การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงเส้นพหุ

นางสาวอุษณีย์ วงศ์อามาตย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING DATA IN MULTIPLE LINEAR REGRESSION

Miss Ausanee Wongarmart

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 2012

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงเส้นพหุ
โดย	นางสาวอุษณีย์ วงศ์อามาตย์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี)

..... กรรมการ
(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

..... กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. ธิดาพร ศุภภากร)

อุษณีย์ วงศ์อามาตย์: การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงเส้นพหุ (COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING DATA IN MULTIPLE LINEAR REGRESSION). อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. อนุภาพ สมบูรณ์สวัสดี, 93 หน้า.

ข้อมูลสูญหายเป็นปัญหาที่พบได้ทั่วไปในงานวิจัยทุกสาขา การสูญหายของข้อมูลที่มีความสัมพันธ์กับพารามิเตอร์อาจนำไปสู่ปัญหาที่ร้ายแรงในการวิเคราะห์ข้อมูล ซึ่งการสูญหายในลักษณะนี้เป็นการสูญหายแบบนอนอิกนอร์เรเบิล วิธีการหนึ่งที่ใช้แก้ไขปัญหาดังกล่าวคือการประมาณค่าที่สูญหายโดยวิธีการต่างๆ งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุ เมื่อข้อมูลตัวแปรตามมีการสูญหายแบบนอนอิกนอร์เรเบิล วิธีการประมาณค่าสูญหายที่ใช้ในงานการวิจัยนี้คือ วิธี EM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM)

ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองโดยมีสัดส่วนของการสูญหาย 3 ระดับคือ 10%, 20%, 30% และมีระดับของการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับคือ ไม่มี, ปานกลาง, สูง จากการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (average mean square error; AMSE) พบว่า i) วิธีการประมาณทุกวิธีสามารถประมาณได้ดีขึ้นเมื่อขนาดตัวอย่างมีขนาดใหญ่ขึ้น ii) วิธีการประมาณทุกวิธีประมาณได้แย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อน สัดส่วนของการสูญหาย และ ระดับของการสูญหายแบบอิกนอร์เรเบิล มีค่าเพิ่มขึ้น iii) โดยรวมแล้ววิธี EM ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าไม่สูง (10-30) และ iv) วิธี KNN ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าสูง (90)

ภาควิชา สถิติ ลายมือชื่อนิสิต.....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2555

5381928426: MAJOR STATISTICS

KEYWORDS: MULTIPLE LINEAR REGRESSION/ NONIGNORABLE MISSING DATA

AUSANEE WONGARMART: COMPARISON OF THE ESTIMATION METHODS FOR NONIGNORABLE MISSING DATA IN MULTIPLE LINEAR REGRESSION. ADVISOR: ANUPAP SOMBOONSAVATDEE, Ph.D., 93 pp.

Problems of missing data are common in all fields of research. When the missingness of data depends on the parameters of interest, this could lead to serious problems. This type of missingness is called “nonignorable”. One remedy to deal with missing data is to estimate or to approximate the missing data by various methods. The purpose of this research is to study and to compare the estimation methods under multiple linear regression settings with nonignorable missing data on the dependent variables. The methods for estimating missing data are EM Algorithm (EM), K-Nearest Neighbor Imputation (KNN) and Predictive Mean Matching Imputation (PMM) method.

Three levels of missing proportion of data of 10%, 20%, 30% and three levels of nonignorable missingness of none, medium, high are studied from the simulations. Based on the size of average mean square error (AMSE), the findings are the followings: i) all estimation methods perform better as the sample size increases, ii) all estimation methods perform worse as the standard deviation of errors, the missing proportion, or level of nonignorable missingness increase, iii) overall, EM method performs best when the standard deviation of errors are not high (10-30) and iv) KNN method performs best when the standard deviation is high (90).

Department: Statistics Student’s Signature……...

Field of Study: Statistics Advisor’s Signature……...

Academic Year : 2012

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความช่วยเหลือ และเอาใจใส่อย่างดียิ่งของ อาจารย์ ดร. อนุภาพ สมบูรณ์สวัสดี อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูง ที่กรุณาให้คำแนะนำ และคำปรึกษาเกี่ยวกับวิทยานิพนธ์ด้วยดีเสมอมา

ผู้วิจัยขอกราบขอบพระคุณรองศาสตราจารย์ ดร. ธีระพร วีระถาวร ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช และอาจารย์ ดร. ธิดาพร ศุภภากร กรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่ช่วยส่งเสริม สนับสนุนและให้กำลังใจเสมอมาจนสำเร็จการศึกษา รวมทั้งขอขอบคุณเพื่อนๆ ทุกคน ที่คอยช่วยให้กำลังใจผู้วิจัยมาโดยตลอด

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การพยากรณ์เป็นเทคนิคหนึ่งที่นิยมนำมาใช้ในงานวิจัยในหลายๆสาขา ไม่ว่าจะเป็นทางด้านธุรกิจ การแพทย์ วิศวกรรม ชีววิทยา ด้านจิตวิทยา ด้านเศรษฐศาสตร์ และด้านสิ่งแวดล้อม เป็นต้น ทั้งนี้เนื่องจากสภาพเศรษฐกิจและสังคมในปัจจุบันมีความยุ่งยากซับซ้อนมากขึ้น ทั้งยังมีความผันผวนและเปลี่ยนแปลงอยู่เสมอ ซึ่งอาจจะส่งผลกระทบต่อแผนงานได้ ดังนั้นหากสามารถพยากรณ์เหตุการณ์หรือทำนายผลที่อาจเกิดขึ้นในอนาคตได้ก็น่าจะส่งผลดีต่อการวางแผนงานมากกว่าการที่ไม่ทราบข้อมูลอะไรเลย เพราะช่วยลดความเสี่ยงในการตัดสินใจดำเนินงานที่ผิดพลาดได้ หรือสามารถเตรียมรับมือกับปัญหาที่อาจเกิดขึ้นได้

การวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple linear Regression) เป็นอีกเทคนิคหนึ่งที่นิยมนำมาใช้ในการสร้างสมการเพื่อการพยากรณ์ เนื่องจากเป็นเทคนิคที่ง่ายต่อการใช้งานและอธิบายผล โดยเป็นสมการที่นำตัวแปรอธิบายหรือตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป มาใช้ในการพยากรณ์ตัวแปรตาม ยกตัวอย่างเช่น รายได้ของแต่ละบริษัท ขึ้นอยู่กับจำนวนผลผลิต ต้นทุนของการผลิต และส่วนแบ่งทางการตลาด เป็นต้น ซึ่งโดยทั่วไปแล้วการวิเคราะห์การถดถอยนั้น จะใช้ข้อมูลของทั้งตัวแปรอิสระและตัวแปรตามที่ครบสมบูรณ์มาใช้ในการสร้างสมการพยากรณ์ แต่ในความเป็นจริงนั้น ในหลายๆงานวิจัยที่ต้องทำการเก็บข้อมูล มักจะเกิดปัญหาข้อมูลของบางตัวแปรมีค่าสูญหายหรือไม่ทราบค่า ซึ่งการสูญหายนี้อาจเกิดจากความตั้งใจหรือไม่ตั้งใจก็ได้ เช่น เกิดจากความผิดพลาดในการเก็บข้อมูล เวลาและค่าใช้จ่ายมีจำนวนจำกัด หรืออาจเกิดจากความจงใจในการไม่ตอบคำถามของผู้ตอบแบบสอบถามก็เป็นไปได้

ทั้งนี้หากละเลยหรือตัดชุดข้อมูลที่เกิดการสูญหายหรือชุดข้อมูลที่ไม่สมบูรณ์ทิ้งไป ก็จะส่งผลกระทบต่อการวิเคราะห์ข้อมูล เพราะจะทำให้ขนาดตัวอย่างที่ใช้ในการพยากรณ์มีขนาดลดลง ซึ่งจะทำให้ความคลาดเคลื่อนของการพยากรณ์มีค่าสูงขึ้น นอกจากนี้ยังอาจจะสูญเสียรายละเอียดบางส่วนที่สำคัญไป และทำให้ได้ผลสรุปที่ผิดพลาดได้ ดังนั้นเพื่อลดความผิดพลาดนี้ Little และ Rubin (1987) ได้นำเสนอวิธีการประมาณค่าสูญหายออกมาหลายวิธี แต่มีเงื่อนไขของการใช้งานคือ การสูญหายของข้อมูลจะต้องเป็นการสูญหายแบบสุ่ม

ในการวิจัยครั้งนี้ผู้วิจัยได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการประมาณค่าสูญหายของตัวแปรตาม ในการวิเคราะห์การถดถอยเชิงพหุ โดยมีงานวิจัยที่ศึกษาดังนี้

วารุณี ตรีบำรุงศักดิ์ (2537) ได้ทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยเชิงเส้นพหุ โดยทำการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบสุ่มด้วยวิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย วิธี EM และวิธีการของฮันท์ ในการเปรียบเทียบทำภายใต้สถานการณ์ของขนาดตัวอย่าง ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน สัดส่วนการสูญหายของตัวแปรตาม และลักษณะของตัวแปรอิสระที่แตกต่างกัน จากการศึกษาพบว่า ในกรณีที่ขนาดตัวอย่างมีขนาดเล็ก ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนไม่สูงมาก และสัดส่วนของการสูญหายของตัวแปรตามมีจำนวนมาก วิธีการของฮันท์จะมีความเหมาะสมที่สุด แต่ถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูง วิธีค่าเฉลี่ยจะมีความเหมาะสมมากกว่าวิธีอื่นๆ และในกรณีที่ขนาดตัวอย่างมีขนาดใหญ่พอ วิธีสูญหายจะเหมาะสมเกือบทุกกรณี

เพียงออ ยีสา (2551) ได้ทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุเพื่อการพยากรณ์ โดยข้อมูลตัวแปรตามมีการสูญหายแบบสุ่ม และข้อมูลที่ทำการศึกษามี 2 ลักษณะคือ ข้อมูลภาคตัดขวาง และข้อมูลอนุกรมเวลา วิธีที่ใช้ในการประมาณค่าสูญหายของตัวแปรตามคือ วิธี Regression Imputation (RI) วิธี Nearest Neighbor Imputation (NNI) วิธี Weighted Nearest Neighbor and Regression Imputation (WNR) และวิธี EM Algorithm (EM) โดยกระทำภายใต้สถานการณ์ของขนาดตัวอย่าง ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน และร้อยละการสูญหายของตัวแปรตามที่แตกต่างกัน ซึ่งจากการศึกษาพบว่า สำหรับข้อมูลภาคตัดขวาง ถ้าส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำถึงปานกลางวิธี RI และ EM จะดีกว่าวิธีอื่น ถ้าส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูงวิธี WNR จะเหมาะสมที่สุด สำหรับข้อมูลอนุกรมเวลา ถ้าข้อมูลมีอิทธิพลของฤดูกาลสูงวิธี WNR จะเหมาะสมที่สุด และถ้าข้อมูลมีอิทธิพลของปัจจัยแนวโน้มสูงวิธี RI และ EM จะเป็นวิธีที่ให้ผลดีกว่าวิธีอื่นๆ

ซึ่งจากงานวิจัยที่กล่าวมาในข้างต้น ตัวแปรตามมีการสูญหายแบบสุ่มเท่านั้น แต่ในสภาพของความเป็นจริงแล้ว การสูญหายของข้อมูลอาจไม่เป็นแบบสุ่มเสมอไป และถึงแม้ว่าจะเป็นการยากที่จะตรวจสอบและระบุชนิดของการเกิดข้อมูลสูญหาย แต่เมื่อใดก็ตามที่ข้อมูลนั้นเกิดการสูญหายแบบ Nonignorable คือ ข้อมูลที่สูญหายจะขึ้นอยู่กับตัวแปรที่เกิดการสูญหาย ยกตัวอย่าง เช่น การสูญหายของข้อมูลรายได้ จะขึ้นอยู่กับระดับของรายได้ หรือการสูญหายของข้อมูล

แอลกอฮอล์จะขึ้นอยู่กับปริมาณแอลกอฮอล์ที่มีอยู่ในร่างกาย ซึ่งการสูญหายในลักษณะนี้จะพบได้บ่อยในงานวิจัยเชิงสำรวจ และจะส่งผลกระทบต่อการวิเคราะห์ข้อมูลมากกว่าการสูญหายแบบสุ่ม

ดังนั้นผู้วิจัยจึงสนใจศึกษาวิธีการประมาณค่าสูญหายของตัวแปรตาม เพื่อใช้ในการพยากรณ์ ในกรณีที่ตัวแปรตามมีการสูญหายแบบ Nonignorable และข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น โดยเทคนิคที่นำมาใช้ในการพยากรณ์คือการวิเคราะห์การถดถอยเชิงเส้นพหุ และทำการหาสัมประสิทธิ์การถดถอยเพื่อใช้ในการสร้างสมการพยากรณ์จากวิธีกำลังสองน้อยสุด และเนื่องจากในขณะนี้ยังไม่มีวิธีการประมาณค่าสูญหายที่มีความเหมาะสมกับกรณีที่เกิดการสูญหายแบบ Nonignorable ดังนั้นจึงสนใจที่จะศึกษาวิธีการประมาณค่าสูญหายที่ใช้นิยมใช้กับข้อมูลที่มีการสูญหายแบบสุ่ม เพื่อดูความเหมาะสมของแต่ละวิธีการประมาณค่าสูญหายเมื่อนำมาประยุกต์ใช้กับข้อมูลที่มีการสูญหายแบบ Nonignorable ซึ่งวิธีการประมาณค่าสูญหายที่สนใจคือ

1. วิธีการประมาณค่าสูญหายโดยวิธี EM Algorithm (EM)
2. วิธีการประมาณค่าสูญหายแบบ K-Nearest Neighbor Imputation (KNN)
3. วิธีการประมาณค่าสูญหายแบบ Predictive Mean Matching Imputation (PMM)

วิธีการประมาณค่าสูญหายใดที่ให้ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงต่ำกว่า จะเป็นวิธีการประมาณค่าสูญหายที่ดีกว่า

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าสูญหายของตัวแปรตามเมื่อมีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น เพื่อใช้ในการพยากรณ์ด้วยการถดถอยเชิงเส้นพหุ
2. เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามเมื่อมีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น ทั้ง 3 วิธี คือวิธี EM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) โดยจะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์ของตัวแปรตามที่ได้จากแต่ละวิธี

1.3 ข้อตกลงเบื้องต้น

1. ตัวแปรตามและตัวแปรอิสระที่สนใจศึกษา มีความสัมพันธ์กันภายใต้การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ซึ่งมีรูปแบบดังนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad ; i = 1, 2, ..., n$$

เมื่อ y_i เป็นค่าสังเกตของตัวแปรตามของข้อมูลตัวที่ i

x_{i1} เป็นค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 1

x_{i2} เป็นค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 2

x_{i3} เป็นค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 3

β_p เป็นสัมประสิทธิ์การถดถอย เมื่อ $p =$ 0, 1, 2, 3

ε_i เป็นค่าความคลาดเคลื่อนสุ่มของข้อมูลตัวที่ i

n เป็นจำนวนค่าสังเกตทั้งหมด

m เป็นจำนวนค่าสังเกตที่ทราบค่า โดยที่ $m \le n$

$n-m$ เป็นจำนวนค่าสังเกตที่สูญหาย

2. ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่มีการแจกแจงปกติและมี $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ สำหรับทุกค่า i
3. $\varepsilon_i, \varepsilon_k$ ไม่มีสหสัมพันธ์กัน นั่นคือ $E(\varepsilon_i \varepsilon_k) =$ 0 เมื่อ $i \ne k$
4. การสูญหายเกิดขึ้นที่ตัวแปรตามเท่านั้นและเป็นการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์กับค่าของตัวแปรตามเท่านั้น โดยจะทำการแบ่งช่วงของตัวแปรตาม y ออกเป็น 3 ช่วง และให้แต่ละช่วงมีความน่าจะเป็นของการสูญหายที่แตกต่างกัน

1.4 คำจำกัดความที่ใช้ในการวิจัย

ในงานวิจัยนี้มีคำจำกัดความที่ใช้ในงานวิจัยดังนี้

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 $P(z < Z < z') \times 100\%$ ของพื้นที่ทั้งหมด

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

ความน่าจะเป็นของการสูญหายในแต่ละช่วง คือ อัตราส่วนระหว่างจำนวนตัวอย่างที่สูญหายในช่วงนั้นกับจำนวนตัวอย่างทั้งหมดที่ตกอยู่ในช่วงนั้น

ร้อยละของการสูญหาย คือ (จำนวนตัวอย่างที่สูญหาย/จำนวนตัวอย่างทั้งหมด) $\times 100\%$

