การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติก

นางสาวภัทฐิดา นิลภัทรฉัตร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

A COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING-DATA IN LOGISTIC REGRESSION ANALYSIS

Miss Pattida Nilpattarachat

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2016

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติก
โดย	นางสาวภัทฐิดา นิลภัทรฉัตร
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

....ประธานกรรมการ

(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี)

....กรรมการ

(อาจารย์ ดร. ณัตติฤดี เจริญรักษ์)

....กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร. มนต์ทิพย์ เทียนสุวรรณ)

ภัทฐิดา นิลภัทรฉัตร : การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์การถดถอยโลจิสติก (A COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING-DATA IN LOGISTIC REGRESSION ANALYSIS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. อนุภาพ สมบูรณ์สวัสดี, หน้า.

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระที่มีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม เมื่อมีตัวแปรอิสระ 3 ตัว และเกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง โดยวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ คือ วิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI) ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองข้อมูล โดยกำหนดขนาดตัวอย่าง 70, 100 และ 200 ตัวแปรอิสระที่เกิดการสูญหายมีร้อยละของการสูญหายโดยเฉลี่ย 3 ระดับ คือ ร้อยละ 10, 20 และ 30 มีระดับการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับ คือ การสูญหายแบบอิกนอร์เรเบิล การสูญหายแบบนอนอิกนอร์เรเบิลระดับปานกลาง และการสูญหายแบบนอนอิกนอร์เรเบิลระดับสูง และกำหนดค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระทั้ง 3 ตัว คือ 0.5, 1 และ 1.5 ตามลำดับ ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ และเกณฑ์ที่ใช้ในการเปรียบเทียบแต่ละวิธี คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) พบว่า i) วิธี MI จะมีประสิทธิภาพ เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายมีค่าต่ำ และขนาดตัวอย่างมีขนาดเล็ก ii) วิธี MEAN และ วิธี MED จะมีประสิทธิภาพ เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายมีค่าสูง และขนาดตัวอย่างมีขนาดใหญ่ iii) ค่า AMSE มีแนวโน้มเพิ่มขึ้น เมื่อร้อยละการสูญหายและสัดส่วนการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น iv) ค่า AMSE มีแนวโน้มเพิ่มขึ้น เมื่อค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหายมีค่าสูง

ภาควิชา สถิติ ลายมือชื่อนิสิต ....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ปีการศึกษา 2559

5781563326 : MAJOR STATISTICS
KEYWORDS: LOGISTIC REGRESSION / NONIGNORABLE MISSING-DATA / IMPUTATION METHODS

PATTIDA NILPATTARACHAT: A COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING-DATA IN LOGISTIC REGRESSION ANALYSIS. ADVISOR: ASST. PROF. ANUPAP SOMBOONSAVATDEE, Ph.D., pp.

The objective of this research is to compare the estimation methods for nonignorable missing-data of the independent variables in binary logistic regression models with three independent variables. The estimation methods considered in study are Mean Imputation (MEAN), Median Imputation (MED), K-Nearest Neighbor (KNN) and Multiple Imputation (MI). Data of this research are simulated with three sample sizes of 70, 100 and 200. Three levels of missing proportion of data are 10%, 20% and 30% and three levels of nonignorable-missingness of data are ignorable, middle nonignorable and high nonignorable. Coefficients of three independent variables in simulation are set to be 0.5, 1 and 1.5, respectively. The comparison of each methods using the average mean square error (AMSE), the findings are as follows: i) MI method yield higher performance when coefficients of independent variables are low and small sample sizes, ii) MEAN method and MED method yield higher performance when coefficients of independent variables are high and large sample sizes. Iii) the AMSE increase when proportion of nonignorable-missingness of data increase, iv) the AMSE increase when coefficients of independent variables increases.

Department:	Statistics	Student's Signature	....
Field of Study:	Statistics	Advisor's Signature	....
Academic Year:	2016		

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยความช่วยเหลือและความเอาใจใส่จากอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี ผู้วิจัยจึงขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่ได้สละเวลาให้คำปรึกษา ข้อคิดเห็น และให้คำแนะนำที่ดีเสมอมา ทั้งนี้ผู้วิจัยขอขอบพระคุณท่านคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. ณัตติฤดี เจริญรักษ์ กรรมการสอบวิทยานิพนธ์ และท่านคณะกรรมการภายนอก รองศาสตราจารย์ ดร. มนต์ทิพย์ เทียนสุวรรณ ที่กรุณาทำการตรวจสอบวิทยานิพนธ์ ให้คำแนะนำ รวมทั้งแนวคิดที่เป็นประโยชน์ต่อการปรับปรุงวิทยานิพนธ์

ผู้วิจัยขอขอบพระคุณคณาจารย์ และเจ้าหน้าที่ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ทุก ๆ ท่าน ที่ได้ให้ความช่วยเหลือในการทำวิทยานิพนธ์ให้สำเร็จลุล่วงมาได้

สุดท้ายนี้ ผู้วิจัยขอขอบพระคุณครอบครัวที่ช่วยส่งเสริม สนับสนุนและเป็นกำลังใจเสมอมาจนสำเร็จการศึกษา รวมทั้งขอขอบคุณเพื่อน ๆ ที่คอยให้กำลังใจมาโดยตลอด

สารบัญ

หน้า

หน้า

สารบัญตาราง

สารบัญรูปภาพ

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญ

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์การถดถอยแบบหนึ่งที่ตัวแปรตามเป็นข้อมูลเชิงคุณภาพ ส่วนตัวแปรอิสระเป็นได้ทั้งข้อมูลเชิงปริมาณและคุณภาพ โดยมีวัตถุประสงค์ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ พร้อมทั้งศึกษาระดับความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ และสามารถนำสมการถดถอยที่ได้ไปพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจศึกษา ซึ่งเทคนิคการวิเคราะห์การถดถอยโลจิสติกนี้เริ่มต้นได้ถูกนำไปใช้อย่างแพร่หลายในทางการแพทย์และสาธารณสุข ต่อมาจึงมีการนำไปใช้ในด้านสังคมศาสตร์ ด้านเศรษฐศาสตร์ ด้านธุรกิจ และอื่น ๆ

ในการวิเคราะห์การถดถอยโลจิสติกนั้นสามารถแบ่งประเภท ตามจำนวนกลุ่มของตัวแปรตามได้เป็น 2 แบบ คือ การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis) และการวิเคราะห์การถดถอยโลจิสติกแบบหลายกลุ่ม (Multinomial Logistic Regression Analysis)

ในงานวิจัยส่วนใหญ่ได้นำเอาเทคนิคการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม มาเป็นเครื่องมือในการวิเคราะห์ข้อมูล เพื่อพยากรณ์โอกาสที่จะเกิดเหตุการณ์ที่สนใจศึกษา ซึ่งการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่มนั้น ตัวแปรตามสามารถแบ่งได้เป็น 2 กลุ่ม คือ กลุ่มที่เกิดเหตุการณ์ที่สนใจ และกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ ตัวอย่างเช่น การป่วยเป็นโรคหัวใจ (เป็น ไม่เป็น) การไปใช้สิทธิ์ออกเสียงเลือกตั้ง (ไป ไม่ไป) หรือการจัดประเภทลูกหนี้ (ดี ไม่ดี) เป็นต้น

ในทางปฏิบัติการดำเนินการสำรวจและเก็บรวบรวมข้อมูลนั้น อาจเกิดความผิดพลาดในการบันทึกข้อมูลหรือผู้ให้ข้อมูลให้ข้อมูลไม่ครบถ้วนทำให้เกิดปัญหาข้อมูลสูญหาย (Missing Data) โดยเมื่อเกิดปัญหาข้อมูลสูญหายและไม่สามารถดำเนินการเก็บรวบรวมข้อมูลเพิ่มได้ เนื่องจากมีระยะเวลาในการดำเนินการที่จำกัด หรือต้องเสียค่าใช้จ่ายในการเก็บข้อมูลเพิ่มขึ้น ทั้งนี้หากใช้วิธีการตัดชุดข้อมูลที่สูญหายหรือชุดข้อมูลที่ไม่สมบูรณ์ทิ้ง จะส่งผลให้ขนาดตัวอย่างที่ใช้ในการวิเคราะห์ข้อมูลมีขนาดลดลง ซึ่งจะทำให้ความคลาดเคลื่อนของการพยากรณ์มีค่าสูงขึ้น และยังอาจจะทำให้สูญเสียรายละเอียดบางส่วนที่สำคัญไป ดังนั้นเพื่อลดความผิดพลาดที่เกิดจากปัญหาข้อมูลสูญหาย จึงมีการคิดค้นวิธีต่าง ๆ เพื่อนำมาใช้ในการประมาณค่าสูญหาย แต่การเลือกใช้วิธีการประมาณค่าสูญหายให้เหมาะสมกับลักษณะของข้อมูลนั้นจะช่วยให้การประมาณค่าสูญหายมีประสิทธิภาพมากขึ้นด้วย

ในการวิจัยครั้งนี้ผู้วิจัยได้ทำการศึกษาเกี่ยวกับวิธีการประมาณค่าสูญหายของตัวแปรอิสระในการวิเคราะห์การถดถอยโลจิสติก ซึ่งพบว่ามีวิธีการต่าง ๆ ที่นักสถิติได้นำมาใช้แก้ปัญหานี้

Wilks (1931) ได้เสนอ วิธี Mean Imputation ซึ่งเป็นวิธีการประมาณที่ไม่ซับซ้อนและเป็นวิธีที่เก่าแก่วิธีหนึ่งสำหรับตัวแปรอิสระที่เป็นแบบต่อเนื่อง โดยมีหลักการคือ ทำการประมาณค่าเฉลี่ยของข้อมูลที่ไม่ได้สูญหายแล้วนำมาแทนค่าให้กับข้อมูลที่สูญหาย และนำข้อมูลที่ได้มาประมาณค่าพารามิเตอร์ (อ้างถึงใน ประลองพล ประสงค์พร, 2551)

Hosmer, Lemeshow และ Sturdivant (2013: 395 - 401) ได้เสนอวิธี Multiple Imputation (MI) เพื่อใช้แก้ปัญหาข้อมูลสูญหายที่ตัวแปรอิสระที่เป็นแบบต่อเนื่องในการวิเคราะห์การถดถอยโลจิสติก

จากงานวิจัยของ ประลองพล ประสงค์พร (2551) ได้ทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ในตัวแบบการถดถอยโลจิสติก เมื่อมีค่าสูญหายเกิดขึ้นในตัวแปรอิสระ และการสูญหายเป็นแบบสุ่ม (Missing at Random) โดยมีการประมาณค่าสูญหายด้วยวิธี Mean Imputation (MEAN) วิธี Maximum Likelihood Estimation (MLE) วิธี Pseudo Maximum Likelihood Estimation (PMLE) และวิธี The Filling Method เมื่อมีค่าสูญหายเกิดขึ้นในกรณีที่มีตัวแปรอิสระ 2 ตัว และเกิดค่าสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง โดยการจำลองข้อมูลกำหนดขนาดตัวอย่าง 40, 70, 90, 100, 200 และ 400 ร้อยละการสูญหายของตัวแปรอิสระ คือ ร้อยละ 5, 10 และ 15 และค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ คือ 0, 0.1 และ 0.2 โดยกำหนดค่าพารามิเตอร์เริ่มต้นแบ่งเป็น 2 กรณี คือ $\beta_0,\ \beta_1,\ \beta_2 = 0.2$ และ $\beta_0,\ \beta_1 = 0.2$ และ $\beta_2 = 1$ พิจารณาเปรียบเทียบค่าเฉลี่ยของความแตกต่างระหว่างค่าจริงและค่าประมาณ (Bias) และระยะห่างมาฮาลาโนบิสเฉลี่ย (Average Mahalanobis Distance: AMH) เป็นเกณฑ์ประกอบในการตัดสินใจ ผลการศึกษาพบว่า ในกรณีขนาดตัวอย่าง น้อยกว่า 90 วิธี MEAN จะให้ค่า Bias และค่า AMH น้อยที่สุด แต่ในกรณีขนาดตัวอย่างมากกว่า 90 วิธี FILL จะให้ค่า Bias และค่า AMH น้อยที่สุด

ซึ่งจากงานวิจัยที่ศึกษาข้างต้นได้ศึกษากรณีที่ตัวแปรอิสระมีการสูญหายแบบสุ่ม แต่ในบางครั้งการสำรวจหรือการเก็บรวมรวบข้อมูล การสูญหายของข้อมูลนั้นอาจจะไม่ได้เกิดขึ้นอย่างสุ่ม (Not Missing at Random) นั้นคือ การสูญหายของข้อมูลขึ้นอยู่กับตัวแปรที่เกิดการสูญหาย หรืออาจขึ้นอยู่กับตัวแปรอื่นร่วมด้วย เรียกการสูญหายประเภทนี้ว่า ข้อมูลสูญหายแบบนอนอิกนอร์เรเบิล (Nonignorable Missing-Data) ตัวอย่างการสูญหาย เช่น การสูญหายของข้อมูลรายได้จะขึ้นอยู่กับระดับของรายได้ การสูญหายของข้อมูลหนี้สินจะขึ้นอยู่กับระดับของหนี้สิน หรือการสูญหายของข้อมูลจำนวนบุหรี่ที่สูบต่อสัปดาห์จะขึ้นอยู่กับพฤติกรรมการสูบบุหรี่ เป็นต้น ซึ่งการสูญหายในลักษณะนี้จะส่งผลกระทบต่อการวิเคราะห์ข้อมูลมากกว่าการสูญหายแบบสุ่ม

ดังนั้นในงานวิจัยนี้จึงสนใจทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis) กรณีที่มีตัวแปรอิสระ 3 ตัว และตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ การสูญหายเกิดขึ้นในตัวแปรอิสระตัวใดตัวหนึ่ง โดยเป็นการสูญหายแบบนอนอิกนอร์เรเบิล วิธีประมาณค่าสูญหายที่สนใจศึกษา คือ

วิธีที่ 1 วิธี Mean Imputation (MEAN)

วิธีที่ 2 วิธี Median Imputation (MED)

วิธีที่ 3 วิธี K-Nearest Neighbor Imputation (KNN)

วิธีที่ 4 วิธี Multiple Imputation (MI)

เมื่อประมาณค่าข้อมูลสูญหายแล้วจะทำการประมาณค่าสัมประสิทธิ์การถดถอยโลจิสติก โดยวิธีประมาณค่าสูญหายที่มีประสิทธิภาพมากที่สุดจะเป็นวิธีการที่ให้ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) น้อยที่สุด

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาวิธีการประมาณค่าสูญหายของตัวแปรอิสระเมื่อมีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่มเมื่อตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ

1.2.2 เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระ เมื่อมีการสูญหายแบบนอนอิกนอร์เรเบิลในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่มกรณีที่มีตัวแปรอิสระ 3 ตัว ภายใต้สถานการณ์ที่แตกต่างกัน โดยวิธีการประมาณค่าสูญหายที่สนใจศึกษา คือวิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI)

1.3 ข้อตกลงเบื้องต้น

การวิจัยครั้งนี้มีข้อตกลงเบื้องต้นสำหรับการดำเนินงานวิจัย ดังนี้

1. ตัวแปรตาม (Y) เป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1

2. ทำการศึกษากรณีที่มีตัวแปรอิสระ 3 ตัวแปร โดยตัวแปรอิสระ (X_1, X_2, X_3) เป็นตัวแปรเชิงปริมาณ และไม่มีความสัมพันธ์กัน

3. การศึกษาในครั้งนี้สนใจศึกษา กรณีที่ตัวแปรอิสระ และตัวแปรตาม มีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Model) โดยมีรูปแบบคือ

$$\boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \pi(\boldsymbol{X}) = \frac{\boldsymbol{e}^{\beta_0 + \beta_1 \boldsymbol{x}_{i1} + \beta_1 \boldsymbol{x}_{i2} + \beta_1 \boldsymbol{x}_{i3}}}{1 + \boldsymbol{e}^{\beta_0 + \beta_1 \boldsymbol{x}_{i1} + \beta_1 \boldsymbol{x}_{i2} + \beta_1 \boldsymbol{x}_{i3}}} \qquad ; \boldsymbol{i} = 1, 2, \ldots, \boldsymbol{m}, \boldsymbol{m}+1, \ldots, \boldsymbol{n}$$

เมื่อ $\pi(X)$ คือ ค่าเฉลี่ยแบบมีเงื่อนไขของ Y ที่กำหนดโดย X

X คือ เวกเตอร์ของค่าสังเกต $x_{i1},\ x_{i2},\ x_{i3}$

x_{i1} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 1

x_{i2} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 2

x_{i3} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 3

β_p คือ สัมประสิทธิ์การถดถอย เมื่อ p = 0, 1, 2, 3

n คือ จำนวนค่าสังเกตทั้งหมด

m คือ จำนวนค่าสังเกตที่ทราบค่า

$n-m$ คือ จำนวนค่าสังเกตที่สูญหาย

4. ความคลาดเคลื่อนเป็นอิสระกัน

5. การสูญหายของข้อมูลเกิดขึ้นที่ตัวแปรอิสระตัวใดตัวหนึ่งเท่านั้น และมีการสูญหายแบบนอนอิกนอร์เรเบิล โดยจะแบ่งช่วงของตัวแปรอิสระออกเป็น 3 ช่วง และกำหนดให้ร้อยละของการสูญหายในแต่ละช่วงแตกต่างกัน

1.4 คำจำกัดความที่ใช้ในการวิจัย

1.4.1 ช่วงต้น คือ ช่วงของพื้นที่ใต้โค้งปกติมาตรฐานที่อยู่ใน $(-\infty, z)$ เมื่อ z มีการแจกแจงปกติมาตรฐาน (N (0, 1)) ดังนั้น ในช่วงนี้จะมีพื้นที่เป็น $P(-\infty < Z < z)\times 100\%$ ของพื้นที่ทั้งหมด

ช่วงกลาง คือ ช่วงของพื้นที่ใต้โค้งปกติมาตรฐานที่อยู่ใน (z, z') เมื่อ z และ z' มีการแจกแจงปกติมาตรฐาน (N(0,1)) และ $z<z'$ ดังนั้น ในช่วงนี้จะมีพื้นที่เป็น $P\left(z < Z < z'\right)\times 100\%$ ของพื้นที่ทั้งหมด

ช่วงปลาย คือ ช่วงของพื้นที่ใต้โค้งปกติมาตรฐานที่อยู่ใน (z', ∞) เมื่อ z' มีการแจกแจงปกติมาตรฐาน (N(0,1)) ดังนั้น ในช่วงนี้จะมีพื้นที่เป็น $1-P\left(Z < z'\right)\times 100\%$ของพื้นที่ทั้งหมด

วิธีในการแบ่งช่วงของตัวแปรอิสระ คือ

ถ้า $x_{ip} \le \mu_{x_p} + z_{\frac{1}{3}}\sigma_{x_p}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงต้น

ถ้า $\mu_{x_p} + z_{\frac{1}{3}}\sigma_{x_p} < x_{ip} \le \mu_{x_p} + z_{\frac{2}{3}}\sigma_{x_p}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงกลาง

ถ้า $\mu_{x_p} + z_{\frac{2}{3}}\sigma_{x_p} < x_{ip}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงปลาย

1.4.2 ความน่าจะเป็นของการสูญหายในแต่ละช่วง คือ อัตราส่วนระหว่างจำนวนตัวอย่างที่สูญหายในช่วงนั้นกับจำนวนตัวอย่างทั้งหมดที่ตกอยู่ในช่วงนั้น

1.4.3 ร้อยละของการสูญหาย คือ (จำนวนตัวอย่างที่สูญหาย/จำนวนตัวอย่างทั้งหมด) × 100%

1.5 ขอบเขตของการวิจัย

การวิจัยครั้งนี้มีขอบเขตการวิจัย ดังนี้

1. ตัวแปรอิสระมีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งมีฟังก์ชันการแจกแจง คือ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \qquad ; \ -\infty < x < \infty$$

มีค่าเฉลี่ยเท่า μ และความแปรปรวนเท่ากับ σ^2 โดยจะศึกษาการสูญหายของข้อมูลในกรณีที่ตัวแปรอิสระ มีการแจกแจงแบบปกติมาตรฐาน (N(0,1))

2. กำหนดให้ตัวแปรอิสระทั้ง 3 ตัว ไม่มีความสัมพันธ์กัน นั้นคือ ค่าสหสัมพันธ์เท่ากับ 0

3. ตัวแปรตาม (Y) เป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1

4. ตัวแปรตาม (Y) มีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution)

โดย $$Y = \begin{cases} 1 & \text{เมื่อเกิดเหตุการณ์ที่สนใจ ด้วยความน่าจะเป็น } \pi(X) \\ 0 & \text{เมื่อไม่เกิดเหตุการณ์ที่สนใจ ด้วยความน่าจะเป็น } 1-\pi(X) \end{cases}$$

