทนคณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

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

ลงชื่อ .....
(นางสาวธัญพิชชา ฤทธิ์เทวา)
นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ .....
(นางสาวธัญพิชชา ฤทธิ์เทวา)
นักศึกษา

ชื่อวิทยานิพนธ์ การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ
ผู้เขียน นางสาวธัญพิชชา ฤทธิ์เทวา
สาขาวิชา สถิติประยุกต์
ปีการศึกษา 2566

บทคัดย่อ

การวิเคราะห์การถดถอยลอจิสติก (Logistic regression analysis) เป็นเทคนิควิธีการที่ใช้พยากรณ์ความน่าจะเป็นที่จะเกิดหรือไม่เกิดเหตุการณ์ที่สนใจ ที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระเป็นได้ทั้งข้อมูลเชิงปริมาณและคุณภาพ ซึ่งได้นำมาประยุกต์หลากหลายศาสตร์ โดยเฉพาะอย่างยิ่งในกรณีของข้อมูลทางการแพทย์ ข้อมูลที่สูญหายอาจส่งผลต่อความน่าเชื่อถือในการประเมินผู้ป่วย และทำให้ไม่สามารถจำแนกบุคคลตามระดับของสุขภาพหรือการเป็นโรคได้ นอกจากนี้ ลักษณะของข้อมูลอาจเกิดความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ระหว่างตัวแปรอิสระ ทำให้ผลลัพธ์ที่ได้ไม่สอดคล้องกับความเป็นจริง ดังนั้นในงานวิจัยนี้จึงสนใจเปรียบเทียบวิธีการประมาณค่าสูญหายของข้อมูลเมื่อมีการสูญหายบนตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) 6 วิธี ได้แก่ วิธี Mean imputation (Mean), Multiple imputation (MI), K-nearest neighbor imputation (KNN), Random forest imputation (RF), Stochastic regression imputation (SRI) และ วิธี Bayesian linear regression imputation (BRI) ที่มีระดับเปอร์เซ็นต์การสูญหายที่ระดับ 10%, 20%, 30% และ 40% โดยมีรูปแบบการสูญหายของข้อมูล 3 แบบ คือ การสูญหายแบบสุ่มสมบูรณ์ (Missing completely at random: MCAR), การสูญหายแบบสุ่ม (Missing at random: MAR) และการสูญหายแบบไม่สุ่ม (Missing not at random: MNAR) เกณฑ์ในการเปรียบเทียบประสิทธิภาพพิจารณาจาก ค่า Estimated mean square error: EMSE โดยวิธีที่ให้ค่า EMSE ต่ำที่สุดคือวิธีที่มีประสิทธิภาพมากที่สุด ผลการวิจัยพบว่าเมื่อตัวอย่างมีขนาดใหญ่ ทุกระดับเปอร์เซ็นต์การสูญหาย วิธี RF มีประสิทธิภาพมากที่สุด นอกจากนี้พบว่า ค่า EMSE เพิ่มขึ้น เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น และลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

Thesis Title Comparison of Missing Data Imputation Methods for Independent Variables in Logistic Regression Model with Multicollinearity Data
Author Miss Thanpitcha Ritthewa
Major Program Applied Statistics
Academic Year 2023

ABSTRACT

Logistic regression analysis is a technique for predicting the probability of an occurrence of a particular event when the dependent variable is qualitative. Data from both quantitative and qualitative sources can be used as the independent variable. It has been used in a wide range of sciences. In particular, in the case of medical data, missing data can lead to a loss of trust in patient evaluation and make it impossible to classify people according to their level of health or disease. Furthermore, multicollinearity between independent variables can lead to misleading results. Therefore, the objective of this research is to study the efficiency of missing data imputation methods for logistic regression when multicollinearity occurs. The missing data imputation methods considered in this research were : mean imputation (MEAN), multiple imputation (MI), k-nearest neighbor imputation (KNN), random forest imputation (RF), stochastic regression imputation (SRI), and bayesian linear regression imputation (BRI). In this study, the simulation was done with sample sizes of 20, 50, 100, 150, 200, 500, and 1000, and the percentages of missing data were 10%, 20%, 30%, and 40%. The estimated mean square error (EMSE) was used to compare efficiency. The results showed that when the sample size is large and there is a high percentage of missing data, the RF method is most effective. The EMSE rises when the percentage of missing data rises and falls when the sample size decreases.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบคุณที่ได้รับความอนุเคราะห์และสนับสนุนเป็นอย่างดีจากอาจารย์ที่ปรึกษา รศ. ดร. ใกล้รุ่ง สามารถ ที่ได้กรุณาให้คำปรึกษา ความรู้ ข้อคิด ปรับปรุงแก้ไขข้อบกพร่องต่างๆ และให้ความเมตตาตลอดมา ทำให้ข้าพเจ้าสำเร็จการศึกษาลุล่วงไปได้ด้วยดี

ขอกราบขอบพระคุณ “ทุนสนับสนุนบัณฑิตศึกษาจากกองทุนวิจัย คณะวิทยาศาสตร์ ประเภททุนตรี-โท ประจำปีการศึกษา 2565” ที่กรุณามอบโอกาสให้เข้าศึกษาในระดับบัณฑิตศึกษา พร้อมทั้งมอบเงินทุนการศึกษาให้แก่ข้าพเจ้า

ขอกราบขอบพระคุณ รศ. ดร. วราฤทธิ์ พานิชกิจโกศลกุล อาจารย์ประจำสาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่ได้กรุณารับเป็นกรรมการสอบวิทยานิพนธ์ และได้ให้คำแนะนำในการแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

ขอขอบคุณอาจารย์ทุกท่านในสาขาวิทยาศาสตร์การคำนวณ มหาวิทยาลัยสงขลานครินทร์ ที่ร่วมกันแบ่งปันความรู้ให้คำปรึกษาและสนับสนุนจนทำให้ข้าพเจ้าได้รับปริญญามหาบัณฑิตนี้

ขอบคุณครอบครัวสำหรับความรักและกำลังใจและขอบคุณเพื่อนๆที่คอยให้กำลังใจและข้อเสนอแนะต่างๆ

สุดท้ายนี้ข้าพเจ้าขอขอบคุณทุกท่านที่สนับสนุนและให้กำลังใจ ที่ไม่ได้กล่าวถึงข้างต้น

ธัญพิชชา ฤทธิ์เทวา

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูปภาพ

สารบัญรูปภาพ (ต่อ)

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

การวิเคราะห์การถดถอยลอจิสติก (Logistic regression analysis) เป็นเทคนิควิธีการที่ใช้พยากรณ์ความน่าจะเป็นที่จะเกิดหรือไม่เกิดเหตุการณ์ที่สนใจ ที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ส่วนตัวแปรอิสระเป็นได้ทั้งข้อมูลเชิงปริมาณและคุณภาพ ซึ่งได้นำมาประยุกต์หลากหลายศาสตร์ โดยเฉพาะอย่างยิ่งทางการแพทย์ เนื่องจากในงานวิจัยนี้ผู้วิจัยได้นำเทคนิคมาวิเคราะห์ข้อมูลเกี่ยวกับปัจจัยเสี่ยงการเป็นโรค ซึ่งจะเป็นประโยชน์ในด้านการแพทย์ในการวินิจฉัยโรคและลดการผิดพลาดในการวินิจฉัยโรคได้ ซึ่งส่งผลให้ผู้ป่วยมีอัตราการรอดชีวิตเพิ่มมากขึ้น รวมไปถึงการนำไปใช้วิเคราะห์อย่างแพร่หลายทางด้านเศรษฐศาสตร์ ด้านสังคมศาสตร์ และอื่น ๆ (Schober & Vetter, 2021)

ข้อมูลสูญหาย เป็นปัญหาที่นักวิจัยพบในการนำข้อมูลไปวิเคราะห์ ด้วยเทคนิควิธีการต่างๆ ในทางปฏิบัติเป็นไปได้ยากที่ข้อมูลจากการเก็บรวบรวมจะมีความสมบูรณ์ถูกต้อง อาจเนื่องมาจากขั้นตอนในการเก็บข้อมูลไม่รัดกุม ผู้ตอบให้ข้อมูลไม่ครบถ้วนหรือผู้ตอบไม่ให้ข้อมูลอาจเนื่องมาจากผู้ตอบไม่ทราบคำตอบ เช่น การสอบถามระดับน้ำตาลในเลือด การสอบถามค่าสายตา (Graham, 2009) ซึ่งปัญหาเหล่านี้อาจทำให้เกิดผลกระทบต่อประสิทธิภาพในการวิเคราะห์ข้อมูล หากนำข้อมูลที่มีข้อมูลสูญหายนั้นมาทำการวิเคราะห์จะทำให้ผลการวิเคราะห์ข้อมูลเกิดความคลาดเคลื่อนไปจากความเป็นจริง จึงมีผลกระทบทำให้การประมาณค่าพารามิเตอร์เกิดความคลาดเคลื่อนมากขึ้น อย่างไรก็ตามนักวิจัยส่วนใหญ่จะให้ความสำคัญตรงนี้น้อย มักจะไม่บอกถึงการมีข้อมูลสูญหายและวิธีการจัดการกับข้อมูลสูญหาย แต่จะตัดทิ้งตัวอย่างที่มีข้อมูลสูญหายออกไปทำให้ตัวอย่างมีขนาดลดลง ซึ่งจะส่งผลต่อการวิเคราะห์และนำไปสู่การลดความน่าเชื่อถือทางสถิติ ทั้งนี้ Little and Rubin (2019) ได้จำแนกประเภทของข้อมูลสูญหายไว้ 3 ประเภท (1) การสูญหายแบบสุ่มสมบูรณ์ (MCAR) เป็นลักษณะการสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด ไม่ขึ้นกับตัวแปรตัวอื่นหรือการสูญหายของตัวเอง (2) การสูญหายแบบสุ่ม (MAR) เป็นลักษณะการสูญหายอย่างสุ่มที่ขึ้นอยู่กับตัวแปรอื่น และ (3) การสูญหายแบบไม่สุ่ม (MNAR) เป็นลักษณะการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มแต่ขึ้นอยู่กับ

ตัวแปรเดียวกันรวมถึงตัวแปรตัวอื่นด้วย ซึ่งในการวิเคราะห์ข้อมูล หากตัดข้อมูลสูญหายทิ้ง อาจทำให้การประมาณค่าพารามิเตอร์ในตัวแบบเกิดการเบี่ยงเบน และนำไปสู่การสรุปผลที่ผิดพลาด

ดังนั้นนักวิจัยจึงจำเป็นต้องมีแนวทางในการจัดการเพื่อลดการเกิดข้อมูลสูญหาย และต้องมีความรู้เกี่ยวกับการดำเนินการเกี่ยวกับข้อมูลสูญหาย เพื่อเลือกใช้วิธีการดำเนินการกับข้อมูลสูญหายได้อย่างเหมาะสม การจัดการกับข้อมูลสูญหายนั้นสามารถทำได้หลายวิธี ซึ่งวิธีที่ง่ายที่สุดคือ วิธี Listwise Deletion ซึ่งเป็นวิธีการตัดค่าของข้อมูลสูญหายทิ้งไป นั่นคือไม่สนใจข้อมูลสูญหายที่เกิดขึ้น โดยจะทำการวิเคราะห์เฉพาะข้อมูลส่วนที่สมบูรณ์ (Roth, 1994) โดยวิธีนี้จะทำให้ขนาดของข้อมูลลดน้อยลง ส่งผลต่อการสูญเสียระดับความเชื่อมั่นและนำไปสู่การสรุปผลที่ผิดพลาด จึงมีนักวิจัยหลายท่านคิดค้นและพัฒนาวิธีการประมาณค่าสูญหายให้มีประสิทธิภาพมากขึ้น และมีความคลาดเคลื่อนน้อยลง นอกจากนี้อีกปัญหาหนึ่งที่มักจะพบ คือ ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งจะนำไปสู่การสรุปผลที่มีความน่าเชื่อถือลดน้อยลง

อย่างไรก็ตามงานวิจัยที่ทำในด้านการประมาณค่าสูญหายในตัวแบบการถดถอยลอจิสติกมีไม่มากนัก โดยเฉพาะในกรณีข้อมูลมีความสัมพันธ์เชิงเส้นพหุ ดังนั้นนักวิจัยจึงมีความสนใจที่จะศึกษาเกี่ยวกับ การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าข้อมูลสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกกรณีที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) 6 วิธี ได้แก่ Mean imputation (Mean), MI, KNN, RF, SRI และ Bayesian linear regression imputation (BRI) เพื่อใช้เป็นแนวทางในการเลือกวิธีการจัดการค่าสูญหายเมื่อตัวแปรอิสระมีการสูญหายแบบ MCAR, MAR และ MNAR รวมถึงแสดงให้เห็นถึงการป้องกันและลดขนาดของข้อมูลสูญหายที่จะเกิดขึ้น เพื่อเป็นแนวทางสำหรับการบริหารจัดการข้อมูลและนำไปเปรียบเทียบได้อย่างเหมาะสมและดีที่สุดกับลักษณะของข้อมูลในกรณีที่เกิดปัญหาข้อมูลสูญหายในงานวิจัยต่อไป

1.2 ขอบเขตการวิจัย

งานวิจัยนี้มุ่งเน้นไปที่การประยุกต์ใช้วิธีการแทนค่าสูญหายในสถานการณ์จริงและเปรียบเทียบวิธีการแทนค่าสูญหายที่แตกต่างกัน 6 วิธี โดยได้ศึกษาจาก 2 ชุดข้อมูล ได้แก่ ข้อมูลจำลองและข้อมูลจริง สำหรับข้อมูลจำลองที่มีการสูญหายของตัวแปรอิสระจำนวน 3 ตัว ในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ผ่านการจำลองด้วยขนาดตัวอย่าง 20, 50, 100, 150, 200, 500 และ 1000 กำหนดเปอร์เซ็นต์การสูญหาย 10%, 20%, 30%, 40% ในรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR นอกจากที่ยังมีการทดลองประยุกต์ใช้กับ