1.5 ขอบเขตของการวิจัย

1. ตัวแปรอิสระที่นำมาใช้ในศึกษามีการแจกแจงปกติ (Normal Distribution) ที่มีฟังก์ชันการแจกแจงคือ

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

โดยมีค่าคาดหวัง $E(x) = \mu$ และความแปรปรวน $Var(x) = \sigma^2$

ซึ่งในการศึกษานี้จะกำหนดให้ตัวแปรอิสระมี 2 ลักษณะ คือ

แบบที่ 1 : $X_1 \sim N(0,300)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$

แบบที่ 2 : $X_1 \sim N(0,100)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$

2. ตัวแปรตามและตัวแปรอิสระที่สนใจศึกษา มีความสัมพันธ์กันภายใต้การถดถอยเชิงเส้นพหุ (Multiple Linear Regression)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad ; i = 1, 2, \ldots, m, m+1, \ldots, n$$

ซึ่งโดยทั่วไปแล้ว $\beta_0, \beta_1, \beta_2$ และ β_3 จะเป็นพารามิเตอร์ที่ไม่ทราบค่า และในงานวิจัยนี้จะกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเป็น $\beta_0 = 42$ และให้ $\beta_1, \beta_2, \beta_3$ มีค่าเท่ากัน คือ $\beta_1 = \beta_2 = \beta_3 = 1$ ทั้งนี้เนื่องจากต้องการเปรียบเทียบชุดของตัวแปรอิสระที่มีรูปแบบของความแปรปรวนแตกต่างกัน ซึ่งเมื่อ $\beta_1, \beta_2, \beta_3$ มีค่าเปลี่ยนไปก็จะส่งผลให้ค่าตัวแปรตามที่ได้จากชุดตัวแปรอิสระแต่ละชุดมีความแปรปรวนแตกต่างกันด้วย ดังนั้นเพื่อควบคุมให้ค่า

ความแปรปรวนของตัวแปรตามมีค่าเท่ากัน จึงกำหนดให้ $\beta_1, \beta_2, \beta_3$ มีค่าเท่ากับ 1 และนอกจากนี้ยังกำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน นั่นคือ กำหนดให้ค่าสหสัมพันธ์มีค่าเป็น 0

3. ค่าความคลาดเคลื่อนของข้อมูลมีการแจกแจงปกติ $(\varepsilon_i \sim N(0,\sigma^2))$ โดยกำหนดให้ $\sigma =$ 10, 30 และ 90 เมื่อพิจารณาจากค่าสัมประสิทธิ์การแปรปรวน(Coefficient of Variation)ที่ 75%, 100% และ 225% ตามลำดับ
4. ขนาดของตัวอย่างที่ใช้ในการศึกษามี 3 ขนาดคือ 50, 100 และ 200
5. พื้นที่ใต้โค้งปกติของข้อมูลตัวแปรตามจะถูกแบ่งเป็น 3 ช่วงโดยกำหนดให้มีการแบ่งช่วงเป็นอัตราส่วนของ

   ช่วงต้น : ช่วงกลาง : ช่วงปลาย เป็น 1 : 1 : 1 ตามลำดับ
6. การสูญหายของข้อมูลเกิดขึ้นกับตัวแปรตามเท่านั้น และเป็นการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น โดยให้ช่วงของตัวแปรตามทั้ง 3 ช่วง มีสัดส่วนการสูญหายของข้อมูลตัวแปรตามที่แตกต่างกันตามระดับของการสูญหายแบบ Nonignorable โดยกำหนดให้ช่วงของตัวแปรตามที่มีค่ามากจะมีสัดส่วนของการสูญหายมากกว่าช่วงของตัวแปรตามที่มีค่าน้อย และก็จะทำให้แต่ละช่วงมีความน่าจะเป็นของการสูญหายที่สูง – ต่ำแตกต่างกันด้วย ซึ่งระดับของการสูญหายแบบ Nonignorable จะแบ่งเป็น 3 ระดับคือ ไม่มี, ปานกลาง และสูง ซึ่งในแต่ละช่วงจะมีอัตราส่วนของการสูญหายดังนี้

   ไม่มี 1 : 1 : 1

   ปานกลาง 7 : 10 : 13

   สูง 4 : 10 : 16
7. สัดส่วนของการสูญหายของตัวแปรตาม คิดเป็นร้อยละโดยเฉลี่ยของทั้ง 3 ช่วงคือร้อยละ 10 , 20 และ 30 ซึ่งกำหนดให้แต่ละช่วงของค่าตัวแปรตามมีร้อยละของการสูญหายดังนี้

ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบ Nonignorable	ร้อยละของการสูญหายในแต่ละช่วง		
		ช่วงต้น	ช่วงกลาง	ช่วงปลาย
10	ไม่มี	10	10	10
	ปานกลาง	7	10	13
	สูง	4	10	16

ร้อยละของการสูญหายโดยเฉลี่ย	ระดับของการสูญหายแบบ Nonignorable	ร้อยละของการสูญหายในแต่ละช่วง		
		ช่วงต้น	ช่วงกลาง	ช่วงปลาย
20	ไม่มี	20	20	20
	ปานกลาง	14	20	26
	สูง	8	20	32
30	ไม่มี	30	30	30
	ปานกลาง	21	30	39
	สูง	12	30	48

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

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่า วิธีการประมาณค่าสูญหายของตัวแปรตามวิธีใดให้ค่าพยากรณ์ของตัวแปรตามใกล้เคียงกับค่าจริงมากที่สุดนั้น จะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง (Average Mean Square Error : AMSE) ซึ่งวิธีการที่ให้ค่า AMSE ต่ำสุดจะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด และใช้ค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency : RE) ซึ่งเป็นอัตราส่วนระหว่างค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี EM กับวิธีการประมาณค่าสูญหายวิธีอื่นๆ ช่วยในการเปรียบเทียบและเพื่อดูแนวโน้มของแต่ละวิธีการประมาณ ซึ่งในการศึกษาครั้งนี้เลือกใช้วิธี EM เป็นเกณฑ์ในการเปรียบเทียบเนื่องจากวิธี EM เป็นวิธีที่นิยมใช้ในการประมาณค่าสูญหายมากที่สุด ซึ่งวิธีใดที่ให้ค่า RE มากกว่า 1 แสดงว่ามีประสิทธิภาพมากกว่าวิธี EM โดยคำนวณได้จากสูตร

$$MSE_t = \frac{\sum_{i=1}^{n}(y_i' - \hat{y}_{ti})^2}{n}$$

$$AMSE = \frac{1}{5{,}000}\sum_{t=1}^{5{,}000} MSE_t$$

$$RE = \frac{AMSE_{EM}}{AMSE_r}; r = 1, 2$$

เมื่อ y_i' แทน ค่าจริงของข้อมูลตัวแปรตามตัวที่ i

$\hat{y}_{ti}$ แทน ค่าพยากรณ์ของข้อมูลตัวแปรตามตัวที่ i จากการทำซ้ำรอบที่ t

MSE_t แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำรอบที่ t

$AMSE$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำทั้งหมด 5,000 รอบ

$AMSE_{EM}$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากการประมาณค่าสูญหายด้วยวิธี EM

$AMSE_r$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของที่ได้จากการประมาณค่าสูญหายแต่ละวิธี

1.7 วิธีดำเนินการวิจัย

1. สร้างข้อมูลของความคลาดเคลื่อนที่มีการแจกแจงปกติ โดยมีพารามิเตอร์ตามที่กำหนด คือ $\sigma =$ 10, 30 และ 90
2. สร้างข้อมูลของตัวแปรอิสระที่มีการแจกแจงปกติตามที่กำหนด และสร้างข้อมูลของตัวแปรตามจากรูปแบบความสัมพันธ์ $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ โดยกำหนดให้ $\underset{\sim}{\beta}$ เป็นพารามิเตอร์ที่มีค่าคงที่ คือ $\beta_0 = 42$, $\beta_1 = 1$, $\beta_2 = 1$ และ $\beta_3 = 1$
3. สร้างข้อมูลของตัวแปรตามให้เกิดการสูญหายแบบ Nonignorable โดยมีความน่าจะเป็นของการสูญหายขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น และมีสัดส่วนของการสูญหายตามที่กำหนด
4. ประมาณค่าข้อมูลเพื่อแทนที่ข้อมูลที่สูญหายในตัวแปรตามด้วยวิธี EM Algorithm (EM), วิธี K-Nearest Neighbor Imputation (KNN) และวิธีการประมาณค่าสูญหายแบบ Predictive Mean Matching Imputation (PMM)
5. ประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method : OLS)
6. สร้างสมการถดถอยเชิงเส้นพหุจากค่าประมาณสัมประสิทธิ์การถดถอย เพื่อใช้ในการพยากรณ์

7. คำนวณหาค่า AMSE และค่า RE ของแต่ละวิธีการประมาณค่าสูญหาย เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี
8. สรุปผลที่ได้จากการทดลอง

1.8 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการเลือกใช้วิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยเชิงเส้นพหุเมื่อตัวแปรตามที่มีการแจกแจงปกติมีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามได้อย่างเหมาะสม
2. เพื่อเป็นแนวทางในการศึกษาเพิ่มเติม และเปรียบเทียบวิธีการประมาณค่าสูญหายในสถานการณ์อื่นๆ หรือในข้อมูลที่มีความสัมพันธ์ในรูปแบบอื่นๆต่อไป

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

เมื่อมีปัญหาข้อมูลตัวแปรตามสูญหายเกิดขึ้นในการวิเคราะห์การถดถอยเชิงเส้นพหุ วิธีที่ง่ายที่สุดที่ใช้จัดการกับปัญหานี้ก็คือ ตัดชุดข้อมูลที่ไม่สมบูรณ์นั้นทิ้ง แล้วหาค่าสัมประสิทธิ์การถดถอยเพื่อสร้างสมการพยากรณ์จากชุดข้อมูลที่สมบูรณ์ที่มีอยู่ในปัจจุบัน แต่การทำเช่นนี้อาจจะทำให้ได้ค่าพยากรณ์ที่มีความคลาดเคลื่อนสูงขึ้นและได้ผลสรุปที่ผิดพลาด ในกรณีที่การสูญหายนั้นไม่ใช่การสูญหายแบบสุ่ม แต่เป็นการสูญหายแบบ Nonignorable ในลักษณะที่ความน่าเป็นที่จะเกิดการสูญหายของตัวแปรตาม จะไม่มีความสัมพันธ์หรือไม่ขึ้นอยู่กับค่าของตัวแปรอื่นๆ แต่การสูญหายนี้จะขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น ซึ่งถ้าข้อมูลเกิดการสูญหายแบบนี้ ก็จะยิ่งส่งผลกระทบต่อการวิเคราะห์ข้อมูลมากขึ้น วิธีการหนึ่งที่จะนำมาใช้จัดการกับปัญหาข้อมูลสูญหายก่อนที่จะนำข้อมูลไปใช้ในการวิเคราะห์การถดถอยก็คือ การประมาณค่าสูญหาย ในบทนี้จะกล่าวถึงวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุจากชุดข้อมูลที่สมบูรณ์ และวิธีการประมาณค่าตัวแปรตามที่เกิดการสูญหาย ซึ่งมีรายละเอียดดังนี้

2.1 การประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method : OLS)

วิธีกำลังสองน้อยสุดเป็นวิธีที่นิยมมากที่สุดที่ใช้ในการประมาณค่าสัมประสิทธิ์ถดถอย โดยจะทำการประมาณค่าจากชุดข้อมูลที่สมบูรณ์ที่มีในปัจจุบัน ในที่นี้ถ้ากำหนดให้ n แทนจำนวนค่าสังเกตทั้งหมด และให้ p แทนจำนวนตัวแปรอิสระ ดังนั้นจะมีพารามิเตอร์ที่ไม่ทราบค่าจำนวน p+1 ตัว ซึ่ง n ควรจะต้องมีค่ามากกว่า p+1 ให้ y_i แทน ค่าสังเกตของตัวแปรตามตัวที่ i x_{ik} แทนค่าสังเกตตัวที่ i ของตัวแปรอิสระตัวที่ k และ ε_i แทน ค่าความคลาดเคลื่อนของข้อมูลตัวที่ i ซึ่งจากสมการ

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \quad ; i = 1, 2, ..., m, m+1, ..., n \qquad .....(1)$$

นำมาเขียนในรูปของเมทริกซ์จะได้

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad .....(2)$$

เมื่อ

$$\underset{\sim}{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad , \quad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad , \quad \underset{\sim}{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad , \quad \underset{\sim}{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

ธีระพร วีระถาวร (2531: 56-59) ได้อธิบายทฤษฎีกำลังสองน้อยสุดไว้ดังนี้ สมมติว่าค่าประมาณตัวแปรตามสามารถประมาณได้จากสมการ $\underset{\sim}{\hat{y}}=X\underset{\sim}{b}$ ซึ่งการที่ค่าประมาณของตัวแปรตามจะมีค่าใกล้เคียงกับค่าสังเกตหรือไม่นั้น ก็ขึ้นอยู่กับเวกเตอร์ของพารามิเตอร์ $\underset{\sim}{b}$ ดังนั้น จะต้องหาค่า $\underset{\sim}{b}$ ที่ทำให้มีค่าระยะห่างกำลังสองระหว่างค่าสังเกตกับค่าประมาณ หรือมีค่าความคลาดเคลื่อนกำลังสองน้อยที่สุด ซึ่งฟังก์ชันกำลังสองคือ

$$f(\underset{\sim}{b}) = \underset{\sim}{e}'\underset{\sim}{e} = (\underset{\sim}{y}-X\underset{\sim}{b})'(\underset{\sim}{y}-X\underset{\sim}{b}) = \underset{\sim}{y}'\underset{\sim}{y}-2\underset{\sim}{b}'X\underset{\sim}{y}+\underset{\sim}{b}'X'X\underset{\sim}{b}$$

ดังนั้น $f(\underset{\sim}{b})\geq 0$ และจะสามารถหาตัวประมาณกำลังสองต่ำสุดได้จากการหาอนุพันธ์ของฟังก์ชันกำลังสองเทียบกับ $\underset{\sim}{b}$ ซึ่งจะได้ดังนี้

$$\begin{aligned} df(\underset{\sim}{b}) &= d(\underset{\sim}{y}'\underset{\sim}{y}-2\underset{\sim}{b}'X\underset{\sim}{y}+\underset{\sim}{b}'X'X\underset{\sim}{b}) \\ &= d(\underset{\sim}{y}'\underset{\sim}{y})-d(2\underset{\sim}{b}'X\underset{\sim}{y})+d(\underset{\sim}{b}'X'X\underset{\sim}{b}) \\ &= -2(\underset{\sim}{y}-X\underset{\sim}{b})'Xd\underset{\sim}{b} \qquad \text{.....(3)} \end{aligned}$$

ซึ่งถ้ากำหนดให้ $df(\underset{\sim}{b})=0$ ที่ค่า $\underset{\sim}{\hat{\beta}}$ ของ $\underset{\sim}{b}$ แสดงว่า $\underset{\sim}{\hat{\beta}}$ จะเป็นค่าที่ทำให้ฟังก์ชันกำลังสองมีค่าน้อยที่สุด ดังนั้นจากสมการ (3) จะได้ว่า

$$(\underset{\sim}{y}-X\underset{\sim}{b})'X=\underset{\sim}{0}'$$

นั่นคือ $$X'X\underset{\sim}{\hat{\beta}}=X'\underset{\sim}{y} \qquad \text{.....(4)}$$

สมการ (4) นี้ก็คือ สมการปกติ และเมื่อแก้สมการปกตินี้ จะได้ค่าประมาณกำลังสองต่ำสุดคือ

$$\underset{\sim}{\hat{\beta}}=(X'X)^{-1}X'\underset{\sim}{y} \qquad \text{.....(5)}$$

จากค่าประมาณที่ได้ในสมการ (5) เมื่อนำมาทดสอบหาความเอนเอียงพบว่า

$$\begin{aligned} E[\underset{\sim}{\hat{\beta}}] &= E[(X'X)^{-1}X'(X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon})] && \text{: เมื่อแทนค่า } \underset{\sim}{y} \text{ ด้วย } X\underset{\sim}{\beta}+\underset{\sim}{\varepsilon} \\ &= \underset{\sim}{\beta}+E[(X'X)^{-1}X'\underset{\sim}{\varepsilon}] \\ &= \underset{\sim}{\beta}+E[E[(X'X)^{-1}X'\underset{\sim}{\varepsilon}|X]] \\ &= \underset{\sim}{\beta}+E\left[(X'X)^{-1}X'E[\underset{\sim}{\varepsilon}|X]\right] && \text{: จากข้อตกลงเบื้องต้น } E[\underset{\sim}{\varepsilon}|X]=\underset{\sim}{0} \\ &= \underset{\sim}{\beta} \end{aligned}$$