เขียนแทนด้วย $Y \sim Ber(\pi(X))$

5. ตัวแปรอิสระ และตัวแปรตาม มีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Model) ซึ่งปกติแล้ว β_0, β_1, β_2 และ β_3 เป็นพารามิเตอร์ที่ไม่ทราบค่า ในการวิจัยนี้กำหนดให้ค่าสัมประสิทธิ์การถดถอย $\beta_0 = 0$ และกำหนดให้ $\beta_1 = 0.5$, $\beta_2 = 1$ และ $\beta_3 = 1.5$ สามารถเขียนตัวแบบได้ดังนี้

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = 0.5x_{i1} + x_{i2} + 1.5x_{i3} \quad ; i = 1, 2, 3, \ldots, n$$

เนื่องจากต้องการเปรียบเทียบกรณีที่ตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามใน 3 ระดับ คือ น้อย $(\beta_1 = 0.5)$, ปานกลาง $(\beta_2 = 1)$ และสูง $(\beta_3 = 1.5)$

6. กำหนดขนาดตัวอย่าง (n) 3 ระดับ คือ 70, 100 และ 200 เนื่องจากต้องการเปรียบเทียบกรณีที่ข้อมูลมีขนาดแตกต่างกัน คือ ข้อมูลขนาดเล็ก (n=70) ข้อมูลขนาดกลาง (n=100) และข้อมูลขนาดใหญ่ (n=200) และเนื่องจากมีตัวแปรอิสระ 3 ตัว ทำให้ได้ว่าขนาดตัวอย่างที่เหมาะสม และสามารถสร้างตัวแบบการถดถอยโลจิสติกได้ ควรมีขนาดตัวอย่างอย่างน้อยเท่ากับ 70 จึงได้ว่าข้อมูลขนาดเล็กจึงเท่ากับ 70

7. การสูญหายแต่ละสถานการณ์เกิดขึ้นในตัวแปรอิสระตัวใดตัวหนึ่ง กล่าวคือ การสูญหายของตัวแปรอิสระทั้ง 3 ตัว เกิดขึ้นไม่พร้อมกัน

8. พื้นที่ใต้โค้งปกติของตัวแปรอิสระ จะถูกแบ่งเป็น 3 ช่วง โดยกำหนดให้มีการแบ่งช่วงเป็นอัตราส่วน ดังนี้

ช่วงต้น : ช่วงกลาง : ช่วงปลาย เป็น 1 : 1 : 1 ตามลำดับ

9. สัดส่วนการสูญหายของตัวแปรอิสระคิดเป็นร้อยละโดยเฉลี่ยของทั้ง 3 ช่วง คือ ร้อยละ 10, 20 และ 30

10. การสูญหายของข้อมูลเกิดขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง และเป็นการสูญหายแบบนอนอิกนอร์เรเบิล โดยกำหนดให้ช่วงของตัวแปรอิสระที่มีค่ามากจะมีสัดส่วนของการสูญหายมากกว่าช่วงของตัวแปรที่มีค่าน้อย ซึ่งระดับของการสูญหายแบบนอนอิกนอร์เรเบิล จะแบ่งเป็น 3 ระดับ คือ

1) Ignorable คือ มีสัดส่วนการสูญหายเท่ากันทั้ง 3 ช่วง
2) Moderate nonignorable คือ มีสัดส่วนการสูญหายแบบนอนอิกนอร์เรเบิล ระดับปานกลาง
3) High nonignorable คือ มีสัดส่วนการสูญหายแบบนอนอิกนอร์เรเบิล ระดับสูง

และกำหนดให้แต่ละช่วงของค่าตัวแปรอิสระมีร้อยละการสูญหายดังตารางที่ 1

11. การวิจัยครั้งนี้จะทำการจำลองข้อมูลให้มีสถานการณ์ที่แตกต่างกันตามข้อกำหนดข้างต้น โดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ

ตารางที่ 1 ร้อยละของการสูญหายแบบนอนอิกนอร์เรเบิล

ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	ร้อยละของการสูญหายในแต่ละช่วง		
		ช่วงต้น	**ช่วงกลาง**	**ช่วงปลาย**
10	Ignorable	10	10	10
	Moderate nonignorable	8	10	12
	High nonignorable	4	10	16
20	Ignorable	20	20	20
	Moderate nonignorable	16	20	24
	High nonignorable	8	20	32
30	Ignorable	30	30	30
	Moderate nonignorable	24	30	36
	High nonignorable	12	30	48

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการประมาณค่าสูญหายวิธีใดให้ค่าประมาณใกล้เคียงค่าจริงมากที่สุด โดยการคำนวณค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) และค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) ซึ่งวิธีการที่ให้ค่า AMSE ต่ำสุดจะเป็นวิธีการประมาณค่าสูญหายที่มีประสิทธิภาพกว่าวิธีอื่น ๆ คำนวณได้จากสูตรดังต่อไปนี้

$$MSE_p = \frac{\sum_{t=1}^{5,000}\left(\hat{\beta}_p^{(t)} - \beta_p\right)^2}{5,000} \qquad ; p = 0,1,2,3$$

$$AMSE = \frac{\sum_{p=0}^{3} MSE_p}{4} \qquad ; p = 0,1,2,3$$

และพิจารณาค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency: RE) ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีการประมาณค่าสูญหายแบบอื่น ๆ เพื่อใช้เปรียบเทียบประสิทธิภาพของแต่ละวิธีให้มีความชัดเจนมากขึ้น โดยการศึกษาครั้งนี้เลือกวิธี MEAN เป็นเกณฑ์ในการเปรียบเทียบ เนื่องจาก วิธี MEAN เป็นวิธีการประมาณค่าสูญหายที่เป็นวิธีพื้นฐาน ซึ่งวิธีการที่ให้ค่า RE มากกว่า 1 แสดงว่าวิธีการนั้นมีประสิทธิภาพมากกว่าวิธี MEAN สามารถคำนวณได้จากสูตรดังต่อไปนี้

$$RE(MSE_p) = \frac{MSE_{MEAN}}{MSE_r} \quad ; \; r = MED,\ KNN,\ MI \;\; , \;\; p = 3$$

$$RE(AMSE) = \frac{AMSE_{MEAN}}{AMSE_r} \quad ; \; r = MED,\ KNN,\ MI$$

เมื่อ $\hat{\beta}_p^{(t)}$ คือ ค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกตัวที่ p ในการจำลองรอบที่ t

β_p คือ ค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกตัวที่ p ที่กำหนดขึ้น

MSE_p คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าสัมประสิทธิ์การถดถอยตัวที่ p จากการทำซ้ำ 5,000 รอบ

MSE_{MEAN} คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MEAN

MSE_r คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MED หรือ วิธี KNN หรือวิธี MI

$AMSE$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

$AMSE_{MEAN}$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MEAN

$AMSE_r$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MED หรือ วิธี KNN หรือวิธี MI

$RE(MSE_p)$ คือ ค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าสัมประสิทธิ์การถดถอยตัวที่ p

$RE(AMSE)$ คือ ค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

1.7 วิธีการดำเนินการวิจัย

การวิจัยครั้งนี้วิธีการดำเนินงานวิจัย ดังนี้

1. สร้างข้อมูลตัวแปรอิสระ 3 ตัว ที่มีการแจกแจงปกติมาตรฐาน ตามขนาดที่กำหนด (n = 70, 100, 200)

2. สร้างชุดข้อมูลที่มีรูปแบบความสัมพันธ์ตามตัวแบบการถดถอยโลจิสติก โดยกำหนดให้ $\beta_0 = 0$ และให้ $\beta_1 = 0.5,\ \beta_2 = 1$ และ $\beta_3 = 1.5$ เขียนตัวแบบได้ดังนี้

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = 0.5x_{i1} + x_{i2} + 1.5x_{i3} \qquad ;\ i = 1, 2, 3, \ldots, n$$

3. สร้างข้อมูลตัวแปรตามที่มีการแจกแจงแบร์นูลลี $Y \sim Ber(\pi(X))$

4. ทำการสุ่มตำแหน่งที่ข้อมูลตัวแปรอิสระจะเกิดการสูญหาย โดยนำข้อมูลตัวแปรอิสระมาแบ่งเป็น 3 ช่วง โดยให้แต่ละช่วงมีสัดส่วนเท่า ๆ กัน และให้แต่ละช่วงมีสัดส่วนการสูญหายแตกต่างกัน โดยสร้างตัวแปรที่มีการแจกแจงแบร์นูลลี จำนวน 3 ช่วง ซึ่งมีขนาดเท่ากับจำนวนตัวแปรอิสระที่อยู่ในแต่ละช่วง และมีความน่าจะเป็นของการสูญหายแตกต่างกันตามที่กำหนดไว้ แล้วทำการจับคู่ข้อมูลตัวแปรอิสระในแต่ละช่วง

5. ประมาณค่าข้อมูลและแทนที่ข้อมูลที่สูญหายของตัวแปรอิสระ ด้วยวิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI)

6. ประมาณค่าพารามิเตอร์ใหม่ด้วยวิธีความควรจะเป็นสูงสุด (MLE)

7. สร้างสมการถดถอยโลจิสติกจากค่าพารามิเตอร์ใหม่ที่ได้ เพื่อใช้ในการพยากรณ์

8. คำนวณค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง และค่าประสิทธิภาพสัมพัทธ์ เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายทั้ง 4 วิธี

9. สรุปผลที่ได้จากการทดลอง

1.8 ประโยชน์ที่คาดว่าจะได้รับ

1.8.1 เพื่อเป็นแนวทางในการเลือกวิธีประมาณค่าสูญหายของตัวแปรอิสระเมื่อมีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม

1.8.2 เพื่อเป็นแนวทางในการศึกษา การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ข้อมูลที่มีความสัมพันธ์กันในรูปแบบอื่น ๆ หรือการสูญหายในสถานการณ์อื่น ๆ

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การวิจัยครั้งนี้ได้ศึกษาเกี่ยวกับการแก้ปัญหาข้อมูลสูญหายในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis) การสูญหายเกิดขึ้นในตัวแปรอิสระ โดยเป็นการสูญหายแบบนอนอิกนอร์เรเบิล ซึ่งการเลือกใช้เฉพาะชุดข้อมูลที่สมบูรณ์จะส่งผลให้ขนาดตัวอย่างลดลง และอาจจะทำให้การพยากรณ์เกิดความคลาดเคลื่อนสูงขึ้น เพื่อให้การวิเคราะห์ข้อมูลมีประสิทธิภาพ ผู้วิจัยจึงสนใจทำการศึกษาหาวิธีการประมาณค่าสูญหายที่เหมาะสม ดังนั้นการศึกษาในครั้งนี้จึงมีทฤษฎีต่าง ๆ ที่เกี่ยวข้อง โดยในบทนี้จะกล่าวถึง การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม และวิธีการประมาณค่าสูญหาย ซึ่งสามารถอธิบายรายละเอียดได้ ดังนี้

2.1 การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis)

การวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis) เป็นการวิเคราะห์หาความสัมพันธ์ของตัวแปรอิสระและตัวแปรตาม เพื่อพยากรณ์เรื่องที่สนใจศึกษา โดยที่ตัวแปรตาม (Y) มีค่าได้ 2 ค่า คือ 0 และ 1 เมื่อ Y = 1 แทน การเกิดเหตุการณ์ที่สนใจศึกษา และ Y = 0 แทนไม่เกิดเหตุการณ์ที่สนใจศึกษา นั่นคือ ตัวแปรตามมีการแจกแจงแบบแบร์นูลลี (Bernoulli Distribution)

โดย $$Y = \begin{cases} 1 & \text{เมื่อเกิดเหตุการณ์ที่สนใจศึกษา ด้วยความน่าจะเป็น } \pi(x) \\ 0 & \text{เมื่อไม่เกิดเหตุการณ์ที่สนใจศึกษา ด้วยความน่าจะเป็น } 1-\pi(x) \end{cases}$$

ส่วนตัวแปรอิสระอาจเป็นข้อมูลเชิงปริมาณ หรือเชิงคุณภาพก็ได้

ในการวิเคราะห์การถดถอยใด ๆ ค่าเฉลี่ยของตัวแปรตาม ถูกกำหนดโดยตัวแปรอิสระ เรียกว่า “ค่าเฉลี่ยแบบมีเงื่อนไข” (Conditional Mean) ใช้สัญลักษณ์ $E(Y|\boldsymbol{X})$ ในกรณีที่ Y มีค่าได้ 2 ค่า (0,1) ทำให้ค่าเฉลี่ยแบบมีเงื่อนไข มีค่าตั้งแต่ 0 ถึง 1 $(0 \le E(Y|\boldsymbol{X}) \le 1)$

กำหนดให้ $E(Y|\boldsymbol{X}) = \pi(\boldsymbol{X})$ แทน ค่าเฉลี่ยแบบมีเงื่อนไขของ Y ที่กำหนดโดย X เมื่อใช้การแจกแจงโลจิสติก เขียนตัวแบบการถดถอยโลจิสติกได้ดังนี้

$$\boldsymbol{E}(\boldsymbol{Y} \mid \boldsymbol{X}) = \pi(\boldsymbol{X}) = \frac{\boldsymbol{e}^{\beta_0 + \beta_1 \boldsymbol{x}_{\boldsymbol{i}1} + \beta_1 \boldsymbol{x}_{\boldsymbol{i}2} + \beta_1 \boldsymbol{x}_{\boldsymbol{i}3}}}{1 + \boldsymbol{e}^{\beta_0 + \beta_1 \boldsymbol{x}_{\boldsymbol{i}1} + \beta_1 \boldsymbol{x}_{\boldsymbol{i}2} + \beta_1 \boldsymbol{x}_{\boldsymbol{i}3}}} \quad ; \ \boldsymbol{i} = 1, 2, \ldots, \boldsymbol{n}$$

เมื่อ $\pi(X)$ คือ ค่าเฉลี่ยแบบมีเงื่อนไขของ Y ที่กำหนดโดย X

X คือ เวกเตอร์ของค่าสังเกต x_{i1}, x_{i2}, x_{i3}

x_{i1} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 1

x_{i2} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 2

x_{i3} คือ ค่าสังเกตของตัวแปรอิสระตัวที่ i ของตัวแปรอิสระตัวที่ 3

β_p คือ สัมประสิทธิ์การถดถอย เมื่อ p = 0, 1, 2, 3

n คือ จำนวนค่าสังเกตทั้งหมด

ซึ่งการแปลงแบบโลจิท (Logit Transformation) จะทำการแปลงค่า $\boldsymbol{\pi(X)}$ จากช่วง (0,1) ให้อยู่ในช่วง $(-\infty, \infty)$ ทำให้ได้ ฟังก์ชันโลจิท หรือ $\boldsymbol{g(X)}$ ที่มีคุณสมบัติเชิงเส้น ทำให้อธิบายได้ง่ายขึ้น เขียนได้ดังนี้

$$g(X) = \ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

2.2 การประมาณค่าสูญหายโดยวิธี Mean Imputation (MEAN)

วิธี Mean Imputation (MEAN) เป็นวิธีการประมาณค่าสูญหายที่จัดอยู่ในวิธี Single Imputation วิธีนี้ใช้หลักการของการหาค่าเฉลี่ย (Mean) เข้ามาช่วยในการแก้ปัญหาค่าสูญหาย โดยมีวิธีการคือ การแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยของข้อมูลค่าสังเกตที่ทราบค่าโดยสมมติให้ ข้อมูลของตัวแปรอิสระตัวใดตัวหนึ่ง $(X_{ip}\,; p = 1, 2, 3)$ เกิดการสูญหาย แต่ตัวแปรตาม (Y) ทราบค่า นั้นคือ ข้อมูลของตัวแปรอิสระ (X) ประกอบด้วย 2 ส่วน คือ $X = (X_{obs}, X_{mis})$ โดยที่มีจำนวนข้อมูลทั้งหมด n ค่า และ X_{obs} แทน ข้อมูลค่าสังเกตที่ทราบค่า มีจำนวน เท่ากับ m ค่า และ X_{mis} แทน ข้อมูลค่าสังเกตที่สูญหาย มีจำนวน เท่ากับ n - m ค่า

คำนวณค่าเฉลี่ยของมูลค่าสังเกตที่ทราบค่า จะได้ว่า $\bar{x}_p$ แทนค่าประมาณค่าสูญหายทุกค่า

$$\bar{x}_p = \frac{\sum_{i=1}^{m} x_{ip}}{m} \qquad ; \qquad i = 1, 2, 3, \ldots, m \quad , p = 1, 2, 3$$

2.3 การประมาณค่าสูญหายโดยวิธี Median Imputation (MED)

ในวิธี Single Imputation มีวิธีในการประมาณค่าสูญหายหลายวิธีด้วยกัน และการแทนที่ค่าสูญหายด้วยค่ากลางของข้อมูลควรเลือกให้เหมาะสมกับลักษณะของข้อมูลเพื่อให้การประมาณมีประสิทธิภาพ และวิธีการประมาณค่าด้วยค่ากลางที่น่าสนใจอีกวิธีหนึ่งคือ วิธี Median Imputation (MED) คือเป็นวิธีการแทนค่าข้อมูลที่สูญหายด้วยค่ามัธยฐานของมูลค่าสังเกตที่ทราบค่า

ค่ามัธยฐาน (Median) คือ ค่าที่มีตำแหน่งอยู่กึ่งกลางของข้อมูลทั้งหมด เมื่อได้เรียงข้อมูลตามลำดับ ทั้งจากน้อยไปมาก หรือมากไปน้อย ค่ามัธยฐานยังสามารถใช้ได้ดีเมื่อข้อมูลเกิดมีข้อมูลบางค่าที่มีค่าสูงหรือต่ำกว่าข้อมูลอื่นมาก ๆ

โดยสมมติให้ ข้อมูลของตัวแปรอิสระตัวใดตัวหนึ่ง $(X_{ip}\,; p = 1, 2, 3)$ เกิดการสูญหาย นั้นคือ ข้อมูลของตัวแปรอิสระ X ประกอบด้วย 2 ส่วน คือ $X = (X_{obs}, X_{mis})$ โดยที่มีจำนวนข้อมูลทั้งหมด n ค่า และ X_{obs} แทน ข้อมูลค่าสังเกตที่ทราบค่า มีจำนวน เท่ากับ m ค่า และ X_{mis} แทน ข้อมูลค่าสังเกตที่สูญหาย มีจำนวน เท่ากับ n-m ค่า

คำนวณค่ามัธยฐานของข้อมูลค่าสังเกตที่ทราบค่า จะได้ว่า $\bar{x}_p^*$ แทนค่าประมาณค่าสูญหายทุกค่าการประมาณค่าสูญหายด้วยวิธี MED มีขั้นตอนดังนี้

1. เรียงลำดับข้อมูลค่าสังเกตที่ทราบค่าจากน้อยไปมากหรือมากไปน้อย
2. หาตำแหน่งของค่ามัธยฐาน

   ตำแหน่งของค่ามัธยฐาน $= \dfrac{\boldsymbol{m}+1}{2}$
3. $\bar{x}_p^* =$ ข้อมูลที่อยู่ ณ ตำแหน่งของค่ามัธยฐาน

2.4 การประมาณค่าสูญหายโดยวิธี K-Nearest Neighbor Imputation (KNN)

วิธี K-Nearest Neighbor Imputation (KNN) เป็นวิธีการประมาณค่าสูญหายที่จัดอยู่ใน Hot Deck Method ซึ่งเป็นวิธีการแทนค่าสูญหายด้วยค่าสังเกตที่ทราบค่าที่มีลักษณะคล้ายคลึงกัน ซึ่งวิธี KNN จะประมาณค่าโดยใช้ค่าใกล้สุด โดยพิจารณาเลือกหน่วยตัวอย่างจากชุดข้อมูล (x_i, y_i) ที่ทราบค่า ที่มีลักษณะคล้ายคลึงกับหน่วยตัวอย่างที่สูญหายมากที่สุด โดยพิจารณาเลือกหน่วยตัวอย่าง K ตัว ซึ่ง Duda และ Hart (1973) เสนอให้ $K \approx \sqrt{m}$ เมื่อ m แทน จำนวนข้อมูลค่าสังเกตที่ทราบค่า (อ้างถึงใน Jonsson และ Wohlin, 2006) จากนั้นแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของหน่วยตัวอย่าง K ตัว

ในการศึกษาครั้งนี้ ได้นำวิธี KNN มาประยุกต์ในกรณีที่ การสูญหายเกิดขึ้นในตัวแปรอิสระ $(X_{ip}\,; p = 1, 2, 3)$ โดยสมมติให้การสูญหายเกิดขึ้นในตัวแปรอิสระตัวที่ 3 และตัวแปรตามทราบค่า โดยที่ตัวแปรตาม (Y) มีได้ 2 ค่า คือ 1 และ 0 และตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ ทำการประมาณค่าสูญหายจากชุดข้อมูลที่สมบูรณ์ โดยพิจารณาจากระยะทางยุคลิด (Euclidean Distance) ดังสมการต่อไปนี้

$$D_{ij} = \sqrt{\sum_{p=1}^{2}\left(x_{ip} - x_{jp}\right)^2 + \left(y_i - y_j\right)^2} \quad ; i = 1, 2, 3, \dots, m \text{ และ } j = m+1, \dots, n$$