ข้อมูลจริงซึ่งเป็นข้อมูลในการศึกษาปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน โดยมีขนาดตัวอย่าง 443 คน กำหนดเปอร์เซ็นต์การสูญหาย 10%, 20%, 30%, 40% ในรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR เช่นกัน เพื่อเปรียบเทียบประสิทธิภาพของวิธีการแทนค่าสูญหายทั้ง 6 วิธี และสามารถนำไปประยุกต์ใช้ได้อย่างเหมาะสม

1.3 วัตถุประสงค์ของการศึกษา

1.3.1 เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าข้อมูลสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ 6 วิธีได้แก่ วิธี Mean imputation (Mean), Multiple imputation (MI), K-nearest neighbor (KNN), Random forest imputation (RF), Stochastic regression imputation (SRI) และ วิธี Bayesian linear regression imputation (BRI)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้รับความรู้ในการเลือกวิธีการประมาณค่าข้อมูลสูญหายที่มีประสิทธิภาพในการประมาณที่ดีและใกล้เคียงกับข้อมูลจริงมากที่สุด

1.4.2. สามารถใช้เป็นแนวทางในการเลือกใช้วิธีการจัดการค่าสูญหายให้เหมาะสมกับลักษณะของข้อมูล

1.4.3. สามารถใช้ในการพิจารณาเปรียบเทียบค่าสูญหายในการวิเคราะห์ข้อมูลที่มีความสัมพันธ์กันในรูปแบบอื่นๆ หรือการสูญหายในสถานการณ์อื่นๆ

1.5 นิยามศัพท์เฉพาะ

1.5.1 Imputation คือ เป็นวิธีการแทนค่าข้อมูลที่มีการสูญหายด้วยค่าอื่นๆแทน วิธีการนี้ถูกนำมาใช้เนื่องจากไม่สามารถลบข้อมูลออกจากชุดข้อมูลได้เพราะอาจนำไปสู่การลดขนาดของชุดข้อมูลและอาจนำไปสู่การวิเคราะห์ที่ผิดพลาดได้ (Singhal, 2021)

1.5.2 การวิเคราะห์การถดถอยลอจิสติก (Logistic regression analysis) คือ การวิเคราะห์ตัวแปรเชิงพหุที่มีวัตถุประสงค์เพื่อทำนายเหตุการณ์ที่สนใจว่าจะเกิดหรือไม่เกิดเหตุการณ์นั้นภายใต้อิทธิพลของตัวปัจจัย แบบจำลองลอจิสติกจะประกอบด้วยตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพ

(Qualitative variable) และตัวแปรอิสระหรือตัวแปรทำนาย อาจมีตัวเดียวหรือหลายตัวที่สามารถเป็นได้ทั้งตัวแปรเชิงคุณภาพ (Qualitative variable) หรือตัวแปรเชิงปริมาณแบบต่อเนื่อง (Continuous variable) การถดถอยลอจิสติกจัดเป็นเครื่องมือวิเคราะห์ข้อมูลในการศึกษาวิจัยที่มีวัตถุประสงค์เพื่อทำนายเหตุการณ์ หรือประเมินความเสี่ยง จึงอาจมีการประยุกต์ใช้ในงานวิจัยหลากหลายสาขาได้ (กาญจน์เขจร ชูชีพ, 2561)

1.5.3 ข้อมูลสูญหาย (Missing data) คือ ข้อมูลที่ค่าสังเกตไม่มีการระบุค่าหรือเกิดค่าว่าง ซึ่งอาจเกิดจากการที่ผู้ให้ข้อมูลไม่ประสงค์จะให้คำตอบหรือข้อมูลที่มีการรวบรวมไว้แล้ว แต่เป็นข้อมูลที่ไม่สมบูรณ์ เช่น เกิดการหายของข้อมูลหรือไม่มีข้อมูลในส่วนนั้น โดยที่ค่านั้นควรจะสามารถทราบค่าได้ หากวิธีการแทนค่าที่ใช้มีประสิทธิภาพและมีความเหมาะสม (พัชนา สุวรรณแสน, 2561)

1.5.4 ความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) คือ การมีสหสัมพันธ์กันเองระหว่างตัวแปรอิสระที่มากกว่า 2 ตัว ซึ่งหากตัวแปรอิสระมีความสัมพันธ์กันในระดับที่สูง อาจส่งผลให้สมการตัวแบบที่ใช้ในการพยากรณ์ตัวแปรตามเกิดความความคลาดเคลื่อนได้ สำหรับสาเหตุของการที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) อาจเกิดขึ้นเพราะตัวแปรอิสระที่นำมาใช้มีความสัมพันธ์กันที่แน่นอนอยู่แล้ว ในบางกรณีอาจเกิดขึ้นเนื่องจากกระบวนการรวบรวมตัวอย่างเชิงสุ่มที่บังเอิญได้ข้อมูลที่มีความสัมพันธ์กันมาก็ได้ (สุฤดี โกศัยเนตร, 2549)

1.5.5 โรคเบาหวาน (Diabetes) โรคที่มีระดับน้ำตาลในเลือดสูงเกินกว่าปกติอย่างต่อเนื่องและเรื้อรัง โดยเกิดจากความผิดปกติของตับอ่อน ก่อให้เกิดการหลั่งฮอร์โมนอินซูลินได้น้อยกว่าปกติหรือเกิดภาวะดื้อต่ออินซูลิน จึงทำให้เซลล์ร่างกายมีความผิดปกติในกระบวนการเปลี่ยนน้ำตาลในเลือดให้เป็นพลังงาน (โรงพยาบาลศิริราช ปิยมหาราชการุณย์, 2553)

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษาเรื่อง การเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องในประเด็นดังต่อไปนี้

1. ทฤษฎีที่เกี่ยวข้องกับข้อมูลสูญหายและวิธีการประมาณค่าสูญหาย
2. งานวิจัยที่เกี่ยวข้องกับวิธีการประมาณค่าสูญหาย

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 รูปแบบการสูญหายของข้อมูล

ในการพิจารณาประเภทของข้อมูลสูญหายจัดเป็นขั้นตอนที่มีความสำคัญ หากสามารถทราบถึงลักษณะของข้อมูลสูญหาย ก็จะช่วยพิจารณาแนวทางสำหรับการจัดการกับการสูญหายของข้อมูลได้อย่างเหมาะสมและมีประสิทธิภาพ ซึ่งจำแนกรูปแบบของข้อมูลสูญหายออกเป็น 3 ประเภท (Little & Rubin, 2019) ดังนี้

1. การสูญหายแบบสุ่มสมบูรณ์ (MCAR) คือ ลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด ข้อมูลสูญหายจะเป็นอิสระจากตัวแปรต่างๆ สำหรับรูปแบบการสูญหายของข้อมูลประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เนื่องจากข้อมูลสูญหายไม่มีความเกี่ยวข้องใดต่อผลลัพธ์ของข้อมูล ดังนั้นข้อมูลสูญหายประเภทนี้สามารถเลือกทำการวิเคราะห์ส่วนของข้อมูลที่สมบูรณ์ได้

2. การสูญหายแบบสุ่ม (MAR) คือ ลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มเพียงบางส่วนของค่าสังเกตเท่านั้น นั่นคือค่าของข้อมูลสูญหายจะขึ้นอยู่กับตัวแปรอื่นๆที่ไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย สำหรับข้อมูลการสูญหายประเภทนี้ยังไม่ได้ส่งผลกระทบในการวิเคราะห์ข้อมูลมากนัก

3. การสูญหายแบบไม่สุ่ม (MNAR) คือ ลักษณะของข้อมูลสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายจะขึ้นอยู่กับค่าของข้อมูลที่สมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรอื่นด้วย ในบางครั้งข้อมูลสูญหายประเภทนี้ ค่าของข้อมูลสูญหายอาจไม่ขึ้นกับตัวแปรใดๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นๆที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษานั้น ดังนั้นข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่อาจส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูลมากที่สุด

2.1.2 วิธีการประมาณค่าสูญหาย

1. Mean imputation (MEAN)

วิธี MEAN ใช้หลักการแทนค่าสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ไม่ได้สูญหาย โดยข้อมูลที่สูญหายในตัวแปรเดียวกันจะถูกแทนด้วยค่าเฉลี่ยที่เท่ากัน โดยมีสูตรคำนวณสำหรับการสูญหายในตัวแปรอิสระ x_i ดังนี้ (Shrive el al., 2006)

$$\overline{X} = \frac{\sum_{i=1}^{c} x_i}{c} \qquad (1)$$

เมื่อ c คือจำนวนค่าสังเกตที่มีข้อมูลครบถ้วน

2. Multiple imputation (MI)

วิธี MI จะมีการนำข้อมูลจากหลาย ๆ ส่วนมาพิจารณาก่อนการแทนค่า Missing values เข้าไป เช่น มีการใช้ข้อมูลจากข้อมูลที่สมบูรณ์ครบถ้วน ที่อาจส่งผลต่อข้อมูลที่หายไป อาจมีการใช้ Machine learning models เป็นตัวช่วยในการหาค่าประมาณที่เหมาะสม
ซึ่งประกอบด้วย 3 ขั้นตอน (Dong & Peng, 2013) ดังนี้

ขั้นตอนที่ 1 คือ ทำการประมาณค่าข้อมูลสูญหายแต่ละวิธี เพื่อให้เป็นชุดข้อมูลที่สมบูรณ์

ขั้นตอนที่ 2 คือ วิเคราะห์ข้อมูลแต่ละชุดแยกกัน เพื่อนำมาประมาณค่าข้อมูลสูญหาย

ขั้นตอนที่ 3 คือ รวบรวมผลลัพธ์มาสรุปค่าที่ใช้แทนค่าข้อมูลสูญหายทั้งหมด
ในการวิจัยในครั้งนี้วิธี MI ได้ใช้ขั้นตอนของ MICE ซึ่งแบ่งออกเป็น 6 ขั้นตอนดังนี้

1. ใช้ Simple imputation แทนค่าสูญหายด้วยวิธี MEAN หรือวิธีการใด ๆ ในชุดข้อมูล ค่าที่ถูกแทนเข้าไปด้วย Simple imputation เรียกว่า “procurator”

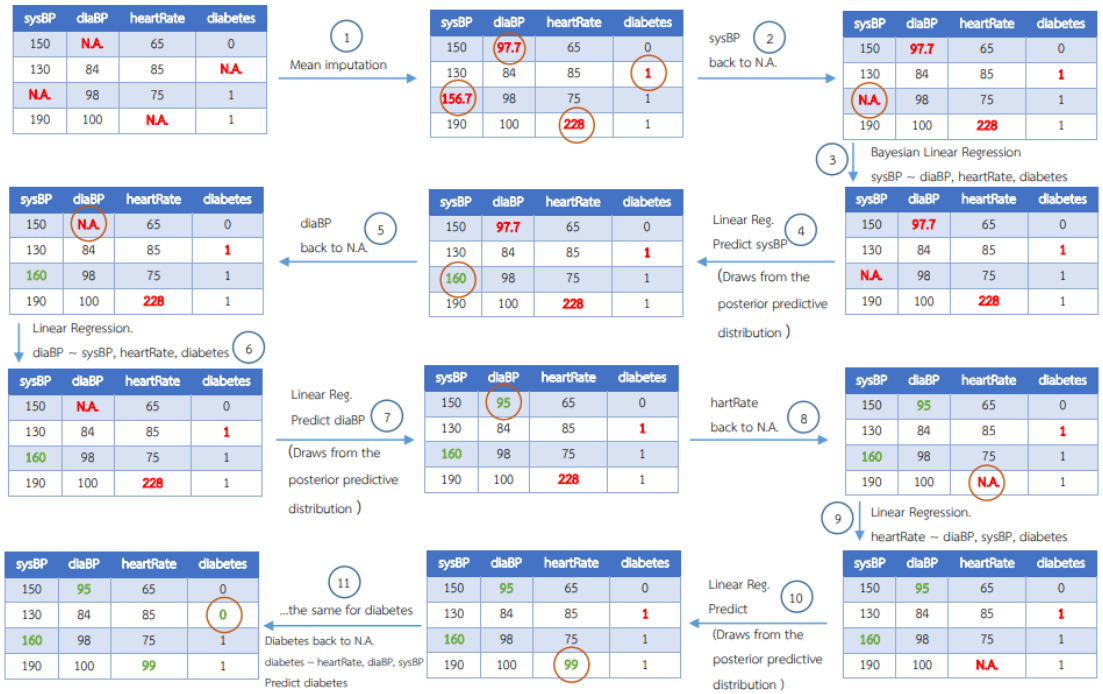
2. จากนั้น “procurator” ของตัวแปรที่ต้องการทำกระบวนการ MICE จะถูกทำให้เป็น missing values อีกครั้ง ตัวแปรนี้เรียกว่า “var”

3. ตัวแปร “var” จะถูกใช้เป็นตัวแปรตาม และตัวแปรอื่น ๆ ทั้งหมดจะเป็นตัวแปรอิสระ ทำการวิเคราะห์การถดถอย เพื่อหาตัวแบบทำนายสำหรับตัวแปร “var”

4. เมื่อได้ตัวแบบทำนายมาแล้ว ค่า “procurator” ของตัวแปร “var” จะถูกแทนที่ด้วย ค่าทำนายที่เกิดจากตัวแบบ และใช้ค่าทั้งหมดของตัวแปร “var” เป็นตัวแปรอิสระสำหรับการหาตัวแบบทำนายของตัวแปร “var” ตัวถัดไป