ดังนั้น ค่าประมาณที่ได้จะเป็นค่าประมาณที่ไม่มีความเอนเอียง และในการหาค่าประมาณจะมีเงื่อนไขคือ $(X'X)^{-1}$ จะต้องหาค่าได้ ซึ่งโดยปกติแล้ว เมทริกซ์ $(X'X)^{-1}$ นี้จะหาค่าได้เสมอถ้าตัวแปรอิสระแต่ละตัวเป็นอิสระต่อกัน

2.2 วิธีการประมาณค่าสูญหายของตัวแปรตาม

การประมาณค่าสูญหายเป็นอีกทางเลือกหนึ่งที่จะใช้จัดการกับปัญหาข้อมูลสูญหาย ซึ่งในหลายๆกรณีก็จะส่งผลดีต่อการวิเคราะห์ข้อมูลและให้ผลสรุปที่ผิดพลาดน้อยกว่าการตัดชุดข้อมูลที่ไม่สมบูรณ์ทิ้งแล้วเลือกพิจารณาเฉพาะชุดข้อมูลที่สมบูรณ์ ดังนั้นจึงได้มีผู้นำเสนอ คิดค้นและพัฒนาวิธีการเพื่อมาแทนค่าที่สูญหายเรื่อยมา โดยทั่วไปแล้ววิธีการประมาณค่าสูญหายจะทำภายใต้เงื่อนไขของข้อมูลที่มีการสูญหายแบบสุ่ม ทั้งนี้เป็นเพราะข้อมูลสูญหายส่วนมากมักจะมีรูปแบบของการสูญหายเป็นแบบสุ่ม และการสูญหายแบบนี้ก็มีความยุ่งยากน้อยกว่าการสูญหายแบบ Nonignorable แต่ถ้าหากพบว่าข้อมูลเกิดการสูญหายแบบ Nonignorable แล้ว ในทางปฏิบัติก็ยังคงนำวิธีการประมาณค่าสูญหายภายใต้เงื่อนไขของการสูญหายแบบสุ่มมาใช้งาน เพราะยังไม่มีวิธีที่เหมาะสมและสะดวกต่อการใช้งานมาใช้ในการประมาณข้อมูลที่เกิดการสูญหายในลักษณะนี้ และในงานวิจัยนี้จะทำการประมาณค่าสูญหายด้วยวิธีที่นิยมนำมาใช้กับข้อมูลที่เกิดการสูญหาย ซึ่งจะมีโครงสร้างหรือรูปแบบของการคำนวณที่แตกต่างกันดังนี้

2.2.1 วิธีการประมาณค่าสูญหายโดยวิธี EM Algorithm (EM)

EM Algorithm เป็นวิธีที่ถูกนำไปใช้งานกันอย่างแพร่หลาย และมีรูปแบบของการคำนวณที่ใช้พารามิเตอร์ โดยจะเป็นกระบวนการวนซ้ำเพื่อใช้สำหรับการหาค่าประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) ของค่าพารามิเตอร์ในกรณีที่มีปัญหาเกี่ยวกับข้อมูลไม่สมบูรณ์

EM Algorithm มีขั้นตอนที่ใช้ในการจัดการกับข้อมูลที่สูญหายคือ 1). ประมาณค่าที่สูญหายโดยใช้ค่าพารามิเตอร์จากข้อมูลที่ทราบค่า 2). ทำการประมาณค่าพารามิเตอร์ใหม่ 3). ประมาณค่าที่สูญหายใหม่อีกรอบโดยการใช้ค่าพารามิเตอร์ใหม่ 4). ทำการประมาณค่าพารามิเตอร์ใหม่อีกครั้ง และทำวนซ้ำเช่นนี้ไปเรื่อยๆ จนกระทั่ง ค่าพารามิเตอร์ลู่เข้าสู่ค่าใดค่าหนึ่งจึงหยุด ซึ่งจากขั้นตอนเหล่านี้ สามารถแบ่งได้เป็น 2 ขั้นหลักๆ คือ ขั้น E-Step จะเป็นการประมาณค่าที่สูญหาย จากค่าคาดหวังของค่าที่สูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ทราบค่าและ

พารามิเตอร์ที่มีในปัจจุบัน และขั้น M-Step จะเป็นการประมาณค่าภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ จากการแทนค่าสูญหายที่ได้จากการประมาณในขั้น E-Step

Little และ Rubin (1987) ได้นำเสนอวิธี EM มาเพื่อใช้ในการประมาณค่าสูญหายของตัวแปรตาม ในการวิเคราะห์การถดถอยเชิงเส้นพหุ ซึ่งมีขั้นตอนดังนี้

1. จัดชุดข้อมูล $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ ออกเป็น 2 ส่วน คือส่วนที่มีข้อมูลครบสมบูรณ์ และส่วนที่มีข้อมูลสูญหาย ดังนี้

$$\begin{bmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}_1$ เป็นเวกเตอร์ของตัวแปรตามที่ทราบค่าขนาด $m \times 1$

$\underset{\sim}{y}_2$ เป็นเวกเตอร์ของตัวแปรตามที่สูญหายขนาด $(n-m) \times 1$

X_1 เป็นเมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามทราบค่า ซึ่งมีขนาด $m \times (p+1)$

X_2 เป็นเมทริกซ์ของตัวแปรอิสระ ที่ชุดข้อมูลของตัวแปรตามสูญหาย ซึ่งมีขนาด $(n-m) \times (p+1)$

2. ทำการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุตัวเริ่มต้นด้วยวิธี OLS จากชุดข้อมูลสมบูรณ์ที่มีอยู่ในปัจจุบัน ดังนี้

$$\underset{\sim}{\hat{\beta}}^{(0)} = (X_1' X_1)^{-1} X_1' \underset{\sim}{y}_1$$

3. เข้าสู่ขั้นตอน E- Step ในการทำซ้ำรอบที่ 1 เพื่อหาค่าคาดหวังและค่าประมาณตัวแปรตามที่สูญหาย โดยใช้ค่าประมาณสัมประสิทธิ์การถดถอยที่ได้จากข้อ 2

$$E(y_i \mid X, \underset{\sim}{y}_1, \underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}^{(0)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \hat{\beta}_0^{(0)} + \sum_{k=1}^{p} \hat{\beta}_k^{(0)} x_{ik} & ;i = m+1,...,n \end{cases}$$

โดยที่ $\underset{\sim}{\hat{\theta}}^{(0)} = (\hat{\beta}_0^{(0)}, \hat{\beta}_1^{(0)}, ..., \hat{\beta}_p^{(0)})$

ดังนั้นจะได้ $y_i^{(1)} = E(y_i \mid X, \underset{\sim}{y}_1, \underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}^{(0)})$

4. เข้าสู่ขั้นตอน M-Step ในการทำซ้ำรอบที่ 1 เพื่อหาค่าประมาณสัมประสิทธิ์การถดถอยตัวใหม่ โดยใช้ชุดข้อมูลตัวแปรตามที่ได้จากการประมาณในข้อ 3

$$\underset{\sim}{\hat{\beta}}^{(1)} = (X'X)^{-1} X \underset{\sim}{y}^{(1)}$$

5. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยเริ่มต้นในข้อ 2 กับค่าสัมประสิทธิ์การถดถอยใหม่รอบที่ 1 ของค่าสัมประสิทธิ์การถดถอยทุกค่า

6. ถ้าค่าสัมบูรณ์ของผลต่างทุกค่าในข้อ 5 มีค่ามากกว่า 0.001 ให้ทำขั้นตอนต่อไป แต่ถ้ามีค่าน้อยกว่าหรือเท่ากับ 0.001 ให้หยุด และจะได้ค่าประมาณข้อมูลสูญหายตามข้อ 3

7. เข้าสู่ขั้นตอน E-Step ในการทำซ้ำรอบที่ t ; t = 2,3,… โดยใช้ค่าประมาณสัมประสิทธิ์การถดถอยใหม่ที่ได้มาประมาณค่าตัวแปรตามที่สูญหาย

$$E(y_i \mid X, \underset{\sim}{y}_1, \underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}^{(t-1)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \hat{\beta}_0^{(t-1)} + \sum_{k=1}^{p} \hat{\beta}_k^{(t-1)} x_{ik} & ;i = m+1,...,n \end{cases}$$

โดยที่ $\underset{\sim}{\hat{\theta}}^{(t-1)} = (\hat{\beta}_0^{(t-1)}, \hat{\beta}_1^{(t-1)}, ..., \hat{\beta}_p^{(t-1)})$

ดังนั้นจะได้ $y_i^{(t)} = E(y_i \mid X, \underset{\sim}{y}_1, \underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}^{(t-1)})$

8. เข้าสู่ขั้นตอน M-Step ในการทำซ้ำรอบที่ t เพื่อหาค่าประมาณสัมประสิทธิ์การถดถอยตัวใหม่ โดยใช้ชุดข้อมูลตัวแปรตามที่ได้จากการประมาณในข้อ 7

$$\underset{\sim}{\hat{\beta}}^{(t)} = (X'X)^{-1} X' \underset{\sim}{y}^{(t)}$$

9. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ t-1 กับค่าสัมประสิทธิ์การถดถอยรอบที่ t ของค่าสัมประสิทธิ์การถดถอยทุกค่า

10. ถ้าค่าสัมบูรณ์ของผลต่างทุกค่าในข้อ 9 มีค่ามากกว่า 0.001 ให้กลับไปทำข้อที่ 7 ถึง9 เรื่อยๆจนกระทั่งค่าสัมบูรณ์ของผลต่างทุกค่ามีค่าน้อยกว่าหรือเท่ากับ 0.001 จึงหยุด และจะได้ค่าประมาณข้อมูลสูญหายจากขั้นตอน E ขั้นสุดท้าย

11. นำค่าประมาณข้อมูลสูญหายที่ได้จากข้อ 6 หรือ 10 มาทำการประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธี OLS เพื่อสร้างสมการถดถอยเชิงเส้นพหุมาใช้ในการพยากรณ์

2.2.2 **วิธีการประมาณค่าสูญหายแบบ** K-Nearest Neighbor Imputation (KNN)

K-Nearest Neighbor Imputation เป็นวิธีที่มีรูปแบบการคำนวณที่ไม่ใช้พารามิเตอร์ เพราะในกระบวนการคำนวณใช้เพียงการหาระยะห่างของข้อมูลที่มีอยู่ ไม่จำเป็นต้องหาค่าประมาณพารามิเตอร์ใดๆ และเป็นวิธีการประมาณที่จัดอยู่ใน Hot Deck Method ที่จะแทนค่าที่สูญหายด้วยค่าสังเกตที่ทราบค่า ซึ่งวิธีการนี้ค่อนค้างที่จะมีประสิทธิภาพมากกว่า Hot Deck Method วิธีอื่นๆ วิธี K-Nearest Neighbor Imputation จะประมาณค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าจำนวน K ตัว ที่ลักษณะของตัวแปรที่ไม่เกิดการสูญหายมีความคล้ายคลึงกันกับหน่วยตัวอย่างที่เกิดการสูญหายมากที่สุด ซึ่งโดยทั่วไปแล้วเพื่อความสะดวกใน

การใช้งานก็มักจะกำหนดให้ K = 1(จะเป็นวิธี Nearest Neighbor Imputation:NNI) แต่ Duda และ Hart (1973 อ้างถึงใน Jösson และ Wohlin, 2006) ได้เสนอว่าควรใช้ $K \approx \sqrt{m}$ โดย K จะเป็นจำนวนคี่ที่มีค่าใกล้เคียงกับ $\sqrt{m}$ มากที่สุด เมื่อ m เป็นจำนวนข้อมูลที่สมบูรณ์

กำหนดให้มีการสูญหายเกิดขึ้นที่ตัวแปรตามและให้ y_j^* โดยที่ $j = m+1,...,n$ เป็นค่าประมาณของข้อมูลที่สูญหายด้วยวิธี KNN และในการพิจารณาความคล้ายของหน่วยตัวอย่างจะพิจารณาจากระยะทางยุคลิด (Euclidean Distance) ของตัวแปรที่ไม่เกิดการสูญหาย ซึ่งก็คือตัวแปรอิสระ ดังนี้

$$D_{ij} = \sqrt{\sum_{p=1}^{3} \left(x_{ip} - x_{jp} \right)^2}$$

สำหรับ $i = 1,2,...,m$ และ $j = m+1,...,n$

ซึ่งวิธี KNN มีขั้นตอนดังนี้

1. หา D_{ij} ที่มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างตัวที่ j
2. หาค่าเฉลี่ยของตัวแปรตาม y_i ที่สอดคล้องกับค่าต่ำสุด K ตัว ของชุดตัวอย่างตัวที่ j โดยกำหนดให้เป็น $\bar{y}^*$
3. จะได้ว่า $y_j^* = \bar{y}^*$

2.2.3 วิธีการประมาณค่าสูญหายแบบ Predictive Mean Matching Imputation (PMM)

Predictive Mean Matching Imputation เป็นวิธีการประมาณค่าสูญหายที่มีรูปแบบการคำนวณที่ทั้งใช้และไม่ใช้พารามิเตอร์ โดยการรวมสองแนวคิดเข้าด้วยกัน คือการหาค่าคาดหวังและการแทนที่ ซึ่งสามารถหาค่าคาดหวังของข้อมูลที่สูญหายได้จากชุดข้อมูลที่ทราบค่าและพารามิเตอร์ที่มีในปัจจุบัน จากนั้นจะทำการแทนที่หรือประมาณค่าข้อมูลที่สูญหายด้วยข้อมูลที่ทราบค่าที่มีค่าคาดหวังใกล้เคียงกับค่าคาดหวังของข้อมูลที่เกิดการสูญหายมากที่สุด

กำหนดให้ $\hat{y}_j^*$ โดยที่ $j = m+1,...,n$ เป็นค่าประมาณของข้อมูลตัวแปรตามที่สูญหายด้วยวิธี PMM และในการพิจารณาความใกล้ จะพิจารณาจากค่าสัมบูรณ์ระหว่างค่าคาดหวังของตัวแปรที่ทราบค่ากับค่าคาดหวังของตัวแปรที่สูญหาย

van Buuren และ Groothuis-Oudshoorn (2011) ได้นำเสนอขั้นตอนของวิธี PMM ดังนี้

1. ทำการหาค่าประมาณสัมประสิทธิ์การถดถอย $\hat{\underset{\sim}{\beta}} = (\hat{\beta}_0, \hat{\beta}_1,..., \hat{\beta}_p)$ จากชุดข้อมูลสมบูรณ์ที่มีอยู่ในปัจจุบันด้วยวิธี OLS

2. ทำการหาค่าประมาณสัมประสิทธิ์การถดถอยใหม่ $\underset{\sim}{\hat{\beta}}^{*} = (\hat{\beta}_0^{*}, \hat{\beta}_1^{*}, ..., \hat{\beta}_p^{*})$ จากตัวประมาณเบส์ (Bayes' Estimators) โดยใช้การแจกแจงโดยหลักเกณฑ์ที่ไม่ทราบข้อมูล (Noninformative Prior Distribution) ที่มีรูปแบบเป็น $p(\underset{\sim}{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$ ซึ่งจะพิจารณาจากค่าความแปรปรวน และค่าประมาณ $\underset{\sim}{\hat{\beta}}$ ที่ได้ในข้อ 1 ดังนี้

$$\underset{\sim}{\hat{\beta}}^{*} = \underset{\sim}{\hat{\beta}} + \hat{\sigma}_{*} V' \underset{\sim}{Z}$$

3. ค่าความแปรปรวน $\hat{\sigma}_{*}^2$ หาได้จาก

$$\hat{\sigma}_{*}^2 = \frac{\hat{\sigma}^2 \left(m - (p+1)\right)}{c}$$

เมื่อ $\hat{\sigma}^2$ เป็นค่าประมาณความแปรปรวนของความคลาดเคลื่อน

m เป็นจำนวนตัวแปรตามที่ทราบค่า

p เป็นจำนวนตัวแปรอิสระ ดังนั้นมีพารามิเตอร์ที่ไม่ทราบจำนวน p+1 ตัว

c เป็นตัวแปรสุ่มที่มีการแจกแจงไคสแควร์ $\left(\chi^2_{m-(p+1)}\right)$ และมีระดับขั้นความเสรี $m-(p+1)$

4. V เป็นเมทริกซ์สามเหลี่ยมที่มีสมาชิกอยู่เหนือแถวทแยงมุม (Upper Triangular Matrix) ที่ได้จากการแยกแบบโซเลสกี้ (Cholesky Decomposition) ของ $V^{*} = V'V$ โดยที่ $V^{*} = (X'X)^{-1}$