เมื่อ D_{ij} คือ ระยะทางยุคลิดของค่าสังเกตตัวที่ i และ j

การประมาณค่าสูญหายด้วยวิธี KNN มีขั้นตอนดังนี้

1. ทำการแปลงข้อมูลตัวแปรอิสระและตัวแปรตามทั้งหมดที่ทราบค่า ให้เป็นค่ามาตรฐาน
2. คำนวณหาค่า D_{ij} สำหรับตัวแปรอิสระ และตัวแปรตามทุกคู่ที่เป็นไปได้แล้วนับจำนวน D_{ij} ที่มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างที่ j
3. คำนวณหาค่าเฉลี่ยของตัวแปรอิสระที่สอดคล้องกับ D_{ij} มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างที่ j โดยกำหนดให้เป็น $\bar{x}_{i3}^*$
4. $x_{i3}^* = \bar{x}_{i3}^*$ เมื่อ x_{i3}^* แทน ค่าประมาณของตัวแปรอิสระตัวที่ 3

2.5 การประมาณค่าสูญหายโดยวิธี Multiple Imputation (MI)

Rubin (1978) ได้เสนอวิธี Multiple Imputation เป็นครั้งแรก โดยมีแนวคิดว่า การแจกแจงของข้อมูลค่าสังเกตที่สูญหายสามารถหาได้จากข้อมูลค่าสังเกตที่มีอยู่ นั้นคือเป็นการหา Predictive Distribution ของข้อมูลที่สูญหายซึ่งวิธี Multiple Imputation นี้ ข้อมูลที่สูญหายจะถูกแทนที่ด้วยชุดข้อมูลของค่าที่เป็นไปได้ มากกว่า 1 (D>1) ซึ่ง Rubin ได้เสนอว่า วิธีนี้จะมีประสิทธิภาพมากขึ้น เมื่อมีการสร้างชุดข้อมูลที่สมบูรณ์ ตั้งแต่ 5 ถึง 10 ชุด แม้ว่าจะมีการสูญหายของข้อมูลมากกว่าร้อยละ 50

วิธี Multiple Imputationนี้ มีข้อดีกว่าวิธี Single Imputation และ Hot Deck Imputation คือ มีการดำเนินการแทนค่ามากกว่า 1 ครั้ง ทำให้เพิ่มประสิทธิภาพของตัวประมาณและทำให้อนุมานเกี่ยวกับการกระจายของค่าสูญหายได้

โดยให้ $X = (X_{obs}, X_{mis})$ เมื่อ X_{obs} แทน ข้อมูลค่าสังเกตที่ทราบค่า และ X_{mis} แทน ข้อมูลค่าสังเกตที่สูญหาย และสมมติให้ X มีการแจกแจงแบบปกติ ซึ่งพารามิเตอร์หาได้จาก Predictive Distribution โดยการจำลองพารามิเตอร์จากค่าสังเกตที่ทราบค่า แล้วประมาณค่าสังเกตที่สูญหายทั้งหมด D ชุด ทำการวิเคราะห์ข้อมูลที่สมบูรณ์ และนำค่าสัมประสิทธิ์การถดถอยทั้ง D ชุด ที่ได้ไปหาค่าเฉลี่ย

บทที่ 3

วิธีดำเนินการวิจัย

ในบทที่ 3 นี้ จะอธิบายถึงรายละเอียดของวิธีการดำเนินการวิจัยในการทำวิจัยครั้งนี้ ซึ่งเป็นการวิจัยเชิงทดลอง เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression Analysis) กรณีที่มีตัวแปรอิสระ 3 ตัว ภายใต้สถานการณ์ที่แตกต่างกัน โดยวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ คือ วิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI) เปรียบเทียบแต่ละวิธีด้วยค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) ดำเนินการจำลองข้อมูลโดยใช้โปรแกรม R version 3.3.0 ในงานวิจัยครั้งนี้จะทำการศึกษาสถานการณ์ที่มีความแตกต่างกันทั้งหมด 81 สถานการณ์ ซึ่งแตกต่างกันตามขนาดของตัวอย่าง และลักษณะการสูญหายของข้อมูล โดยจะทำการจำลองแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ มีขั้นตอนในการวิจัย ดังนี้

3.1 การจำลองข้อมูล

3.1.1 สร้างชุดข้อมูลตัวแปรอิสระ (X_1, X_2, X_3) ที่ตัวแปรอิสระแต่ละตัวมีการแจกแจงแบบปกติมาตรฐาน (Standard Normal Distribution) โดยกำหนดให้ตัวแปรอิสระทั้ง 3 ตัว ไม่มีความสัมพันธ์กัน

3.1.2 สร้างตัวแปรตาม ที่มีความสัมพันธ์กับตัวแปรอิสระภายใต้ตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม โดยกำหนดให้ค่าสัมประสิทธิ์การถดถอย $\beta_0 = 0, \beta_1 = 0.5, \beta_2 = 1$ และ $\beta_3 = 1.5$ เขียนตัวแบบได้ดังนี้

$$\ln\left(\frac{\pi(X)}{1-\pi(X)}\right) = 0.5x_{i1} + x_{i2} + 1.5x_{i3} \; ; \; i = 1, 2, 3, \ldots, n$$

3.1.3 ขนาดตัวอย่าง (n) มี 3 ขนาด คือ 70, 100 และ 200

3.2 การสุ่มตำแหน่งที่ข้อมูลตัวแปรอิสระเกิดการสูญหาย

3.2.1 การสูญหายแต่ละสถานการณ์เกิดขึ้นในตัวแปรอิสระตัวใดตัวหนึ่ง โดยให้การสูญหายเกิดขึ้นที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอยที่ $\beta_1 = 0.5$ หรือ $\beta_2 = 1$ หรือ $\beta_3 = 1.5$ สลับกันในทุกสถานการณ์

3.2.2 ทำการสุ่มตำแหน่งที่ข้อมูลตัวแปรอิสระเกิดการสูญหาย โดยนำข้อมูลตัวแปรอิสระมาแบ่งเป็น 3 ช่วง โดยให้แต่ละช่วงมีสัดส่วนเท่า ๆ กัน

วิธีในการแบ่งช่วงของตัวแปรอิสระ คือ

ถ้า $x_{ip} \le \mu_{x_p} + z_{\frac{1}{3}}\sigma_{x_p}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงต้น

ถ้า $\mu_{x_p} + z_{\frac{1}{3}}\sigma_{x_p} < x_{ip} \le \mu_{x_p} + z_{\frac{2}{3}}\sigma_{x_p}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงกลาง

ถ้า $\mu_{x_p} + z_{\frac{2}{3}}\sigma_{x_p} < x_{ip}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ในช่วงปลาย

โดยที่ $p = 1,2,3$, $i = 1,2,3,\ldots,n$ และ $z_{\frac{1}{3}} = -0.43$, $z_{\frac{2}{3}} = 0.43$

3.2.3 จากนั้นทำให้ข้อมูลอิสระเกิดการสูญหาย โดยสร้างตัวแปรที่มีการแจกแจงแบร์นูลลี ซึ่งมีขนาดเท่ากับจำนวนตัวแปรอิสระที่อยู่ในแต่ละช่วง และมีความน่าจะเป็นของการสูญหายแตกต่างกันตามที่กำหนดไว้ในขอบเขตของงานวิจัย จับคู่ข้อมูลตัวแปรอิสระกับตัวแปรที่มีการแจกแจงแบร์นูลลี่ ถ้าตัวแปรที่มีการแจกแจงแบร์นูลลีมีค่าเท่ากับ 1 แสดงว่าตัวแปรอิสระที่ถูกจับคู่เกิดการสูญหาย

3.3 การประมาณค่าสังเกตที่สูญหาย

เมื่อได้จำลองการสูญหายของข้อมูลแล้ว ต่อมาทำการประมาณค่าข้อมูลและแทนที่ข้อมูลที่สูญหายของตัวแปรอิสระ ด้วยวิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI)

3.4 การประมาณค่าสัมประสิทธิ์การถดถอย

เมื่อแทนค่าข้อมูลที่สูญหายแล้วทำการประมาณค่าสัมประสิทธิ์การถดถอยใหม่ด้วยวิธีความควรจะเป็นสูงสุด (MLE) จะได้ค่าสัมประสิทธิ์การถดถอยที่แตกต่างกันในแต่ละสถานการณ์

3.5 การเปรียบเทียบวิธีการประมาณค่าสูญหาย

3.5.1 คำนวณค่า MSE และ AMSE

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) และค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) ในการศึกษาครั้งนี้ใช้ค่าสัมประสิทธิ์การถดถอยในการคำนวณค่า AMSE ซึ่งวิธีการที่ให้ค่า AMSE ต่ำสุดจะเป็นวิธีการประมาณค่าสูญหายที่มีประสิทธิภาพกว่าวิธีอื่น ๆ คำนวณได้จากสูตรดังต่อไปนี้

$$MSE_p = \frac{\sum_{t=1}^{5,000}\left(\hat{\beta}_p^{(t)} - \beta_p\right)^2}{5,000} \qquad ; p = 0, 1, 2, 3$$

$$AMSE = \frac{\sum_{p=0}^{3} MSE_p}{4} \qquad ; p = 0, 1, 2, 3$$

3.5.2 คำนวณค่า RE

ค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency: RE) ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีการประมาณค่าสูญหายแบบอื่น ๆ เพื่อใช้เปรียบเทียบประสิทธิภาพของแต่ละวิธีให้มีความชัดเจนมากขึ้น โดยการศึกษาครั้งนี้เลือกวิธี MEAN เป็นเกณฑ์ในการเปรียบเทียบ เนื่องจากวิธี MEAN เป็นวิธีการประมาณค่าสูญหายที่เป็นวิธีพื้นฐาน โดยวิธีการที่ให้ค่า RE มากกว่า 1 แสดงว่าวิธีการนั้นมีประสิทธิภาพมากกว่าวิธี MEAN สามารถคำนวณได้จากสูตรดังต่อไปนี้

$$RE(MSE_p) = \frac{MSE_{MEAN}}{MSE_r} \qquad ; \; r = MED,\ KNN,\ MI \;,\; p=3$$

$$RE(AMSE) = \frac{AMSE_{MEAN}}{AMSE_r} \qquad ; \; r = MED, KNN, MI$$

เมื่อ $\hat{\beta}_p^{(t)}$ คือ ค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกตัวที่ p ในการจำลองรอบที่ t

β_p คือ ค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกตัวที่ p ที่กำหนดขึ้น

MSE_p คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าสัมประสิทธิ์การถดถอยตัวที่ p จากการทำซ้ำ 5,000 รอบ

MSE_{MEAN} คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MEAN

MSE_r คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MED หรือ วิธี KNN หรือวิธี MI

$AMSE$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

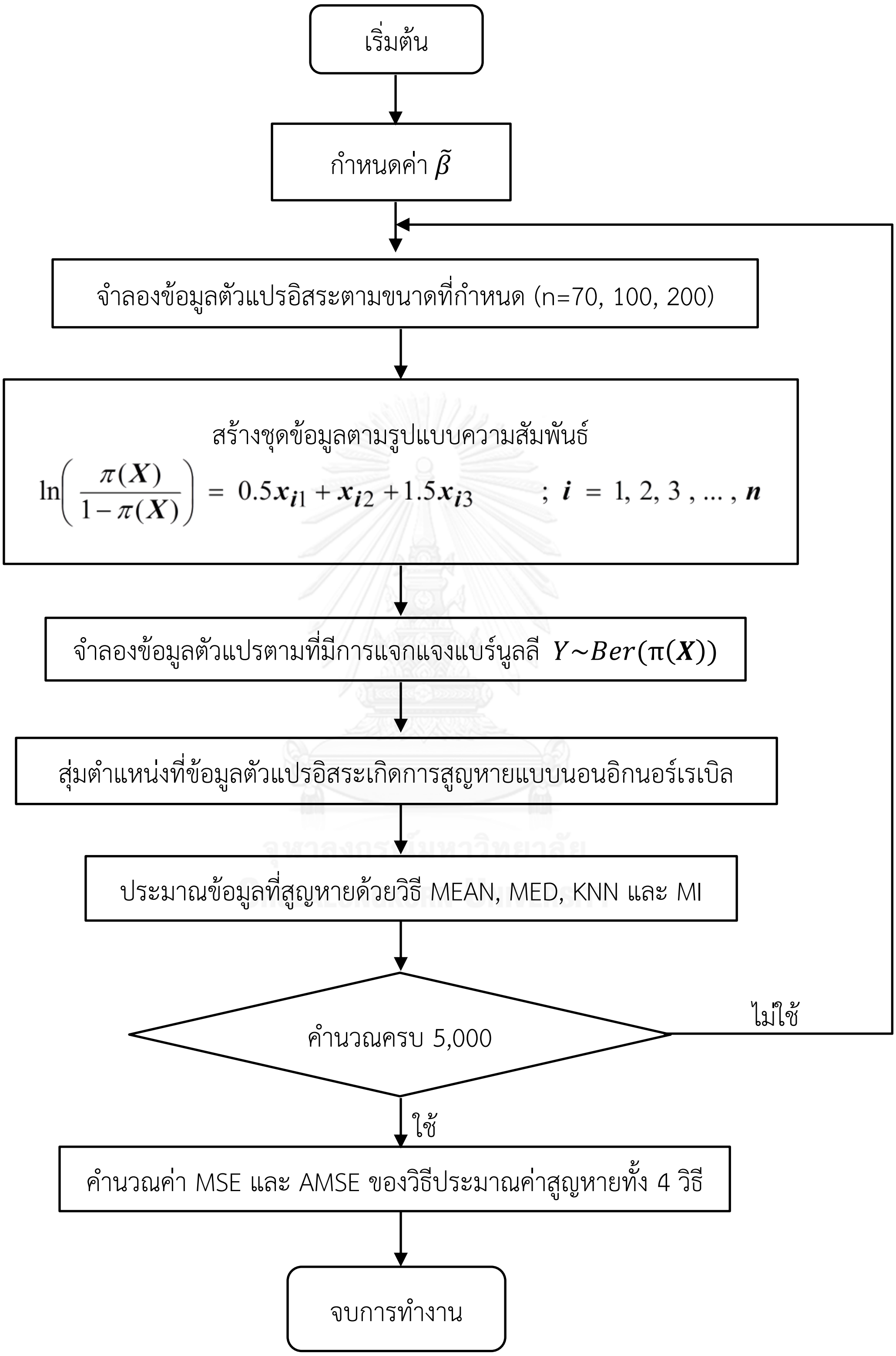
$AMSE_{MEAN}$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MEAN

$AMSE_r$ คือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี MED หรือ วิธี KNN หรือวิธี MI

$RE(MSE_p)$ คือ ค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าสัมประสิทธิ์การถดถอยตัวที่ p

$RE(AMSE)$ คือ ค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

รูปที่ 1 แผนผังขั้นตอนการดำเนินการวิจัย



บทที่ 4

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระที่มีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม เมื่อมีตัวแปรอิสระ 3 ตัว และเกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง โดยวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ คือ วิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI) ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองข้อมูล โดยกำหนดขนาดตัวอย่าง เท่ากับ 70, 100 และ 200 โดยเกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง มีร้อยละของการสูญหายโดยเฉลี่ย 3 ระดับ คือ ร้อยละ 10, 20 และ 30 มีระดับการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับ คือ การสูญหายแบบอิกนอร์เรเบิล การสูญหายแบบนอนอิกนอร์เรเบิลระดับปานกลาง และ การสูญหายแบบนอนอิกนอร์เรเบิลระดับสูง ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ และวิเคราะห์ผลเปรียบเทียบแต่ละวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) ระหว่างค่าสัมประสิทธิ์การถดถอยที่กำหนดกับค่าสัมประสิทธิ์การถดถอยที่ได้จากการประมาณ และค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error: AMSE) รวมทั้งค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency: RE)

การนำเสนอผลการวิจัยจะแสดงในรูปแบบของตารางและกราฟ เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายในแต่ละสถานการณ์ โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่าง ๆ ดังนี้

n	แทน	ขนาดตัวอย่าง
Level of nonignorability	แทน	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล
Ignorable	แทน	การสูญหายแบบอิกนอร์เรเบิล
Moderate	แทน	การสูญหายแบบนอนอิกนอร์เรเบิลระดับปานกลาง
High	แทน	การสูญหายแบบนอนอิกนอร์เรเบิลระดับสูง
10%	แทน	การสูญหายโดยเฉลี่ยร้อยละ 10
20%	แทน	การสูญหายโดยเฉลี่ยร้อยละ 20
30%	แทน	การสูญหายโดยเฉลี่ยร้อยละ 30
MEAN	แทน	การประมาณค่าสูญหายด้วยวิธี Mean Imputation
MED	แทน	การประมาณค่าสูญหายด้วยวิธี Median Imputation
KNN	แทน	การประมาณค่าสูญหายด้วยวิธี K-Nearest Neighbor Imputation
MI	แทน	การประมาณค่าสูญหายด้วยวิธี Multiple Imputation

MSE	แทน	ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง
AMSE	แทน	ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง
RE	แทน	ค่าประสิทธิภาพสัมพัทธ์

โดยการวิเคราะห์ผลการวิจัยจะแบ่งออกเป็น 3 ส่วน ตามลักษณะการสูญหายของตัวแปรอิสระ ดังนี้

4.1 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5

4.2 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1

4.3 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5

4.1 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5

ในส่วนนี้การสูญหายเกิดที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 และมีข้อมูลตามขอบเขตการวิจัยที่กำหนด โดยแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายแต่ละวิธีด้วยค่า MSE ของสัมประสิทธิ์การถดถอยของตัวแปรอิสระที่สูญหาย, RE(MSE), AMSE และ RE(AMSE) ดังแสดงในตารางที่ 2 - 5 ตามลำดับ และนำเสนอกราฟแสดงค่า RE(MSE) และ RE(AMSE) เพื่อให้เข้าใจชัดเจนมากขึ้น

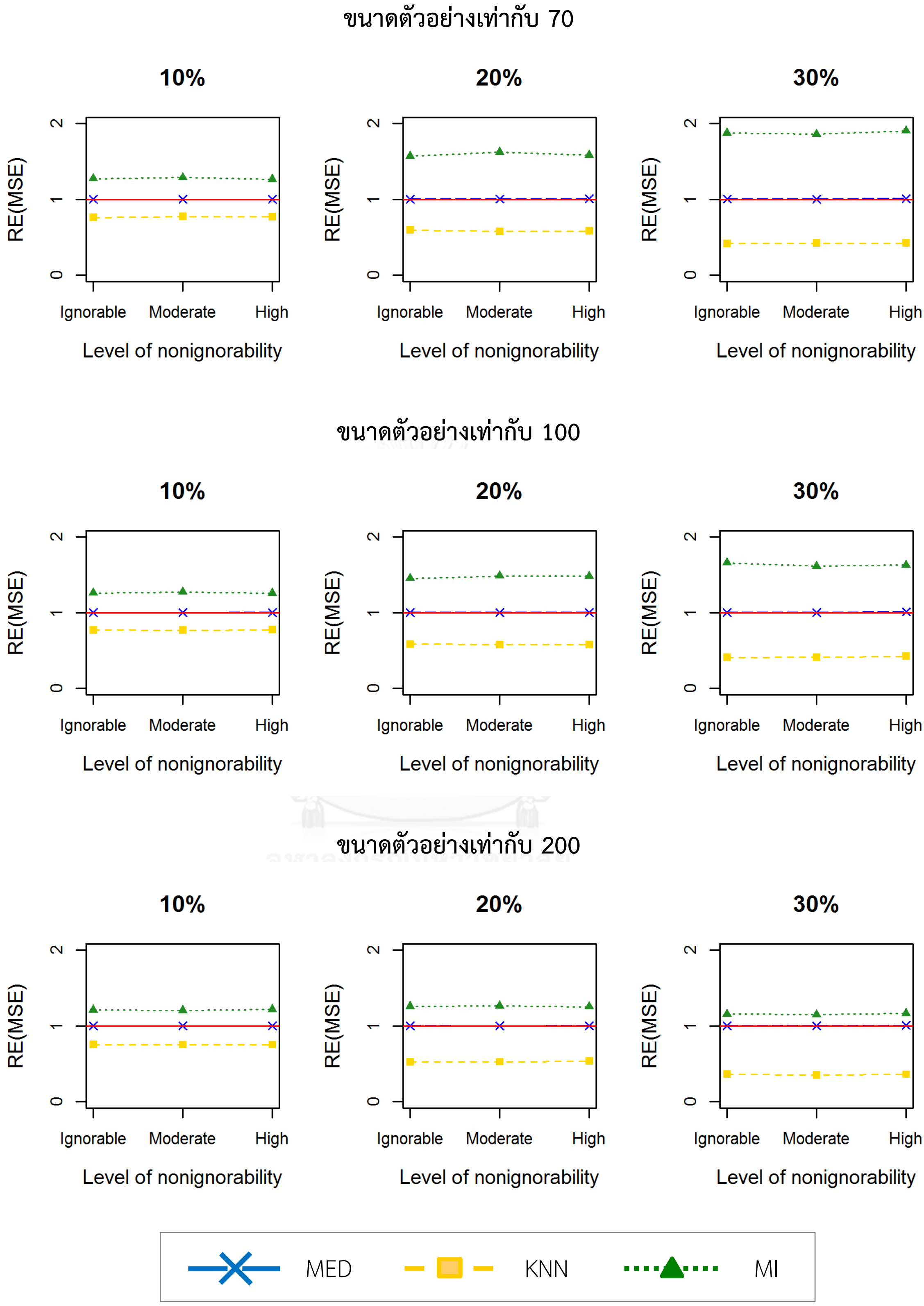
ตารางที่	ค่าที่แสดงในตาราง
ตารางที่ 2	MSE
ตารางที่ 3	RE(MSE)
ตารางที่ 4	AMSE
ตารางที่ 5	RE(AMSE)