5. ทำซ้ำตั้งแต่ขั้นตอนที่ 2–4 เพื่อหาตัวแบบทำนายของทุกตัวแปร “var” ที่มีค่าสูญหายจนทุกตัวแปร “var” ถูกเติมเต็มทั้งหมด

6. ทำซ้ำตั้งแต่ขั้นตอนที่ 2–4 เพื่อหาตัวแบบทำนายในแต่ละรอบ เพราะในแต่ละรอบตัวแบบอาจมีค่าทำนายที่เปลี่ยนแปลงไป



รูปที่ 2.1 ตัวอย่างการคำนวณวิธี Multiple imputation

MI เป็นวิธีการที่มีศักยภาพช่วยให้ผลการศึกษาทางด้านการแพทย์และสาธารณสุขที่มีข้อมูลสูญหายมีความถูกต้องมากยิ่งขึ้น เนื่องจากวิธีการนี้เป็นการสร้างชุดข้อมูลขึ้นมาหลายๆชุด และแทนค่าข้อมูลสูญหายในแต่ละชุดข้อมูลด้วยค่าทำนายจากตัวแบบทางสถิติที่เหมาะสม และเนื่องจากในงานวิจัยนี้ประกอบไปด้วยข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ ดังนั้นในการจัดการด้วยวิธี MI จึงอาจเป็นวิธีการที่ให้ประสิทธิภาพและให้ความถูกต้องเหมาะสมของตัวแบบทางสถิติที่ใช้ทำนายข้อมูลสูญหายอีกด้วย (Azur et al. 2011)

3. K-nearest neighbor imputation (KNN)

วิธี KNN เป็นวิธีที่ได้รับความนิยมอย่างมาก เนื่องจากเป็นวิธีที่ง่ายและมีประสิทธิภาพอีกวิธีหนึ่งที่นำมาใช้ประมาณ โดยข้อมูลที่จะนำมาประมาณค่ากับข้อมูลที่สูญหายต้องมีความสัมพันธ์กันเพื่อนำมาสร้างเป็นแบบจำลอง ผู้วิจัยต้องกำหนดค่า k เพื่อใช้ในการพิจารณาข้อมูลที่อยู่ใกล้ที่สุด ซึ่งจะต้องเป็นจำนวนเต็มบวก โดย $k=\sqrt{c}$ เมื่อ c คือจำนวนค่าสังเกตที่มีข้อมูลครบถ้วน เช่น $k=3$ คือพิจารณาเฉพาะข้อมูล 3 ตัวแรกที่มีค่า Euclidian distance น้อยที่สุด ซึ่งมีขั้นตอนดังนี้ (Troyanskaya et al., 2001)

1. กำหนดค่า k

sysBP	diaBP	heartRate	diabetes
132	83.5	75	0
206	92	76	1
96	63	65	0
179.5	114	90	0
N.A.	71	88	0
114	76	88	1
143.5	81	75	0
190	99	100	1
123	76.5	60	0
134	80	88	1

ทำการหาค่า k คือ

$$k = \sqrt{c} = \sqrt{9} = 3$$

รูปที่ 2.2 ตัวอย่างการคำนวณวิธี K-nearest neighbor imputation

2. คำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidian distance ระหว่างข้อมูลที่เกิดค่าสูญหายที่ต้องการพิจารณากับข้อมูลที่มีความสมบูรณ์ ดังสมการ (Jösson & Wohlin, 2006)

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

โดยที่

$dist(R_i, R_j)$	แทน	ระยะห่างระหว่างข้อมูลแถวที่ i และข้อมูลแถวที่ j
p	แทน	จำนวนตัวแปรอิสระที่ข้อมูลมีความสมบูรณ์
$x_{i,s}$	แทน	ค่าข้อมูลที่เกิดการสูญหาย แถวที่ i คอลัมน์ที่ s
$x_{j,s}$	แทน	ค่าข้อมูลที่มีความสมบูรณ์ แถวที่ i คอลัมน์ที่ s

$$dist(R_5, R_1) = \sqrt{(71-83.5)^2 + (88-75)^2} = 18.0347$$

$$dist(R_5, R_2) = \sqrt{(71-92)^2 + (88-76)^2} = 24.1868$$

$$dist(R_5, R_3) = \sqrt{(71-63)^2 + (88-65)^2} = 24.3516$$

$$dist(R_5, R_4) = \sqrt{(71-114)^2 + (88-90)^2} = 43.0465$$

$$dist(R_5, R_6) = \sqrt{(71-76)^2 + (88-88)^2} = 5.0000$$

$$dist(R_5, R_7) = \sqrt{(71-81)^2 + (88-75)^2} = 16.4012$$

$$dist(R_5, R_8) = \sqrt{(71-99)^2 + (88-100)^2} = 30.4630$$

$$dist(R_5, R_9) = \sqrt{(71-76.5)^2 + (88-60)^2} = 28.5350$$

R	sysBP	diaBP	heartRate	diabetes
R1	132	83.5	75	0
R2	206	92	76	1
R3	96	63	65	0
R4	179.5	114	90	0
R5	N.A.	71	88	0
R6	114	76	88	1
R7	143.5	81	75	0
R8	190	99	100	1
R9	123	76.5	60	0
R10	134	80	88	1

รูปที่ 2.3 ตัวอย่างการคำนวณวิธี K-nearest neighbor imputation (ต่อ)

3. เรียงลำดับระยะห่างระหว่างจุดซึ่งพิจารณาจากข้อมูลที่ใกล้ที่สุดตามจำนวน k

R	sysBP	diaBP	heartRate	diabetes	Dist	Sort
R6	114	76	88	1	5.0000	1
R7	143.5	81	75	0	16.4012	3
R10	134	80	88	1	9.0000	2

รูปที่ 2.4 ตัวอย่างการคำนวณวิธี K-nearest neighbor imputation (ต่อ)

4. ประมาณค่าข้อมูลสูญหายจากค่าเฉลี่ยของข้อมูลที่อยู่ใกล้ที่สุด ดังสมการ

$$\hat{x}_i = \frac{\sum_{i=1}^{k} x_i}{k} \tag{3}$$

โดยที่

$\hat{x}_i$ แทน ค่าของข้อมูลสูญหายที่ได้จากการประมาณค่าใหม่

k แทน จำนวนที่กำหนดไว้เพื่อพิจารณาค่าที่อยู่ใกล้ที่สุด

x_i แทน ค่าข้อมูลสมบูรณ์ที่ระยะห่างใกล้ที่สุดของตัวแปรที่มีการสูญหาย

R	sysBP	diaBP	heartRate	diabetes
R1	132	83.5	75	0
R2	206	92	76	1
R3	96	63	65	0
R4	179.5	114	90	0
R5	130.5	71	88	0
R6	114	76	88	1
R7	143.5	81	75	0
R8	190	99	100	1
R9	123	76.5	60	0
R10	134	80	88	1

ระยะห่างที่ใกล้ที่สุด 3 ตัว คือ

$dist(R_5, R_6)$

$dist(R_5, R_7)$

$dist(R_5, R_{10})$

นั่นคือ ค่าประมาณที่ได้จากวิธี KNN คือ

$$\hat{x}_i = \frac{\sum_{i=1}^{k} x_i}{k} = \frac{114 + 143.5 + 134}{3} = 130.5$$

รูปที่ 2.5 ตัวอย่างการคำนวณวิธี K-nearest neighbor imputation (ต่อ)

4. Random forest imputation (RF)

วิธี RF สามารถใช้ในการประมาณค่าสูญหายของข้อมูลที่มีความสัมพันธ์เชิงเส้นและความสัมพันธ์ไม่เชิงเส้นได้และสามารถใช้ได้ทั้งข้อมูลเชิงปริมาณและคุณภาพรวมทั้งยังใช้ได้ดีในกรณีที่ข้อมูลมีความสัมพันธ์เชิงพหุ ในโปรแกรม R อยู่ในแพ็กเกจที่มีชื่อว่า missForest ซึ่งมีประสิทธิภาพในการประมาณค่าสูญหายอยู่ในระดับที่ดีภายใต้การสูญหายของข้อมูลที่ระดับปานกลางจนถึงระดับสูง

โดยที่อธิบายถึงหลักการ ของวิธี RF (Stekhoven & Bühlmann, 2012; Thongsri & Samart, 2022a, 2022b) ดังนี้

1. ตำแหน่งที่มีข้อมูลสูญหายจะถูกแทนค่าด้วยค่าเฉลี่ย (Mean) สำหรับข้อมูลเชิงปริมาณ และฐานนิยม (Mode) สำหรับข้อมูลเชิงคุณภาพ

sysBP	diaBP	heartRate	diabetes
132	83.5	75	0
206	92	76	1
96	63	65	0
179.5	114	90	0
N.A.	71	88	0
114	76	88	1
143.5	81	75	0
190	99	100	1
123	76.5	60	0
134	80	88	1

sysBP	diaBP	heartRate	diabetes
132	83.5	75	0
206	92	76	1
96	63	65	0
179.5	114	90	0
146.44	71	88	0
114	76	88	1
143.5	81	75	0
190	99	100	1
123	76.5	60	0
134	80	88	1

$$Mean = \frac{132 + 206 + 96 + ... + 123 + 134}{9} = 146.44$$

รูปที่ 2.6 ตัวอย่างการคำนวณวิธี Random forest imputation

2. กระบวนการแทนค่าจะทำตามลำดับโดยเรียงจากน้อยไปมากหรือจากมากไปน้อย ของค่าตัวแปรข้อมูลสูญหายแต่ละตัวแปร ตัวแปรที่อยู่ภายใต้การแทนค่าจะใช้เป็นตัวแปรตามในการสร้างตัวแบบ Random forest ซึ่งข้อมูลสูญหายจะถูกแทนค่าจากการทำนายด้วยตัวแบบ Random forest โดยทั่วไปจะสุ่มตัวอย่างที่มีขนาด 2 ใน 3 ของประชากร

sysBP	diaBP	heartRate	diabetes
132	83.5	75	0
206	92	76	1
96	63	65	0
179.5	114	90	0
146.44	71	88	0
114	76	88	1
143.5	81	75	0
190	99	100	1
123	76.5	60	0
134	80	88	1

Random forest model

sysBP	diaBP	heartRate	diabetes
190	99	100	1
96	63	65	0
123	76.5	60	0
146.44	71	88	0
206	92	76	1
143.5	81	75	0

รูปที่ 2.7 ตัวอย่างการคำนวณวิธี Random forest imputation (ต่อ)

3. โดยชุดข้อมูลจะแบ่งออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลที่ไม่พบค่าสูญหาย เรียกว่า “ชุดข้อมูลฝึกฝน” และชุดข้อมูลที่พบค่าสูญหาย เรียกว่า “ชุดข้อมูลทำนาย”

sysBP	diaBP	heartRate	diabetes
132	83.5	75	0
179.5	114	90	0
114	76	88	1
143.5	81	75	0
123	76.5	60	0
134	80	88	1

ชุดข้อมูลฝึกฝน

sysBP	diaBP	heartRate	diabetes
190	99	100	1
96	63	65	0
123	76.5	60	0
146.44	71	88	0
206	92	76	1
143.5	81	75	0

ชุดข้อมูลทำนาย

รูปที่ 2.8 ตัวอย่างการคำนวณวิธี Random forest imputation (ต่อ)

4. กระบวนการแทนค่าจะหยุดก็ต่อเมื่อตัวแปรที่มีข้อมูลสูญหายได้ถูกแทนค่าเข้าไปก็จะจบการวนซ้ำ 1 ครั้ง ดังนั้นกระบวนการแทนค่าจะทำซ้ำไปซ้ำมาจนกระทั่งค่า Normalized root mean squared error (NRMSE) ของการแทนค่าครั้งล่าสุดและครั้งก่อนหน้ามีค่าเพิ่มขึ้น นั่นคือค่า $NRMSE_t$ มีค่ามากกว่าค่า $NRMSE_{t-1}$ (โดยที่ t คือ ชุดข้อมูลที่ได้จากแบบจำลอง Random forest) กระบวนการถึงจะหยุด โดยปกติจะมีการวนซ้ำ 5-6 ครั้ง จึงจะได้ชุดข้อมูลที่ได้รับการแทนค่าที่ดีที่สุด (Oba et al., 2003)

$$NRMSE = \sqrt{\frac{mean((x_{true} - x_{imp})}{\text{var}(x_{true})}} \quad (4)$$

โดยที่

x_{true} คือ ค่า x ที่สูญหายและได้รับการแทนค่าสูญหายด้วยค่าเฉลี่ย

x_{imp} คือ ค่าประมาณของตัวแปร x ที่ได้จากแบบจำลอง Random forest

$\text{var}(x_{true})$ คือ ค่าความแปรปรวนของ x_{true}

ซึ่งการดำเนินการกับชุดข้อมูลที่ใส่เข้าไปในแต่ละชุด โดยผ่านการถดถอยเชิงเส้นหรือการถดถอยลอจิสติกก็ขึ้นอยู่กับประเภทของแต่ละตัวแปร ซึ่งกระบวนการเหล่านี้ถูกเปรียบเทียบโดยใช้ค่า NRMSE

5. Stochastic regression imputation (SRI)

วิธี SRI เป็นวิธีการที่ใช้สมการถดถอย จากชุดข้อมูลที่ทราบค่าหรือชุดข้อมูลที่สมบูรณ์ ซึ่งจะมีการเพิ่มพจน์ของค่าคลาดเคลื่อนสุ่ม (Residual term) เข้ามาในตัวแบบการถดถอยเพื่อมาประมาณ

ค่าสูญหายของตัวแปร เพื่อให้ความแปรปรวนมีค่าใกล้เคียงกับความเป็นจริงมากขึ้น และ SRI มีตัวแบบการถดถอย ดังนี้ (Enderers, 2022; Thongsri & Samart, 2022a)