5. $\underset{\sim}{Z}$ เป็นเวกเตอร์ของตัวแปรสุ่มที่มีการแจกแจงปกติมาตรฐาน $\left(N(0,1)\right)$ ขนาด $(p+1)\times 1$

6. นำค่าประมาณสัมประสิทธิ์การถดถอยใหม่ที่ได้ และค่าตัวแปรอิสระ มาประมาณค่าตัวแปรตามทั้งที่เกิดการสูญหายและที่มีข้อมูลสมบูรณ์ ดังสมการนี้

$$\hat{y}_i = \hat{\beta}_0^{*} + \hat{\beta}_1^{*} x_{i1} + ... + \hat{\beta}_p^{*} x_{ip}$$

7. ในการพิจารณาว่าควรจะใช้ y_i ที่ทราบค่าตัวใดมาแทนที่ค่า y_j ที่สูญหายนั้น จะใช้การหาค่าสัมบูรณ์ระหว่างค่าคาดหวังของข้อมูลตัวแปรตามที่ทราบค่า กับค่าคาดหวังของตัวแปรตามที่สูญหายที่ได้ในข้อ 6 ดังนี้

$$D_{ij} = \left| E(\hat{y}_i \mid X) - E(\hat{y}_j \mid X) \right|$$

สำหรับทุกๆ $i; 1 \le i \le m$ และ $j = m+1, ..., n$

8. y_i ตัวใดที่ให้ค่า D_{ij} ต่ำสุด จะได้ว่า $\hat{y}_j^{*} = y_i$

บทที่ 3

วิธีดำเนินการวิจัย

ในงานวิจัยครั้งนี้จะเป็นการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable ในการวิเคราะห์การถดถอยเชิงเส้นพหุ โดยข้อมูลที่ใช้ศึกษาจะเป็นข้อมูลภาคตัดขวาง (Cross-Section Data)และวิธีที่ใช้ในการประมาณค่าสูญหายคือ วิธี EM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) ในการเปรียบเทียบว่าวิธีการประมาณใดเป็นวิธีที่ดีกว่า จะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง(AMSE)ระหว่างค่าจริงกับค่าพยากรณ์ และในการศึกษาจะทำการจำลองข้อมูลให้มีสถานการณ์ที่แตกต่างกันด้วยเทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) โดยใช้โปรแกรม R เวอร์ชั่น 2.13.0 ซึ่งมีแผนการจำลองข้อมูลและขั้นตอนในการวิจัยดังนี้

3.1 แผนการจำลองข้อมูล

ในงานวิจัยนี้จะทำการศึกษาสถานการณ์จำลองที่มีความแตกต่างกันทั้งหมด 162 สถานการณ์ เพื่อใช้ในการเปรียบวิธีการประมาณค่าสูญหาย โดยจะมีรูปแบบเงื่อนไขของการจำลองดังนี้

1. ตัวแปรอิสระที่นำมาศึกษามีการแจกแจงปกติและมี 2 ลักษณะ คือ
   1). $X_1 \sim N(0,300)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$
   2). $X_1 \sim N(0,100)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$
2. ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์กันภายใต้การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ดังนี้

   $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad ; i = 1,2,...,m,m+1,...,n$$

   และกำหนดให้ค่าสัมประสิทธิ์การถดถอยมีค่าเป็น $\beta_0 = 42$ และให้ $\beta_1, \beta_2, \beta_3$ มีค่าเท่ากัน คือ $\beta_1 = \beta_2 = \beta_3 = 1$ และกำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน นั่นคือมีค่าสหสัมพันธ์มีค่าเป็น 0

3. ความคลาดเคลื่อนสุ่มมีการแจกแจงปกติที่มีค่าเฉลี่ยเป็นศูนย์ และมีส่วนเบี่ยงเบนมาตรฐานเป็น 10, 30 และ 90
4. ขนาดของตัวอย่างเท่ากับ 50, 100 และ 200
5. พื้นที่ใต้โค้งปกติของข้อมูลตัวแปรตามจะถูกแบ่งเป็น 3 ช่วงโดยกำหนดให้มีการแบ่งช่วงเป็นอัตราส่วนของ

   ช่วงต้น : ช่วงกลาง : ช่วงปลาย เป็น 1 : 1 : 1 ตามลำดับ
6. ตัวแปรตามมีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น ซึ่งในแต่ละช่วงจะมีสัดส่วนการสูญหายของข้อมูลที่แตกต่างกันตามระดับของการสูญหายแบบ Nonignorable โดยแบ่งเป็น 3 ระดับ คือ ไม่มี, ปานกลาง และสูง ซึ่งในแต่ละช่วงจะมีอัตราส่วนของการสูญหายดังนี้

ไม่มี	1	:	1	:	1
ปานกลาง	7	:	10	:	13
สูง	4	:	10	:	16
7. สัดส่วนของการสูญหายของตัวแปรตามเท่ากับ 10%, 20% และ 30%
8. ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ

3.2 ขั้นตอนในการวิจัย

1. สร้างข้อมูลของตัวแปรอิสระและความคลาดเคลื่อนที่มีการแจกแจงปกติ โดยมีพารามิเตอร์ตามที่กำหนด
2. สร้างข้อมูลของตัวแปรตามจากรูปแบบความสัมพันธ์ $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ โดยกำหนดให้ $\underset{\sim}{\beta}$ เป็นพารามิเตอร์ที่มีค่าคงที่
3. สร้างข้อมูลของตัวแปรตามให้เกิดการสูญหายโดยเกิดขึ้นแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น และมีสัดส่วนของการสูญหายตามที่กำหนด
4. ประมาณค่าตัวแปรตามที่เกิดการสูญหายด้วยวิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM)
5. ประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method : OLS)

6. สร้างสมการถดถอยเชิงเส้นพหุจากค่าประมาณสัมประสิทธิ์การถดถอย เพื่อใช้ในการพยากรณ์
7. คำนวณหาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) ซึ่งเป็นอัตราส่วนระหว่างค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี EM กับวิธีการประมาณค่าสูญหายวิธีอื่นๆ เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย

ซึ่งในแต่ละขั้นตอนมีรายละเอียดดังนี้

1. การสร้างตัวแปรอิสระและความคลาดเคลื่อน

สร้างชุดข้อมูลตัวแปรอิสระจำนวน 3 ตัว ที่มีขนาดตัวอย่างเท่ากับ n โดยที่ตัวแปรแต่ละตัวไม่มีความสัมพันธ์กัน ทุกตัวมีการแจกแจงปกติ ด้วยค่าเฉลี่ย $\mu = 0$ และมีความแปรปรวน σ^2 ที่แตกต่างกันตามแต่ลักษณะของชุดตัวแปรอิสระที่ทำการศึกษาในการจำลองข้อมูลรอบนั้นๆ โดยตัวแปรอิสระแบบที่ 1 จะมีความแปรปรวนเท่ากันคือ $\sigma_k^2 = 300 \quad ; k = 1,2,3$ และตัวแปรอิสระแบบที่ 2 จะมีความแปรปรวนเป็น $\sigma_1^2 = 100, \sigma_2^2 = 300, \sigma_3^2 = 500$

สร้างชุดข้อมูลความคลาดเคลื่อนที่มีขนาดตัวอย่างเท่ากับ n และมีการแจกแจงปกติ ด้วยค่าเฉลี่ย $\mu = 0$ มีส่วนเบี่ยงเบนมาตรฐาน σ = 10 , 30 และ 90

ดังนั้นจะได้ฟังก์ชันความหนาแน่นของการแจกแจงปกติเป็น

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

2. การสร้างข้อมูลตัวแปรตาม

สร้างตัวแปรตามจากรูปแบบความสัมพันธ์ของการถดถอยเชิงเส้นพหุ โดยใช้ข้อมูลตัวแปรอิสระและข้อมูลความคลาดเคลื่อนที่ได้จากข้อ 1 ตามสมการต่อไปนี้

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \qquad ; i = 1,2,...,n$$

โดยมีค่าสัมประสิทธิ์การถดถอยเป็น $\beta_0 = 42$ และให้ $\beta_1, \beta_2, \beta_3$ มีค่าเท่ากัน คือ $\beta_1 = \beta_2 = \beta_3 = 1$

3. การสร้างข้อมูลตัวแปรตามให้เกิดการสูญหาย

สร้างข้อมูลตัวแปรตามให้เกิดการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น โดยนำข้อมูลตัวแปรตามที่มีการแจกแจงปกติที่ได้จากข้อ 2 มาแบ่งออกเป็น 3 ช่วง โดยให้แต่ละช่วงมีอัตราส่วนที่เท่ากันๆคือ 1 : 1 : 1 ซึ่งวิธีการที่ใช้ในการแบ่งช่วงของตัวแปรตามคือ

ถ้า $y_i \leq \bar{y} - z_{\frac{1}{3}}\hat{\sigma}_y$ ตัวแปรตาม y_i นี้จะถูกจัดให้อยู่ในช่วงต้น

ถ้า $\bar{y} - z_{\frac{1}{3}}\hat{\sigma}_y < y_i \leq \bar{y} + z_{\frac{1}{3}}\hat{\sigma}_y$ ตัวแปรตาม y_i นี้จะถูกจัดให้อยู่ในช่วงกลาง

และถ้า $\bar{y} + z_{\frac{1}{3}}\hat{\sigma}_y < y_i$ ตัวแปรตาม y_i นี้จะถูกจัดให้อยู่ในช่วงปลาย

จากนั้นจะทำให้แต่ละช่วงของตัวแปรตามมีการสูญหาย โดยจะสร้างตัวแปรสุ่มที่มีค่าเป็น (0 , 1) ที่มีการแจกแจงทวินามจำนวน 3 ชุด มีขนาดเท่ากับจำนวนตัวแปรตามที่ตกอยู่ในแต่ละช่วง และมีความน่าจะเป็นที่จะเกิดข้อมูลสูญหายตามที่กำหนดไว้ในขอบเขตการวิจัย ซึ่งถ้าเป็นการสูญหายแบบ Nonignorable นั้น ในแต่ละช่วงจะมีความน่าจะเป็นที่จะเกิดการสูญหายที่แตกต่างกัน แล้วทำการจับคู่ตัวแปรตามกับชุดของตัวแปรสุ่มที่สร้างขึ้นมาสำหรับแต่ละช่วง ซึ่งตัวแปรตามใดที่มีตัวแปรสุ่มมีค่าเป็น 1 ตัวแปรตามนั้นก็จะต้องเกิดการสูญหาย

4. การประมาณค่าตัวแปรตามที่เกิดการสูญหาย

เมื่อได้ข้อมูลตัวแปรอิสระตามข้อ 1 และข้อมูลตัวแปรตามที่เกิดการสูญหายตามข้อ 3 แล้ว ขั้นต่อไปคือทำการประมาณค่าข้อมูลที่สูญหายด้วยวิธี EM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) โดยวิธี KNN จะเป็นวิธีที่มีรูปแบบการคำนวณที่ไม่ใช้พารามิเตอร์ วิธี EM จะเป็นวิธีที่มีรูปแบบการคำนวณที่ใช้พารามิเตอร์ และวิธี PMM จะเป็นวิธีที่มีรูปแบบการคำนวณกึ่งพารามิเตอร์ ซึ่งแต่ละวิธีก็จะมีรายละเอียดขั้นตอนตามที่ได้กล่าวมาแล้วในบทที่ 2 และเมื่อเสร็จสิ้นกระบวนการคำนวณแล้ว จะได้ชุดข้อมูลที่สมบูรณ์ที่ค่าสูญหายจะถูกแทนที่ด้วยค่าประมาณ

5. การประมาณค่าสัมประสิทธิ์การถดถอย

เมื่อได้ค่าประมาณข้อมูลสูญหายจากทั้ง 3 วิธีและได้ชุดข้อมูลที่สมบูรณ์จากข้อ 4 แล้ว ก็จะทำการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (OLS) ตามทฤษฎีที่ได้กล่าวมาแล้วในบทที่ 2 ซึ่งวิธี OLS จะไม่พิจารณาหรือตัดชุดข้อมูลที่สูญหายทิ้ง และนำเฉพาะชุดข้อมูลที่สมบูรณ์มาใช้ในการคำนวณ ดังนั้นจึงต้องประมาณค่าที่สูญหายก่อนแล้วจึงหาค่าสัมประสิทธิ์การถดถอย และสำหรับแต่ละวิธีการประมาณค่าสูญหายก็จะให้ค่าประมาณสัมประสิทธิ์การถดถอยที่แตกต่างกัน

6. การสร้างสมการพยากรณ์

สร้างสมการเพื่อใช้ในการพยากรณ์ค่าตัวแปรตาม โดยใช้ค่าประมาณสัมประสิทธิ์การถดถอยที่ได้จากข้อ 5 และสมการพยากรณ์จะมีรูปแบบดังนี้

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} \quad ;i = 1,2,...,n$$

ซึ่งแต่ละวิธีการประมาณค่าสูญหายก็จะให้สมการพยากรณ์ออกมาวิธีละหนึ่งสมการ

7. การหาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE)

คำนวณหาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ของแต่ละวิธีประมาณค่าสูญหายที่มีการทำซ้ำทั้งหมด 5,000 รอบในแต่ละสถานการณ์จำลอง ซึ่งมีสูตรดังนี้

$$MSE_t = \frac{\sum_{i=1}^{n} (y_i' - \hat{y}_{ti})^2}{n}$$

$$AMSE = \frac{1}{5,000} \sum_{t=1}^{5,000} MSE_t$$

เมื่อ y_i' แทน ค่าจริงของข้อมูลตัวแปรตามตัวที่ i
โดยค่าจริงนี้เป็นค่าสังเกตของตัวแปรตามที่ไม่รวมค่าความคลาดเคลื่อน
ซึ่งก็คือ $y_i' = y_i - \varepsilon_i$

$\hat{y}_{ti}$ แทน ค่าพยากรณ์ของข้อมูลตัวแปรตามตัวที่ i จากการทำซ้ำรอบที่ t

MSE_t แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำรอบที่ t

$AMSE$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำทั้งหมด 5,000 รอบ

ในงานวิจัยนี้ใช้ค่าจริงในการหาค่า AMSE แทนการใช้ค่าสังเกต ทั้งนี้เป็นเพราะต้องการเปรียบเทียบว่าวิธีการประมาณค่าสูญหายใดจะดีกว่า ถ้าพิจารณาจากผลต่างระหว่างค่าพยากรณ์กับค่าจริงที่เป็นค่าของตัวแปรตามที่ควรจะได้จริงๆโดยไม่มีค่าความคลาดเคลื่อนมารบกวน