ตารางที่ 2 การเปรียบเทียบค่า MSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 (β = 0.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	MSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.164	0.164	0.216	**0.129**
		Moderate nonignorable	0.162	0.162	0.210	**0.126**
		High nonignorable	0.163	0.162	0.211	**0.128**
	ร้อยละ 20	Ignorable	0.181	0.181	0.305	**0.115**
		Moderate nonignorable	0.188	0.187	0.325	**0.116**
		High nonignorable	0.185	0.184	0.318	**0.117**
	ร้อยละ 30	Ignorable	0.207	0.206	0.494	**0.110**
		Moderate nonignorable	0.214	0.213	0.505	**0.115**
		High nonignorable	0.220	0.218	0.520	**0.115**
n=100	ร้อยละ 10	Ignorable	0.100	0.100	0.130	**0.080**
		Moderate nonignorable	0.100	0.100	0.130	**0.078**
		High nonignorable	0.101	0.101	0.131	**0.081**
	ร้อยละ 20	Ignorable	0.112	0.112	0.191	**0.077**
		Moderate nonignorable	0.114	0.113	0.197	**0.076**
		High nonignorable	0.113	0.113	0.195	**0.076**
	ร้อยละ 30	Ignorable	0.130	0.130	0.317	**0.078**
		Moderate nonignorable	0.131	0.130	0.316	**0.081**
		High nonignorable	0.134	0.132	0.316	**0.082**
n=200	ร้อยละ 10	Ignorable	0.045	0.045	0.059	**0.037**
		Moderate nonignorable	0.044	0.044	0.059	**0.037**
		High nonignorable	0.046	0.046	0.061	**0.037**
	ร้อยละ 20	Ignorable	0.050	0.049	0.095	**0.039**
		Moderate nonignorable	0.050	0.050	0.095	**0.040**
		High nonignorable	0.051	0.051	0.095	**0.040**
	ร้อยละ 30	Ignorable	0.056	0.056	0.156	**0.049**
		Moderate nonignorable	0.056	0.056	0.160	**0.049**
		High nonignorable	0.058	0.058	0.161	**0.050**

ตารางที่ 3 การเปรียบเทียบค่า RE(MSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 (β = 0.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(MSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	1.000	0.760	**1.272**
		Moderate nonignorable	1.000	0.999	0.773	**1.288**
		High nonignorable	1.000	1.001	0.772	**1.265**
	ร้อยละ 20	Ignorable	1.000	1.002	0.595	**1.569**
		Moderate nonignorable	1.000	1.005	0.579	**1.622**
		High nonignorable	1.000	1.006	0.581	**1.583**
	ร้อยละ 30	Ignorable	1.000	1.005	0.419	**1.875**
		Moderate nonignorable	1.000	1.002	0.424	**1.858**
		High nonignorable	1.000	1.008	0.422	**1.903**
n=100	ร้อยละ 10	Ignorable	1.000	1.001	0.771	**1.260**
		Moderate nonignorable	1.000	1.000	0.769	**1.272**
		High nonignorable	1.000	1.002	0.774	**1.257**
	ร้อยละ 20	Ignorable	1.000	1.003	0.586	**1.451**
		Moderate nonignorable	1.000	1.002	0.577	**1.485**
		High nonignorable	1.000	1.002	0.578	**1.479**
	ร้อยละ 30	Ignorable	1.000	1.001	0.410	**1.658**
		Moderate nonignorable	1.000	1.002	0.414	**1.612**
		High nonignorable	1.000	1.011	0.423	**1.627**
n=200	ร้อยละ 10	Ignorable	1.000	1.000	0.756	**1.210**
		Moderate nonignorable	1.000	1.001	0.752	**1.205**
		High nonignorable	1.000	1.000	0.753	**1.218**
	ร้อยละ 20	Ignorable	1.000	1.002	0.525	**1.258**
		Moderate nonignorable	1.000	1.001	0.527	**1.264**
		High nonignorable	1.000	1.003	0.535	**1.253**
	ร้อยละ 30	Ignorable	1.000	1.001	0.362	**1.154**
		Moderate nonignorable	1.000	1.001	0.354	**1.149**
		High nonignorable	1.000	1.006	0.361	**1.165**



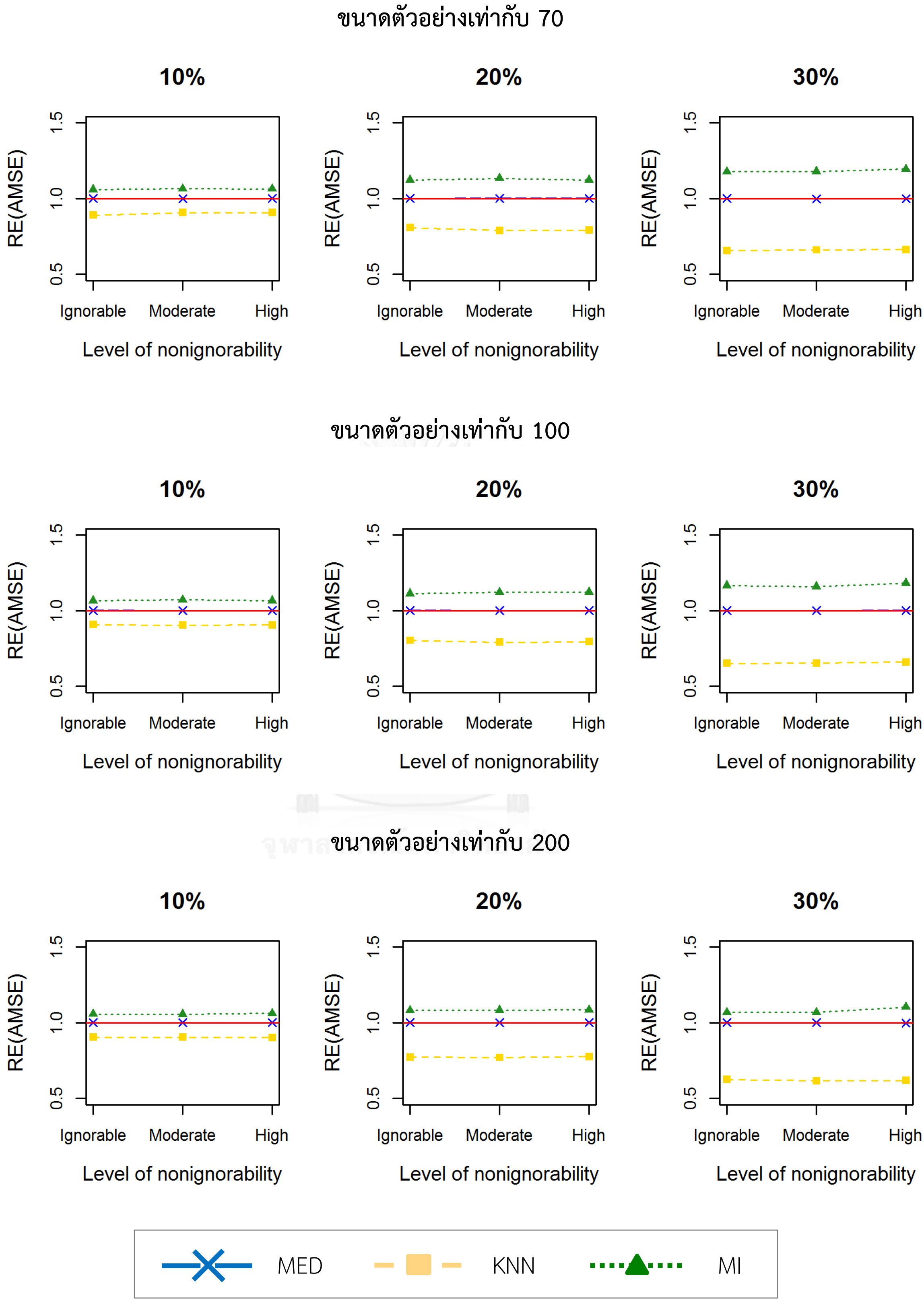
รูปที่ 2 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(MSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5

ตารางที่ 4 การเปรียบเทียบค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 (β = 0.5)

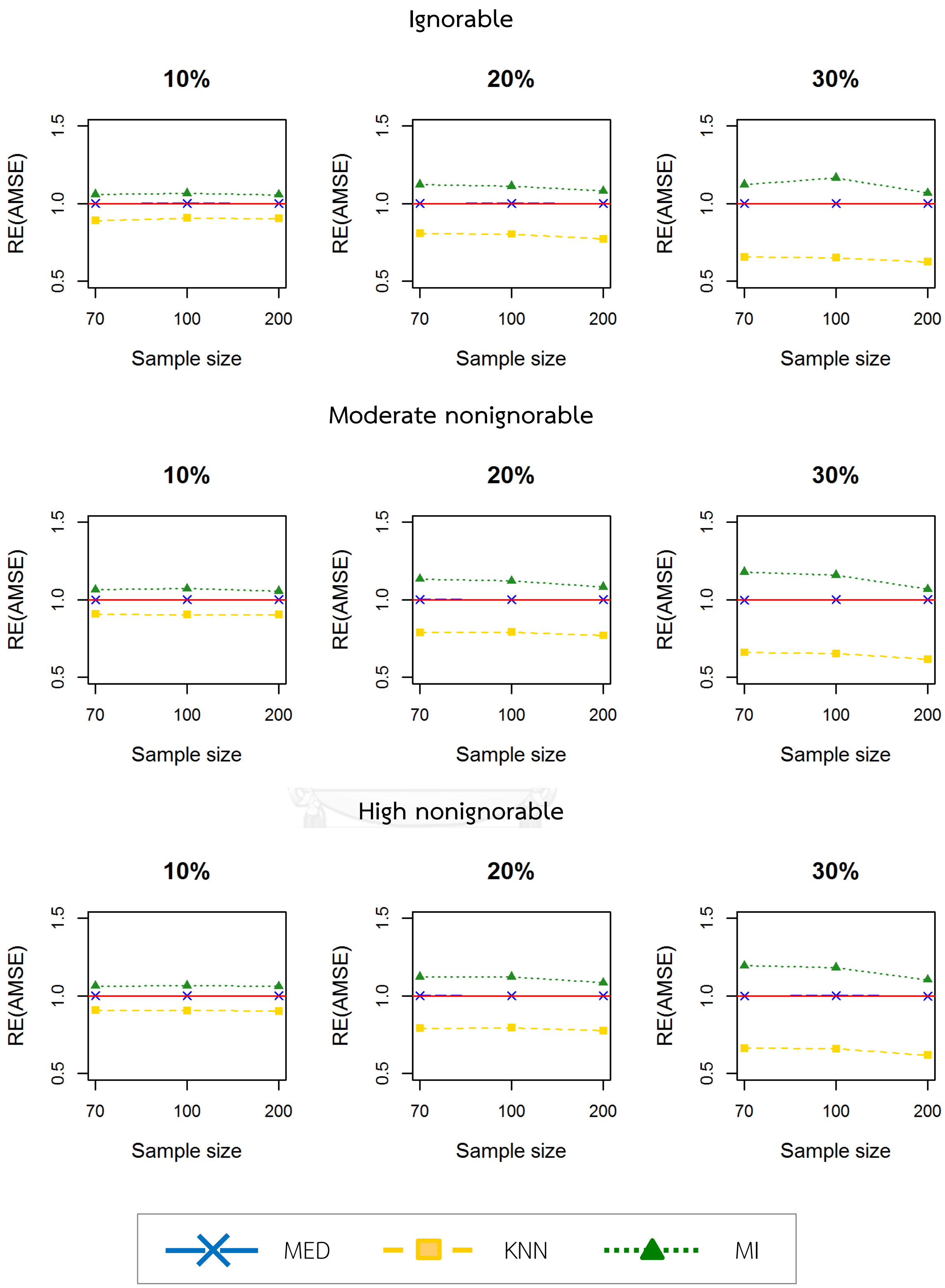
ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	AMSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.210	0.210	0.235	**0.198**
		Moderate nonignorable	0.208	0.208	0.229	**0.195**
		High nonignorable	0.209	0.209	0.230	**0.196**
	ร้อยละ 20	Ignorable	0.210	0.210	0.260	**0.187**
		Moderate nonignorable	0.212	0.211	0.268	**0.187**
		High nonignorable	0.211	0.211	0.267	**0.188**
	ร้อยละ 30	Ignorable	0.212	0.212	0.324	**0.180**
		Moderate nonignorable	0.215	0.215	0.325	**0.182**
		High nonignorable	0.219	0.219	0.329	**0.183**
n=100	ร้อยละ 10	Ignorable	0.116	0.116	0.128	**0.109**
		Moderate nonignorable	0.116	0.116	0.129	**0.109**
		High nonignorable	0.116	0.116	0.129	**0.109**
	ร้อยละ 20	Ignorable	0.118	0.118	0.147	**0.106**
		Moderate nonignorable	0.119	0.119	0.150	**0.106**
		High nonignorable	0.119	0.119	0.150	**0.106**
	ร้อยละ 30	Ignorable	0.121	0.122	0.186	**0.104**
		Moderate nonignorable	0.121	0.121	0.186	**0.105**
		High nonignorable	0.126	0.126	0.191	**0.106**
n=200	ร้อยละ 10	Ignorable	0.050	0.050	0.056	**0.048**
		Moderate nonignorable	0.050	0.050	0.056	**0.048**
		High nonignorable	0.051	0.051	0.056	**0.048**
	ร้อยละ 20	Ignorable	0.051	0.051	0.066	**0.047**
		Moderate nonignorable	0.051	0.051	0.067	**0.047**
		High nonignorable	0.052	0.052	0.067	**0.048**
	ร้อยละ 30	Ignorable	0.052	0.052	0.083	**0.049**
		Moderate nonignorable	0.052	0.052	0.085	**0.049**
		High nonignorable	0.056	0.056	0.090	**0.051**

ตารางที่ 5 การเปรียบเทียบค่า RE(AMSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 (β = 0.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(AMSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	1.000	0.891	**1.059**
		Moderate nonignorable	1.000	0.999	0.908	**1.064**
		High nonignorable	1.000	1.000	0.907	**1.064**
	ร้อยละ 20	Ignorable	1.000	1.000	0.808	**1.122**
		Moderate nonignorable	1.000	1.002	0.789	**1.133**
		High nonignorable	1.000	1.001	0.791	**1.122**
	ร้อยละ 30	Ignorable	1.000	1.001	0.656	**1.177**
		Moderate nonignorable	1.000	0.998	0.661	**1.179**
		High nonignorable	1.000	0.999	0.664	**1.194**
n=100	ร้อยละ 10	Ignorable	1.000	1.001	0.907	**1.066**
		Moderate nonignorable	1.000	1.000	0.904	**1.071**
		High nonignorable	1.000	1.000	0.905	**1.065**
	ร้อยละ 20	Ignorable	1.000	1.001	0.803	**1.112**
		Moderate nonignorable	1.000	1.000	0.791	**1.121**
		High nonignorable	1.000	1.000	0.794	**1.122**
	ร้อยละ 30	Ignorable	1.000	0.999	0.652	**1.165**
		Moderate nonignorable	1.000	1.000	0.653	**1.158**
		High nonignorable	1.000	1.001	0.660	**1.182**
n=200	ร้อยละ 10	Ignorable	1.000	1.000	0.904	**1.056**
		Moderate nonignorable	1.000	1.000	0.903	**1.056**
		High nonignorable	1.000	1.000	0.902	**1.061**
	ร้อยละ 20	Ignorable	1.000	1.000	0.772	**1.082**
		Moderate nonignorable	1.000	1.000	0.770	**1.083**
		High nonignorable	1.000	1.000	0.776	**1.085**
	ร้อยละ 30	Ignorable	1.000	1.000	0.625	**1.067**
		Moderate nonignorable	1.000	1.000	0.616	**1.068**
		High nonignorable	1.000	0.998	0.619	**1.103**



รูปที่ 3 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5



รูปที่ 4 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ขนาดตัวอย่าง เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5

สรุปผลส่วนที่ 4.1

ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 เมื่อพิจาณาค่า AMSE พบว่า การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า AMSE น้อยที่สุดในทุกสถานการณ์ รองลงมา คือ วิธี MEAN และ วิธี MED ซึ่งให้ค่า AMSE ใกล้เคียงกัน และวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า AMSE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มเพิ่มขึ้น เพราะเมื่อข้อมูลมีการสูญหายเพิ่มขึ้นจะทำให้เกิดความคลาดเคลื่อนจากการประมาณค่าสูญหาย และส่งผลให้เกิดความคลาดเคลื่อนต่อการประมาณค่าพารามิเตอร์อีกด้วย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีค่าใกล้เคียงกัน เนื่องจากการสูญหายเกิดขึ้นที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 นั้นคือตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามน้อย การสูญหายของตัวแปรอิสระจึงส่งผลต่อการประมาณค่าพารามิเตอร์น้อย

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีประมาณค่าสูญหายวิธีอื่น ๆ พบว่า การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพมากที่สุด เมื่อเทียบกับวิธี MEAN และวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า RE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างใหญ่ขึ้น พบว่า ค่า RE ของวิธี MI และวิธี KNN มีแนวโน้มลดลง สำหรับวิธี MED การเปลี่ยนแปลงของขนาดตัวอย่างมีผลต่อค่า RE เพียงเล็กน้อย

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า RE ของวิธี KNN มีแนวโน้มลดลงอย่างชัดเจน แต่สำหรับวิธี MI จะมีประสิทธิภาพเพิ่มขึ้น

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มเพิ่มขึ้นเล็กน้อย ในส่วนของวิธี MED และวิธี KNN ค่า RE ไม่มีการเปลี่ยนแปลงที่ชัดเจน

4.2 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1

ในส่วนนี้การสูญหายเกิดที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 และมีข้อมูลตามขอบเขตการวิจัยที่กำหนด โดยแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายแต่ละวิธีด้วยค่า MSE, RE(MSE), AMSE และ RE(AMSE) ดังแสดงในตารางที่ 6 - 9 ตามลำดับ และนำเสนอกราฟแสดงค่า RE(MSE) และ RE(AMSE) เพื่อให้เข้าใจชัดเจนมากขึ้น