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + ... + \hat{\beta}_p x_p^* + e \quad (5)$$

โดยที่

$\hat{y}^*$	คือ	ค่าทำนายจากตัวแบบการถดถอย (Regression model) ของชุดข้อมูลที่สมบูรณ์
$\hat{\beta}_0 ... \hat{\beta}_p$	คือ	ค่าประมาณของสัมประสิทธิ์การถดถอย
$x_1^* ... x_p^*$	คือ	ตัวแปรทำนายของข้อมูลสมบูรณ์
e	คือ	ค่าประมาณของค่าคลาดเคลื่อนสุ่ม โดยมีการแจกแจงปรกติ ค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับความแปรปรวนของความคลาดเคลื่อนในสมการถดถอย

6. Bayesian linear regression imputation (BRI)

การถดถอยเชิงเส้นแบบเบย์ (Bayesian linear regression imputation) เป็นอีกตัวแบบหนึ่งที่สามารถนำมาใช้ได้ โดยจุดมุ่งหมายของการถดถอยเชิงเส้นแบบเบย์ไม่ใช่เป็นการหาผลลัพธ์เพียง 1 ค่าจากตัวแบบ แต่เป็นการกำหนดการแจกแจงภายหลัง (Posterior distribution) ของพารามิเตอร์จากตัวแบบที่เป็นไปได้ ประมาณค่าจาก Probabilistic distribution กำหนดสมการถดถอยเชิงเส้นโดยใช้การแจกแจงความน่าจะเป็น ซึ่งจะแตกต่างจากตัวแบบการถดถอยเชิงเส้น (Linear regression) คือเป็นการประมาณค่าจุดเดียวสำหรับผลลัพธ์ซึ่งนั่นคือค่าประมาณที่สามารถเป็นไปได้มากที่สุด อย่างไรก็ตามหากมีชุดข้อมูลขนาดเล็กอาจจะต้องแสดงค่าประมาณเป็นการกระจายค่าที่เป็นไปได้ (Jadhav et al., 2019)

ขั้นตอนการดำเนินงาน วิธี BRI ดำเนินการดังนี้

1.กำหนด Prior distribution ให้ $\hat{\beta} \sim N(\hat{\beta}_{com}, \text{var}(\hat{\beta}_{com}))$

โดย $\hat{\beta}_{com}$ คือ สัมประสิทธิ์การถดถอยประมาณจากชุดข้อมูลที่ไม่มีค่าสูญหาย

2. สุ่มซ้ำ $\hat{\beta}$ 1000 ครั้งโดยใช้วิธี Markov Chain Monte Carlo (MCMC) เพื่อสร้าง Posterior distribution

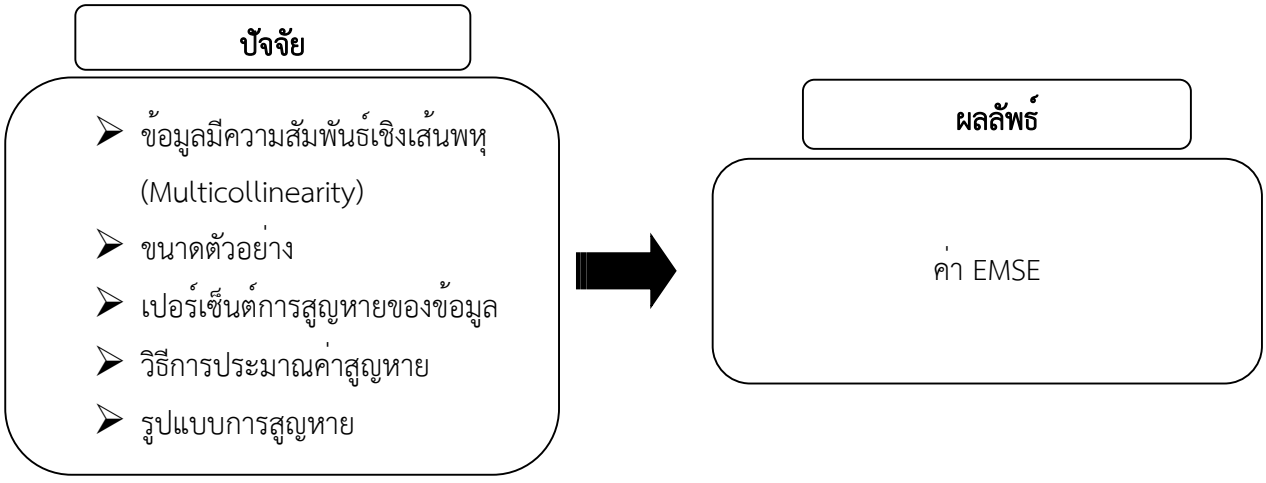
3. นำค่าเฉลี่ยของ $\hat{\beta}$ ที่ได้จากการสุ่มซ้ำ มาใช้ในสมการถดถอยเชิงเส้นเพื่อหาค่าประมาณของค่าสูญหาย โดยการแทนค่าข้อมูลสูญหายสามารถทำได้โดยอัตโนมัติในโปรแกรม R ใช้คำสั่ง MICE ซึ่งการวนซ้ำจะอยู่ที่ 5 ครั้ง

2.2 งานวิจัยที่เกี่ยวข้อง

ในการจัดการกับข้อมูลสูญหายมีหลากหลายวิธีให้เลือกใช้ การพิจารณาว่าจะเลือกใช้วิธีใดขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้น หากเลือกวิธีที่ไม่เหมาะสมอาจนำไปสู่การเพิ่มค่าความคลาดเคลื่อนและสูญเสียระดับความเชื่อมั่นได้ โดยในช่วงปีที่ผ่านๆมา ได้มีงานวิจัยในลักษณะการเปรียบเทียบประสิทธิภาพวิธีการแทนค่าสูญหายที่แตกต่างกันไป ซึ่งก่อนหน้านี้ได้มีงานวิจัยในลักษณะการเปรียบเทียบประสิทธิภาพวิธีการแทนค่าสูญหายในตัวแบบการถดถอยลอจิสติกที่ตัวแปรอิสระไม่มีความสัมพันธ์กันโดย Meeyai (2016) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 5 วิธี ได้แก่ วิธี Listwise deletion, Mean substitution, Regression imputation (RI), SRI และ MI โดยการจำลองข้อมูลตัวอย่างขนาด 10, 20, 30, 40, 50,100, 250, 500, 1000 และ 2500 โดยมีตัวแปรอิสระ 2 ตัว กำหนดให้เกิดข้อมูลสูญหายบนตัวแปรอิสระ ที่ระดับเปอร์เซ็นต์การสูญหาย 10%, 20%, 30%, 40%, 50%, 60%, 70% และ 80% ในรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR ผลการวิจัยพบว่าวิธี MI ให้ประสิทธิภาพดีที่สุด ซึ่งสอดคล้องกับงานวิจัยของ Tsiampalis and Panagiotakos (2020) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 7 วิธี ได้แก่ วิธี Complete case analysis (CCA), Proration, Score mean imputation (SMI), Item mean imputation (IMI), Person mean imputation (PMI), SRI และ MI โดยการใช้ข้อมูลขนาดตัวอย่างจาก ATTICA epidemiological study ซึ่งเป็นมาตรวัดระดับการรับประทานอาหารและภาวะเป็นโรคซึมเศร้า โดยในแต่ละมาตรวัดจะมีปัจจัยต่าง ๆ ที่ใช้ในการประเมิน ซึ่งงานวิจัยนี้กำหนดให้ปัจจัยในแต่ละมาตรวัดเป็นตัวแปรอิสระที่มีการสูญหายของข้อมูล และการเป็นโรคความดันโลหิตสูง (เป็น/ไม่เป็น) เป็นตัวแปรตาม ผลการวิจัยพบว่าวิธี MI ให้ประสิทธิภาพดีที่สุด และยังสอดคล้องกับงานวิจัยของ Xu et al. (2020) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 4 วิธีได้แก่ วิธี Direct deletion method, Mode imputation, Hot-deck imputation (HD) และ MI ในชุดข้อมูลแบบสอบถามทางจิตวิทยา 3 ชุด โดย ชุดข้อมูล Self-acceptance scale (SAQ), ชุดข้อมูล Activities of daily living scale (ADL) และ ชุดข้อมูล Self-esteem scale (RSES) โดยให้ชุดข้อมูล SAQ และ ADL เป็นกลุ่มจำลอง (Simulation group) กำหนดเปอร์เซ็นต์การสูญหาย 5%, 10%, 15% และ 20% และชุดข้อมูล RSES ให้เป็นกลุ่มตรวจสอบ (Validation group) กำหนดรูปแบบการสูญหายแบบ MAR ผลการวิจัยพบว่าวิธี MI ให้ประสิทธิภาพดีที่สุดเช่นกัน ในขณะที่งานวิจัยของ Kokla et al. (2019) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 9 วิธีได้แก่ วิธี Metabolomics: ZERO, Mean, Minimum value (MIN), Half minimum ($\frac{1}{2}$MIN), Singular value decomposition (SVD), Probabilistic principal component analysis (PPCA), Bayesian principal component analysis (BPCA), RF และ KNN ในการแทนค่าสูญหายของจำนวนคุณลักษณะของโมเลกุล ที่มีผลต่อการป้อนข้อมูลเมแทบอโลมิกส์

(Metabolomics data-analysis) ในเซลล์ ผลการวิจัยพบว่าวิธี RF ให้ประสิทธิภาพดีที่สุด ซึ่งสอดคล้องกับงานวิจัยของ Hong and Lynn (2020) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 2 วิธีได้แก่ วิธี Predictive mean matching (PMM) และ RF โดยใช้ข้อมูลจำลองทางคลินิกในการศึกษาผู้ป่วยมะเร็งตับ ในรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR ผลการวิจัยพบว่าวิธี RF ให้ประสิทธิภาพดีที่สุดในรูปแบบ MAR อีกทั้งยังสอดคล้องกับงานวิจัยของ Epp-Stobbe et al. (2022) ได้ใช้เทคนิควิธีการแทนค่าสูญหาย 10 วิธีได้แก่ วิธี Daily team mean substitution, KNN, RF, Support vector machine (SVM), Neural network, Linear, Stepwise, Lasso, Ridge, และ Elastic net ในการแทนค่าสูญหายของตัวแปรอิสระได้แก่ Match number, Player, Opponent, Total distance, Player time และ Contact count ที่มีอิทธิพลในการวัดค่า Rate of perceived exertion (RPE) ของนักกีฬารักบี้ ที่ระดับเปอร์เซ็นต์การสูญหาย 5%, 10%, 15%, 20%, 25% และ 30% ในรูปแบบ MAR ผลการวิจัยพบว่าวิธี RF ให้ประสิทธิภาพดีที่สุดเช่นกัน

2.3 กรอบแนวคิดการวิจัย



รูปที่ 2.9 กรอบแนวคิดการวิจัย

บทที่ 3
ระเบียบวิธีวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ ผู้วิจัยจึงได้กำหนดวิธีดำเนินการวิจัย ซึ่งมีรายละเอียดในการดำเนินการศึกษา ดังนี้

3.1 ขั้นตอนการวิเคราะห์ข้อมูล

3.2 เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล

3.1 ขั้นตอนการวิเคราะห์ข้อมูล

3.1.1 ตัวอย่างและการสุ่มตำแหน่งของข้อมูลสูญหายในข้อมูลจำลอง (Simulation data)

ผู้วิจัยได้จำลองข้อมูลตามสถานการณ์ต่างๆ โดยทำซ้ำ 1000 ครั้งในแต่ละสถานการณ์ด้วยโปรแกรม RStudio ภายใต้สถานการณ์ดังนี้

3.1.1.1 สร้างตัวแปรอิสระสามตัว (X_1, X_2, X_3) ที่มีการแจกแจงปรกติมาตรฐาน ซึ่งก็คือ $X_1 \sim \mathrm{N}(0,1)$, $X_2 \sim \mathrm{N}(0,1)$ และ $X_3 \sim \mathrm{N}(0,1)$ โดยจะสร้างตัวแปรอิสระ X_1 และ X_2 มีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน โดย X_2 เกิดจากการสุ่มตัวแปร $X_2 \sim \mathrm{N}(0,1) + (10\mathrm{X}_1)$ โดยค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ระหว่างตัวแปรอิสระ X_1 และ X_2 มีค่าเท่ากับ 0.95

3.1.1.2 สร้างชุดข้อมูลที่มีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยลอจิสติกทวิภาค โดยกำหนดค่าสัมประสิทธิ์การถดถอย $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$ และ $\beta_3 = 1$ มีตัวแบบดังนี้ (Boateng & Abaye, 2019)

$$\pi_x = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)} \qquad (6)$$

โดยที่

$\beta_0, \beta_1, \beta_2$ และ β_3 คือ สัมประสิทธิ์การถดถอย

π_x คือ ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

3.1.1.3 สร้างตัวแปรตามที่มีการแจกแจงแบร์นูลลี $Y \sim Ber(\pi_x)$

3.1.1.4 เลือกตัวอย่างจากการการสุ่มอย่างง่าย (Simple random sampling) กำหนดขนาดตัวอย่าง (n) เท่ากับ 20 50 100 150 200 500 และ 1000

3.1.1.5 สร้างสมการถดถอยลอจิสติกประมาณเพื่อหาค่า $\hat{\beta}_{b(t)}$ โดย $\hat{\beta}_{b(t)}$ คือ สัมประสิทธิ์การถดถอยประมาณของข้อมูลตัวอย่างที่ไม่มีค่าสูญหายในรอบที่ t

3.1.1.6 กำหนดเปอร์เซ็นต์การสูญหาย 10%, 20%, 30% และ 40% บนตัวแปรอิสระ โดยรูปแบบการสูญหายแบบ MCAR, MAR และ MNAR ตามลำดับ