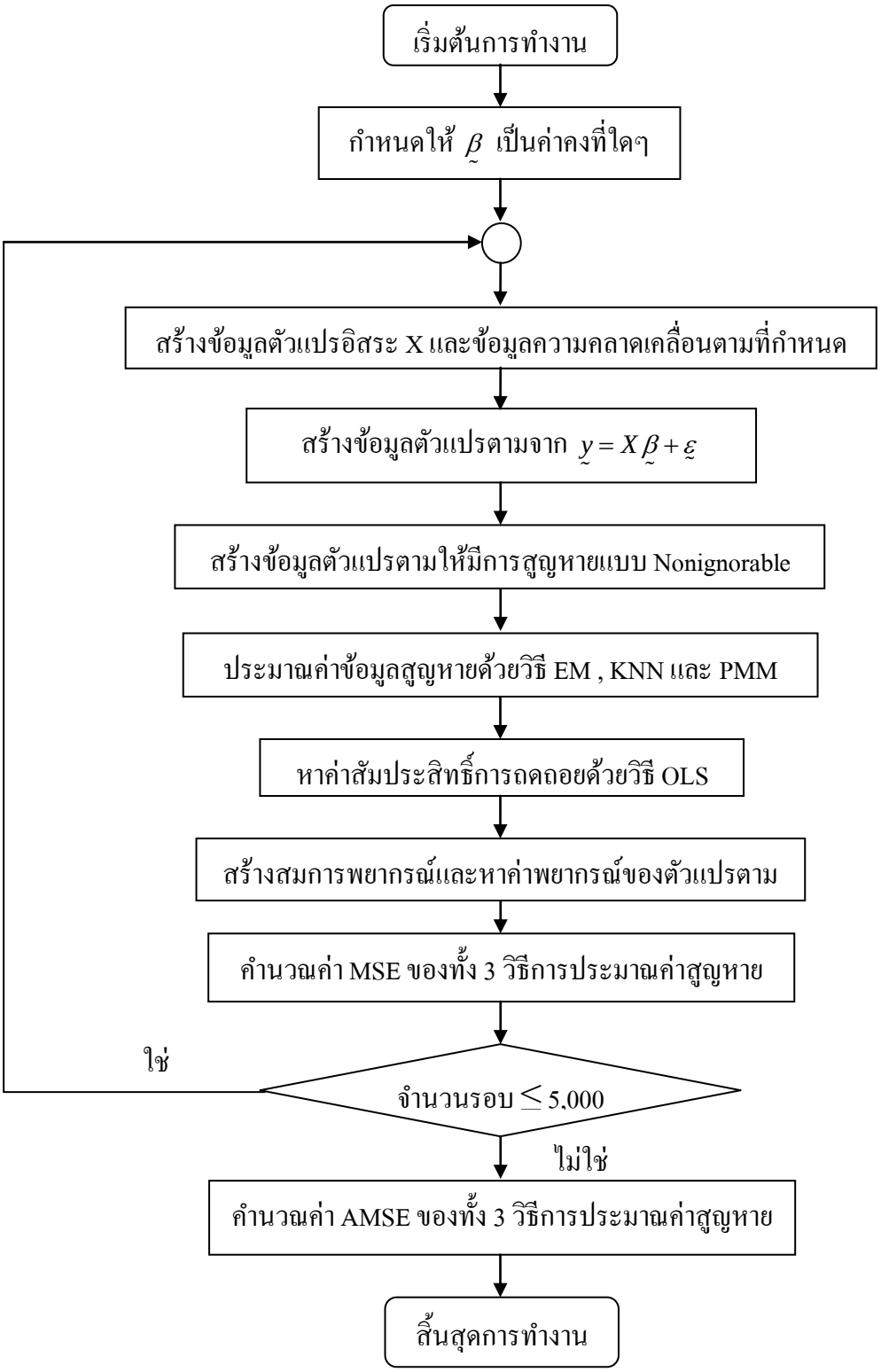
คำนวณค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency : RE) ซึ่งเป็นอัตราส่วนระหว่างค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธี EM กับวิธีการประมาณค่าสูญหายวิธีอื่นๆ ซึ่งในการศึกษาครั้งนี้เลือกใช้วิธี EM เป็นเกณฑ์ในการเปรียบเทียบเนื่องจากวิธี EM เป็นวิธีที่นิยมใช้ในการประมาณค่าสูญหายมากที่สุด ซึ่งวิธีใดที่ให้ค่า RE มากกว่า 1 แสดงว่ามีประสิทธิภาพมากกว่าวิธี EM

$$RE = \frac{AMSE_{EM}}{AMSE_r}, r = 1, 2$$

เมื่อ $AMSE_{EM}$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากการประมาณค่าสูญหายด้วยวิธี EM

$AMSE_r$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของที่ได้จากการประมาณค่าสูญหายแต่ละวิธี

<u>ภาพที่ 3.1</u> แผนผังการเขียนโปรแกรม



บทที่ 4

ผลการวิจัย

ในงานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable ในการวิเคราะห์การถดถอยเชิงเส้นพหุ โดยมีวิธีที่ใช้ในการประมาณค่าสูญหายทั้งหมด 3 วิธีคือ วิธีEM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) เกณฑ์ที่ใช้ในการเปรียบเทียบว่าวิธีการประมาณใดจะเป็นวิธีที่ดีกว่า จะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ระหว่างค่าจริงกับค่าพยากรณ์ ซึ่งวิธีการประมาณค่าสูญหายวิธีใดที่ให้ค่า AMSE ต่ำกว่าจะเป็นวิธีที่ดีกว่า

ในการนำเสนอผลการวิจัยจะแสดงในรูปแบบของตารางและกราฟ โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่างๆดังนี้

n	แทน ขนาดของตัวอย่าง
σ	แทน ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
X_1	แทน ชุดตัวแปรอิสระแบบที่ 1: $X_1 \sim N(0,300)$, $X_2 \sim N(0,300)$, $X_3 \sim N(0,300)$
X_2	แทน ชุดตัวแปรอิสระแบบที่ 2: $X_1 \sim N(0,100)$, $X_2 \sim N(0,300)$, $X_3 \sim N(0,500)$
None	แทน ระดับของการสูญหายแบบ Nonignorable ในระดับไม่มี
Medium	แทน ระดับของการสูญหายแบบ Nonignorable ในระดับปานกลาง
High	แทน ระดับของการสูญหายแบบ Nonignorable ในระดับสูง
EM	แทน การประมาณค่าสูญหายของตัวแปรตามด้วยวิธี EM Algorithm
KNN	แทน การประมาณค่าสูญหายของตัวแปรตามด้วยวิธี KNN
PMM	แทน การประมาณค่าสูญหายของตัวแปรตามด้วยวิธี PMM
AMSE	แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์
RE	แทน ค่าประสิทธิภาพสัมพัทธ์ ซึ่งเป็นอัตราส่วนระหว่างค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์ของวิธี EM Algorithm กับวิธีการประมาณค่าสูญหายวิธีอื่นๆ

ในการนำเสนอผลการเปรียบเทียบวิธีวิธีการประมาณค่าสูญหายทั้ง 3 วิธีนั้น จะแบ่งการนำเสนอออกเป็น 3 ส่วน โดยใน 2 ส่วนแรกจะใช้ลักษณะของชุดตัวแปรอิสระเป็นเกณฑ์ในการแบ่ง และในแต่ละส่วนจะถูกแบ่งออกเป็น 3 ส่วนย่อย โดยเกณฑ์ที่ใช้แบ่งคือค่าของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน และในส่วนสุดท้ายจะเป็นการเปรียบเทียบผลการวิจัยที่ได้จากส่วนที่ 1 และส่วนที่ 2 ซึ่งจะมีการนำเสนอทั้งหมดดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 1 : $X_1 \sim N(0,300), X_2 \sim N(0,300), X_3 \sim N(0,300)$

1.1 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 10

1.2 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 30

1.3 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 90

<u>ส่วนที่ 2</u> ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300), X_3 \sim N(0,500)$

2.1 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 10

2.2 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 30

2.3 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 90

<u>ส่วนที่ 3</u> การเปรียบเทียบผลการวิจัยของชุดตัวแปรอิสระแบบที่ 1 กับชุดตัวแปรอิสระแบบที่ 2

การนำเสนอรายละเอียดจะประกอบด้วย ตารางแสดงผล รูปภาพ และการอธิบายผลที่ได้จากตารางและรูปภาพที่แสดง เมื่อจบการแสดงผลของแต่ละส่วนแล้วจะมีการอธิบายผลการวิจัยโดยรวมสำหรับส่วนนั้นๆในตอนท้ายอีกครั้ง และเมื่อนำเสนอผลการวิจัยครบทุกส่วนแล้วจะมีการอธิบายและสรุปผลการวิจัยทั้งหมดอีกครั้งในตอนท้ายบท

4.1 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 1

ในส่วนนี้ผู้วิจัยทำการศึกษาในกรณีที่ตัวแปรอิสระเป็นแบบที่ 1 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 10, 30 และ 90 ขนาดของตัวอย่างเท่ากับ 50, 100 และ 200 สัดส่วนของการสูญหายของข้อมูลตัวแปรตามเท่ากับ 10%, 20% และ 30% ระดับของการสูญหายแบบ Nonignorable เป็น ไม่มี, ปานกลาง และสูง ซึ่งผลการวิจัยในส่วนนี้จะนำเสนอในตารางที่ 4.1.1 – 4.1.3

ตารางที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐาน
4.1.1	แบบที่ 1	10
4.1.2	แบบที่ 1	30
4.1.3	แบบที่ 1	90

<u>ตารางที่ 4.1.1</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

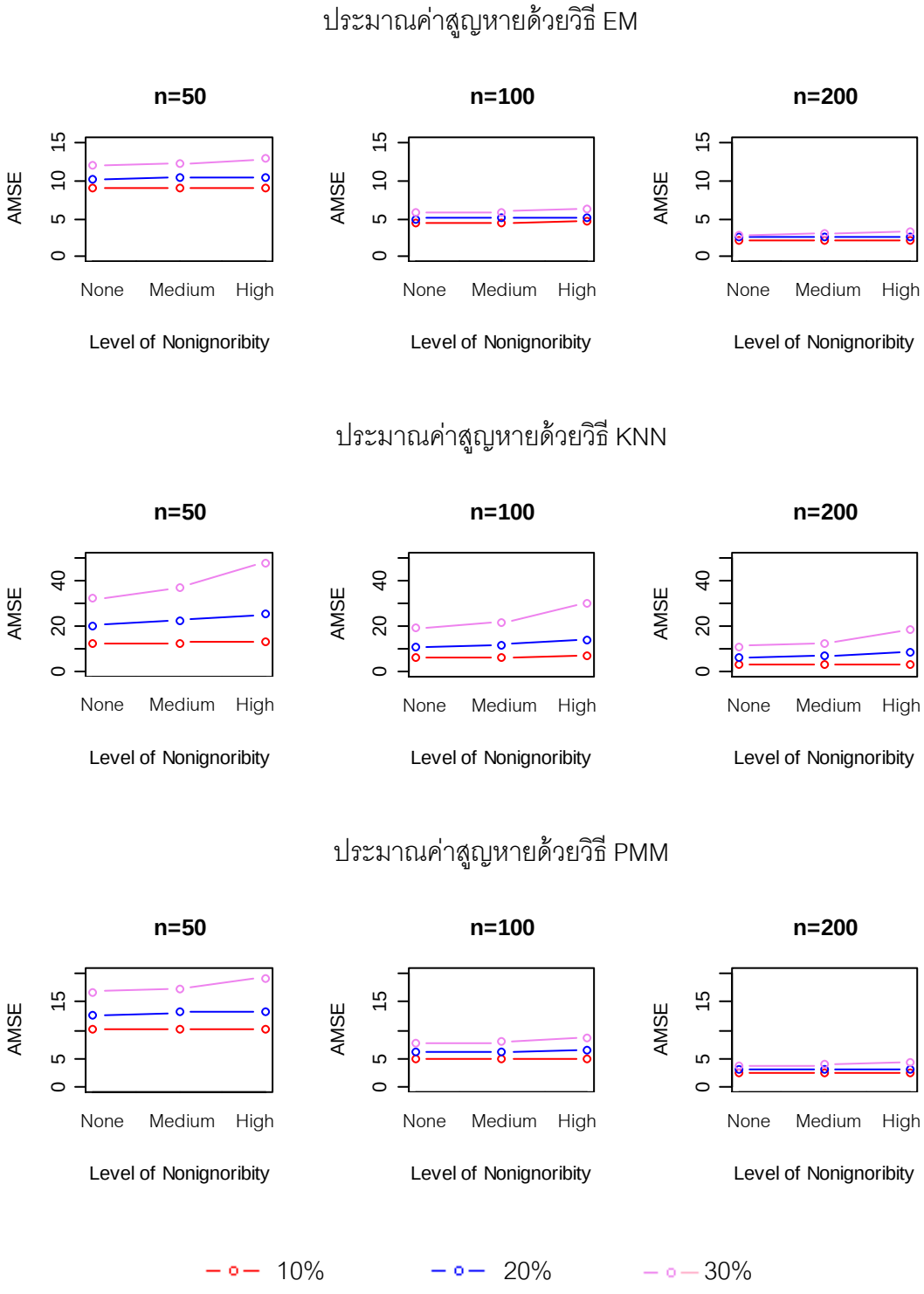
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	9.0237	12.6320	10.1861
			RE	1.0000	0.7144	0.8859
		Medium	AMSE	9.0933	12.8318	10.2751
			RE	1.0000	0.7087	0.8850
		High	AMSE	8.9210	13.4588	10.1717
			RE	1.0000	0.6628	0.8770
	20	None	AMSE	10.1741	20.4897	12.6822
			RE	1.0000	0.4965	0.8022
		Medium	AMSE	10.3638	22.7250	13.1582
			RE	1.0000	0.4561	0.7876
		High	AMSE	10.3063	25.3228	13.1950
			RE	1.0000	0.4070	0.7811
	30	None	AMSE	12.0703	32.0767	16.7169
			RE	1.0000	0.3763	0.7220
		Medium	AMSE	12.2072	36.8893	17.3599
			RE	1.0000	0.3309	0.7032
		High	AMSE	12.8524	48.0367	19.2496
			RE	1.0000	0.2676	0.6677

<u>ตารางที่ 4.1.1(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	4.4956	6.1594	5.0095
			RE	1.0000	0.7299	0.8974
		Medium	AMSE	4.4541	6.3886	4.9630
			RE	1.0000	0.6972	0.8975
		High	AMSE	4.5744	6.8747	5.1038
			RE	1.0000	0.6654	0.8963
	20	None	AMSE	5.0012	11.0722	6.1101
			RE	1.0000	0.4517	0.8185
		Medium	AMSE	5.1163	12.0137	6.2445
			RE	1.0000	0.4259	0.8193
		High	AMSE	5.2303	14.0539	6.5654
			RE	1.0000	0.3722	0.7967
	30	None	AMSE	5.8412	19.3745	7.7537
			RE	1.0000	0.3015	0.7533
		Medium	AMSE	5.8759	21.7031	7.9074
			RE	1.0000	0.2707	0.7431
		High	AMSE	6.2953	30.1336	8.6923
			RE	1.0000	0.2089	0.7242

<u>ตารางที่ 4.1.1(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	2.1950	3.1957	2.4284
			RE	1.0000	0.6868	0.9039
		Medium	AMSE	2.2293	3.3055	2.4675
			RE	1.0000	0.6744	0.9035
		High	AMSE	2.2097	3.5298	2.4727
			RE	1.0000	0.6260	0.8936
	20	None	AMSE	2.5081	6.2789	3.0415
			RE	1.0000	0.3994	0.8246
		Medium	AMSE	2.5045	6.9342	3.0722
			RE	1.0000	0.3612	0.8152
		High	AMSE	2.6476	8.4815	3.2250
			RE	1.0000	0.3122	0.8210
	30	None	AMSE	2.8697	11.1112	3.7936
			RE	1.0000	0.2583	0.7564
		Medium	AMSE	3.0064	12.8433	3.9234
			RE	1.0000	0.2341	0.7663
		High	AMSE	3.3023	18.5775	4.3850
			RE	1.0000	0.1778	0.7531



<u>ภาพที่</u> 4.1.1 แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

จากตารางที่ 4.1.1 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 พบว่า สำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี PMM และวิธี KNN ตามลำดับ ดังนั้นวิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธี PMM และวิธี KNN เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างใหญ่ขึ้น ค่า AMSE จะลดลง และจากภาพที่ 4.1.1 จะเห็นว่า เมื่อสัดส่วนของการสูญหายสูงขึ้น ค่า AMSE จะมีค่าเพิ่มขึ้น โดยที่สัดส่วนของการสูญหายเท่ากับ 30% จะมีค่า AMSE มากที่สุด ส่วนการเพิ่มระดับของการสูญหายแบบ Nonignorable จะทำให้ค่า AMSE มีแนวโน้มเพิ่มขึ้นเพียงเล็กน้อย โดยเฉพาะกรณีที่สัดส่วนของการสูญหายเท่ากับ 10% จะเห็นว่าค่า AMSE มีอัตราการเพิ่มขึ้นไม่ถึง 1 หน่วย และที่ขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 10% ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี จะมีค่า AMSE ต่ำที่สุด เพราะขนาดตัวอย่างที่เพิ่มขึ้นจะทำให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง ข้อมูลที่มีสัดส่วนของการสูญหายน้อยจะทำให้สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงมากขึ้น และการที่ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มีก็คือการสูญหายแบบสุ่ม ซึ่งวิธี EM วิธี PMM และวิธี KNN ก็เป็นวิธีการประมาณค่าสูญหายที่ทำภายใต้เงื่อนไขของข้อมูลที่มีการสูญหายแบบสุ่มจึงสามารถประมาณค่าสูญหายได้ใกล้เคียงมากกว่ากรณีอื่น

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี EM กับค่า AMSE ของวิธีการประมาณค่าสูญหายวิธีอื่นๆ พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN และ วิธี PMM มีแนวโน้มที่จะลงลงเรื่อยๆ ถ้าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable สูงขึ้น โดยค่า RE ของวิธี PMM จะมีค่ามากกว่าวิธี KNN และมีค่าน้อยกว่า 1 เพราะในกรณีนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ ซึ่งที่สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มีจะให้ค่า RE มากที่สุด แสดงว่า ณ จุดนี้ ค่า AMSE ของวิธี KNN และวิธี PMM จะมีค่าใกล้เคียงค่า AMSE ของวิธี EM มากที่สุด เมื่อพิจารณาที่ขนาดตัวอย่างต่างกัน พบว่าค่า RE จะไม่มีความแตกต่างกันมาก ดังนั้นปัจจัยที่มีอิทธิพลต่อค่า RE ของแต่ละวิธีการประมาณคือสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable