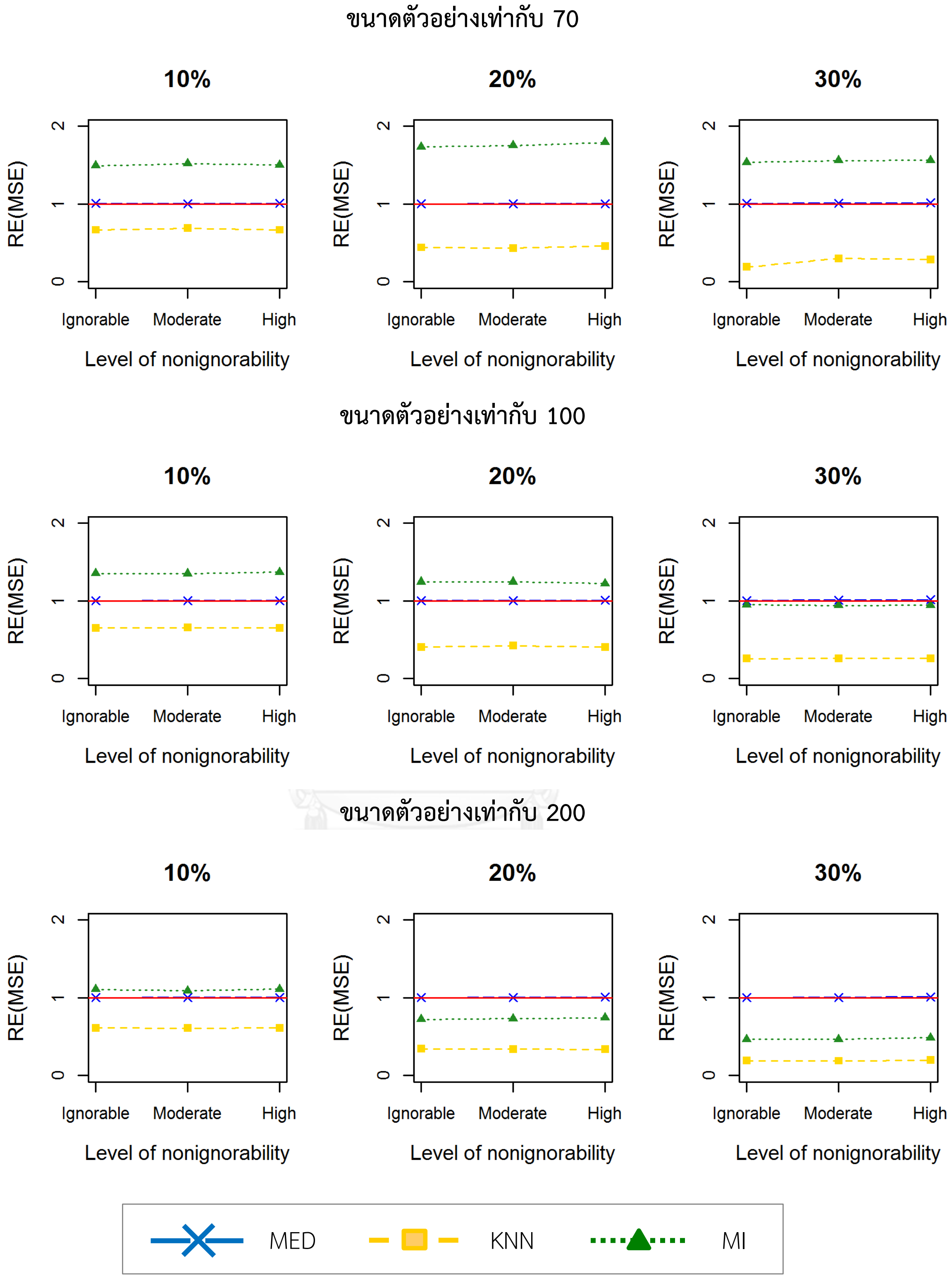
ตารางที่	ค่าที่แสดงในตาราง
ตารางที่ 6	MSE
ตารางที่ 7	RE(MSE)
ตารางที่ 8	AMSE
ตารางที่ 9	RE(AMSE)

ตารางที่ 6 การเปรียบเทียบค่า MSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 (β = 1)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	MSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.236	0.234	0.354	**0.158**
		Moderate nonignorable	0.238	0.237	0.345	**0.157**
		High nonignorable	0.226	0.225	0.340	**0.151**
	ร้อยละ 20	Ignorable	0.256	0.255	0.580	**0.148**
		Moderate nonignorable	0.253	0.252	0.584	**0.144**
		High nonignorable	0.276	0.275	0.601	**0.154**
	ร้อยละ 30	Ignorable	0.273	0.272	1.428	**0.178**
		Moderate nonignorable	0.276	0.273	0.918	**0.177**
		High nonignorable	0.279	0.275	0.970	**0.179**
n=100	ร้อยละ 10	Ignorable	0.124	0.124	0.191	**0.092**
		Moderate nonignorable	0.121	0.121	0.185	**0.089**
		High nonignorable	0.126	0.126	0.193	**0.092**
	ร้อยละ 20	Ignorable	0.134	0.133	0.330	**0.107**
		Moderate nonignorable	0.137	0.136	0.323	**0.110**
		High nonignorable	0.133	0.132	0.328	**0.109**
	ร้อยละ 30	Ignorable	0.149	**0.148**	0.577	0.157
		Moderate nonignorable	0.145	**0.144**	0.559	0.154
		High nonignorable	0.154	**0.152**	0.594	0.163
n=200	ร้อยละ 10	Ignorable	0.055	0.055	0.090	**0.050**
		Moderate nonignorable	0.055	0.055	0.090	**0.050**
		High nonignorable	0.056	0.056	0.091	**0.050**
	ร้อยละ 20	Ignorable	**0.060**	**0.060**	0.174	0.083
		Moderate nonignorable	**0.060**	**0.060**	0.178	0.082
		High nonignorable	0.061	**0.060**	0.180	0.082
	ร้อยละ 30	Ignorable	**0.065**	**0.065**	0.342	0.141
		Moderate nonignorable	**0.065**	**0.065**	0.344	0.139
		High nonignorable	**0.069**	**0.069**	0.353	0.143

ตารางที่ 7 การเปรียบเทียบค่า RE(MSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 (β = 1)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(MSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	1.005	0.666	**1.492**
		Moderate nonignorable	1.000	1.002	0.689	**1.517**
		High nonignorable	1.000	1.006	0.666	**1.499**
	ร้อยละ 20	Ignorable	1.000	1.001	0.441	**1.732**
		Moderate nonignorable	1.000	1.002	0.433	**1.750**
		High nonignorable	1.000	1.003	0.459	**1.790**
	ร้อยละ 30	Ignorable	1.000	1.006	0.191	**1.534**
		Moderate nonignorable	1.000	1.011	0.300	**1.557**
		High nonignorable	1.000	1.014	0.288	**1.558**
n=100	ร้อยละ 10	Ignorable	1.000	1.001	0.651	**1.352**
		Moderate nonignorable	1.000	1.002	0.654	**1.351**
		High nonignorable	1.000	1.003	0.653	**1.369**
	ร้อยละ 20	Ignorable	1.000	1.003	0.406	**1.244**
		Moderate nonignorable	1.000	1.003	0.423	**1.244**
		High nonignorable	1.000	1.007	0.407	**1.221**
	ร้อยละ 30	Ignorable	1.000	**1.003**	0.258	0.949
		Moderate nonignorable	1.000	**1.010**	0.260	0.941
		High nonignorable	1.000	**1.013**	0.259	0.944
n=200	ร้อยละ 10	Ignorable	1.000	1.000	0.612	**1.107**
		Moderate nonignorable	1.000	1.002	0.610	**1.090**
		High nonignorable	1.000	1.002	0.611	**1.108**
	ร้อยละ 20	Ignorable	**1.000**	1.000	0.345	0.722
		Moderate nonignorable	1.000	1.004	0.339	0.731
		High nonignorable	1.000	1.006	0.336	0.742
	ร้อยละ 30	Ignorable	1.000	1.000	0.192	0.465
		Moderate nonignorable	1.000	1.003	0.188	0.466
		High nonignorable	1.000	1.008	0.196	0.484



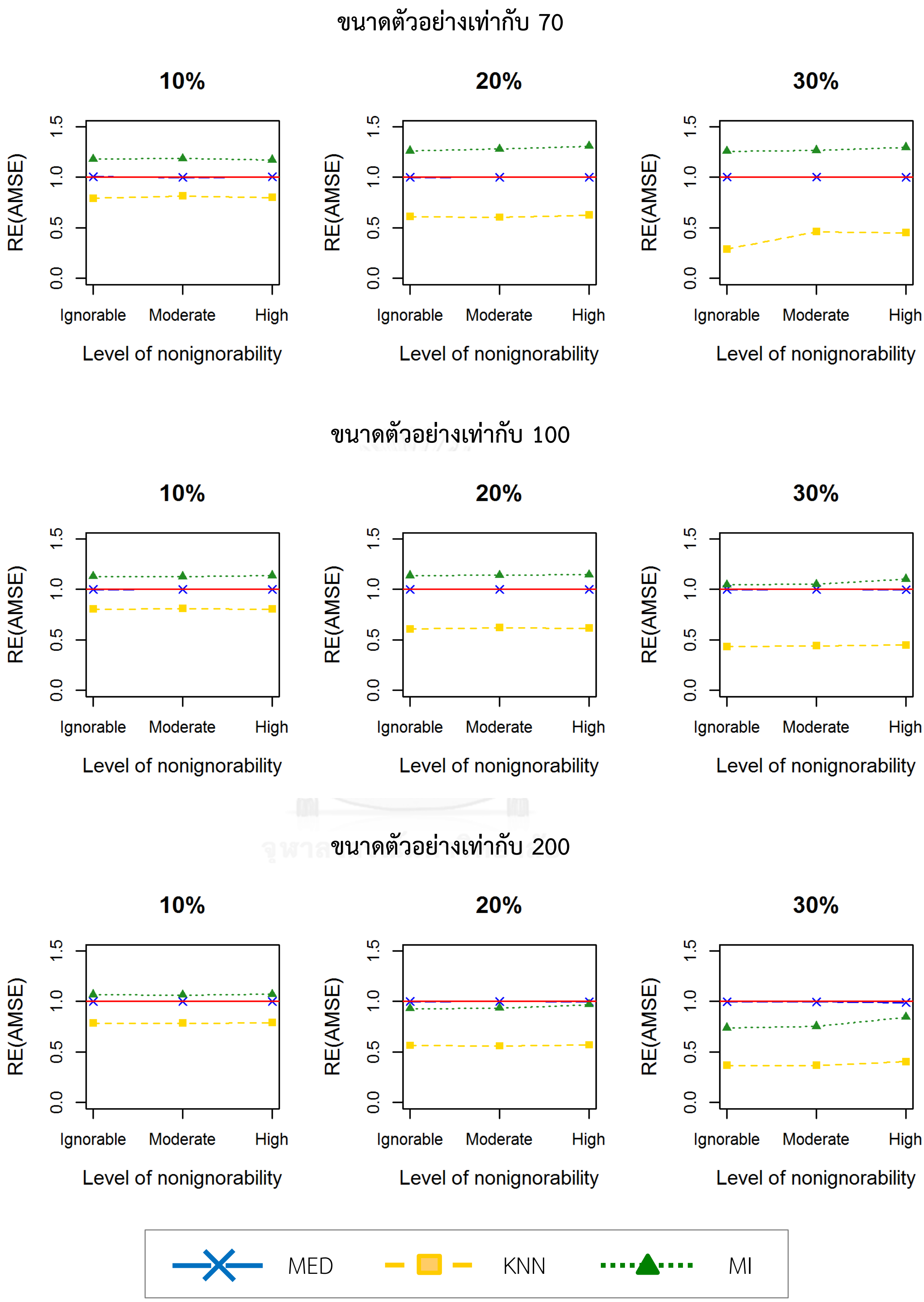
รูปที่ 5 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(MSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1

ตารางที่ 8 การเปรียบเทียบค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 (β = 1)

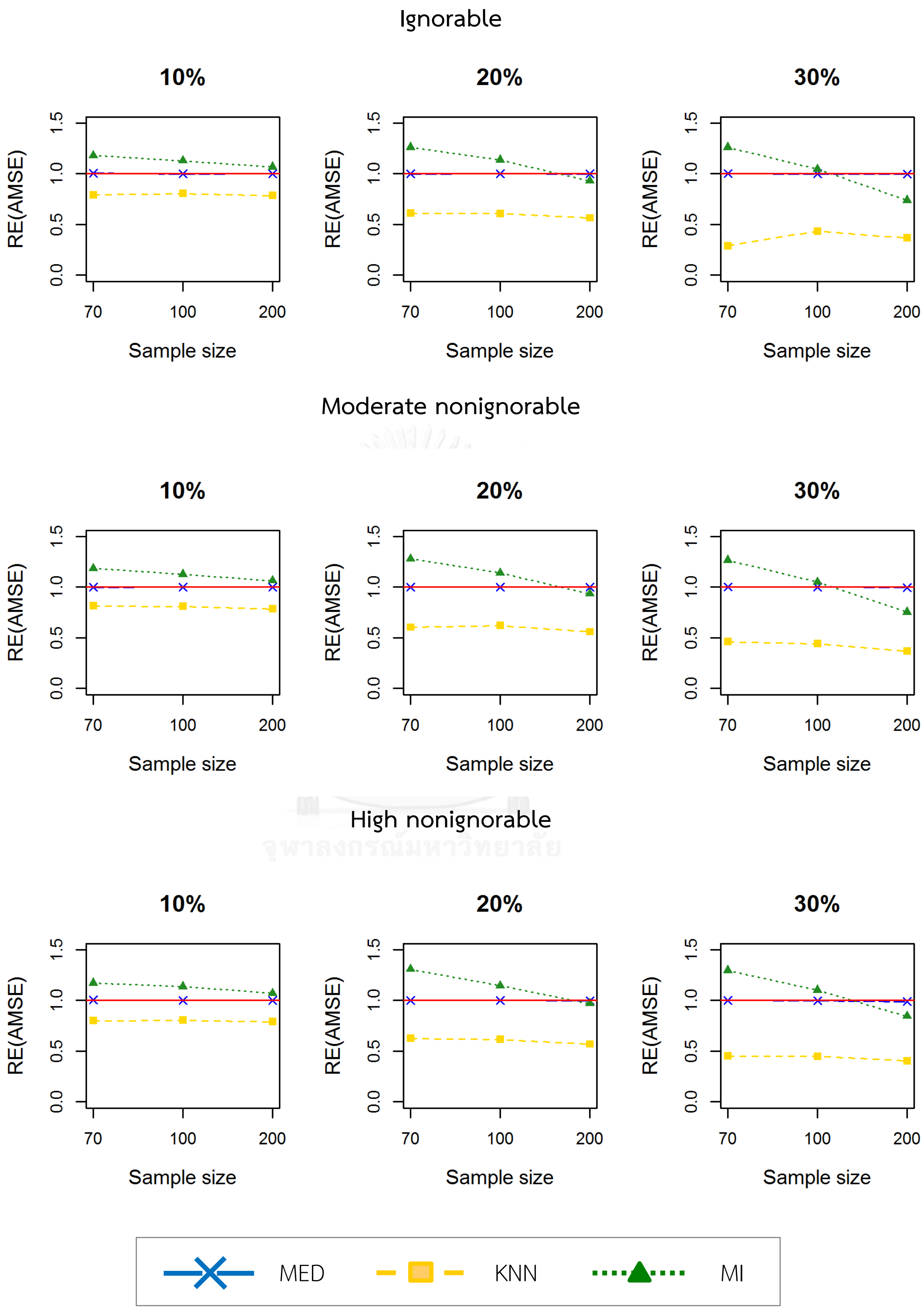
ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	AMSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.205	0.204	0.260	**0.174**
		Moderate nonignorable	0.205	0.205	0.252	**0.173**
		High nonignorable	0.198	0.198	0.248	**0.169**
	ร้อยละ 20	Ignorable	0.203	0.203	0.332	**0.161**
		Moderate nonignorable	0.199	0.199	0.330	**0.156**
		High nonignorable	0.214	0.214	0.341	**0.163**
	ร้อยละ 30	Ignorable	0.199	0.198	0.685	**0.158**
		Moderate nonignorable	0.200	0.200	0.434	**0.158**
		High nonignorable	0.210	0.210	0.465	**0.162**
n=100	ร้อยละ 10	Ignorable	0.113	0.113	0.141	**0.100**
		Moderate nonignorable	0.113	0.113	0.139	**0.100**
		High nonignorable	0.115	0.114	0.143	**0.101**
	ร้อยละ 20	Ignorable	0.112	0.112	0.185	**0.099**
		Moderate nonignorable	0.114	0.114	0.183	**0.100**
		High nonignorable	0.115	0.115	0.188	**0.101**
	ร้อยละ 30	Ignorable	0.113	0.113	0.260	**0.108**
		Moderate nonignorable	0.113	0.113	0.255	**0.108**
		High nonignorable	0.125	0.126	0.280	**0.114**
n=200	ร้อยละ 10	Ignorable	0.049	0.049	0.062	**0.046**
		Moderate nonignorable	0.049	0.049	0.062	**0.046**
		High nonignorable	0.050	0.050	0.063	**0.047**
	ร้อยละ 20	Ignorable	**0.049**	**0.049**	0.087	0.053
		Moderate nonignorable	**0.050**	**0.050**	0.089	0.053
		High nonignorable	**0.053**	**0.053**	0.093	0.055
	ร้อยละ 30	Ignorable	**0.049**	0.050	0.135	0.067
		Moderate nonignorable	**0.051**	**0.051**	0.138	0.067
		High nonignorable	**0.061**	0.062	0.152	0.072

ตารางที่ 9 การเปรียบเทียบค่า RE(AMSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 (β = 1)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(AMSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	1.005	0.791	**1.180**
		Moderate nonignorable	1.000	1.000	0.814	**1.184**
		High nonignorable	1.000	1.003	0.799	**1.171**
	ร้อยละ 20	Ignorable	1.000	1.000	0.611	**1.262**
		Moderate nonignorable	1.000	1.000	0.604	**1.279**
		High nonignorable	1.000	1.001	0.627	**1.308**
	ร้อยละ 30	Ignorable	1.000	1.002	0.290	**1.258**
		Moderate nonignorable	1.000	1.002	0.461	**1.267**
		High nonignorable	1.000	1.000	0.452	**1.296**
n=100	ร้อยละ 10	Ignorable	1.000	1.000	0.804	**1.129**
		Moderate nonignorable	1.000	1.000	0.810	**1.127**
		High nonignorable	1.000	1.001	0.804	**1.137**
	ร้อยละ 20	Ignorable	1.000	1.000	0.607	**1.137**
		Moderate nonignorable	1.000	1.000	0.621	**1.140**
		High nonignorable	1.000	1.001	0.615	**1.146**
	ร้อยละ 30	Ignorable	1.000	0.998	0.434	**1.046**
		Moderate nonignorable	1.000	1.000	0.443	**1.050**
		High nonignorable	1.000	0.997	0.448	**1.102**
n=200	ร้อยละ 10	Ignorable	1.000	1.000	0.785	**1.067**
		Moderate nonignorable	1.000	1.000	0.785	**1.064**
		High nonignorable	1.000	1.000	0.789	**1.070**
	ร้อยละ 20	Ignorable	**1.000**	0.999	0.565	0.929
		Moderate nonignorable	1.000	**1.001**	0.560	0.935
		High nonignorable	**1.000**	0.997	0.569	0.970
	ร้อยละ 30	Ignorable	**1.000**	0.998	0.367	0.739
		Moderate nonignorable	**1.000**	0.995	0.368	0.754
		High nonignorable	**1.000**	0.988	0.403	0.845



รูปที่ 6 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1



รูปที่ 7 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ขนาดตัวอย่าง เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1

สรุปผลส่วนที่ 4.2

ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 เมื่อพิจารณาค่า AMSE พบว่า การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า AMSE น้อยที่สุด แต่เมื่อขนาดตัวอย่างเท่ากับ 200 และร้อยละการสูญหายเท่ากับ 20 และ 30 พบว่า วิธี MEAN และ วิธี MED ให้ค่า AMSE น้อยกว่าวิธีอื่น ๆ และวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า AMSE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มเพิ่มขึ้น เพราะเมื่อข้อมูลมีการสูญหายเพิ่มขึ้นจะทำให้เกิดความคลาดเคลื่อนจากการประมาณค่าสูญหาย และส่งผลให้เกิดความคลาดเคลื่อนต่อการประมาณค่าพารามิเตอร์อีกด้วย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อพิจารณาระดับการสูญหายแบบนอนอิกนอร์เรเบิลทั้ง 3 ระดับ ในกรณีที่ขนาดตัวอย่างเท่ากับ 70 และ 100 ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีค่าใกล้เคียงกัน แต่เมื่อขนาดตัวอย่างเท่ากับ 200 พบว่า ค่า AMSE ของทั้ง 4 วิธี มีแนวโน้มเพิ่มขึ้น เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น นั้นคือ ระดับการสูญหายแบบนอนอิกนอร์เรเบิลมีผลต่อค่า AMSE เมื่อขนาดตัวอย่างมีขนาดใหญ่

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีประมาณค่าสูญหายวิธีอื่น ๆ พบว่า การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพมากที่สุด และที่ขนาดตัวอย่างเท่ากับ 200 ร้อยละการสูญหายเพิ่มขึ้นเป็นร้อยละ 20 และ 30 วิธี MEAN และ วิธี MED จะมีประสิทธิภาพมากกว่า และวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า RE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างใหญ่ขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มลดลงอย่างมาก สำหรับวิธี KNN และวิธี MED ค่า RE จะมีแนวโน้มลดลงเล็กน้อย

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า RE ของวิธี KNN มีแนวโน้มลดลงอย่างชัดเจน และวิธี MI จะมีแนวโน้มลดลงเล็กน้อย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มเพิ่มขึ้นเล็กน้อย ในส่วนของวิธี MED และวิธี KNN ค่า RE ไม่มีการเปลี่ยนแปลงที่ชัดเจน

4.3 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5

ในส่วนนี้การสูญหายเกิดที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 และมีข้อมูลตามขอบเขตการวิจัยที่กำหนด โดยแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายแต่ละวิธีด้วยค่า MSE, RE(MSE), AMSE และ RE(AMSE) ดังแสดงในตารางที่ 10 – 13 ตามลำดับ และนำเสนอกราฟแสดงค่า RE(MSE) และ RE(AMSE) เพื่อให้เข้าใจชัดเจนมากขึ้น