3.1.1.7 ประมาณค่าข้อมูลและแทนค่าข้อมูลสูญหายจากวิธีการประมาณค่าสูญหายทั้ง 6 วิธี

3.1.1.8 สร้างสมการถดถอยลอจิสติกประมาณเพื่อหาค่า $\hat{\beta}^{*}{}_{b(t)}$ โดย $\hat{\beta}^{*}{}_{b(t)}$ คือ สัมประสิทธิ์การถดถอยประมาณของข้อมูลตัวอย่างที่มีการแทนค่าสูญหายแล้วในรอบที่ t

3.1.1.9 คำนวณค่า EMSE จากวิธีการประมาณค่าสูญหายทั้ง 6 วิธี โดยการทำซ้ำ 1000 ครั้ง ในแต่ละสถานการณ์

3.1.2 การสุ่มตำแหน่งของข้อมูลสูญหายในข้อมูลจริง (Real life data)

ข้อมูลที่นำมาศึกษาเป็นชุดข้อมูลจริงที่เผยแพร่ต่อสาธารณะจาก www.kaggle.com ที่มาจากการศึกษาโรคหัวใจและโรคหลอดเลือดของประชาชนในเมืองฟรามิงแฮม รัฐแมสซาชูเซตส์ ซึ่งในงานวิจัยนี้ต้องการศึกษาปัจจัยที่มีผลต่อการเป็นโรคเบาหวาน โดยมีขนาดตัวอย่างเท่ากับ 443 คน สำหรับงานวิจัยนี้นำตัวแปรอิสระจำนวน 3 ตัว ที่มีอิทธิพลต่อการเป็นโรคเบาหวาน คือ ความดันของเลือดสูงสุดขณะหัวใจห้องล่างบีบตัว (Systolic blood pressure), ความดันเลือดที่ต่ำสุดขณะหัวใจห้องล่างคลายตัว (Diastolic blood pressure) และอัตราการเต้นของหัวใจ (Heart rate) ซึ่งความดันเลือดที่ต่ำสุดขณะหัวใจห้องล่างคลายตัว (Diastolic blood pressure) และอัตราการเต้นของหัวใจ (Heart rate) เป็นตัวแปรที่มีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน โดยค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation coefficient) ระหว่างความดันเลือดที่ต่ำสุดขณะหัวใจห้องล่างคลายตัวและอัตราการเต้นของหัวใจมีค่าเท่ากับ 0.82 ซึ่งการเป็นโรคเบาหวานและไม่เป็นโรคเบาหวานเป็นตัวแปรตามเชิงคุณภาพมีเพียงสองค่า กำหนดให้เป็น 1 เมื่อเป็นโรคเบาหวาน และเป็น 0 เมื่อไม่ได้เป็นโรคเบาหวาน กำหนดให้มีการสูญหายรูปแบบ MCAR MAR และ MNAR ในตัวแปรอิสระทั้ง 3 ตัว ที่ระดับเปอร์เซ็นต์การสูญหาย 10% 20% 30% และ 40% ของขนาดตัวอย่าง จากนั้นประมาณค่าสูญหายจากวิธีการประมาณค่าสูญหายทั้ง 6 วิธี

3.1.3 การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย

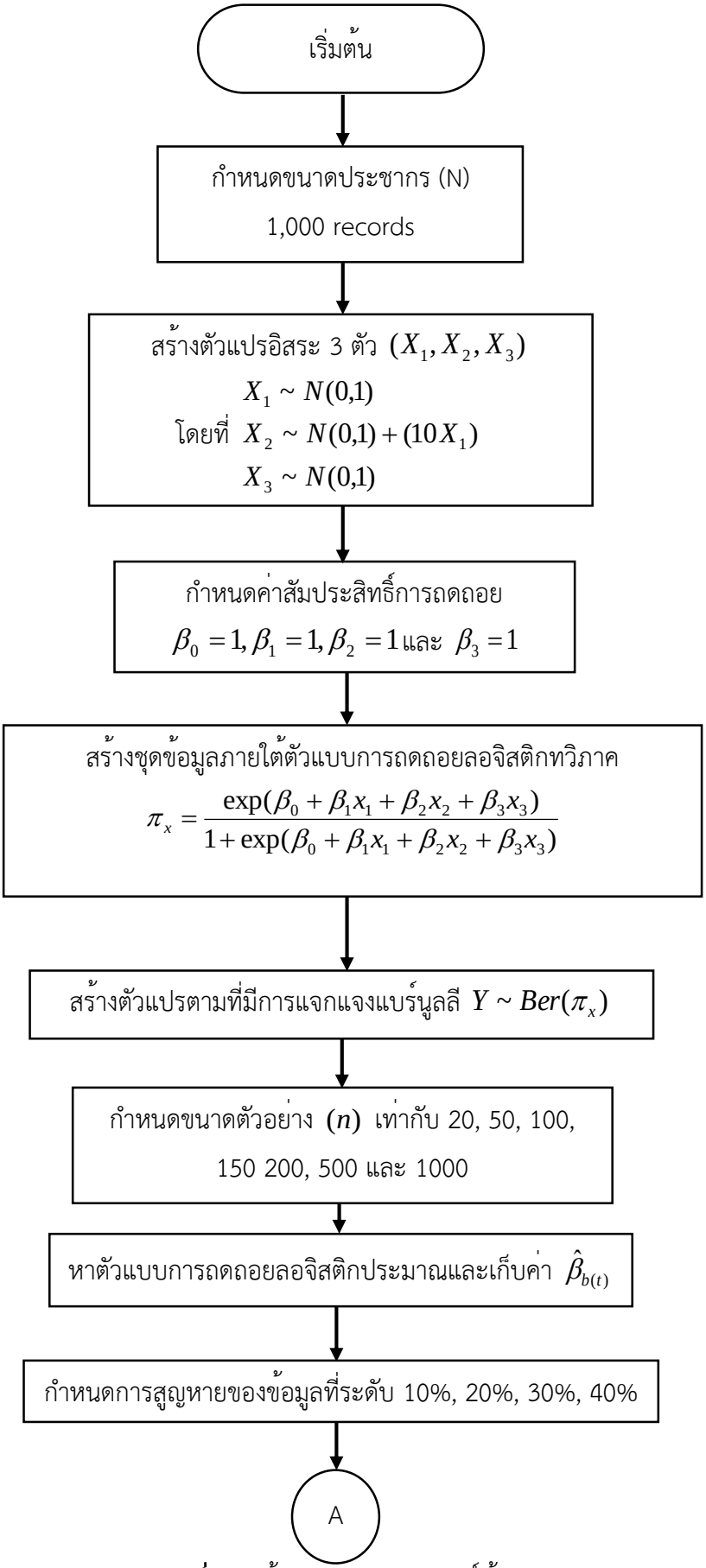
คำนวณค่า EMSE จากวิธีการประมาณค่าสูญหายทั้ง 6 วิธี โดยการทำซ้ำ 1000 ครั้งในแต่ละสถานการณ์ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจัดการค่าสูญหาย ดังสูตรต่อไปนี้ (Torabi & Rao, 2013)

$$EMSE(\tilde{\beta}) = \frac{1}{1000}\sum_{t=1}^{1000}\sum_{b=0}^{3}(\hat{\beta}_{b(t)} - \hat{\beta}^{*}_{b(t)})^2 \qquad (7)$$

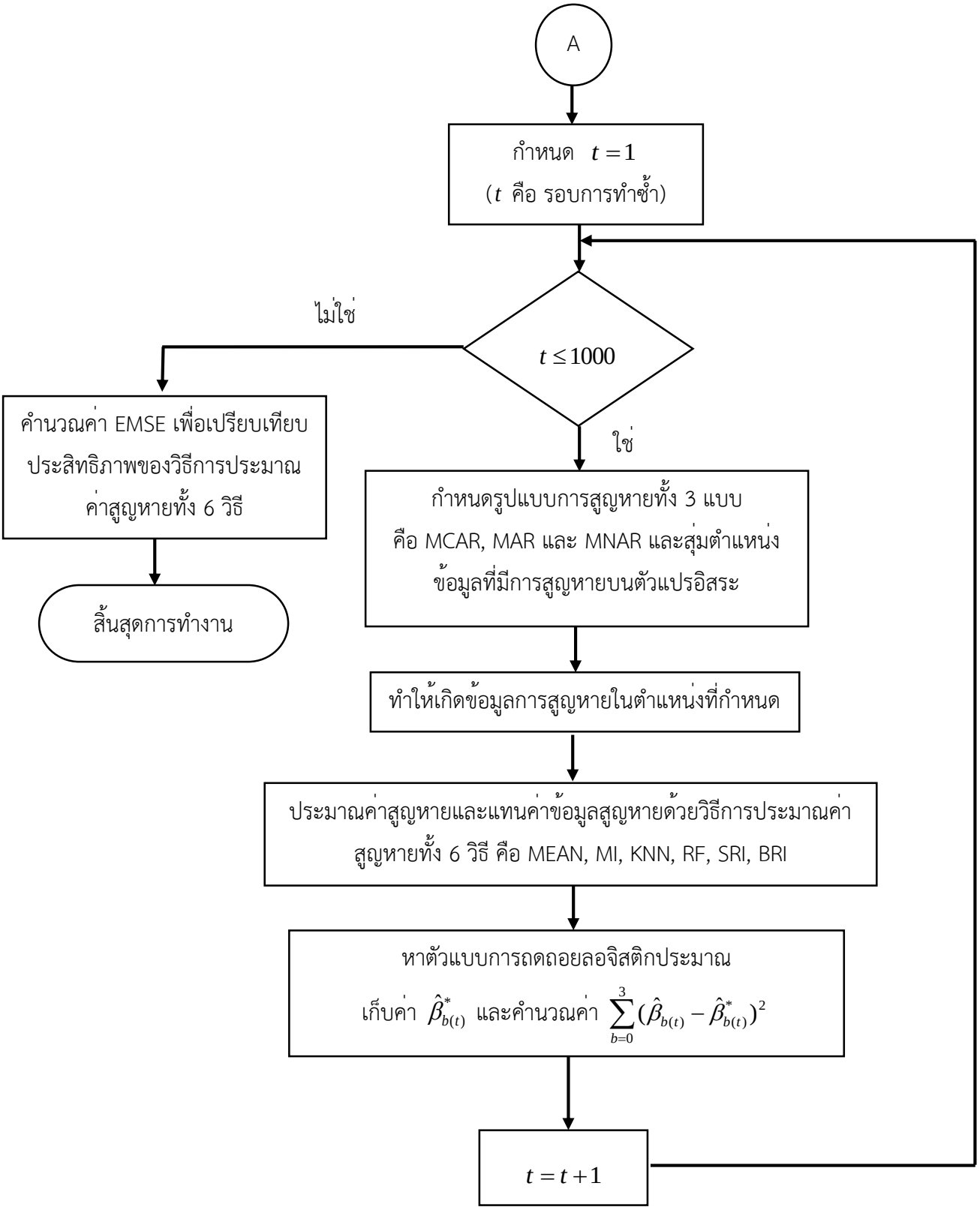
เปรียบเทียบประสิทธิภาพ โดยพิจารณาจากค่า EMSE วิธีการประมาณค่าสูญหายวิธีใดมีค่า EMSE ต่ำกว่า แสดงว่าวิธีการประมาณค่าสูญหายนั้นมีประสิทธิภาพดีกว่า

เนื่องจากการที่ชุดข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน อาจส่งผลต่อตัวแบบการถดถอย ซึ่งนั่นคือ จะส่งผลต่อค่าสัมประสิทธิ์การถดถอย ดังนั้น การเลือกใช้ค่า EMSE ซึ่งเป็นสูตรที่มีการคำนวณผลต่างระหว่างค่าสัมประสิทธิ์ของข้อมูลที่ไม่มีค่าสูญหายและค่าสัมประสิทธิ์ของข้อมูลที่มีการแทนค่าสูญหายแล้ว จึงมีความเหมาะสมและดีที่สุดในการเลือกใช้สำหรับงานวิจัยนี้

3.1.4 ขั้นตอนการวิเคราะห์ข้อมูลแสดงโดยแผนผังการดำเนินการ ดังนี้



รูปที่ 3.1 ขั้นตอนการวิเคราะห์ข้อมูล



รูปที่ 3.2 ขั้นตอนการวิเคราะห์ข้อมูล (ต่อ)

3.2 เครื่องมือที่ใช้ในการวิเคราะห์ข้อมูล

วิเคราะห์ข้อมูลโดยใช้โปรแกรม RStudio Version 4.2.0 ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 6 วิธี กรณีข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน

บทที่ 4
ผลการศึกษา

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ผลการศึกษาในงานวิจัยนี้มีการศึกษาเปรียบเทียบ 2 ชุดข้อมูลได้แก่ ชุดข้อมูลจำลองและชุดข้อมูลจริง โดยในแต่ละชุดข้อมูลจะจำแนกตามรูปแบบการสูญหายของข้อมูล 3 รูปแบบ (MCAR MAR และ MNAR) ใน**ตารางที่ 4.1–4.3** ตามลำดับ จะแสดงค่า EMSE ของตัวแบบการถดถอยลอจิสติกที่มีการแทนค่าสูญหายด้วย 6 วิธีการของข้อมูลจำลองที่ขนาดตัวอย่าง เปอร์เซ็นต์การสูญหายและรูปแบบการสูญหายของข้อมูลที่แตกต่างกัน และ**ตารางที่ 4.4–4.6** ตามลำดับ จะแสดงค่า EMSE ของตัวแบบการถดถอยลอจิสติกที่มีการแทนค่าสูญหายด้วย 6 วิธีการของข้อมูลจริงที่ขนาดตัวอย่าง เปอร์เซ็นต์การสูญหายและรูปแบบการสูญหายของข้อมูลที่แตกต่างกัน โดยผู้วิจัยได้แบ่งผลการปรียบเทียบเป็น 4 ส่วน ดังนี้