<u>ตารางที่ 4.1.2</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

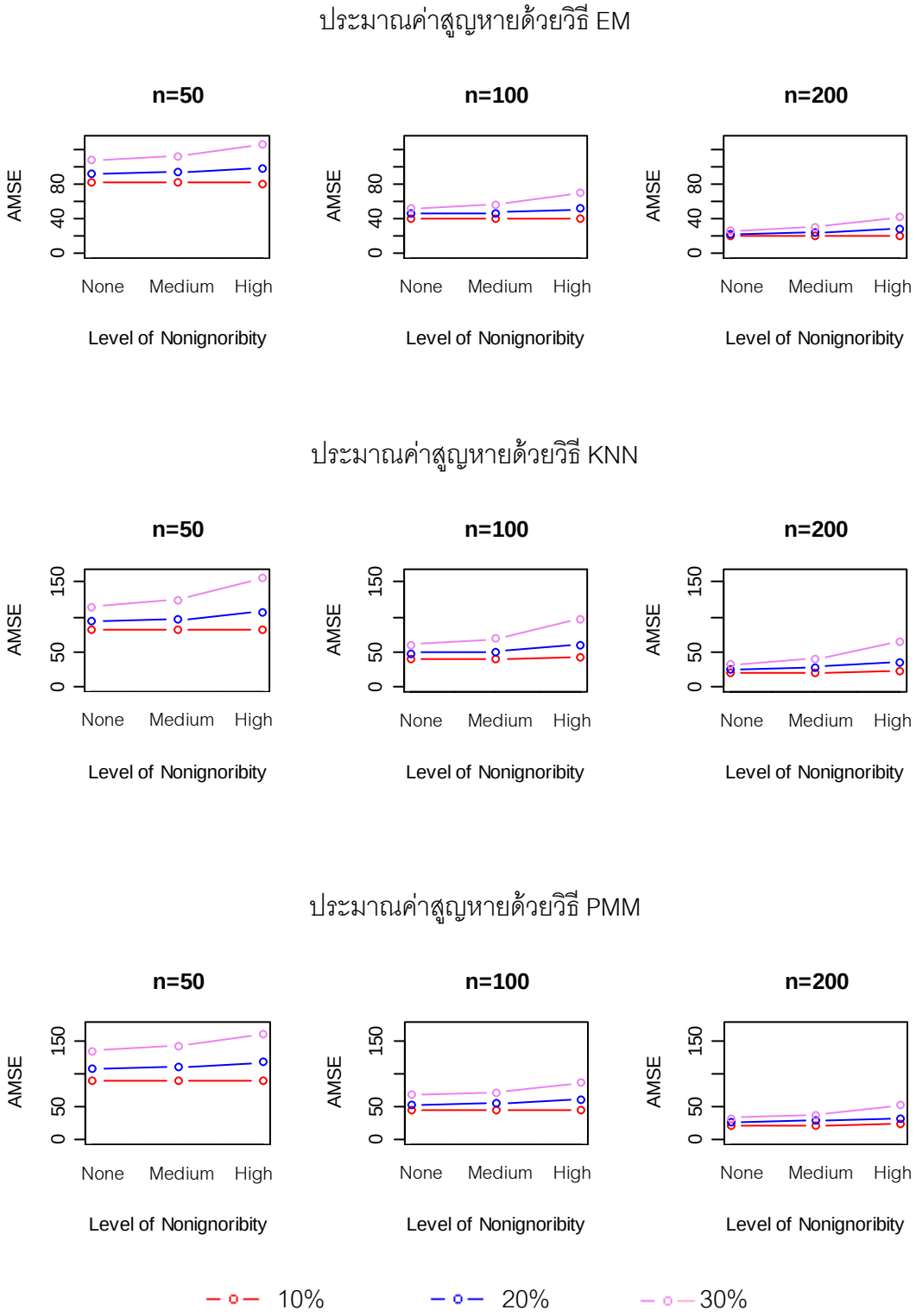
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	81.7730	81.2173	89.4617
			RE	1.0000	1.0068	0.9141
		Medium	AMSE	83.1204	82.6773	90.3901
			RE	1.0000	1.0054	0.9196
		High	AMSE	81.1049	81.5568	88.9791
			RE	1.0000	0.9945	0.9115
	20	None	AMSE	93.1806	94.0119	107.8598
			RE	1.0000	0.9912	0.8639
		Medium	AMSE	93.4735	95.7582	109.3554
			RE	1.0000	0.9761	0.8548
		High	AMSE	98.6827	106.8010	117.2276
			RE	1.0000	0.9240	0.8418
	30	None	AMSE	108.3765	113.5465	134.8306
			RE	1.0000	0.9545	0.8038
		Medium	AMSE	112.8985	123.7336	141.5796
			RE	1.0000	0.9124	0.7974
		High	AMSE	125.7497	155.1707	160.1182
			RE	1.0000	0.8104	0.7854

<u>ตารางที่ 4.1.2(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	40.3393	40.6085	44.5548
			RE	1.0000	0.9934	0.9054
		Medium	AMSE	39.9717	40.8297	44.0669
			RE	1.0000	0.9790	0.9071
		High	AMSE	40.8240	41.9513	44.8888
			RE	1.0000	0.9731	0.9094
	20	None	AMSE	45.4164	48.2635	53.8285
			RE	1.0000	0.9410	0.8437
		Medium	AMSE	46.8172	50.7981	55.3487
			RE	1.0000	0.9216	0.8459
		High	AMSE	51.6191	60.2308	61.1555
			RE	1.0000	0.8570	0.8441
	30	None	AMSE	52.8838	59.9974	67.4639
			RE	1.0000	0.8814	0.7839
		Medium	AMSE	55.7447	68.2236	70.9079
			RE	1.0000	0.8171	0.7862
		High	AMSE	69.9477	97.0956	87.3451
			RE	1.0000	0.7204	0.8008

<u>ตารางที่ 4.1.2(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	20.1262	20.6195	22.0387
			RE	1.0000	0.9761	0.9132
		Medium	AMSE	20.2733	20.7705	22.2247
			RE	1.0000	0.9761	0.9122
		High	AMSE	21.0192	22.1423	23.0023
			RE	1.0000	0.9493	0.9138
	20	None	AMSE	22.3600	24.6512	26.3819
			RE	1.0000	0.9071	0.8476
		Medium	AMSE	24.3507	27.9733	28.8341
			RE	1.0000	0.8705	0.8445
		High	AMSE	28.1676	35.0636	33.1959
			RE	1.0000	0.8033	0.8485
	30	None	AMSE	26.2913	32.3563	32.8274
			RE	1.0000	0.8126	0.8009
		Medium	AMSE	30.1697	39.8052	37.9859
			RE	1.0000	0.7579	0.7942
		High	AMSE	43.0112	63.8134	51.7062
			RE	1.0000	0.6740	0.8318



<u>ภาพที่</u> 4.1.2 แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

จากตารางที่ 4.1.2 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 พบว่า เมื่อขนาดตัวอย่างเท่ากับ 50 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ ยกเว้นที่สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี – ปานกลาง ที่วิธี KNN จะมีค่า AMSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 100 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ ยกเว้นที่สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ที่วิธี PMM จะมีค่า AMSE ต่ำกว่าวิธี KNN และเมื่อขนาดตัวอย่างเท่ากับ 200 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ แต่ถ้าสัดส่วนของการสูญหายเพิ่มขึ้น และระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น วิธี PMM จะมีค่า AMSE ต่ำกว่าวิธี KNN เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างลดลง สัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น ค่า AMSE จะมีค่าเพิ่มขึ้น และจากภาพที่ 4.1.2 จะเห็นว่าที่สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ค่า AMSE จะมีค่าสูงที่สุด แต่อย่างไรก็ตาม จะเห็นได้ว่าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable ค่อนข้างมีอิทธิพลต่อค่า AMSE มากกว่าขนาดตัวอย่าง เพราะถึงแม้ว่าตัวอย่างจะมีขนาดใหญ่ แต่ถ้ามีสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ค่า AMSE ที่ได้ก็จะสูงกว่าด้วย

เมื่อพิจารณาค่า RE พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN และ วิธี PMM มีแนวโน้มที่จะลงลงเรื่อยๆ ถ้าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น โดยค่า RE ของวิธี KNN และวิธี PMM ที่ได้จะมีค่าน้อยกว่า 1 เพราะในกรณีนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ ยกเว้นที่ขนาดตัวอย่างเท่ากับ 50 สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี – ปานกลางที่ค่า RE ของวิธี KNN จะมีค่ามากกว่า 1 เนื่องจาก ณ จุดนี้วิธี KNN เป็นวิธีประมาณค่าสูญหายที่ดีที่สุด เมื่อพิจารณาที่ขนาดตัวอย่าง พบว่าการเพิ่มขึ้นของขนาดตัวอย่างจะไม่ทำให้ค่า RE ของแต่ละวิธีแตกต่างกันมาก

ตารางที่ 4.1.3 แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

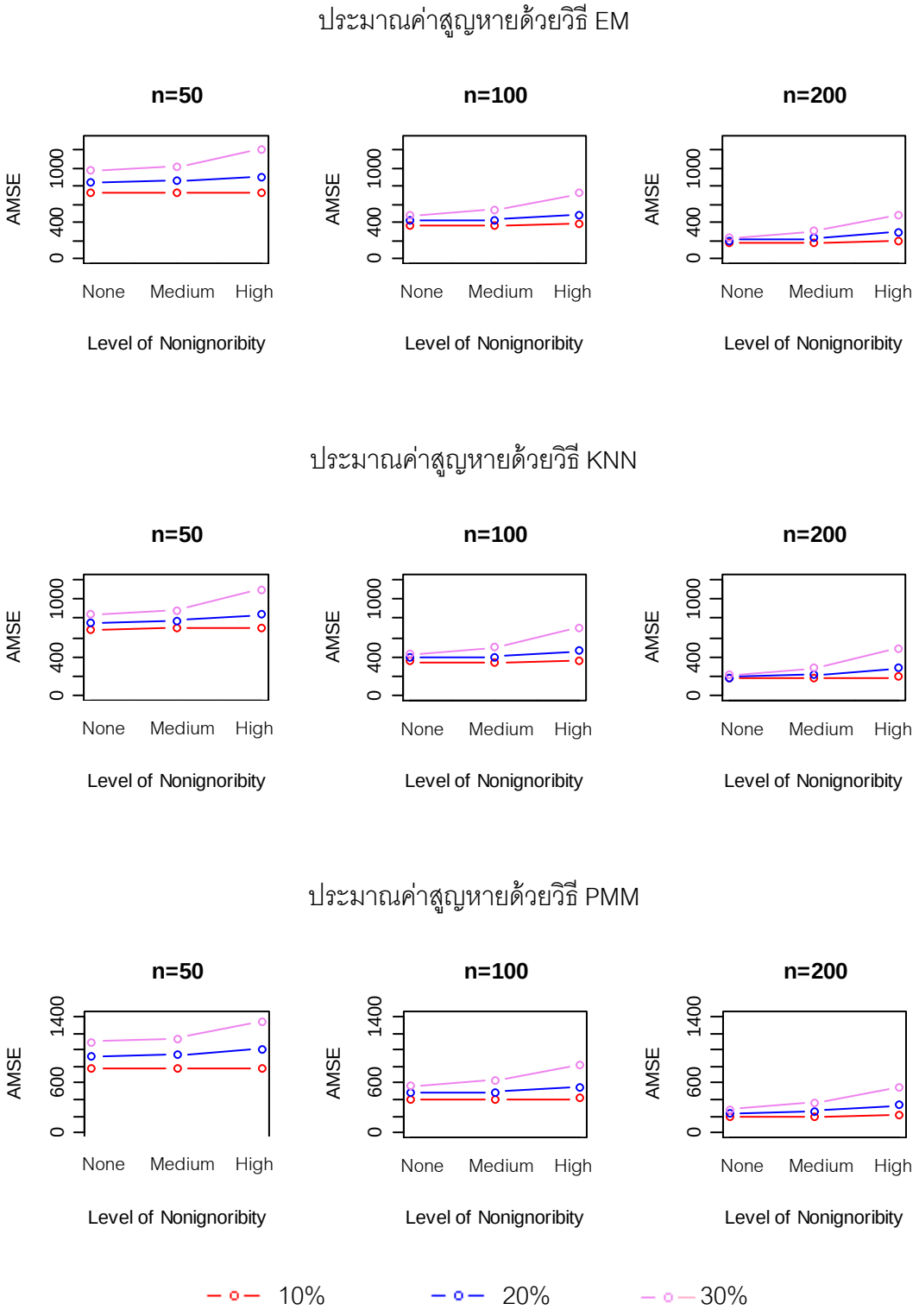
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	723.1410	685.5151	768.9203
			RE	1.0000	1.0549	0.9405
		Medium	AMSE	736.1830	699.2490	778.5785
			RE	1.0000	1.0528	0.9455
		High	AMSE	737.8014	702.7598	781.9020
			RE	1.0000	1.0499	0.9436
	20	None	AMSE	834.9311	747.3564	922.3917
			RE	1.0000	1.1172	0.9052
		Medium	AMSE	857.3068	775.7086	940.4895
			RE	1.0000	1.1052	0.9116
		High	AMSE	909.5812	837.0822	1004.505
			RE	1.0000	1.0866	0.9055
	30	None	AMSE	980.3351	843.1639	1098.473
			RE	1.0000	1.1627	0.8925
		Medium	AMSE	1011.390	888.2242	1137.105
			RE	1.0000	1.1387	0.8894
		High	AMSE	1204.290	1099.207	1341.930
			RE	1.0000	1.0956	0.8974

<u>ตารางที่ 4.1.3(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	367.3659	352.5124	397.0479
			RE	1.0000	1.0421	0.9252
		Medium	AMSE	362.6913	349.1303	392.0174
			RE	1.0000	1.0388	0.9252
		High	AMSE	380.9417	368.0581	411.4513
			RE	1.0000	1.0350	0.9258
	20	None	AMSE	423.5191	392.5287	477.4507
			RE	1.0000	1.0790	0.8870
		Medium	AMSE	433.1827	406.0306	493.3501
			RE	1.0000	1.0669	0.8780
		High	AMSE	488.1020	467.4740	553.4344
			RE	1.0000	1.0441	0.8820
	30	None	AMSE	474.3056	425.7903	563.5465
			RE	1.0000	1.1139	0.8416
		Medium	AMSE	538.1941	496.3531	634.3100
			RE	1.0000	1.0843	0.8485
		High	AMSE	723.2525	704.7652	827.8240
			RE	1.0000	1.0262	0.8737

<u>ตารางที่ 4.1.3(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	181.0283	174.9084	196.6516
			RE	1.0000	1.0350	0.9206
		Medium	AMSE	182.2302	177.2552	198.4730
			RE	1.0000	1.0281	0.9182
		High	AMSE	196.6496	192.9311	214.7278
			RE	1.0000	1.0193	0.9158
	20	None	AMSE	201.8056	191.3193	233.0671
			RE	1.0000	1.0548	0.8659
		Medium	AMSE	225.8031	216.7850	261.3875
			RE	1.0000	1.0416	0.8639
		High	AMSE	291.3785	287.7619	329.4321
			RE	1.0000	1.0126	0.8845
	30	None	AMSE	229.7940	214.4574	285.5349
			RE	1.0000	1.0715	0.8048
		Medium	AMSE	301.1511	289.7244	361.7260
			RE	1.0000	1.0394	0.8325
		High	AMSE	484.6579	493.1248	551.6971
			RE	1.0000	0.9828	0.8785



<u>ภาพที่</u> 4.1.3 แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

จากตารางที่ 4.1.3 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 พบว่า สำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี KNN วิธี EM และวิธี PMM ตามลำดับ ดังนั้นวิธี KNN เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธี EM และวิธี PMM ยกเว้นในกรณีขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ที่วิธี EM จะสามารถประมาณค่าสูญหายได้ดีกว่าวิธี KNN และวิธี PMM ตามลำดับ เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างใหญ่ขึ้น ค่า AMSE จะลดลง และจากภาพที่ 4.1.3 จะเห็นว่าเมื่อข้อมูลมีสัดส่วนของการสูญหายและมีระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น จะทำให้ค่า AMSE มีค่าสูงขึ้นด้วย ซึ่งที่ขนาดตัวอย่างเท่ากับ 50 สัดส่วนของการสูญหายเท่ากับ 30% ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง จะมีค่า AMSE สูงที่สุด แต่อย่างไรก็ตาม จะเห็นได้ว่าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable จะมีอิทธิพลต่อค่า AMSE มากกว่าขนาดตัวอย่าง เพราะถึงแม้ว่าตัวอย่างจะมีขนาดใหญ่ แต่ถ้ามีสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable สูง ค่า AMSE ที่ได้ก็จะสูงกว่าด้วย