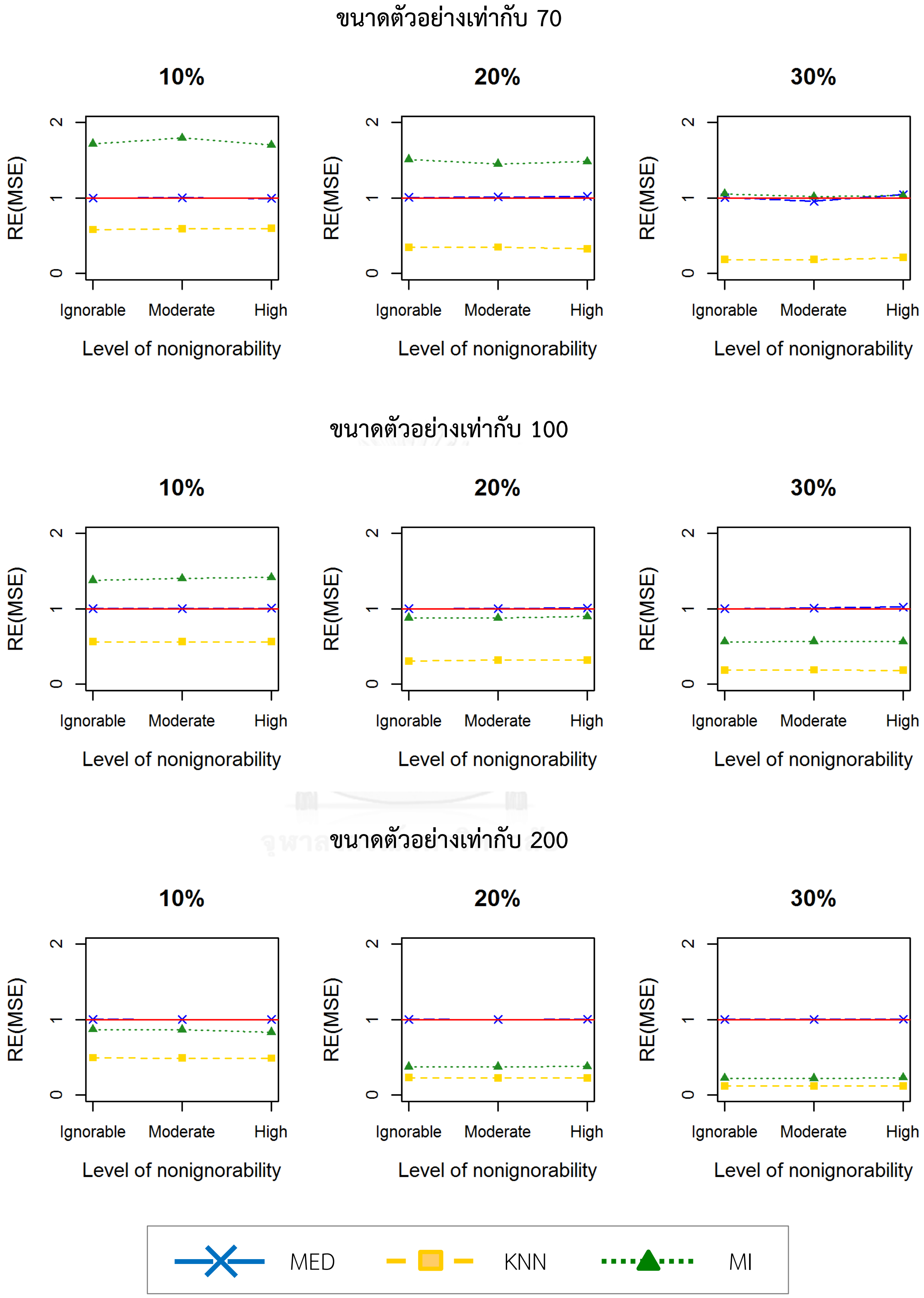
ตารางที่	ค่าที่แสดงในตาราง
ตารางที่ 10	MSE
ตารางที่ 11	RE(MSE)
ตารางที่ 12	AMSE
ตารางที่ 13	RE(AMSE)

ตารางที่ 10 การเปรียบเทียบค่า MSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 (β = 1.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	MSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.330	0.330	0.568	**0.192**
		Moderate nonignorable	0.336	0.335	0.568	**0.187**
		High nonignorable	0.338	0.340	0.570	**0.199**
	ร้อยละ 20	Ignorable	0.347	0.344	1.000	**0.230**
		Moderate nonignorable	0.341	0.336	0.980	**0.235**
		High nonignorable	0.341	0.335	1.043	**0.230**
	ร้อยละ 30	Ignorable	0.390	0.388	2.131	**0.370**
		Moderate nonignorable	0.380	0.398	2.059	**0.374**
		High nonignorable	0.388	**0.371**	1.843	0.377
n=100	ร้อยละ 10	Ignorable	0.178	0.177	0.317	**0.129**
		Moderate nonignorable	0.182	0.182	0.325	**0.130**
		High nonignorable	0.187	0.186	0.332	**0.133**
	ร้อยละ 20	Ignorable	**0.187**	**0.187**	0.611	0.213
		Moderate nonignorable	**0.190**	**0.190**	0.597	0.217
		High nonignorable	0.195	**0.193**	0.610	0.218
	ร้อยละ 30	Ignorable	**0.209**	**0.209**	1.118	0.371
		Moderate nonignorable	0.210	**0.208**	1.110	0.370
		High nonignorable	0.216	**0.211**	1.174	0.383
n=200	ร้อยละ 10	Ignorable	**0.074**	**0.074**	0.151	0.086
		Moderate nonignorable	**0.074**	**0.074**	0.152	0.086
		High nonignorable	**0.072**	**0.072**	0.148	0.087
	ร้อยละ 20	Ignorable	0.077	**0.076**	0.332	0.205
		Moderate nonignorable	**0.077**	**0.077**	0.339	0.205
		High nonignorable	**0.078**	**0.078**	0.348	0.207
	ร้อยละ 30	Ignorable	**0.085**	**0.085**	0.701	0.382
		Moderate nonignorable	**0.085**	**0.085**	0.713	0.382
		High nonignorable	**0.089**	**0.089**	0.747	0.385

ตารางที่ 11 การเปรียบเทียบค่า RE(MSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 (β = 1.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(MSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	0.998	0.580	**1.714**
		Moderate nonignorable	1.000	1.004	0.592	**1.792**
		High nonignorable	1.000	0.994	0.594	**1.698**
	ร้อยละ 20	Ignorable	1.000	1.008	0.346	**1.509**
		Moderate nonignorable	1.000	1.012	0.348	**1.449**
		High nonignorable	1.000	1.018	0.327	**1.480**
	ร้อยละ 30	Ignorable	1.000	1.005	0.183	**1.054**
		Moderate nonignorable	1.000	0.955	0.185	**1.017**
		High nonignorable	1.000	**1.045**	0.211	1.029
n=100	ร้อยละ 10	Ignorable	1.000	1.002	0.562	**1.377**
		Moderate nonignorable	1.000	1.001	0.562	**1.402**
		High nonignorable	1.000	1.007	0.564	**1.413**
	ร้อยละ 20	Ignorable	**1.000**	**1.000**	0.306	0.880
		Moderate nonignorable	1.000	**1.002**	0.318	0.876
		High nonignorable	1.000	**1.010**	0.320	0.897
	ร้อยละ 30	Ignorable	**1.000**	**1.000**	0.187	0.563
		Moderate nonignorable	1.000	**1.009**	0.189	0.567
		High nonignorable	1.000	**1.022**	0.184	0.565
n=200	ร้อยละ 10	Ignorable	1.000	**1.003**	0.494	0.868
		Moderate nonignorable	1.000	**1.001**	0.490	0.865
		High nonignorable	1.000	**1.000**	0.488	0.833
	ร้อยละ 20	Ignorable	1.000	**1.004**	0.231	0.375
		Moderate nonignorable	1.000	**1.001**	0.227	0.376
		High nonignorable	1.000	**1.004**	0.226	0.379
	ร้อยละ 30	Ignorable	1.000	**1.002**	0.122	0.224
		Moderate nonignorable	1.000	**1.003**	0.120	0.223
		High nonignorable	1.000	**1.005**	0.119	0.232



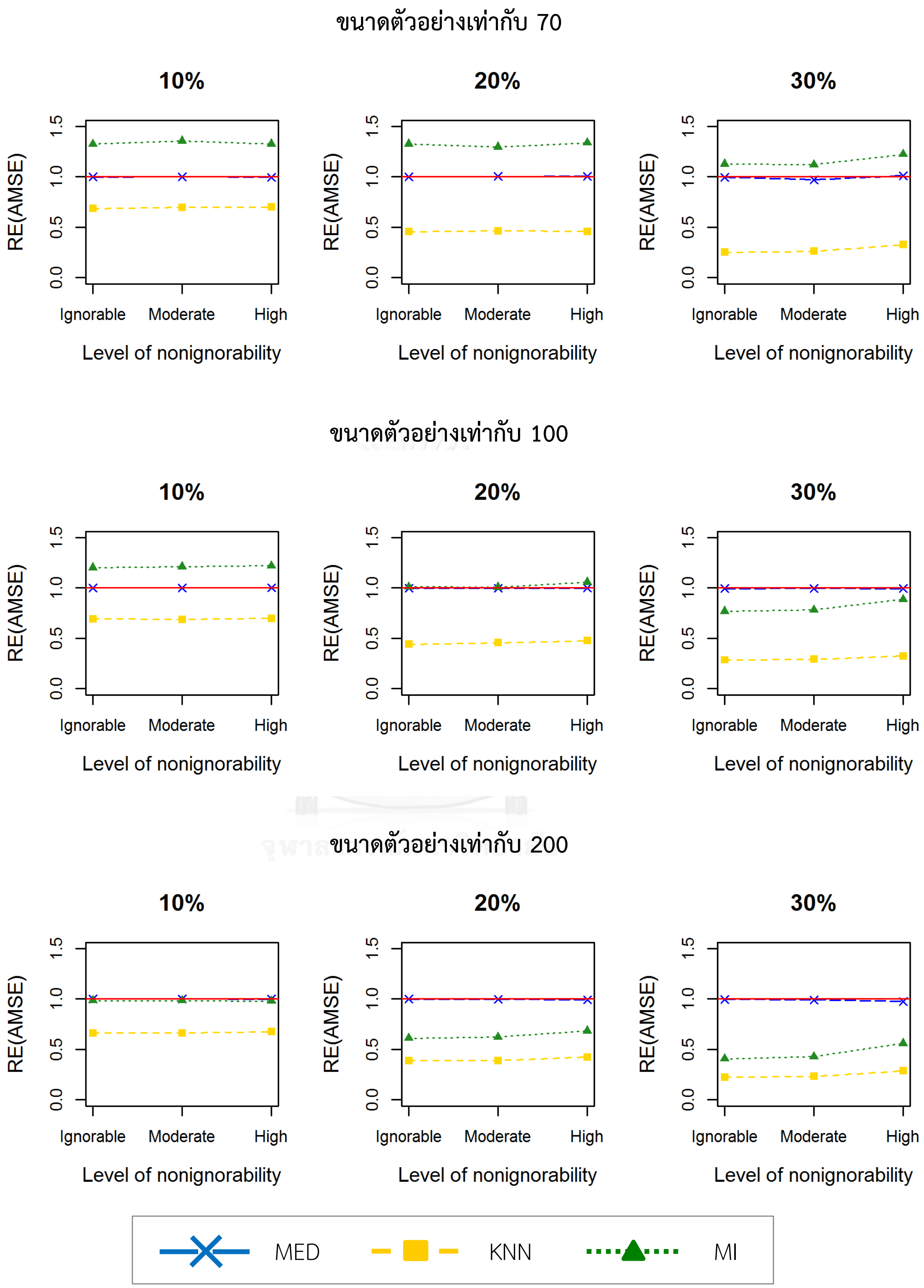
รูปที่ 8 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(MSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5

ตารางที่ 12 การเปรียบเทียบค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 (β = 1.5)

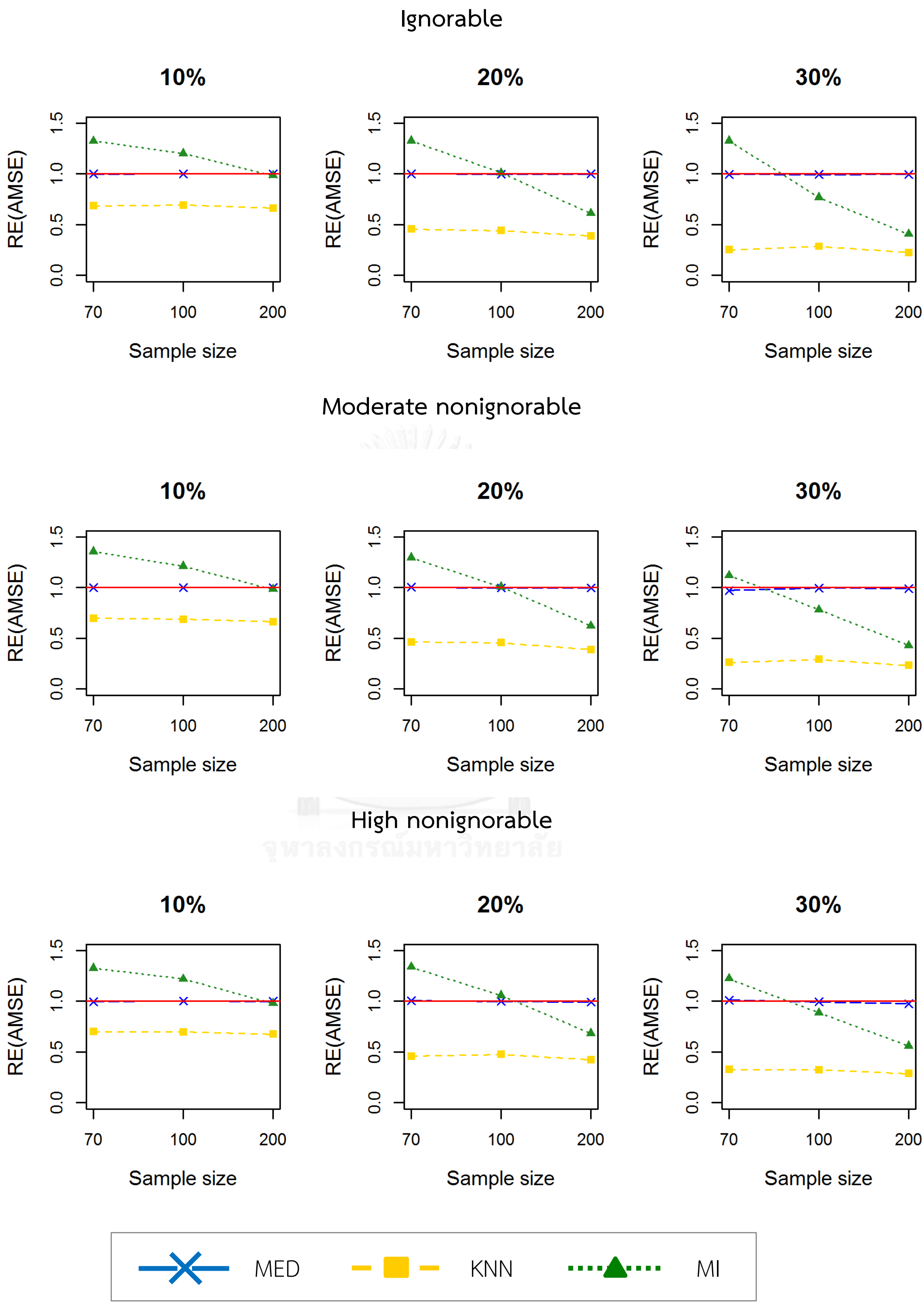
ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	AMSE			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	0.195	0.195	0.285	**0.147**
		Moderate nonignorable	0.197	0.197	0.283	**0.146**
		High nonignorable	0.198	0.198	0.282	**0.149**
	ร้อยละ 20	Ignorable	0.191	0.191	0.419	**0.144**
		Moderate nonignorable	0.191	0.190	0.410	**0.147**
		High nonignorable	0.201	0.200	0.437	**0.150**
	ร้อยละ 30	Ignorable	0.197	0.197	0.776	**0.174**
		Moderate nonignorable	0.197	0.203	0.752	**0.176**
		High nonignorable	0.227	0.225	0.696	**0.186**
n=100	ร้อยละ 10	Ignorable	0.110	0.110	0.159	**0.091**
		Moderate nonignorable	0.111	0.112	0.162	**0.092**
		High nonignorable	0.114	0.114	0.163	**0.093**
	ร้อยละ 20	Ignorable	0.109	0.109	0.247	**0.108**
		Moderate nonignorable	0.110	0.110	0.241	**0.109**
		High nonignorable	0.119	0.119	0.250	**0.113**
	ร้อยละ 30	Ignorable	**0.113**	**0.113**	0.395	0.147
		Moderate nonignorable	**0.115**	0.116	0.392	0.147
		High nonignorable	**0.140**	0.141	0.430	0.157
n=200	ร้อยละ 10	Ignorable	**0.049**	**0.049**	0.074	0.050
		Moderate nonignorable	**0.049**	**0.049**	0.074	0.050
		High nonignorable	**0.050**	**0.050**	0.074	0.051
	ร้อยละ 20	Ignorable	**0.048**	**0.048**	0.125	0.079
		Moderate nonignorable	**0.050**	**0.050**	0.128	0.080
		High nonignorable	**0.057**	0.058	0.135	0.084
	ร้อยละ 30	Ignorable	**0.051**	**0.051**	0.227	0.125
		Moderate nonignorable	**0.054**	0.055	0.234	0.126
		High nonignorable	**0.075**	0.077	0.261	0.134

ตารางที่ 13 การเปรียบเทียบค่า RE(AMSE) ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 (β = 1.5)

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	RE(AMSE)			
			MEAN	MED	KNN	MI
n=70	ร้อยละ 10	Ignorable	1.000	0.999	0.685	**1.324**
		Moderate nonignorable	1.000	1.002	0.698	**1.355**
		High nonignorable	1.000	0.997	0.700	**1.326**
	ร้อยละ 20	Ignorable	1.000	1.001	0.457	**1.326**
		Moderate nonignorable	1.000	1.005	0.464	**1.296**
		High nonignorable	1.000	1.005	0.459	**1.337**
	ร้อยละ 30	Ignorable	1.000	0.996	0.253	**1.128**
		Moderate nonignorable	1.000	0.971	0.262	**1.120**
		High nonignorable	1.000	1.010	0.326	**1.223**
n=100	ร้อยละ 10	Ignorable	1.000	1.000	0.692	**1.202**
		Moderate nonignorable	1.000	1.000	0.688	**1.212**
		High nonignorable	1.000	1.002	0.696	**1.219**
	ร้อยละ 20	Ignorable	1.000	0.997	0.442	**1.010**
		Moderate nonignorable	1.000	0.997	0.456	**1.008**
		High nonignorable	1.000	0.999	0.476	**1.057**
	ร้อยละ 30	Ignorable	**1.000**	0.994	0.286	0.770
		Moderate nonignorable	**1.000**	0.995	0.293	0.783
		High nonignorable	**1.000**	0.993	0.325	0.887
n=200	ร้อยละ 10	Ignorable	1.000	**1.001**	0.663	0.984
		Moderate nonignorable	**1.000**	**1.000**	0.663	0.984
		High nonignorable	**1.000**	0.999	0.675	0.979
	ร้อยละ 20	Ignorable	**1.000**	0.999	0.388	0.612
		Moderate nonignorable	**1.000**	0.998	0.389	0.622
		High nonignorable	**1.000**	0.992	0.425	0.684
	ร้อยละ 30	Ignorable	**1.000**	0.997	0.225	0.407
		Moderate nonignorable	**1.000**	0.992	0.232	0.428
		High nonignorable	**1.000**	0.977	0.288	0.560



รูปที่ 9 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5



รูปที่ 10 การเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE(AMSE) กับ ขนาดตัวอย่าง เมื่อตัวแปรอิสระที่สูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5

สรุปผลส่วนที่ 4.3

ผลการเปรียบเทียบวิธีการประมาณค่าสูญหาย กรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 เมื่อพิจารณาค่า AMSE พบว่า กรณีที่ขนาดตัวอย่างเท่ากับ 70 การประมาณค่าสูญหายด้วยวิธี MI ให้ค่า AMSE น้อยที่สุด กรณีที่ขนาดตัวอย่างเท่ากับ 100 มีร้อยละการสูญหายเท่ากับ 10 วิธีการประมาณค่าสูญหายด้วยวิธี MI ให้ค่า AMSE น้อยที่สุด แต่เมื่อร้อยละการสูญหายเท่ากับ 20 และ 30 พบว่า วิธี MEAN และ วิธี MED ให้ค่า AMSE น้อยกว่า และกรณีที่ขนาดตัวอย่างเท่ากับ 200 พบว่า วิธี MEAN และ วิธี MED ให้ค่า AMSE น้อยกว่าวิธีอื่น ๆ และวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า AMSE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มเพิ่มขึ้น เพราะเมื่อข้อมูลมีการสูญหายเพิ่มขึ้นจะทำให้เกิดความคลาดเคลื่อนจากการประมาณค่าสูญหาย และส่งผลให้เกิดความคลาดเคลื่อนต่อการประมาณค่าพารามิเตอร์อีกด้วย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มเพิ่มขึ้น เนื่องจากการสูญหายเกิดขึ้นที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 นั้นคือ ตัวแปรอิสระมีความสัมพันธ์กับตัวแปรตามสูง การสูญหายของตัวแปรอิสระจึงส่งผลต่อการประมาณค่าพารามิเตอร์มากขึ้น