4.1 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MCAR สำหรับข้อมูลจำลอง

4.2 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MAR สำหรับข้อมูลจำลอง

4.3 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MNAR สำหรับข้อมูลจำลอง

4.4 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MCAR สำหรับข้อมูลจริง

4.5 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MAR สำหรับข้อมูลจริง

4.6 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MNAR สำหรับข้อมูลจริง

ตารางที่ 4.1 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MCAR สำหรับข้อมูลจำลอง

Sample size	Percent of missing data	Imputation methods					
		KNN	RF	MEAN	MI	SRI	BRI
20	10	23.7992	**19.0738**	36.2594	29.0452	30.0160	31.7090
	20	44.1405	**40.0106**	55.3267	47.7033	50.6724	49.2388
	30	47.5078	49.9433	53.8835	52.8398	**47.3946**	47.8936
	40	**54.4982**	56.0811	61.9076	58.6243	54.6611	55.5599
50	10	25.6163	**14.245**	45.5481	42.7178	41.8258	44.2428
	20	40.0348	**32.6905**	57.6051	50.1667	53.4080	53.3692
	30	47.2899	**43.0832**	56.6310	58.2593	52.6011	50.8315
	40	**62.5970**	74.1813	64.6013	69.7725	64.9811	67.7484
100	10	9.8350	**5.9855**	22.9239	20.9177	21.3363	23.0791
	20	14.8219	**11.4742**	25.8157	22.9490	24.7963	25.2837
	30	20.4867	**17.5985**	26.3160	27.3076	26.7810	25.7673
	40	**25.7382**	27.2526	31.1672	29.4120	31.2357	29.4678
150	10	4.6693	**2.8057**	12.8839	10.8782	12.0460	13.5447
	20	8.6411	**5.4395**	16.2057	14.6634	15.2777	15.2273
	30	11.6875	**10.9347**	17.8989	15.2296	17.9940	16.6729
	40	14.8529	**14.7937**	19.7010	17.3445	17.9651	18.7976
200	10	3.2251	**1.8566**	7.7747	5.7208	8.7534	7.7362
	20	6.4149	**4.1875**	11.3141	8.2469	10.8207	11.3962
	30	9.3625	**6.6097**	14.1751	9.5665	12.1479	12.2570
	40	11.3769	**9.3209**	17.2734	11.3919	14.2646	15.1533
500	10	1.2412	**0.5629**	3.9471	1.4473	4.7114	4.9135
	20	2.8403	**1.3680**	6.7615	2.7573	6.0749	5.8981
	30	4.4444	**2.2932**	8.8240	4.1184	5.8422	5.8068
	40	6.5146	**3.7695**	10.4504	5.1197	6.5382	7.2261
1000	10	0.6087	**0.2527**	2.8179	0.6991	4.0415	3.9572
	20	1.3364	**0.6690**	4.9908	1.8289	4.0494	4.0136
	30	2.4895	**1.2404**	7.0393	2.7869	3.7180	3.7411
	40	3.9724	**2.2417**	8.7888	3.5876	3.8731	4.1157

หมายเหตุ ตัวหนาขีดเส้นใต้ หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ผลลัพธ์ที่กำหนดไว้ดัง**ตารางที่ 4.1** ระบุค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MCAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 ให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ กรณีขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี RF ให้ค่า EMSE ต่ำที่สุดที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% แต่เมื่อเปอร์เซ็นต์การสูญหายระดับ 40% พบว่า วิธี KNN ให้ค่า EMSE ต่ำที่สุด และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

ตารางที่ 4.2 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MAR สำหรับข้อมูลจำลอง

Sample size	Percent of missing data	Imputation methods					
		KNN	RF	MEAN	MI	SRI	BRI
20	10	18.7952	**15.2368**	28.8982	27.0524	27.6328	27.5901
	20	43.8965	**37.5208**	53.9992	54.0936	52.5584	53.6018
	30	47.7754	**46.5357**	57.4059	54.3205	51.4940	53.2179
	40	**49.0639**	50.7413	55.6581	52.3826	50.5957	54.7931
50	10	38.1548	**23.7419**	62.3981	55.3606	49.4996	55.1880
	20	51.2667	**40.7564**	59.3836	57.3272	54.2419	51.3839
	30	54.8953	59.4692	58.7250	62.5505	**53.8829**	56.1250
	40	57.1284	67.4308	54.9728	65.7695	54.3825	**53.0079**
100	10	11.7494	**4.7266**	24.3449	22.6977	23.2914	20.9613
	20	19.6381	**14.2282**	27.1495	26.8689	24.5959	26.7726
	30	29.1441	**20.3513**	29.0366	30.2585	30.8276	29.5962
	40	**28.9198**	29.9500	29.7600	31.4140	29.3700	30.4777
150	10	6.3521	**2.6211**	13.1986	12.0279	14.2599	14.1335
	20	9.8149	**7.0891**	14.5959	14.2905	15.3269	14.8952
	30	14.7641	**9.8820**	20.2289	17.3501	17.7943	18.0733
	40	19.0537	**17.5198**	21.9783	20.8537	20.6663	20.4178
200	10	3.6471	**1.7436**	7.2631	7.3248	8.3804	9.5107
	20	7.1236	**3.6389**	10.8924	8.2083	10.8845	11.7971
	30	9.9043	**7.6237**	14.0613	10.3610	12.4227	12.8825
	40	13.0364	**10.8108**	17.1177	12.7730	14.7210	15.3003
500	10	1.3214	**0.4918**	3.3937	1.4092	4.5123	4.7623
	20	2.7116	**1.3131**	6.0363	2.6069	5.4169	5.3499
	30	4.6640	**2.4311**	8.9755	4.0063	5.8207	5.8094
	40	7.3555	**4.4660**	11.1641	5.4559	7.2836	7.6494
1000	10	0.6089	**0.2528**	2.8196	0.7000	4.0591	3.9601
	20	1.3383	**0.6699**	4.9986	1.8319	4.0498	4.0141
	30	2.3143	**1.2576**	6.8773	2.6548	3.6586	3.6928
	40	4.6559	**2.2455**	9.1584	3.5185	4.1853	4.1807

หมายเหตุ ตัวหนาขีดเส้นใต้ หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ผลลัพธ์ที่กำหนดไว้ดัง**ตารางที่ 4.2** ระบุค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 และ 50 ให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ กรณีขนาดตัวอย่างเท่ากับ 100 วิธี RF ให้ค่า EMSE ต่ำที่สุดที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% แต่เมื่อเปอร์เซ็นต์การสูญหายระดับ 40% พบว่า วิธี KNN ให้ค่า EMSE ต่ำที่สุด และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

ตารางที่ 4.3 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MNAR สำหรับข้อมูลจำลอง

Sample size	Percent of missing data	Imputation methods					
		KNN	RF	MEAN	MI	SRI	BRI
20	10	30.4115	**23.5956**	45.0423	40.1513	39.8826	38.5856
	20	43.4004	**40.0082**	49.5811	46.5454	46.0165	44.3187
	30	**46.8688**	48.1851	53.0948	50.9659	49.6905	46.4016
	40	**49.5928**	54.0385	54.9641	53.4350	51.3271	50.5633
50	10	39.7437	**24.9551**	51.3273	45.8444	47.8821	44.0631
	20	50.8431	**38.8564**	52.4758	57.3129	56.0780	53.4832
	30	53.2794	**47.6151**	58.5351	58.8863	56.0243	55.8453
	40	62.7924	74.3806	63.9437	71.2654	**59.1911**	62.8359
100	10	16.7178	**6.7997**	25.6581	27.5058	27.0861	26.1148
	20	22.3454	**12.9727**	27.6822	26.8885	25.5714	28.1499
	30	28.3361	**23.7718**	31.2251	31.3516	28.2189	29.3928
	40	**29.9636**	31.7309	32.9629	30.5474	31.9060	30.3313
150	10	7.9016	**2.3055**	12.3153	12.7527	13.8179	12.4018
	20	10.5149	**5.0916**	15.8391	13.8246	14.9143	15.9368
	30	12.9308	**9.7000**	19.1751	15.9949	16.8595	18.1101
	40	15.6077	**15.4223**	20.3321	19.2694	18.4300	18.1700
200	10	3.3703	**1.1990**	8.4543	7.7112	9.2146	9.5024
	20	5.8584	**3.0795**	11.0765	7.5056	10.7692	10.9667
	30	8.8357	**5.7662**	13.6558	9.1957	11.5676	11.7631
	40	11.8552	**9.3823**	16.6508	12.2990	13.4326	14.3862
500	10	0.9810	**0.4278**	3.2943	1.1632	4.2109	4.0283
	20	2.4021	**1.0299**	6.4871	2.4096	4.8893	5.0114
	30	3.8118	**1.9546**	8.3650	3.3145	4.7881	4.9159
	40	6.4030	**3.9154**	10.8168	4.4675	7.1657	7.3207
1000	10	0.4828	**0.1829**	2.3484	0.6060	3.4012	3.4870
	20	1.2448	**0.4812**	4.9152	1.6313	2.8883	2.7817
	30	2.5035	**1.0752**	7.1803	2.3412	3.2087	2.8612
	40	4.1130	**2.0263**	9.2254	2.9977	4.4062	4.6635

หมายเหตุ ตัวหนาขีดเส้นใต้ หมายถึง ค่า EMSE ต่ำที่สุดในสถานการณ์นั้น

ผลลัพธ์ที่กำหนดไว้ดัง**ตารางที่ 4.3** ระบุค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MNAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 50 และ 100 ที่ระดับเปอร์เซ็นต์การสูญหายต่ำ (10% 20% และ 30%) วิธี RF ให้ค่า EMSE ต่ำที่สุด แต่เมื่อระดับเปอร์เซ็นต์การสูญหายที่สูง (40%) จะให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

ตารางที่ 4.4 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MCAR สำหรับข้อมูลวัดการเป็นโรคเบาหวาน

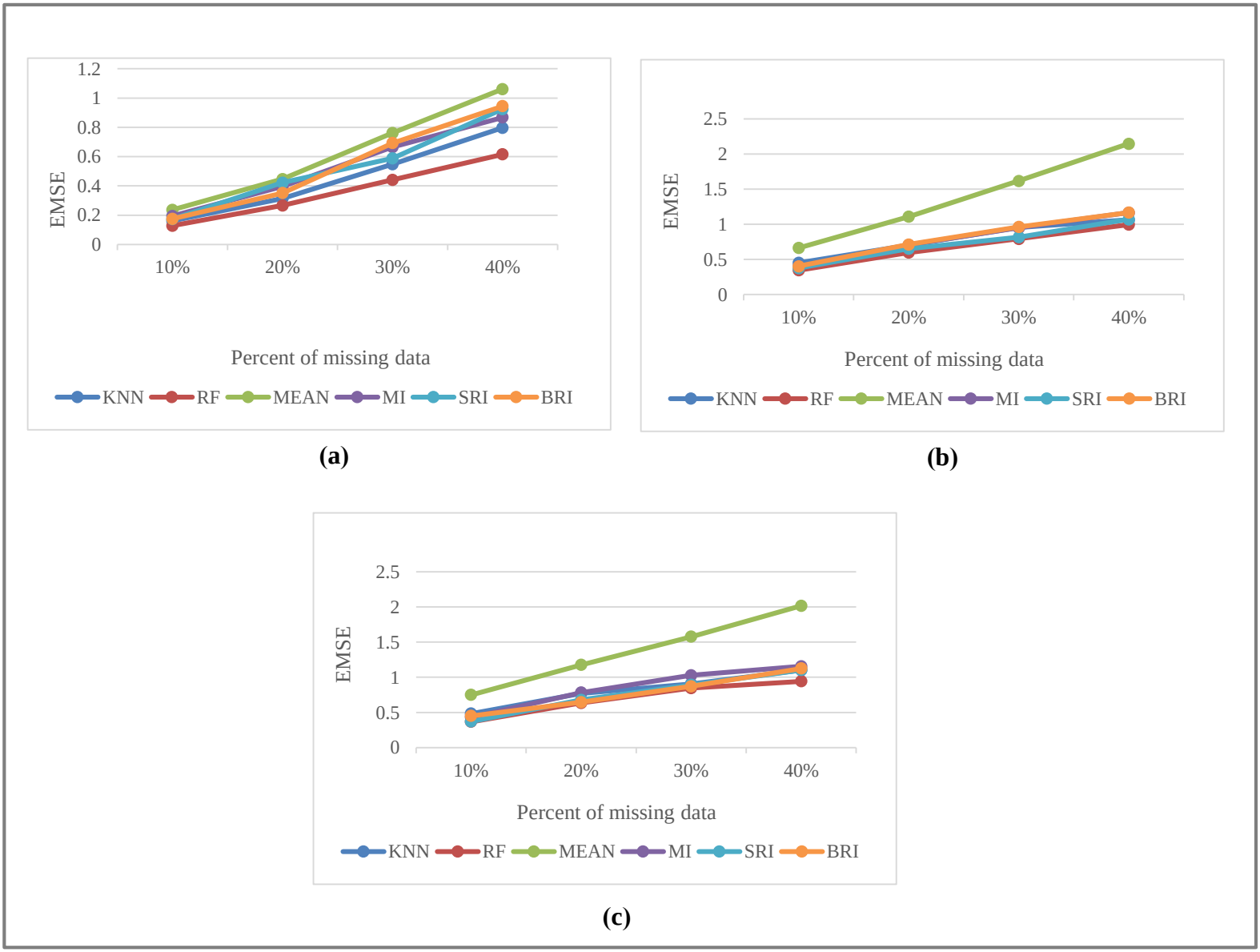
Percent of missing data	Imputation methods					
	KNN	RF	MEAN	MI	SRI	BRI
10	0.1611	**0.1276**	0.2361	0.1948	0.1734	0.1751
20	0.3145	**0.2666**	0.4462	0.3961	0.4221	0.3511
30	0.5483	**0.4409**	0.7619	0.6655	0.5887	0.6928
40	0.7968	**0.6158**	1.0610	0.8675	0.9242	0.9447