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี EM กับค่า AMSE ของวิธีการประมาณค่าสูญหายวิธีอื่นๆ พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN มีค่ามากกว่า 1 เพราะในกรณีนี้วิธี KNN เป็นวิธีประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุด ยกเว้นที่ขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูงที่ค่า RE จะมีค่าน้อยกว่า 1 เนื่องจากที่จุดนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ และที่ระดับของการสูญหายแบบ Nonignorable ระดับเดียวกัน ค่า RE มีแนวโน้มเพิ่มขึ้นเมื่อมีสัดส่วนของการสูญหายเพิ่มขึ้น แต่วิธี PMM จะมีค่า RE ต่ำกว่า 1 สำหรับทุกขนาดตัวอย่าง และมีแนวโน้มลดลงเมื่อสัดส่วนของการสูญหายเพิ่มขึ้น เมื่อพิจารณาที่ขนาดตัวอย่าง พบว่าการเพิ่มขึ้นของขนาดตัวอย่างจะไม่ทำให้ค่า RE ของแต่ละวิธีประมาณค่าสูญหายแตกต่างกันมาก

สรุปส่วนที่ 4.1 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 1

เมื่อพิจารณาผลการวิจัยที่ได้ดังตารางที่ 4.1.1 – 4.1.3 ซึ่งเป็นผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10, 30 และ 90 ตามลำดับ พบว่าสำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable ถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูงขึ้น ค่า AMSE ของแต่ละวิธีประมาณค่าสูญหายจะมีค่าสูงขึ้นด้วย เพราะการเพิ่มส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน จะทำให้ข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งจะส่งผลให้มีค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการประมาณค่าสูญหายมากขึ้น ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 – 30 โดยส่วนมากวิธีการประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุดคือวิธี EM ส่วนกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 โดยส่วนมากวิธีการประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุดคือวิธี KNN

จากผลการวิจัยในส่วนที่ 4.1 สามารถสรุปปัจจัยที่ส่งผลต่อค่า AMSE ได้ดังนี้

1. **ขนาดตัวอย่าง** เมื่อตัวอย่างมีขนาดใหญ่ขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะทำให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง
2. **สัดส่วนของการสูญหาย** เมื่อข้อมูลมีสัดส่วนของการสูญหายเพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะสูงขึ้น เพราะเมื่อข้อมูลมีการสูญหายมากขึ้น จะมีความคลาดเคลื่อนจากการประมาณค่าสูญหายและการประมาณค่าพารามิเตอร์มากขึ้นด้วย
3. **ระดับของการสูญหายแบบ Nonignorable** เมื่อระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะมีแนวโน้มสูงขึ้น เพราะวิธี KNN วิธี EM และวิธี PMM เป็นวิธีการประมาณค่าสูญหายที่ทำภายใต้เงื่อนไขของข้อมูลที่มีการสูญหายแบบสุ่ม ดังนั้นเมื่อมีการเพิ่มระดับความสัมพันธ์ของ

ข้อมูลสูญหายกับตัวแปรตาม จะส่งผลให้มีค่าความคลาดเคลื่อนจากการประมาณค่าสูญหายมากขึ้น

4. **ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน** เมื่อความคลาดเคลื่อนมีความแปรปรวนมากขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะสูงขึ้น เพราะข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งส่งผลให้ความสามารถในการประมาณค่าสูญหายลดลง ค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการพยากรณ์เพิ่มขึ้น

4.2 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 2

ในส่วนนี้ผู้วิจัยทำการศึกษาในกรณีที่ตัวแปรอิสระเป็นแบบที่ 2 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 10, 30 และ 90 ขนาดของตัวอย่างเท่ากับ 50, 100 และ 200 สัดส่วนของการสูญหายของข้อมูลตัวแปรตามเท่ากับ 10%, 20% และ 30% ระดับของการสูญหายแบบ Nonignorable เป็น ไม่มี, ปานกลาง และสูง ซึ่งผลการวิจัยในส่วนนี้จะนำเสนอในตารางที่ 4.2.1 – 4.2.3

ตารางที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐาน
4.2.1	แบบที่ 2	10
4.2.2	แบบที่ 2	30
4.2.3	แบบที่ 2	90

<u>ตารางที่ 4.2.1</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

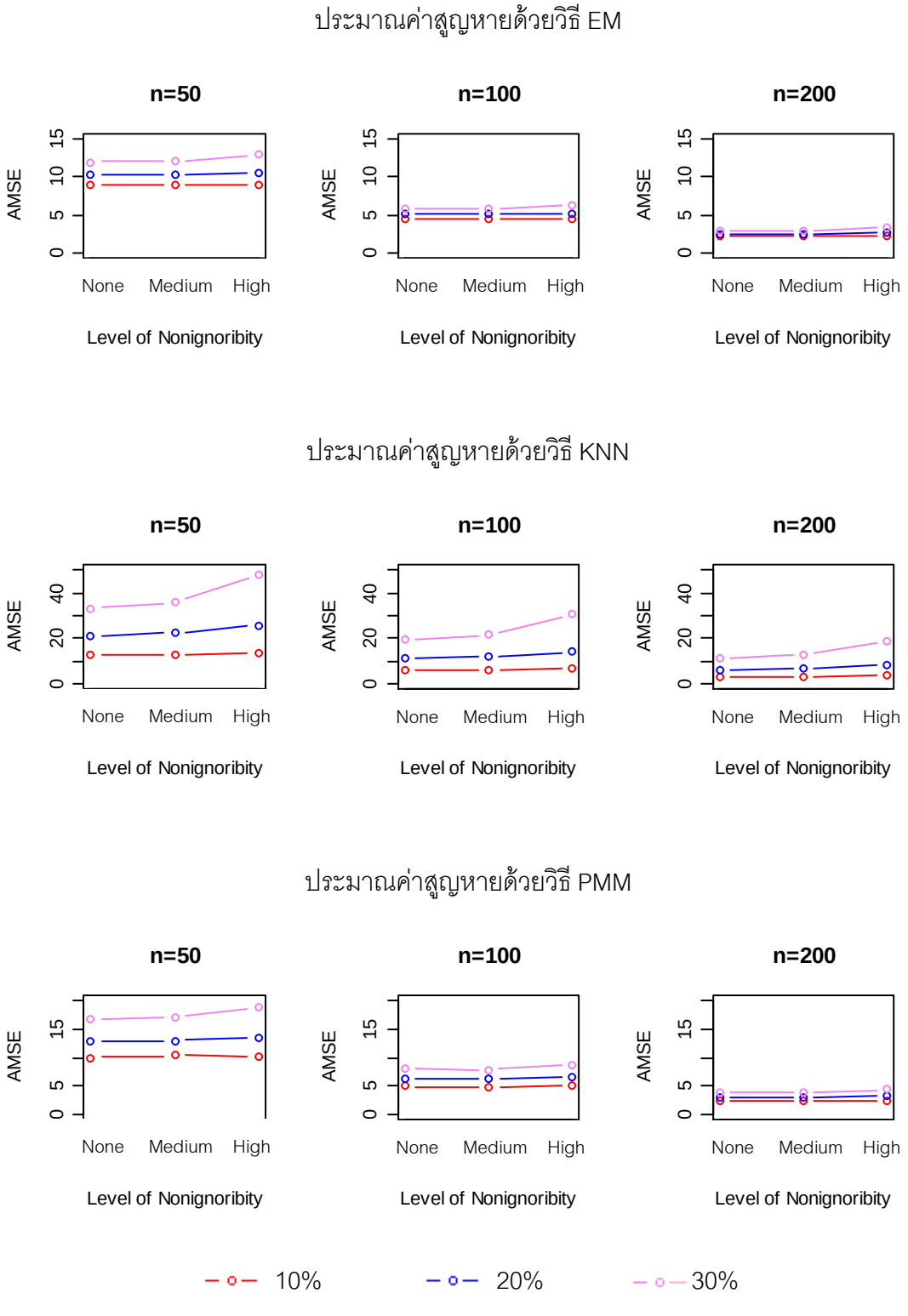
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	8.9603	12.7995	10.0098
			RE	1.0000	0.7000	0.8952
		Medium	AMSE	9.0842	12.9477	10.3870
			RE	1.0000	0.7016	0.8746
		High	AMSE	9.0156	13.3298	10.2437
			RE	1.0000	0.6763	0.8801
	20	None	AMSE	10.2383	20.9559	12.9312
			RE	1.0000	0.4886	0.7918
		Medium	AMSE	10.2807	22.4611	13.0314
			RE	1.0000	0.4577	0.7889
		High	AMSE	10.6238	25.7499	13.4337
			RE	1.0000	0.4126	0.7908
	30	None	AMSE	11.9962	33.0074	16.8116
			RE	1.0000	0.3634	0.7136
		Medium	AMSE	12.1204	35.6668	17.0003
			RE	1.0000	0.3398	0.7130
		High	AMSE	12.9138	48.1226	18.8703
			RE	1.0000	0.2684	0.6843

<u>ตารางที่ 4.2.1(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	4.4394	6.3134	4.9795
			RE	1.0000	0.7032	0.8915
		Medium	AMSE	4.4309	6.3009	4.9103
			RE	1.0000	0.7032	0.9024
		High	AMSE	4.4652	6.7116	4.9895
			RE	1.0000	0.6653	0.8949
	20	None	AMSE	5.0830	11.1853	6.1984
			RE	1.0000	0.4544	0.8201
		Medium	AMSE	5.1168	11.8822	6.2990
			RE	1.0000	0.4306	0.8123
		High	AMSE	5.2821	14.0355	6.5258
			RE	1.0000	0.3763	0.8094
	30	None	AMSE	5.9128	19.3992	7.9648
			RE	1.0000	0.3048	0.7424
		Medium	AMSE	5.8501	21.6679	7.9486
			RE	1.0000	0.2700	0.7360
		High	AMSE	6.3923	30.5063	8.7632
			RE	1.0000	0.2095	0.7294

<u>ตารางที่ 4.2.1(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	2.2491	3.2332	2.4816
			RE	1.0000	0.6956	0.9063
		Medium	AMSE	2.2116	3.2465	2.4365
			RE	1.0000	0.6812	0.9077
		High	AMSE	2.2545	3.5740	2.4878
			RE	1.0000	0.6308	0.9062
	20	None	AMSE	2.5458	6.3746	3.0456
			RE	1.0000	0.3994	0.8359
		Medium	AMSE	2.5455	6.8487	3.1042
			RE	1.0000	0.3717	0.8200
		High	AMSE	2.6836	8.5920	3.2184
			RE	1.0000	0.3123	0.8338
	30	None	AMSE	2.9021	11.0061	3.8115
			RE	1.0000	0.2637	0.7614
		Medium	AMSE	2.9961	12.6621	3.9296
			RE	1.0000	0.2366	0.7624
		High	AMSE	3.3478	18.4147	4.4116
			RE	1.0000	0.1818	0.7588



<u>ภาพที่</u> 4.2.1 แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

จากตารางที่ 4.2.1 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 พบว่า สำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี PMM และวิธี KNN ตามลำดับ ดังนั้นวิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธี PMM และวิธี KNN เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างใหญ่ขึ้น ค่า AMSE จะลดลง และจากภาพที่ 4.2.1 จะเห็นว่า เมื่อสัดส่วนของการสูญหายสูงขึ้น ค่า AMSE จะมีค่าเพิ่มขึ้น โดยที่สัดส่วนของการสูญหายเท่ากับ 30% จะมีค่า AMSE มากที่สุด ส่วนการเพิ่มระดับของการสูญหายแบบ Nonignorable จะทำให้ค่า AMSE มีแนวโน้มเพิ่มขึ้นเพียงเล็กน้อย โดยเฉพาะกรณีที่สัดส่วนของการสูญหายเท่ากับ 10% จะเห็นว่าค่า AMSE มีอัตราการเพิ่มขึ้นไม่ถึง 1 หน่วย และที่ขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 10% ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี จะมีค่า AMSE ต่ำที่สุด เพราะขนาดตัวอย่างที่เพิ่มขึ้นจะทำให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง ข้อมูลที่มีสัดส่วนของการสูญหายน้อยจะทำให้สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงมากขึ้น และการที่ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มีก็คือการสูญหายแบบสุ่ม ซึ่งวิธี EM วิธี PMM และวิธี KNN ก็เป็นวิธีการประมาณค่าสูญหายที่ทำภายใต้เงื่อนไขของข้อมูลที่มีการสูญหายแบบสุ่มจึงสามารถประมาณค่าสูญหายได้ใกล้เคียงมากกว่ากรณีอื่น

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี EM กับค่า AMSE ของวิธีการประมาณค่าสูญหายวิธีอื่นๆ พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN และ วิธี PMM มีแนวโน้มที่จะลงลงเรื่อยๆ ถ้าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable สูงขึ้น โดยค่า RE ของวิธี PMM จะมีค่ามากกว่าวิธี KNN และมีค่าน้อยกว่า 1 เพราะในกรณีนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ ซึ่งที่สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มีจะให้ค่า RE มากที่สุด แสดงว่า ณ จุดนี้ ค่า AMSE ของวิธี KNN และวิธี PMM จะมีค่าใกล้เคียงค่า AMSE ของวิธี EM มากที่สุด เมื่อพิจารณาที่ขนาดตัวอย่างต่างกัน พบว่าค่า RE จะไม่มีความแตกต่างกันมาก ดังนั้นปัจจัยที่มีอิทธิพลต่อค่า RE ของแต่ละวิธีการประมาณคือสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable

ตารางที่ 4.2.2 แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

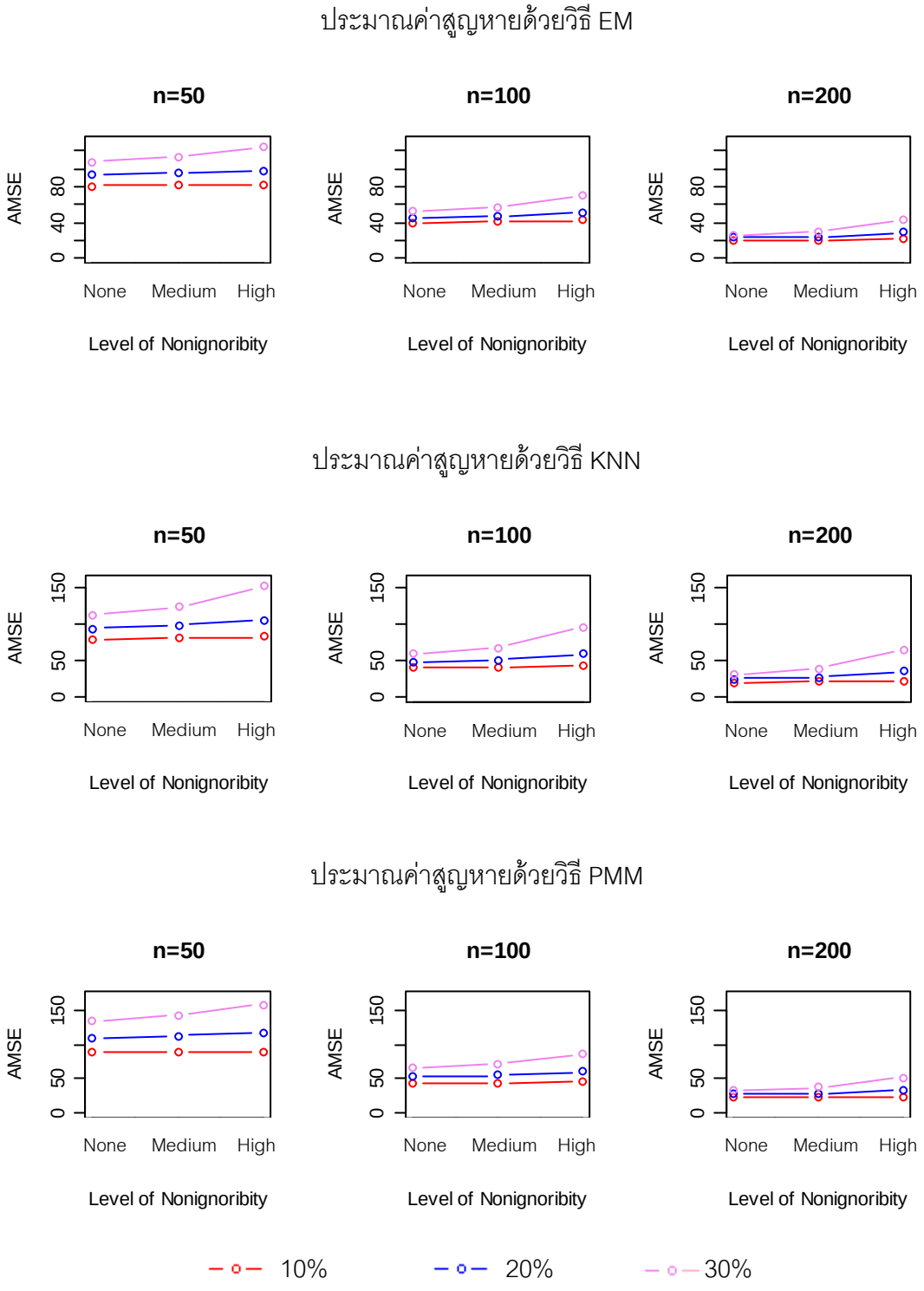
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	80.7732	80.1879	88.0119
			RE	1.0000	1.0073	0.9178
		Medium	AMSE	81.3242	81.2320	88.4885
			RE	1.0000	1.0011	0.9190
		High	AMSE	82.3387	82.8624	90.2384
			RE	1.0000	0.9937	0.9125
	20	None	AMSE	92.9237	94.2756	109.9458
			RE	1.0000	0.9857	0.8452
		Medium	AMSE	95.8172	98.6829	112.9840
			RE	1.0000	0.9710	0.8481
		High	AMSE	97.9113	105.2268	117.4256
			RE	1.0000	0.9305	0.8338
	30	None	AMSE	107.8976	112.8449	134.0065
			RE	1.0000	0.9562	0.8052
		Medium	AMSE	112.5758	124.3366	143.0527
			RE	1.0000	0.9054	0.7870
		High	AMSE	124.9795	152.5111	158.6027
			RE	1.0000	0.8195	0.7880