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีประมาณค่าสูญหายวิธีอื่น ๆ พบว่า การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพมากที่สุด ที่ขนาดตัวอย่างเท่ากับ 70 และที่ขนาดตัวอย่างเท่ากับ 100 ร้อยละการสูญหายเท่ากับร้อยละ 10 วีธี MI ก็ยังมีประสิทธิภาพมากกว่าวิธี MEAN แต่เมื่อร้อยละการสูญหายเพิ่มขึ้นเป็นร้อยละ 20 และ 30 วิธี MEAN จะมีประสิทธิภาพมากกว่า และที่ขนาดตัวอย่างเท่ากับ 200 วิธี MEAN มีประสิทธิภาพมากที่สุด และสำหรับวิธี KNN มีประสิทธิภาพน้อยที่สุดในทุกสถานการณ์

การสรุปปัจจัยที่มีผลต่อค่า RE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างใหญ่ขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มลดลงอย่างมาก สำหรับวิธี KNN และวิธี MED ค่า RE จะมีแนวโน้มลดลงเล็กน้อย

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า RE ของวิธี KNN มีแนวโน้มลดลงอย่างชัดเจน เมื่อร้อยละการสูญหายเพิ่มขึ้น และวิธี MI จะมีแนวโน้มลดลงเล็กน้อย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มเพิ่มขึ้นเล็กน้อย ในส่วนของวิธี MED และวิธี KNN ค่า RE ไม่มีการเปลี่ยนแปลงที่ชัดเจน

จากผลการวิจัยในส่วนที่ 4.1 – 4.3 สามารถสรุปผล ได้ดังนี้

ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า AMSE สามารถสรุปผลได้ดังนี้

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 การประมาณค่าสูญหายด้วยวิธี MI จะให้ประสิทธิภาพดีที่สุด รองลงมาคือ วิธี MEAN และ วิธี MED

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 ขนาดตัวอย่างเท่ากับ 70 และ 100 การประมาณค่าสูญหายด้วยวิธี MI จะเป็นวิธีที่ดี แต่เมื่อมีขนาดตัวอย่างเท่ากับ 200 วิธี MEAN และ วิธี MED จะให้ประสิทธิภาพที่ดีกว่า

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพที่ขนาดตัวอย่าง เท่ากับ 70 และ 100 แต่กรณีที่ขนาดตัวอย่างเท่ากับ 100 ร้อยละการสูญหายสูงขึ้นที่ร้อยละ 30 วิธี MEAN และ วิธี MED จะให้ค่า AMSE ต่ำกว่า และกรณีที่ขนาดตัวอย่างเท่ากับ 200 วิธี MEAN และ วิธี MED จะมีประสิทธิภาพกว่าวิธีอื่น ๆ

การสรุปปัจจัยที่มีผลต่อค่า AMSE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า AMSE มีแนวโน้มเพิ่มขึ้น เพราะเมื่อข้อมูลมีการสูญหายเพิ่มขึ้นจะทำให้เกิดความคลาดเคลื่อนจากการประมาณค่าสูญหาย และส่งผลให้เกิดความคลาดเคลื่อนต่อการประมาณค่าพารามิเตอร์อีกด้วย

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น ค่า AMSE มีแนวโน้มเพิ่มขึ้น แสดงว่าการสูญหายแบบนอนอิกนอร์เรเบิลส่งผลต่อการประมาณค่าสัมประสิทธิ์การถดถอยหรือการวิเคราะห์ข้อมูลสูงกว่าการสูญหายแบบสุ่ม

4. ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระ

เมื่อข้อมูลสูญหายเกิดขึ้นที่ตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยสูง จะทำให้ค่า AMSE มีแนวโน้มเพิ่มขึ้น เนื่องจากค่าสัมประสิทธิ์การถดถอยที่เพิ่มขึ้นแสดงถึงความสัมพันธ์ของตัวแปรอิสระต่อตัวแปรตาม หรือแสดงถึงน้ำหนักของตัวแปรอิสระที่เพิ่มขึ้น จึงส่งผลให้เกิดความคลาดเคลื่อนในการประมาณค่าสัมประสิทธิ์การถดถอยด้วย

ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยค่า RE สามารถสรุปผลได้ดังนี้

ค่า RE เป็นอัตราส่วนระหว่างค่า AMSE ของวิธี MEAN กับวิธีประมาณค่าสูญหายวิธีอื่น ๆ นั้นคือ เป็นการเปรียบเทียบว่าการประมาณค่าสูญหายแต่ละวิธีมีประสิทธิภาพเป็นอย่างไรเมื่อเทียบกับการประมาณค่าสูญหายด้วยวิธี MEAN สามารถสรุปผลได้ดังนี้

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 การประมาณค่าสูญหายด้วยวิธี MI จะให้ประสิทธิภาพดีที่สุด ในทุกสถานการณ์

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 ขนาดตัวอย่างเท่ากับ 70 และ 100 การประมาณค่าสูญหายด้วยวิธี MI จะเป็นวิธีที่ดี แต่เมื่อมีขนาดตัวอย่างเท่ากับ 200 วิธี MEAN และ วิธี MED จะให้ประสิทธิภาพที่ดีกว่า

กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพที่ขนาดตัวอย่าง เท่ากับ 70 และ 100 แต่กรณีที่ขนาดตัวอย่างเท่ากับ 100 ร้อยละการสูญหายสูงขึ้นที่ร้อยละ 20 และ 30 วิธี MEAN จะมีประสิทธิภาพมากกว่า และกรณีที่ขนาดตัวอย่างเท่ากับ 200 วิธี MEAN จะมีประสิทธิภาพกว่าวิธีอื่น

ในทุกกรณี วิธี MED จะมีค่าใกล้เคียงกับวิธี MEAN เนื่องจากเป็นการแทนที่ค่าสูญหายด้วยค่าเฉลี่ยเหมือนกัน และสำหรับ วิธี KNN นั้น มีค่า RE น้อยกว่า 1 ในทุกสถานการณ์ นั้นคือมีประสิทธิภาพน้อยที่สุด

การสรุปปัจจัยที่มีผลต่อค่า RE ได้ดังนี้

1. ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างใหญ่ขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มลดลงอย่างมาก และสำหรับวิธี KNN และวิธี MED นั้นค่า RE มีแนวโน้มลดลงเล็กน้อย กล่าวได้ว่า ขนาดตัวอย่างมีผลต่อการประมาณค่าสูญหายด้วยวิธี MI มากที่สุด

2. ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า RE ของวิธี KNN มีแนวโน้มลดลงอย่างชัดเจน และวิธี MI จะมีแนวโน้มลดลงเล็กน้อย ในส่วนของวิธี MED ไม่มีการเปลี่ยนแปลงที่ชัดเจน

3. ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น พบว่า ค่า RE ของวิธี MI มีแนวโน้มเพิ่มขึ้น ในส่วนของวิธี MED และวิธี KNN ค่า RE ไม่มีการเปลี่ยนแปลงที่ชัดเจน

4. ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระ

เมื่อข้อมูลสูญหายเกิดขึ้นที่ตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยสูงขึ้น พบว่า ค่า RE ของวิธี MI และวิธี KNN มีแนวโน้มลดลงเล็กน้อย ส่วนวิธี MED ค่า RE ไม่มีการเปลี่ยนแปลงที่ชัดเจน

บทที่ 5

สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรอิสระที่มีการสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม ที่มีตัวแปรอิสระ 3 ตัว โดยวิธีการประมาณค่าสูญหายที่ใช้ในงานวิจัยนี้ คือ วิธี Mean Imputation (MEAN) วิธี Median Imputation (MED) วิธี K-Nearest Neighbor (KNN) และ วิธี Multiple Imputation (MI) เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการประมาณค่าสูญหายคือ ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) โดยมีการจำลองสถานการณ์ทั้งหมด 81 สถานการณ์ ตามขอบเขตการวิจัยดังนี้

1. ตัวแปรอิสระมีการแจกแจงแบบปกติมาตรฐาน (N(0,1)) และตัวแปรอิสระทั้ง 3 ตัว ไม่มีความสัมพันธ์กัน
2. กำหนดค่าสัมประสิทธิ์การถดถอยเริ่มต้น $\beta_1 = 0.5,\ \beta_2 = 1$ และ $\beta_3 = 1.5$ โดยมีการสูญหายทั้ง 3 กรณี การสูญหายเกิดขึ้นที่ตัวแปรอิสระตัวใดตัวหนึ่ง
3. ขนาดตัวอย่าง (n) คือ 70, 100 และ 200
4. การสูญหายของข้อมูลเกิดขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง และเป็นการสูญหายแบบนอนอิกนอร์เรเบิล โดยกำหนดสัดส่วนการสูญหายของตัวแปรอิสระ คิดเป็นร้อยละโดยเฉลี่ย คือ ร้อยละ 10 ร้อยละ 20 และ ร้อยละ 30
5. ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล แบ่งเป็น 3 ระดับ คือ การสูญหายแบบอิกนอร์เรเบิล การสูญหายแบบนอนอิกนอร์เรเบิลระดับปานกลาง และ การสูญหายแบบนอนอิกนอร์เรเบิลระดับสูง
6. การจำลองข้อมูลให้มีสถานการณ์ที่แตกต่างกันตามข้อกำหนดข้างต้น ใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ โดยใช้โปรแกรม R

5.1 ผลการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

จากการเปรียบเทียบวิธีการประมาณค่าสูญหายทั้ง 4 วิธี โดยพิจารณาจากค่า AMSE พบว่า สำหรับ กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 0.5 ทุกขนาดตัวอย่าง ทุกระดับการสูญหาย การประมาณค่าสูญหายด้วยวิธี MI จะให้ประสิทธิภาพดีที่สุด กรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1 ขนาดตัวอย่างเท่ากับ 70 และ 100 การประมาณค่าสูญหายด้วยวิธี MI จะเป็นวิธีที่ดี แต่เมื่อมีขนาดตัวอย่างเท่ากับ 200 วิธี MEAN และ วิธี MED จะให้ประสิทธิภาพที่ดีกว่ากรณีที่การสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอย เท่ากับ 1.5 การประมาณค่าสูญหายด้วยวิธี MI จะมีประสิทธิภาพที่ขนาด

ตัวอย่าง เท่ากับ 70 และ 100 แต่กรณีที่ขนาดตัวอย่างเท่ากับ 100 ร้อยละการสูญหายสูงขึ้นที่ร้อยละ 30 วิธี MEAN และ วิธี MED จะให้ค่า AMSE ต่ำกว่า และกรณีที่ขนาดตัวอย่างเท่ากับ 200 วิธี MEAN และ วิธี MED จะมีประสิทธิภาพกว่าวิธีอื่น ๆ

5.2 ปัจจัยที่มีผลต่อค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

5.2.1 ขนาดตัวอย่าง

เมื่อขนาดตัวอย่างเพิ่มขึ้น ส่งผลให้ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี มีแนวโน้มลดลง เนื่องจากขนาดตัวอย่างที่เพิ่มขึ้นจะช่วยเพิ่มประสิทธิภาพในการประมาณค่าพารามิเตอร์ได้

5.2.2 ร้อยละการสูญหาย

เมื่อร้อยละการสูญหายเพิ่มขึ้น ส่งผลให้ค่า AMSE มีแนวโน้มเพิ่มขึ้น เพราะเมื่อข้อมูลมีการสูญหายเพิ่มขึ้นจะทำให้เกิดความคลาดเคลื่อนจากการประมาณค่าสูญหาย และส่งผลให้เกิดความคลาดเคลื่อนต่อการประมาณค่าพารามิเตอร์อีกด้วย

5.2.3 ระดับการสูญหายแบบนอนอิกนอร์เรเบิล

เมื่อระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มขึ้น ค่า AMSE มีแนวโน้มเพิ่มขึ้น แสดงว่าการสูญหายแบบนอนอิกนอร์เรเบิลส่งผลต่อการประมาณค่าสัมประสิทธิ์การถดถอยหรือการวิเคราะห์ข้อมูลสูงกว่าการสูญหายแบบสุ่ม

5.2.4 ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระ

เมื่อข้อมูลสูญหายเกิดขึ้นที่ตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยสูง จะทำให้ค่า AMSE มีแนวโน้มเพิ่มขึ้น เนื่องจากค่าสัมประสิทธิ์การถดถอยที่เพิ่มขึ้นแสดงถึงความสัมพันธ์ของตัวแปรอิสระต่อตัวแปรตาม หรือแสดงถึงน้ำหนักของตัวแปรอิสระที่เพิ่มขึ้น จึงส่งผลให้เกิดความคลาดเคลื่อนในการประมาณค่าสัมประสิทธิ์การถดถอย

5.3 สรุปความแตกต่างของแต่ละวิธีการประมาณค่าสูญหาย

5.3.1 วิธี Mean Imputation

วิธี MEAN เป็นวิธีการประมาณค่าสูญหายที่เหมาะกับข้อมูลที่มีขนาดใหญ่ และจะใช้ได้ดีเมื่อการสูญหายเป็นแบบอิกนอร์เรเบิล

5.3.2 วิธี Median Imputation

วิธี MED เป็นวิธีการประมาณค่าสูญหายที่เหมาะกับข้อมูลที่มีขนาดใหญ่ และจะใช้ได้ดีเมื่อการสูญหายเป็นแบบนอนอิกนอร์เรเบิล

5.3.3 วิธี K-Nearest Neighbor Imputation (KNN)

วิธี KNN เป็นวิธีการประมาณค่าที่มีการแทนที่ค่าสูญหายมากกว่า 1 ค่า จึงมีความซับซ้อนกว่าวิธี MEAN และวิธี MED จากผลการทดลองพบว่า วิธี KNN เหมาะสำหรับกรณีที่การสูญหาย

เกิดขึ้นที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอยต่ำ $(\beta=0.5)$ แต่ก็ยังมีประสิทธิภาพน้อยกว่าอีก 3 วิธี ดังนั้น ในทางปฏิบัติจึงควรเลือกใช้วิธีการที่มีความยุ่งยากน้อยกว่า

5.3.4 วิธี Multiple Imputation

วิธี MI เป็นวิธีการประมาณค่าสูญหายที่ยุ่งยากที่สุดจากทั้ง 4 วิธี ผลการทดลอง พบว่า วิธี MI จะมีประสิทธิภาพที่สุด เมื่อการสูญหายเกิดขึ้นในตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยต่ำ $(\beta=0.5)$ และจะมีประสิทธิภาพเมื่อข้อมูลมีขนาดเล็ก – กลาง

ผลการเปรียบเทียบประสิทธิภาพของแต่ละวิธีสรุปได้ดังตารางที่ 5.1

ตารางที่ 14 สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี กรณีที่การสูญหายเกิดขึ้นที่ตัวแปรมีค่าสัมประสิทธิ์การถดถอยระดับต่าง ๆ

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	กรณีที่การสูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ		
			0.5	1	1.5
n=70	ร้อยละ 10	Ignorable	MI	MI	MI
		Moderate nonignorable	MI	MI	MI
		High nonignorable	MI	MI	MI
	ร้อยละ 20	Ignorable	MI	MI	MI
		Moderate nonignorable	MI	MI	MI
		High nonignorable	MI	MI	MI
	ร้อยละ 30	Ignorable	MI	MI	MI
		Moderate nonignorable	MI	MI	MI
		High nonignorable	MI	MI	MI
n=100	ร้อยละ 10	Ignorable	MI	MI	MI
		Moderate nonignorable	MI	MI	MI
		High nonignorable	MI	MI	MI
	ร้อยละ 20	Ignorable	MI	MI	MI
		Moderate nonignorable	MI	MI	MI
		High nonignorable	MI	MI	MI
	ร้อยละ 30	Ignorable	MI	MI	MEAN, MED
		Moderate nonignorable	MI	MI	MEAN, MED
		High nonignorable	MI	MI	MEAN, MED

ตารางที่ 14 (ต่อ) สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี กรณีที่การสูญหายเกิดขึ้นที่ตัวแปรมีค่าสัมประสิทธิ์การถดถอยระดับต่าง ๆ

ขนาดตัวอย่าง (n)	ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบนอนอิกนอร์เรเบิล	กรณีที่การสูญหายมีค่าสัมประสิทธิ์การถดถอย เท่ากับ		
			0.5	1	1.5
n=200	ร้อยละ 10	Ignorable	MI	MI	MEAN, MED
		Moderate nonignorable	MI	MI	MEAN, MED
		High nonignorable	MI	MI	MEAN, MED
	ร้อยละ 20	Ignorable	MI	MEAN, MED	MEAN, MED
		Moderate nonignorable	MI	MEAN, MED	MEAN, MED
		High nonignorable	MI	MEAN, MED	MEAN, MED
	ร้อยละ 30	Ignorable	MI	MEAN, MED	MEAN, MED
		Moderate nonignorable	MI	MEAN, MED	MEAN, MED
		High nonignorable	MI	MEAN, MED	MEAN, MED

5.4 ข้อเสนอแนะ

5.4.1 ด้านการนำไปใช้ประโยชน์

เพื่อเป็นแนวทางในการเลือกใช้วิธีการประมาณค่าสูญหายที่เหมาะสม เมื่อเกิดปัญหาข้อมูลสูญหายที่ตัวแปรอิสระที่เป็นข้อมูลเชิงปริมาณ และเป็นการสูญหายแบบนอนอิกนอร์เรเบิล สำหรับการวิเคราะห์การถดถอยโลจิสติกแบบ 2 กลุ่ม ในการทดลองได้จำลองสถานการณ์ที่แตกต่างกัน เพื่อวิเคราะห์ว่าวิธีการใดเหมาะสมกับสถานการณ์ใด สามารถสรุปได้ว่า เมื่อการสูญหายเกิดขึ้นที่ตัวแปรอิสระมีค่าสัมประสิทธิ์การถดถอยต่ำ ควรเลือกใช้การประมาณค่าสูญหายด้วยวิธี MI และเมื่อพิจารณาจากขนาดตัวอย่าง พบว่า วิธี MI จะมีประสิทธิภาพที่ดีมากกว่าวิธีอื่น ๆ เมื่อขนาดตัวอย่างเล็ก และกรณีที่ข้อมูลมีขนาดใหญ่ ควรเลือกใช้การประมาณค่าสูญหายด้วยวิธี MED เนื่องจากกรณีที่ตัวอย่างมีขนาดใหญ่ วิธีการประมาณค่าสูญหายทั้ง 4 วิธีให้ผลไม่แตกต่างกัน จึงสามารถเลือกใช้วิธีการที่ไม่

ยุ่งยากซับซ้อน ซึ่งจะง่ายต่อการนำไปใช้วิเคราะห์ในสถานการณ์จริง และวิธี MED นั้นให้ประสิทธิภาพที่ดีกว่าวิธีอื่น ๆ เมื่อข้อมูลมีการสูญหายแบบนอนอิกนอร์เรเบิล ซึ่งเป็นปัญหาที่สำคัญของการศึกษาวิจัยครั้งนี้

ในการวิเคราะห์ข้อมูลจริงนั้น ก่อนที่จะนำข้อมูลมาวิเคราะห์ด้วยตัวแบบการถดถอยโลจิสติกแบบ 2 กลุ่ม ควรทำการแปลงค่าข้อมูลให้เป็นค่ามาตรฐานก่อน และตรวจสอบข้อมูลว่ามีความสัมพันธ์กันหรือไม่ ก่อนการวิเคราะห์ข้อมูลตามแนวทางการเลือกใช้วิธีการประมาณค่าสูญหายของงานวิจัยชิ้นนี้

5.4.2 ด้านการศึกษาวิจัย

เพื่อเป็นแนวทางในการทำการศึกษาเพิ่มเติมนอกเหนือจากขอบเขตของการวิจัยในครั้งนี้ โดยอาจทำการศึกษากรณีอื่น ๆ ดังนี้

1. งานวิจัยครั้งนี้ศึกษากรณีที่ตัวแปรอิสระมีการแจกแจงแบบปกติมาตรฐาน สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ตัวแปรอิสระมีการแจงแจงแบบอื่น ๆ
2. การสูญหายเกิดขึ้นที่ตัวแปรอิสระที่เป็นข้อมูลเชิงปริมาณ ควรทำการศึกษาเพิ่มเติมกรณีที่ข้อมูลที่สูญหายเป็นข้อมูลเชิงคุณภาพ หรือกรณีที่ตัวแปรอิสระมีทั้งข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ
3. ศึกษาเพิ่มเติมในกรณีที่ตัวแปรอิสระเกิดการสูญหายมากกว่าหนึ่งตัว
4. ศึกษาเพิ่มเติมในกรณีที่ตัวแปรอิสระมีมากกว่า 3 ตัว
5. ศึกษาเพิ่มเติมในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันในระดับต่าง ๆ
6. ในการศึกษาครั้งนี้ การประมาณค่าสูญหายด้วยวิธี MI ได้กำหนดให้ตัวแปรอิสระที่เกิดการสูญหายมีการแจกแจงแบบปกติ แต่ในความเป็นจริงตัวแปรอิสระอาจมีการแจกแจงลักษณะอื่น ๆ ซึ่งอาจจะทำให้ผลลัพธ์ที่ได้แตกต่างจากผลการวิเคราะห์ในครั้งนี้
7. ในการวิจัยครั้งนี้วิธี MEAN และวิธี MED ให้ค่า AMSE ใกล้เคียงกัน อาจจะเป็นผลมากจากการกำหนดให้ตัวแปรอิสระมีการแจกแจงแบบปกติมาตรฐาน ทำให้ค่าเฉลี่ยเลขคณิตและค่ามัธยฐานมีค่าใกล้เคียงกัน ดังนั้นถ้าตัวแปรอิสระไม่ได้มีการแจงแจงแบบปกติมาตรฐาน อาจจะทำให้ได้ผลลัพธ์ที่เปลี่ยนไป

รายการอ้างอิง

ภาษาไทย

กัลยา วานิชย์บัญชา. (2550). *การวิเคราะห์ข้อมูลหลายตัวแปร*. พิมพ์ครั้งที่ 2. กรุงเทพฯ: ธรรมสาร.