ตารางที่ 4.5 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MAR สำหรับข้อมูลวัดการเป็นโรคเบาหวาน

Percent of missing data	Imputation methods					
	KNN	RF	MEAN	MI	SRI	BRI
10	0.4523	**0.3471**	0.6644	0.4026	0.3783	0.4042
20	0.7041	**0.5984**	1.1086	0.7061	0.6542	0.7132
30	0.9571	**0.7940**	1.6179	0.9506	0.8170	0.9630
40	1.0634	**0.9966**	2.1473	1.1666	1.0721	1.1678

ตารางที่ 4.6 ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในรูปแบบการสูญหาย MNAR สำหรับข้อมูลวัดการเป็นโรคเบาหวาน

Percent of missing data	Imputation methods					
	KNN	RF	MEAN	MI	SRI	BRI
10	0.4835	0.3670	0.7500	0.4314	0.3701	0.4493
20	0.7697	0.6346	1.1767	0.7813	0.6775	0.6449
30	0.9063	0.8463	1.5766	1.0284	0.8915	0.8714
40	1.1012	0.9418	2.0164	1.1562	1.1034	1.1244



รูปที่ 4.1 การเปรียบเทียบค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี สำหรับข้อมูลวัดการเป็นโรคเบาหวาน ในรูปแบบการสูญหาย (a) MCAR, (b) MAR และ (c) MNAR

ผลลัพธ์ที่กำหนดไว้ดัง**ตารางที่ 4.4, 4.5, 4.6 และรูปที่ 4.1** แสดงให้เห็นว่า ค่า EMSE ของวิธีการประมาณค่าสูญหาย 6 วิธี ในชุดข้อมูลวัดการเป็นโรคเบาหวาน พบว่าทุกระดับเปอร์เซ็นต์การสูญหาย วิธี RF ให้ค่า EMSE น้อยที่สุด และวิธี MEAN ให้ค่า EMSE สูงที่สุด ซึ่งผลที่ได้สอดคล้องกับผลการวิจัยชุดข้อมูลจำลอง นั่นคือ เมื่อขนาดตัวอย่างตั้งแต่ 150 ขึ้นไป วิธี RF มีแนวโน้มให้ค่า EMSE ต่ำที่สุด หรือจัดว่าเป็นวิธีที่มีประสิทธิภาพมากที่สุด นอกจากนี้พบว่า เมื่อระดับเปอร์เซ็นต์การสูญหายที่เพิ่มขึ้น ค่า EMSE มีแนวโน้มที่จะเพิ่มขึ้นด้วย

บทที่ 5
สรุปและวิจารณ์ผลการศึกษา

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระที่มีการสูญหายสำหรับการวิเคราะห์การถดถอยลอจิสติกทวิภาค ที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) เมื่อมีตัวแปรอิสระ 3 ตัว ที่ตัวแปรอิสระที่มีการสูญหาย 3 รูปแบบได้แก่ MCAR, MAR และ MNAR โดยมีวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ คือ วิธี MEAN, MI, วิธี KNN, RF, SRI และ BRI ข้อมูลที่ใช้ในการศึกษามี 2 ชุดข้อมูลได้แก่ ชุดข้อมูลจำลอง (Simulation data) และชุดข้อมูลจริง (Real life data) ที่ใช้ในการวิจัยเป็นข้อมูลจาก www.kaggle.com ซึ่งประกอบไปด้วยตัวแปรอิสระ 3 ตัว ที่มีอิทธิพลต่อการเป็นโรคเบาหวาน ซึ่งในข้อมูลจำลองจะกำหนดขนาดตัวอย่าง 20, 50, 100, 150, 200, 500 และ 1000 ตัวแปรอิสระที่เกิดการสูญหายมีเปอร์เซ็นต์การสูญหาย 4 ระดับ คือ 10%, 20%, 30% และ 40% ใช้ค่า EMSE ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย ซึ่งหัวข้อที่จะกล่าวในบทนี้ มีดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 6 วิธี ในรูปแบบการสูญหาย MCAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 ให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ กรณีขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี RF ให้ค่า EMSE ต่ำที่สุดที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% แต่เมื่อเปอร์เซ็นต์การสูญหายระดับ 40% พบว่า วิธี KNN ให้ค่า EMSE ต่ำที่สุด และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

สำหรับรูปแบบการสูญหาย MAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 และ 50 ให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ กรณีขนาดตัวอย่างเท่ากับ 100 วิธี RF ให้ค่า EMSE ต่ำที่สุดที่เปอร์เซ็นต์การสูญหายระดับ 10%, 20% และ 30% แต่เมื่อเปอร์เซ็นต์การสูญหายระดับ 40% พบว่า วิธี KNN ให้ค่า EMSE

ต่ำที่สุด และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

สำหรับรูปแบบการสูญหาย MNAR แสดงให้เห็นว่า กรณีที่ขนาดตัวอย่างเท่ากับ 20 50 และ 100 ที่ระดับเปอร์เซ็นต์การสูญหายต่ำ (10% 20% และ 30%) วิธี RF ให้ค่า EMSE ต่ำที่สุด แต่เมื่อระดับเปอร์เซ็นต์การสูญหายที่สูง (40%) จะให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ และกรณีที่ขนาดตัวอย่างเท่ากับ 150, 200, 500 และ 1000 ที่ทุกระดับเปอร์เซ็นต์การสูญหาย พบว่า วิธี RF ให้ค่า EMSE ต่ำที่สุดในทุกสถานการณ์

สำหรับการเปรียบเทียบวิธีการประมาณค่าสูญหายทั้ง 6 วิธีของข้อมูลจริง (Real life data) ในชุดข้อมูลวัดการเป็นโรคเบาหวาน พบว่าทุกระดับเปอร์เซ็นต์การสูญหาย วิธี RF ให้ค่า EMSE น้อยที่สุด และวิธี MEAN ให้ค่า EMSE สูงที่สุด ซึ่งผลที่ได้สอดคล้องกับผลการวิจัยชุดข้อมูลจำลอง นั่นคือ เมื่อขนาดตัวอย่างตั้งแต่ 150 ขึ้นไป วิธี RF มีแนวโน้มให้ค่า EMSE ต่ำที่สุด หรือจัดว่าเป็นวิธีที่มีประสิทธิภาพมากที่สุด นอกจากนี้พบว่า เมื่อระดับเปอร์เซ็นต์การสูญหายที่เพิ่มขึ้น ค่า EMSE มีแนวโน้มที่จะเพิ่มขึ้นด้วย

ซึ่งผลลัพธ์ในข้อมูลจริงและข้อมูลจำลองไปในทิศทางเดียวกัน นั้นคือ เมื่อข้อมูลมีขนาดใหญ่ในทุก ๆ เปอร์เซ็นต์การสูญหายและมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน พบว่าการเลือกใช้วิธี RF จะให้ประสิทธิภาพดีที่สุดในทุกสถานการณ์ ดังนั้นหากจะเลือกวิธีการแทนค่าสูญหายสำหรับข้อมูลขนาดใหญ่ควรเลือกใช้วิธี RF

นอกจากนี้ จะเห็นได้ว่า ปัจจัยที่ส่งผลต่อประสิทธิภาพในการประมาณค่าสูญหายทั้ง 6 วิธี ได้แก่ เปอร์เซ็นต์การสูญหายและขนาดตัวอย่างของข้อมูล ซึ่งผลการวิจัย พบว่า ในการสูญหายทุกรูปแบบที่ข้อมูลมีขนาดเล็ก ในทุกระดับเปอร์เซ็นต์การสูญหาย ให้ผลลัพธ์ของวิธีการแทนค่าที่ดีที่สุดที่ไม่แน่นอน ค่า EMSE ที่ต่ำที่สุดในแต่ละระดับเปอร์เซ็นต์การสูญหายไม่คงที่ แต่เมื่อข้อมูลมีขนาดใหญ่ในทุกระดับเปอร์เซ็นต์การสูญหาย วิธี RF มีแนวโน้มให้ค่า EMSE ต่ำที่สุด หรือจัดว่าเป็นวิธีที่มีประสิทธิภาพมากที่สุด นอกจากนี้พบว่า เมื่อระดับเปอร์เซ็นต์การสูญหายที่เพิ่มขึ้น ค่า EMSE มีแนวโน้มที่จะเพิ่มขึ้นด้วย

5.2 อภิปรายผลการวิจัย

จากการศึกษางานวิจัยที่ผ่านๆมา พบว่ายังไม่มีงานวิจัยที่ทดลองการแทนค่าสูญหายในชุดข้อมูลการทำนายการเป็นโรคเบาหวานด้วยวิธีการเหล่านี้ ดังนั้นผู้วิจัยจึงมีความสนใจในการศึกษาข้อมูลนี้ เพื่อเป็นประโยชน์ในการวินิจฉัยโรคและตัดสินใจในการรักษาผู้ป่วย และ จากผลการวิจัย

พบว่า วิธี RF จะมีประสิทธิภาพดีที่สุดในข้อมูลที่มีขนาดใหญ่ในทุกระดับเปอร์เซ็นต์การสูญหาย ซึ่งสอดคล้องกับงานวิจัยของ Kokla et al. (2019) ที่กล่าวว่า วิธี RF จะมีประสิทธิภาพมากที่สุดเมื่อข้อมูลมีขนาดใหญ่ขึ้น อีกทั้งยังสอดคล้องกับงานวิจัยของ Hong and Lynn (2020) ซึ่งใช้ข้อมูลจำลองทางคลินิกในการศึกษาผู้ป่วยมะเร็งตับในรูปแบบการสูญหายทั้ง 3 รูปแบบ (MCAR, MAR, MNAR) กล่าวว่า วิธี RF มีประสิทธิภาพมากที่สุดเมื่อระดับเปอร์เซ็นต์การสูญหายเพิ่มมากขึ้น และยังสอดคล้องกับงานวิจัยของ Epp-Stobbe et al. (2022) ที่กล่าวว่า วิธี RF จะให้ประสิทธิภาพที่ดีที่สุดเช่นกัน กล่าวคือ ระดับเปอร์เซ็นต์การสูญหายของข้อมูลเพิ่มขึ้น จะส่งผลให้ค่า EMSE เพิ่มขึ้น และขนาดตัวอย่างเพิ่มขึ้น จะส่งผลให้ค่า EMSE ลดลง แต่ในงานวิจัยเหล่านี้ไม่ได้ศึกษาในชุดข้อมูลในการทำนายการเป็นโรคเบาหวาน ดังนั้นในการศึกษางานวิจัยนี้อาจช่วยตัดสินเลือกวิธีการในการแทนค่าสูญหายให้มีความน่าเชื่อถือมากขึ้นทั้งในข้อมูลที่เกี่ยวข้องทางการแพทย์และสาขาอื่น ๆ ได้

ดังนั้นการเลือกวิธีการแทนค่าสูญหายจะส่งผลกระทบอย่างมากต่อผลลัพธ์ของเรา โดยเฉพาะอย่างยิ่งเมื่อเป็นข้อมูลทางการแพทย์ เพราะทำให้ประสิทธิภาพในการประเมินผู้ป่วยลดลงและอาจส่งผลต่อชีวิตของผู้ป่วยได้

จากสถานการณ์จริงเมื่อปัจจัยในการทดสอบที่มีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) มีการสูญหายของข้อมูลในทุกระดับเปอร์เซ็นต์การสูญหาย ก็พบว่า วิธี RF จะให้ประสิทธิภาพมากที่สุดในทุกสถานการณ์ ซึ่งก็แสดงถึงการประมาณค่าสัมประสิทธิ์การถดถอยลอจิสติกใกล้เคียงกับความเป็นจริงมากที่สุด ส่งผลให้วิธี RF มีความน่าเชื่อถือในการแทนค่าการสูญหายของข้อมูลมากที่สุด

5.3 ข้อเสนอแนะ

5.3.1 ด้านการนำไปใช้ประโยชน์

ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 6 วิธี ที่มีรูปแบบการสูญหาย MCAR, MAR และ MNAR สามารถสรุปได้ว่า วิธี RF มีประสิทธิภาพดีที่สุดในกรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุทั้งในชุดข้อมูลจำลองและชุดข้อมูลจริง ดังนั้นผู้วิจัยสามารถนำวิธีการประมาณค่าสูญหายไปใช้กับข้อมูลวิจัยในการวิเคราะห์การถดถอยลอจิสติกทวิภาค โดยในการทดลองได้จำลองสถานการณ์ที่แตกต่างกัน เพื่อวิเคราะห์ว่าวิธีการใดเหมาะสมกับสถานการณ์ใด สามารถสรุปได้ว่า ที่การสูญหายบนตัวแปรอิสระเมื่อขนาดตัวอย่างมากกว่า 150 ควรเลือกใช้วิธี RF แต่ถ้าขนาดตัวอย่างน้อยกว่า 150 ควรเลือกใช้วิธี KNN

5.3.2 ด้านการวิจัยครั้งต่อไป

1. ในการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ในการศึกษาครั้งต่อไปอาจจะศึกษาในขอบเขตงานวิจัยที่หลากหลายขึ้น เช่น การเพิ่มจำนวนตัวแปรอิสระและกรณีที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) มากกว่า 1 คู่

2. ในการศึกษาครั้งต่อไปอาจจะศึกษาและเปรียบเทียบรูปแบบการสูญหายแบบสุ่มทั้งบนตัวแปรอิสระและตัวแปรตามที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งกันและกัน

3. ในการศึกษาครั้งนี้สูญหายเกิดขึ้นที่ตัวแปรอิสระที่เป็นข้อมูลเชิงปริมาณ ควรทำการศึกษาเพิ่มเติมกรณีที่ข้อมูลที่สูญหายเป็นข้อมูลเชิงคุณภาพ หรือกรณีที่ตัวแปรอิสระมีทั้งข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ

4. ในการศึกษาครั้งต่อไปอาจพิจารณาการสูญหายที่เกิดขึ้นเฉพาะตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ที่รูปแบบการสูญหายทั้ง 3 รูปแบบ

5. ในการศึกษาครั้งต่อไปอาจพิจารณาขนาดตัวอย่างที่ใหญ่มากขึ้นเพื่อให้สามารถเปรียบเทียบวิธีการแทนค่าสูญหายได้แม่นยำและน่าเชื่อถือมากยิ่งขึ้น

6. ในกรณีของข้อมูลจำลองอาจมีการศึกษาเปรียบเทียบเพิ่มเติมหากข้อมูลไม่ได้มีการแจกแจงปรกติ หรืออาจทำการศึกษากรณีที่ตัวแปรอิสระมีการแจกแจงแบบอื่นๆ ซึ่งอาจจะทำให้ผลลัพธ์ที่ได้แตกต่างจากผลการวิเคราะห์ในครั้งนี้

บรรณานุกรม

กาญจน์เขจร ชูชีพ. (2561). *Remote Sensing Technical*. มหาวิทยาลัยเกษตรศาสตร์. จาก https://forestadmin.forest.ku.ac.th/304xxx/?q=system/files/book/5%282018%29%20Logistic%20Regression.pdf

พัชนา สุวรรณแสน. (2561). การจัดการข้อมูลสูญหาย: วิธีเคเนียร์เรสเนเบอร์ Management Approach of Missing Data: K- Nearest Neighbor Imputation. *วารสารวิจัยวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา, 4*(1), 1-9.

โรงพยาบาลศิริราช ปิยมหาราชการุณย์. (2553). ศูนย์กลางการแพทย์ ที่ให้บริการดูแลรักษาคำปรึกษาโรคเฉพาะทางด้วยมาตรฐานระดับสากล JCI (Joint Commision International) สืบค้น 23 พฤศจิกายน 2565, จาก www.siphhospital.com

สุฤดี โกศัยเนตร. (2549). MULTICOLINEARITY: EXAMPLES IN BINARY LOGISTIC REGRESSION. *DMBN E Journal, 2*(1), 9-17.

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.

Boateng, E.Y., & Abaye, D.A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing,* 7, 190-207.

Dong, Y., & Peng, C.Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), Article 222: 365-366.

Enderers, C.K. (2022). *Applied Missing Data Analysis* (2^{nd} ed.). Guilford Publications.

Epp-Stobbe, A., Tsai, M.C., & Klimstra, M. (2022). Comparison of Imputation Methods for Missing Rate of Perceived Exertion Data in Rugby. *Machine Learning and Knowledge Extraction*, *4*(4), 827-838.

Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology,* 60, 549-576.

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. *BMS Medical Research Methodology*, *20*(1), Article 199: 1-12.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33, Article 10: 913-933.

Jösson, P., & Wohin, C. (2006). Benchmarking K Nearest neighbor imputation with Homogeneous Likert Data. *Empirical Software Engineering*, 11, 463-489.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC bioinformatics*, 20, Article 492: 1-11.

Little, R.J., & Rubin, D.B. (2019). *Statistical Analysis with Missing Data* (2^{nd} ed.). John Wiley & Sons.

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.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., & Lshii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, *19*(16), 2088-2096.

Roth, P.L. (1994). Missing Data: A conceptual Review for Applied Psychologists. *Personnel psychology*, *47*(3), 537-560.

Schober, P., & Vetter, T.R. (2021). Logistic Regression in Medical research. *Anesthesia & Analgesia Statistical Minute*, *132*(2), 365-366.

Shrive, F.M., Stuart, H., Quan, H. & Ghali, W.A. (2006). Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMS Medical Research Methodology*, 6, Article 57: 1-10.

Singhal, S. (2021). Defining, Analysing, and Implementing Imputation Techniques. Analyticsvidhya.com. Retrieved October 29, 2022, from https://www.analyticsvidhya.com/blog/author/shashank_singhal_1/

Stekhoven, D.J., & Bühlmann, P. (2012). MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Thongsri, T., & Samart, K. (2022a). Composite imputation method for the multiple linear regression with missing at random data. *International Journal of Mathematics and Computer Science*, *17*(1), 51-62.

Thongsri, T., & Samart, K. (2022b). Development of Imputation Methods for Missing Data in Multiple Linear Regression Analysis. *Lobachevskii Journal of Mathematics, 43*(11), 3390-3399.

Torabi, M., & Rao, J.N.K. (2013). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. *Journal of Multivariate Analysis*, 117, 76-87.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstien, D., & Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525.

Tsiampalis, T., & Panagiotakos, D.B. (2020). Missing analysis: socio-demographic, clinical and lifestyle determinants of low response rate on self-reported psychological and nutrition related multi-item instrument in the context of the ATTICA epidemiological study. *BMS Medical Research Methodology*, *20*(1), 1-13.

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. *BMS Medical Research Methodology*, 20, Article 42: 1-9.

ภาคผนวก

คำสั่งโปรแกรม R-Studio

ชุดคำสั่งที่ใช้ในการวิเคราะห์การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรอิสระในตัวแบบการถดถอยลอจิสติกที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุ

```
>set.seed(80)
>x1 <- rnorm(1000)
>x2 <- rnorm(1000)+x1*10
>x3 <- rnorm(1000)
>z = 1 + 1*x1 + 1*x2 + 1*x3
>pr = 1/(1+exp(-z))
>y <- rbinom(n = 1000, size = 1, prob = pr)
>mydata <- data.frame(y=y, x1=x1, x2=x2, x3=x3)
>SampleSize <- 20 # 20 50 100 150 200 #
>PercenMissing <- 0.1 # 0.1 0.2 0.3 0.4 #
>Repeat <- 1000
>Sum_Cal_KNN <- 0
>Sum_Cal_Forest <- 0
>Sum_Cal_MEAN <- 0
>Sum_Cal_MI <- 0
>Sum_Cal_SR <- 0
>Sum_Cal_BR <- 0

>i <- 1
>while (i <= Repeat) {
#---------------------------- Missing ----------------------------#
>Sample <- mydata[sample(nrow(mydata),SampleSize),]; #Sample
>SP <- data.frame(Sample[,1],Sample[,2],Sample[,3],Sample[,4])
>#install.packages("logistf")
>#library(logistf)
>full <- logistf(SP[,1]~SP[,2]+SP[,3]+SP[,4])
```

```
> full$coefficients
> beta0full <- full$coefficients[1]
>beta1full <- full$coefficients[2]
>beta2full <- full$coefficients[3]
>beta3full <- full$coefficients[4]
>Samplenoty <- Sample[-1]
>Missing <- ampute(Samplenoty,prop = PercenMissing,mech = "MCAR")
>MissingData <- Missing$amp; #MissingData
>Missinghasy <- data.frame(Sample[1],MissingData)

#---------------------------- Method ----------------------------#
#------------------------------ KNN ----------------------------#

>#install.packages("VIM")
>#library(VIM)
>Sqrt_M <- round(sqrt(SampleSize-(PercenMissing*SampleSize)))
>KNN <- kNN(Missinghasy,variable = c("x1", "x2", "x3"),k=Sqrt_M)
>fit_KNN <- logistf(KNN[,1] ~ KNN[,2] + KNN[,3] + KNN[,4])
>fit_KNN$coefficients
>beta0fit_KNN <- fit_KNN$coefficients[1]
>beta1fit_KNN <- fit_KNN$coefficients[2]
>beta2fit_KNN <- fit_KNN$coefficients[3]
>beta3fit_KNN <- fit_KNN$coefficients[4]
>Cal_KNN  <-  (((beta0full-beta0fit_KNN)^2)+((beta1full-beta1fit_KNN)^2)+((beta2full-
beta2fit_KNN)^2)+((beta3full-beta3fit_KNN)^2))
>Sum_Cal_KNN <- Sum_Cal_KNN + Cal_KNN
```

```
#---------------------------- Method -------------------------------#
#---------------------------- MissForest ---------------------------#

>#install.packages("missForest")
>#library(missForest)

>MissForest <- missForest(Missinghasy)
>ComForest <-  MissForest$ximp
>fit_Forest <- logistf(ComForest[,1] ~ ComForest[,2] + ComForest[,3] + ComForest[,4])
>fit_Forest$coefficients
>beta0fit_Forest <- fit_Forest$coefficients[1]
>beta1fit_Forest <- fit_Forest$coefficients[2]
>beta2fit_Forest <- fit_Forest$coefficients[3]
>beta3fit_Forest <- fit_Forest$coefficients[4]
>Cal_Forest<-(((beta0full-beta0fit_Forest)^2)+((beta1full-
beta1fit_Forest)^2)+((beta2full-beta2fit_Forest)^2)+((beta3full-beta3fit_Forest)^2))
>Sum_Cal_Forest <- Sum_Cal_Forest + Cal_Forest

#---------------------------- Method -------------------------------#
#---------------------------- Mean ---------------------------------#
>#install.packages("missMethods")
>#library(missMethods)
>ComMEAN <- impute_mean(Missinghasy)
>fit_MEAN <- logistf(ComMEAN[,1] ~ ComMEAN[,2] + ComMEAN[,3] + ComMEAN[,4])
>fit_MEAN$coefficients
beta0fit_MEAN <- fit_MEAN$coefficients[1]
beta1fit_MEAN <- fit_MEAN$coefficients[2]
beta2fit_MEAN <- fit_MEAN$coefficients[3]
beta3fit_MEAN <- fit_MEAN$coefficients[4]
```

```
>Cal_MEAN<-(((beta0full-beta0fit_MEAN)^2)+((beta1full-
beta1fit_MEAN)^2)+((beta2full-beta2fit_MEAN)^2)+((beta3full-beta3fit_MEAN)^2))
>Sum_Cal_MEAN <- Sum_Cal_MEAN + Cal_MEAN

#---------------------------- Method -----------------------------#
#------------------------------ MI -------------------------------#
>#install.packages("mice")
>#library(mice)
>imp <- mice(Missinghasy)
>Mi <- complete(imp)
>fit_MI <- logistf(Mi[,1] ~ Mi[,2] + Mi[,3] + Mi[,4])
>beta0fit_MI <- fit_MI$coefficients[1]
>beta1fit_MI <- fit_MI$coefficients[2]
>beta2fit_MI <- fit_MI$coefficients[3]
>beta3fit_MI <- fit_MI$coefficients[4]
>Cal_MI <- (((beta0full-beta0fit_MI)^2)+((beta1full-beta1fit_MI)^2)+((beta2full-
beta2fit_MI)^2)+((beta3full-beta3fit_MI)^2))
>Sum_Cal_MI <- Sum_Cal_MI + Cal_MI

#---------------------------- Method -----------------------------#
#------------------------------ SRI ------------------------------#

>imp_SR <- mice( Missinghasy , method = "norm.nob")
>SR <- complete(imp_SR)
>fit_SR <- logistf(SR[,1] ~ SR[,2] + SR[,3] + SR[,4])
>beta0fit_SR <- fit_SR$coefficients[1]
>beta1fit_SR <- fit_SR$coefficients[2]
>beta2fit_SR <- fit_SR$coefficients[3]
>beta3fit_SR <- fit_SR$coefficients[4]
```

```
>Cal_SR <- (((beta0full-beta0fit_SR)^2)+((beta1full-beta1fit_SR)^2)+((beta2full-beta2fit_SR)^2)+((beta3full-beta3fit_SR)^2))
>Sum_Cal_SR <- Sum_Cal_SR + Cal_SR

#---------------------------- Method -----------------------------#
#------------------------------ BRI ------------------------------#

>imp_bay <- mice( Missinghasy , method = "norm")
>bay <- complete(imp_bay)
>fit_BR <- logistf(bay[,1] ~ bay[,2] + bay[,3] + bay[,4])
>beta0fit_BR <- fit_BR$coefficients[1]
>beta1fit_BR <- fit_BR$coefficients[2]
>beta2fit_BR <- fit_BR$coefficients[3]
>beta3fit_BR <- fit_BR$coefficients[4]
>Cal_BR <- (((beta0full-beta0fit_BR)^2)+((beta1full-beta1fit_BR)^2)+((beta2full-beta2fit_BR)^2)+((beta3full-beta3fit_BR)^2))
>Sum_Cal_BR <- Sum_Cal_BR + Cal_BR
#---------------------------- Repeat -----------------------------#

i = i+1
}

>EMSE_KNN <- Sum_Cal_KNN/Repeat;round(EMSE_KNN,4)
>EMSE_Forest <- Sum_Cal_Forest/Repeat;round(EMSE_Forest,4)
>EMSE_MEAN <- Sum_Cal_MEAN/Repeat;round(EMSE_MEAN,4)
>EMSE_MI <- Sum_Cal_MI/Repeat;round(EMSE_MI,4)
>EMSE_SR <- Sum_Cal_SR/Repeat;round(EMSE_SR,4)
>EMSE_BR <- Sum_Cal_BR/Repeat;round(EMSE_BR,4)
```

ประวัติผู้เขียน

ชื่อ สกุล นางสาวธัญพิชชา ฤทธิ์เทวา

รหัสประจำตัวนักศึกษา 6510220004

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วิทยาศาสตรบัณฑิต (สถิติ)	มหาวิทยาลัยสงขลานครินทร์	2564

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ทุนสนับสนุนบัณฑิตศึกษาจากกองทุนวิจัย คณะวิทยาศาสตร์ ประเภททุนตรี-โท สัญญาเลขที่ 1-2565-02-007