ประลองพล ประสงค์พร. (2551). *การประมาณค่าพารามิเตอร์ในตัวแบบความถดถอยโลจิสติกเมื่อมีค่าสูญหาย*. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R., X. (2013). *Applied logistic regression*. Third Edition, New Jersey: John Wiley & Sons.

Jönsson, P., & Wohlin, C. (2006). Benchmarking k-nearest neighbor imputation with homogeneous likert data. *Empirical Software Engineering,* 11, 463-489.

Little, R., J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Second Edition, New Jersey: John Wiley & Sons.

ภาคผนวก

รายละเอียดของโปรแกรมที่ใช้ในงานวิจัย

การศึกษาวิจัยในครั้งนี้ได้ทำการจำลองข้อมูลตามสถานการณ์ที่กำหนด และประมาณค่าสูญหายแต่ละวิธี โดยใช้โปรแกรม R เวอร์ชัน 3.3.0 ซึ่งมีคำสั่ง ดังต่อไปนี้

```
###Function for mean imputation method###
	MEAN.impute<-function(Xmiss)
	{	NEW_X_MEAN<-Xmiss
		Xbar<-mean(NEW_X_MEAN,na.rm=TRUE)
		NEW_X_MEAN[is.na(NEW_X_MEAN)]<-Xbar
		return(NEW_X_MEAN)
	}
###Function for median imputation method ###
	MED.impute<-function(Xmiss)
	{	NEW_X_MED<-Xmiss
		Xmedian<-median(NEW_X_MED,na.rm=TRUE)
		NEW_X_MED[is.na(NEW_X_MED)]<-Xmedian
		return(NEW_X_MED)
	}
###Standardize X & Y###
	std_obs<-function(Y,X1,X2)
	{
		Z_X1<-(X1-mean(X1))/sd(X1)
		Z_X2<-(X2-mean(X2))/sd(X2)
		Z_Y<-(Y-mean(Y))/sqrt(mean(Y)*(1-mean(Y)))
		Z_data<-data.frame(Z_Y,Z_X1,Z_X2)
		return(Z_data)
	}
###Function for KNN imputation method ###
	KNN.impute<-function(Std_obs,X_obs,No_Obs,No_mis)
	{	k<-round(sqrt(No_Obs),0)
		KNN_bar<-c()
		for(nn in No_Obs+1:No_mis)
		{	Ynonmis<-Std_obs[1:No_Obs,1]
			Xnonmis<-Std_obs[1:No_Obs,2:3]
```

```
                    X_nonmiss<-X_obs[1:No_Obs]
                    Ymiss<-Std_obs[nn,1]
                    Xmiss<-Std_obs[nn,2:3]
                    Y<-rep(Ymiss,No_Obs)
                    X1<-rep(Xmiss[1,1],No_Obs)
                    X2<-rep(Xmiss[1,2],No_Obs)
                    var_miss<-cbind(X1,X2,Y)
                    Distance<-sqrt(((Xnonmis[,1]-var_miss[,1])^2)+
                                  ((Xnonmis[,2]-var_miss[,2])^2)+
                                  ((Ynonmis-var_miss[,3])^2))
                    ix<-sort.int(Distance,index.return=TRUE)$ix
                    Xix<-X_nonmiss[ix]
                    knn_bar<-sum(Xix[1:k])/k
                    KNN_bar<-c(KNN_bar,knn_bar)
            }
            return(KNN_bar)
    }
###Function for multiple imputation method ###
    MI.impute<-function(Xmiss,No_mis,No_Obs,Y,X1,X2)
    {       Xbar<-mean(Xmiss,na.rm=TRUE)
            Xsd<-sd(Xmiss,na.rm=TRUE)
            X_nonmiss<-Xmiss[1:No_Obs]
            y<-Y
            x1<-X1
            x2<-X2
            n<-No_mis
            BETA<-c()
            for(D in 1:5)
            {       repeat
                    {
                    predict_x<-rnorm(n,Xbar,Xsd)
                    x3_complete<-c(X_nonmiss,predict_x)
                    new_data<-data.frame(y,x1,x2,x3_complete)
                    Logit<-
glm(y~x1+x2+x3_complete,data=new_data,family=binomial)
                    it<-Logit$iter
```

```
                if(it!=25)
                {
                beta<-coef(Logit)
                break
                } }
                BETA<-c(BETA,list(beta))
            }
            M1<-as.matrix(BETA[[1]])
            M2<-as.matrix(BETA[[2]])
            M3<-as.matrix(BETA[[3]])
            M4<-as.matrix(BETA[[4]])
            M5<-as.matrix(BETA[[5]])
            Beta<-cbind(M1,M2,M3,M4,M5)
            Mean_Beta<-rowMeans(Beta)
            return(Mean_Beta)
        }
###Function for MSE###
        impute.MSE <-function(Y,X1,X2,X3_impute,true_Bbeta)
        {
            Impute_data<-data.frame(Y,X1,X2,X3_impute)
            I_logit<-glm(Y~X1+X2+X3_impute,data=Impute_data,family=binomial)
            it<-I_logit$iter
            if(it!=25)
            {Bbeta<-as.matrix(coef(I_logit))
            bias<-( Bbeta -true_Bbeta)^2}
            else
            { bias <-matrix(c(NA,NA,NA,NA),nrow=4,byrow=TRUE)}
            return(bias)
        }
```

####Main code###

```
library(MASS)
N<-5000
n<-70 ###n=70,100,200###

###Simulate Data###
#####Missing at initial coefficient of independent variables is 1.5 #####

rawdata<-c()
for(t in 1:N){
	X1<-rnorm(n)
	X2<-rnorm(n)
	X3<-rnorm(n)

	logitP<-(0.5*X1)+(1*X2)+(1.5*X3)
	P<-plogis(logitP)
	Y<-rbinom(n,1,P)
	data<-data.frame(Y,X1,X2,X3)

	rawdata<-c(rawdata,list(data))
	}

mm.p1.mmean<-matrix(nrow=4,ncol=N)
mm.p2.mmean<-matrix(nrow=4,ncol=N)
mm.p3.mmean<-matrix(nrow=4,ncol=N)
mm.p4.mmean<-matrix(nrow=4,ncol=N)
mm.p5.mmean<-matrix(nrow=4,ncol=N)
mm.p6.mmean<-matrix(nrow=4,ncol=N)
mm.p7.mmean<-matrix(nrow=4,ncol=N)
mm.p8.mmean<-matrix(nrow=4,ncol=N)
mm.p9.mmean<-matrix(nrow=4,ncol=N)
mm.p1.mmedian<-matrix(nrow=4,ncol=N)
mm.p2.mmedian<-matrix(nrow=4,ncol=N)
mm.p3.mmedian<-matrix(nrow=4,ncol=N)
```

```
mm.p4.mmedian<-matrix(nrow=4,ncol=N)
mm.p5.mmedian<-matrix(nrow=4,ncol=N)
mm.p6.mmedian<-matrix(nrow=4,ncol=N)
mm.p7.mmedian<-matrix(nrow=4,ncol=N)
mm.p8.mmedian<-matrix(nrow=4,ncol=N)
mm.p9.mmedian<-matrix(nrow=4,ncol=N)

mm.p1.kknn<-matrix(nrow=4,ncol=N)
mm.p2.kknn<-matrix(nrow=4,ncol=N)
mm.p3.kknn<-matrix(nrow=4,ncol=N)
mm.p4.kknn<-matrix(nrow=4,ncol=N)
mm.p5.kknn<-matrix(nrow=4,ncol=N)
mm.p6.kknn<-matrix(nrow=4,ncol=N)
mm.p7.kknn<-matrix(nrow=4,ncol=N)
mm.p8.kknn<-matrix(nrow=4,ncol=N)
mm.p9.kknn<-matrix(nrow=4,ncol=N)

mm.p1.mmi<-matrix(nrow=4,ncol=N)
mm.p2.mmi<-matrix(nrow=4,ncol=N)
mm.p3.mmi<-matrix(nrow=4,ncol=N)
mm.p4.mmi<-matrix(nrow=4,ncol=N)
mm.p5.mmi<-matrix(nrow=4,ncol=N)
mm.p6.mmi<-matrix(nrow=4,ncol=N)
mm.p7.mmi<-matrix(nrow=4,ncol=N)
mm.p8.mmi<-matrix(nrow=4,ncol=N)
mm.p9.mmi<-matrix(nrow=4,ncol=N)

missingcase<-c()
Output<-c()
out1<-c()
out2<-c()
out3<-c()
out4<-c()

### Initial coefficient of independent variables ###
bbeta<-matrix(c(0,0.5,1,1.5),byrow=TRUE)
```

```
### Simulate missing data ###
For (tt in 1:N) {

## Prob. Missing ##
	Prob_mis1<-c(0.1,0.08,0.04,0.2,0.16,0.08,0.3,0.24,0.12)
	Prob_mis2<-c(0.1,0.1,0.1,0.2,0.2,0.2,0.3,0.3,0.3)
	Prob_mis3<-c(0.1,0.12,0.16,0.2,0.24,0.32,0.3,0.36,0.48)

	X3<-rawdata[[tt]]$X3

	missinglist<-c()
	out_MEAN<-c()
	out_MEDIAN<-c()
	out_KNN<-c()
	out_MI<-c()

	out_Bbeta_MEAN<-c()
	out_Bbeta_MEDIAN<-c()
	out_Bbeta_KNN<-c()
	out_Bbeta_MI<-c()

	out_BIAS_MEAN<-c()
	out_BIAS_MEDIAN<-c()
	out_BIAS_KNN<-c()
	out_BIAS_MI<-c()

	for(i in 1:9) {
	cut1<-qnorm(1/3)
	cut2<-qnorm(2/3)

	X3[X3<=cut1]->part1
	X3[X3>cut1&X3<=cut2]->part2
	X3[X3>cut2]->part3

	rbinom(length(part1),1,Prob_mis1[i])->a
```

```
        rbinom(length(part2),1,Prob_mis2[i])->b
        rbinom(length(part3),1,Prob_mis3[i])->c

        cbind(part1,a)->miss_point1
        cbind(part2,b)->miss_point2
        cbind(part3,c)->miss_point3

        rbind(miss_point1,miss_point2,miss_point3)->miss
        colnames(miss)[c(1,2)]<-c("X3","Miss")
        merge(rawdata[[tt]],miss,by='X3',all.X3=TRUE)->missdata
        missdata$X3[which(missdata$Miss==1)]=NA
        merge(rawdata[[tt]],missdata,by=c("Y","X1","X2"))->DATA
        colnames(DATA)[c(4,5)]<-c("X3.Full","X3.Miss")
        DATA[with(DATA,order(DATA$X3.Miss,na.last=TRUE)),]->DATA
        missinglist<-c(missinglist,list(DATA))
        missingcase[[tt]]<-missinglist

        Y_Obs<-DATA[,1]
        x1<-DATA[,2]
        x2<-DATA[,3]
        x3<-DATA[,4]
        X_Missing<-DATA[,5]
        m<-sum(DATA$Miss==0)
        n_m<-sum(DATA$Miss==1)
if(m==n){
        Output[[t]]<-missinglist
        Bbeta_complete<-impute.Bbeta(Y_Obs,x1,x2,x3,bbeta)
        Bbeta_MEAN<-Bbeta_complete
        Bbeta_MEDIAN<-Bbeta_complete
        Bbeta_KNN<-Bbeta_complete
        Bbeta_MI<-Bbeta_complete
        out_KNN<-x3
        }
if(m!=n){
#####MEAN IMPUTATION#####
        X_MEAN<-MEAN.impute(X_Missing)
```

```
        out_MEAN<-c(out_MEAN,list(X_MEAN))

#####MEDIAN IMPUTATION#####
        X_MEDIAN<-MED.impute(X_Missing)
        out_MEDIAN<-c(out_MEDIAN,list(X_MEDIAN))
###Standardize X&Y###
        STD_OBS<-std_obs(Y_Obs,x1,x2)
#####KNN IMPUTATION#####
        KNN<-X_Missing[1:m]
        X_KNN<-c(KNN,KNN.impute(STD_OBS,X_Missing,m,n_m))
        out_KNN<-c(out_KNN,list(X_KNN))
#####MI IMPUTATION#####
        Bbeta_MI<-MI.impute(X_Missing,n_m,m,Y_Obs,x1,x2)
        out_MI<-c(out_MI,list(Bbeta_MI))
        }
####Function for MSE####
        MMSE_MEAN<-impute.MSE(Bbeta_MEAN,bbeta)
        out_BIAS_MEAN<-c(out_BIAS_MEAN,list(MMSE_MEAN))
        MMSE_MEDIAN<-impute.MSE(Bbeta_MEDIAN,bbeta)
        out_BIAS_MEDIAN<-c(out_BIAS_MEDIAN,list(MMSE_MEDIAN))
        MMSE_KNN<-impute.MSE(Bbeta_KNN,bbeta)
        out_BIAS_KNN<-c(out_BIAS_KNN,list(MMSE_KNN))
        MMSE_MI<-impute.MSE(Bbeta_MI,bbeta)
        out_BIAS_MI<-c(out_BIAS_MI,list(MMSE_MI))
                }
#####Average MSE#####
        mm.p1.mmean[,tt]<-out_BIAS_MEAN[[1]]
        mm.p2.mmean[,tt]<-out_BIAS_MEAN[[2]]
        mm.p3.mmean[,tt]<-out_BIAS_MEAN[[3]]
        mm.p4.mmean[,tt]<-out_BIAS_MEAN[[4]]
        mm.p5.mmean[,tt]<-out_BIAS_MEAN[[5]]
        mm.p6.mmean[,tt]<-out_BIAS_MEAN[[6]]
        mm.p7.mmean[,tt]<-out_BIAS_MEAN[[7]]
        mm.p8.mmean[,tt]<-out_BIAS_MEAN[[8]]
        mm.p9.mmean[,tt]<-out_BIAS_MEAN[[9]]
        out1<-c(out1,list(out_BIAS_MEAN))
```

```
mmean<-
cbind(rowMeans(mm.p1.mmean,na.rm=TRUE),rowMeans(mm.p2.mmean,
na.rm=TRUE),rowMeans(mm.p3.mmean,na.rm=TRUE),
rowMeans(mm.p4.mmean,na.rm=TRUE),rowMeans(mm.p5.mmean,na.rm=TRU
E),rowMeans(mm.p6.mmean,na.rm=TRUE),
rowMeans(mm.p7.mmean,na.rm=TRUE),rowMeans(mm.p8.mmean,na.rm=TRU
E),rowMeans(mm.p9.mmean,na.rm=TRUE))
mmean_bbeta<-colMeans(mmean,na.rm=TRUE)

mm.p1.mmedian[,tt]<-out_BIAS_MEDIAN[[1]]
mm.p2.mmedian[,tt]<-out_BIAS_MEDIAN[[2]]
mm.p3.mmedian[,tt]<-out_BIAS_MEDIAN[[3]]
mm.p4.mmedian[,tt]<-out_BIAS_MEDIAN[[4]]
mm.p5.mmedian[,tt]<-out_BIAS_MEDIAN[[5]]
mm.p6.mmedian[,tt]<-out_BIAS_MEDIAN[[6]]
mm.p7.mmedian[,tt]<-out_BIAS_MEDIAN[[7]]
mm.p8.mmedian[,tt]<-out_BIAS_MEDIAN[[8]]
mm.p9.mmedian[,tt]<-out_BIAS_MEDIAN[[9]]
out2<-c(out2,list(out_BIAS_MEDIAN))
mmedian<-
cbind(rowMeans(mm.p1.mmedian,na.rm=TRUE),rowMeans(mm.p2.mmedian,
na.rm=TRUE),rowMeans(mm.p3.mmedian,na.rm=TRUE),
rowMeans(mm.p4.mmedian,na.rm=TRUE),rowMeans(mm.p5.mmedian,na.rm=T
RUE),rowMeans(mm.p6.mmedian,na.rm=TRUE),
rowMeans(mm.p7.mmedian,na.rm=TRUE),rowMeans(mm.p8.mmedian,na.rm=T
RUE),rowMeans(mm.p9.mmedian,na.rm=TRUE))
mmedian_bbeta<-colMeans(mmedian,na.rm=TRUE)

mm.p1.kknn[,tt]<-out_BIAS_KNN[[1]]
mm.p2.kknn[,tt]<-out_BIAS_KNN[[2]]
mm.p3.kknn[,tt]<-out_BIAS_KNN[[3]]
mm.p4.kknn[,tt]<-out_BIAS_KNN[[4]]
mm.p5.kknn[,tt]<-out_BIAS_KNN[[5]]
mm.p6.kknn[,tt]<-out_BIAS_KNN[[6]]
mm.p7.kknn[,tt]<-out_BIAS_KNN[[7]]
mm.p8.kknn[,tt]<-out_BIAS_KNN[[8]]
```

```
        mm.p9.kknn[,tt]<-out_BIAS_KNN[[9]]
        out3<-c(out3,list(out_BIAS_KNN))
        kknn<-cbind(rowMeans(mm.p1.kknn,na.rm=TRUE),rowMeans(mm.p2.kknn,
na.rm=TRUE),rowMeans(mm.p3.kknn,na.rm=TRUE),
        rowMeans(mm.p4.kknn,na.rm=TRUE),rowMeans(mm.p5.kknn,na.rm=TRUE),row
Means(mm.p6.kknn,na.rm=TRUE),
        rowMeans(mm.p7.kknn,na.rm=TRUE),rowMeans(mm.p8.kknn,na.rm=TRUE),row
Means(mm.p9.kknn,na.rm=TRUE))
        kknn_bbeta<-colMeans(kknn,na.rm=TRUE)

        mm.p1.mmi[,tt]<-out_BIAS_MI[[1]]
        mm.p2.mmi[,tt]<-out_BIAS_MI[[2]]
        mm.p3.mmi[,tt]<-out_BIAS_MI[[3]]
        mm.p4.mmi[,tt]<-out_BIAS_MI[[4]]
        mm.p5.mmi[,tt]<-out_BIAS_MI[[5]]
        mm.p6.mmi[,tt]<-out_BIAS_MI[[6]]
        mm.p7.mmi[,tt]<-out_BIAS_MI[[7]]
        mm.p8.mmi[,tt]<-out_BIAS_MI[[8]]
        mm.p9.mmi[,tt]<-out_BIAS_MI[[9]]
        out4<-c(out4,list(out_BIAS_MI))
        mmi<-cbind(rowMeans(mm.p1.mmi,na.rm=TRUE),rowMeans(mm.p2.mmi,
na.rm=TRUE),rowMeans(mm.p3.mmi,na.rm=TRUE),
        rowMeans(mm.p4.mmi,na.rm=TRUE),rowMeans(mm.p5.mmi,na.rm=TRUE),row
Means(mm.p6.mmi,na.rm=TRUE),
        rowMeans(mm.p7.mmi,na.rm=TRUE),rowMeans(mm.p8.mmi,na.rm=TRUE),row
Means(mm.p9.mmi,na.rm=TRUE))
        mmi_bbeta<-colMeans(mmi,na.rm=TRUE)
}

Total_AMSE<-cbind(mmean_bbeta,mmedian_bbeta,kknn_bbeta,mmi_bbeta)
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวภัทฐิดา นิลภัทรฉัตร เกิดวันที่ 11 กรกฎาคม พ.ศ. 2533 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล ในปีการศึกษา 2555 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557