<u>ตารางที่ 4.2.2(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	39.8741	40.2185	43.5897
			RE	1.0000	0.9914	0.9148
		Medium	AMSE	40.8328	41.3907	44.5153
			RE	1.0000	0.9865	0.9173
		High	AMSE	42.1234	43.1347	46.1820
			RE	1.0000	0.9766	0.9121
	20	None	AMSE	45.1355	48.0445	53.0157
			RE	1.0000	0.9395	0.8514
		Medium	AMSE	47.0907	51.0743	55.2434
			RE	1.0000	0.9220	0.8524
		High	AMSE	50.9951	58.9454	60.1115
			RE	1.0000	0.8651	0.8483
	30	None	AMSE	52.8011	59.7110	66.0309
			RE	1.0000	0.8843	0.7996
		Medium	AMSE	56.2586	68.4053	70.9601
			RE	1.0000	0.8224	0.7928
		High	AMSE	70.0595	96.6463	86.5096
			RE	1.0000	0.7249	0.8098

<u>ตารางที่ 4.2.2(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	20.1893	20.4820	22.0777
			RE	1.0000	0.9857	0.9145
		Medium	AMSE	20.4397	20.9941	22.4960
			RE	1.0000	0.9736	0.9086
		High	AMSE	21.4111	22.4463	23.4544
			RE	1.0000	0.9539	0.9129
	20	None	AMSE	22.9433	25.5022	27.1919
			RE	1.0000	0.8997	0.8438
		Medium	AMSE	24.0150	27.6602	28.3469
			RE	1.0000	0.8682	0.8472
		High	AMSE	28.4863	35.3448	33.1000
			RE	1.0000	0.8060	0.8606
	30	None	AMSE	25.9397	31.5240	33.0514
			RE	1.0000	0.8229	0.7848
		Medium	AMSE	30.0464	39.3797	37.4755
			RE	1.0000	0.7630	0.8018
		High	AMSE	43.4116	64.5806	52.2824
			RE	1.0000	0.6722	0.8303



<u>ภาพที่ 4.2.2</u> แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

จากตารางที่ 4.2.2 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 พบว่า เมื่อขนาดตัวอย่างเท่ากับ 50 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ ยกเว้นที่สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี – ปานกลาง ที่วิธี KNN จะมีค่า AMSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 100 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ ยกเว้นที่สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ที่วิธี PMM จะมีค่า AMSE ต่ำกว่าวิธี KNN และเมื่อขนาดตัวอย่างเท่ากับ 200 วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี EM วิธี KNN และวิธี PMM ตามลำดับ แต่ถ้าสัดส่วนของการสูญหายเพิ่มขึ้น และระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น วิธี PMM จะมีค่า AMSE ต่ำกว่าวิธี KNN เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างลดลง สัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น ค่า AMSE จะมีค่าเพิ่มขึ้น และจากภาพที่ 4.2.2 จะเห็นว่าที่สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ค่า AMSE จะมีค่าสูงที่สุด แต่อย่างไรก็ตาม จะเห็นได้ว่าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable ค่อนข้างมีอิทธิพลต่อค่า AMSE มากกว่าขนาดตัวอย่าง เพราะถึงแม้ว่าตัวอย่างจะมีขนาดใหญ่ แต่ถ้ามีสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ค่า AMSE ที่ได้ก็จะสูงกว่าด้วย

เมื่อพิจารณาค่า RE พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN และ วิธี PMM มีแนวโน้มที่จะลงลงเรื่อยๆ ถ้าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น โดยค่า RE ของวิธี KNN และวิธี PMM ที่ได้จะมีค่าน้อยกว่า 1 เพราะในกรณีนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ ยกเว้นที่ขนาดตัวอย่างเท่ากับ 50 สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี – ปานกลางที่ค่า RE ของวิธี KNN จะมีค่ามากกว่า 1 เนื่องจาก ณ จุดนี้วิธี KNN เป็นวิธีประมาณค่าสูญหายที่ดีที่สุด เมื่อพิจารณาที่ขนาดตัวอย่าง พบว่าการเพิ่มขึ้นของขนาดตัวอย่างจะไม่ทำให้ค่า RE ของแต่ละวิธีแตกต่างกันมาก

ตารางที่ 4.2.3 แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

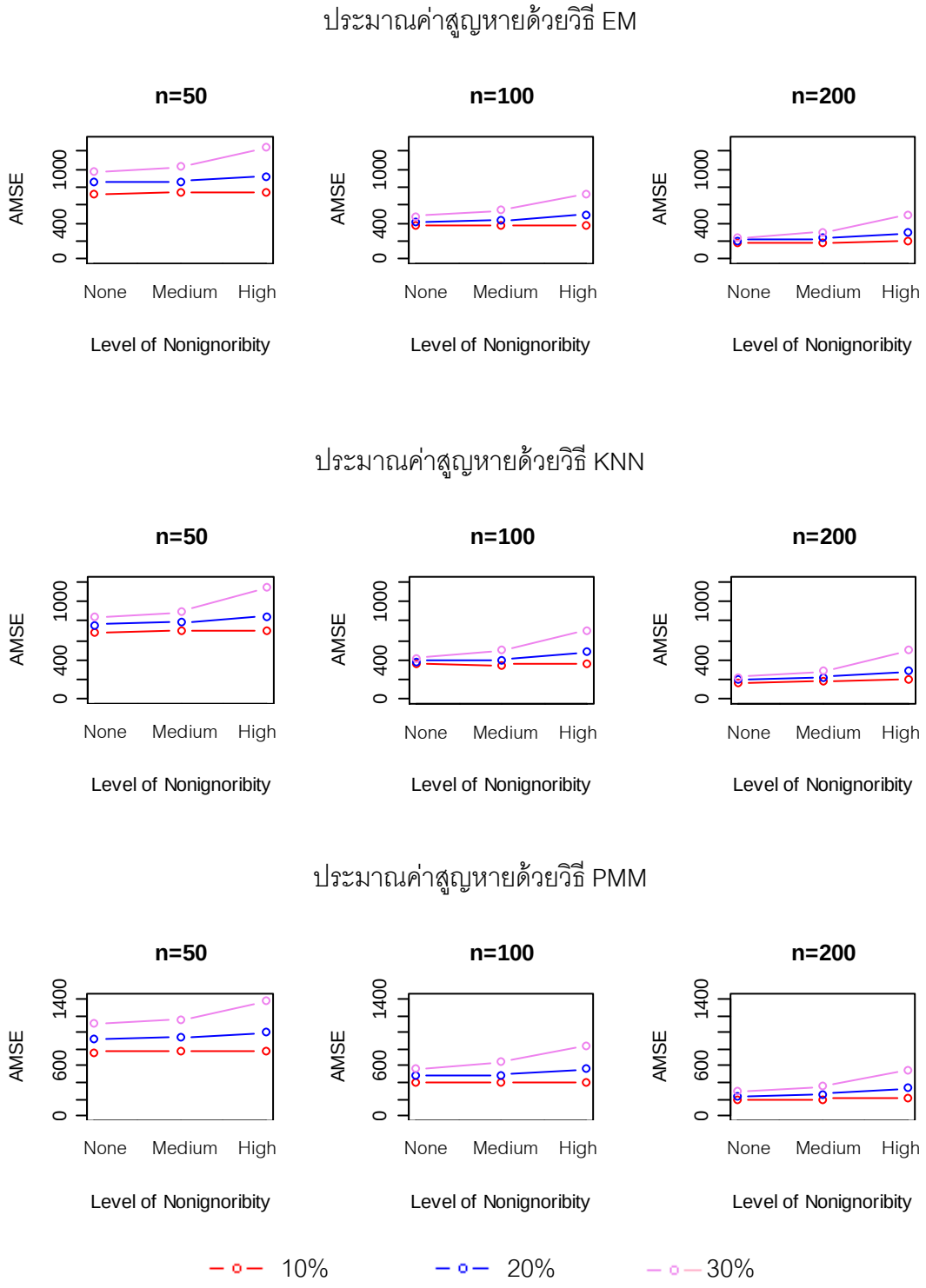
n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
50	10	None	AMSE	718.4833	681.0959	764.0945
			RE	1.0000	1.0549	0.9403
		Medium	AMSE	738.2511	698.4681	783.3801
			RE	1.0000	1.0570	0.9424
		High	AMSE	737.2791	700.3486	784.5244
			RE	1.0000	1.0527	0.9398
	20	None	AMSE	848.2031	761.8839	923.8217
			RE	1.0000	1.1133	0.9181
		Medium	AMSE	866.4152	785.2746	944.4503
			RE	1.0000	1.1033	0.9174
		High	AMSE	918.6444	844.2104	996.8802
			RE	1.0000	1.0882	0.9215
	30	None	AMSE	973.8808	836.3431	1102.0860
			RE	1.0000	1.1645	0.8837
		Medium	AMSE	1024.2350	895.7299	1152.6770
			RE	1.0000	1.1435	0.8886
		High	AMSE	1240.9830	1140.0550	1373.9490
			RE	1.0000	1.0885	0.9032

ตารางที่ 4.2.3(ต่อ) แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
100	10	None	AMSE	368.8769	353.5693	396.0302
			RE	1.0000	1.0433	0.9314
		Medium	AMSE	364.9632	351.2091	392.5729
			RE	1.0000	1.0392	0.9297
		High	AMSE	380.1461	368.6480	408.5005
			RE	1.0000	1.0312	0.9306
	20	None	AMSE	417.7663	387.6718	480.4662
			RE	1.0000	1.0776	0.8695
		Medium	AMSE	426.6401	400.1958	485.1197
			RE	1.0000	1.0661	0.8795
		High	AMSE	497.7226	478.1298	559.4248
			RE	1.0000	1.0410	0.8897
	30	None	AMSE	471.6143	421.7938	563.3280
			RE	1.0000	1.1181	0.8372
		Medium	AMSE	542.3335	499.7658	645.5859
			RE	1.0000	1.0852	0.8401
		High	AMSE	728.9369	707.9358	838.5846
			RE	1.0000	1.0297	0.8692

<u>ตารางที่ 4.2.3(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย		EM	KNN	PMM
200	10	None	AMSE	176.6349	171.3506	193.2558
			RE	1.0000	1.0308	0.9140
		Medium	AMSE	185.0598	180.2110	202.1706
			RE	1.0000	1.0269	0.9154
		High	AMSE	202.5161	198.4883	218.3544
			RE	1.0000	1.0203	0.9275
	20	None	AMSE	204.3925	192.4717	240.2811
			RE	1.0000	1.0619	0.8506
		Medium	AMSE	227.6552	218.9905	262.1459
			RE	1.0000	1.0396	0.8684
		High	AMSE	291.2158	289.1493	329.5798
			RE	1.0000	1.0071	0.8836
	30	None	AMSE	233.7713	218.3699	288.6732
			RE	1.0000	1.0705	0.8098
		Medium	AMSE	296.6359	287.7691	354.6376
			RE	1.0000	1.0308	0.8364
		High	AMSE	490.4257	497.4042	555.0483
			RE	1.0000	0.9860	0.8836



<u>ภาพที่</u> 4.2.3 แสดงการเปรียบเทียบสัดส่วนของการสูญหายของข้อมูลตัวแปรตามระหว่างค่า AMSE กับระดับของการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

จากตารางที่ 4.2.3 ซึ่งแสดงผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 พบว่า สำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable วิธีการประมาณค่าสูญหายที่มีค่า AMSE ต่ำสุดคือ วิธี KNN วิธี EM และวิธี PMM ตามลำดับ ดังนั้นวิธี KNN เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธี EM และวิธี PMM ยกเว้นในกรณีขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ที่วิธี EM จะสามารถประมาณค่าสูญหายได้ดีกว่าวิธี KNN และวิธี PMM ตามลำดับ เมื่อพิจารณาค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย พบว่า เมื่อขนาดตัวอย่างใหญ่ขึ้น ค่า AMSE จะลดลง และจากภาพที่ 4.2.3 จะเห็นว่าเมื่อข้อมูลมีสัดส่วนของการสูญหายและมีระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น จะทำให้ค่า AMSE มีค่าสูงขึ้นด้วย ซึ่งที่ขนาดตัวอย่างเท่ากับ 50 สัดส่วนของการสูญหายเท่ากับ 30% ระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง จะมีค่า AMSE สูงที่สุด แต่อย่างไรก็ตาม จะเห็นได้ว่าสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable จะมีอิทธิพลต่อค่า AMSE มากกว่าขนาดตัวอย่าง เพราะถึงแม้ว่าตัวอย่างจะมีขนาดใหญ่ แต่ถ้ามีสัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable สูง ค่า AMSE ที่ได้ก็จะสูงกว่าด้วย

เมื่อพิจารณาค่า RE ซึ่งเป็นอัตราส่วนระหว่างค่า AMSE ของวิธี EM กับค่า AMSE ของวิธีการประมาณค่าสูญหายวิธีอื่นๆ พบว่า สำหรับทุกขนาดตัวอย่าง ค่า RE ของวิธี KNN มีค่ามากกว่า 1 เพราะในกรณีนี้วิธี KNN เป็นวิธีประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุด ยกเว้นที่ขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูงที่ค่า RE จะมีค่าน้อยกว่า 1 เนื่องจากที่จุดนี้วิธี EM เป็นวิธีประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ และที่ระดับของการสูญหายแบบ Nonignorable ระดับเดียวกัน ค่า RE มีแนวโน้มเพิ่มขึ้นเมื่อมีสัดส่วนของการสูญหายเพิ่มขึ้น แต่วิธี PMM จะมีค่า RE ต่ำกว่า 1 สำหรับทุกขนาดตัวอย่าง และมีแนวโน้มลดลงเมื่อสัดส่วนของการสูญหายเพิ่มขึ้น เมื่อพิจารณาที่ขนาดตัวอย่าง พบว่าการเพิ่มขึ้นของขนาดตัวอย่างจะไม่ทำให้ค่า RE ของแต่ละวิธีประมาณค่าสูญหายแตกต่างกันมาก

สรุปส่วนที่ 4.2 ผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม เมื่อตัวแปรอิสระเป็นแบบที่ 2

เมื่อพิจารณาผลการวิจัยที่ได้ดังตารางที่ 4.2.1 – 4.2.3 ซึ่งเป็นผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable ในลักษณะที่ข้อมูลที่เกิดการสูญหายมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น เมื่อตัวแปรอิสระเป็นแบบที่ 2 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 , 30 และ 90 ตามลำดับ พบว่าสำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable ถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูงขึ้น ค่า AMSE ของแต่ละวิธีประมาณค่าสูญหายจะมีค่าสูงขึ้นด้วย เพราะการเพิ่มส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน จะทำให้ข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งจะส่งผลให้มีค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการประมาณค่าสูญหายมากขึ้น ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 – 30 โดยส่วนมากวิธีการประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุดคือวิธี EM ส่วนกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 โดยส่วนมากวิธีการประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำสุดคือวิธี KNN

จากผลการวิจัยในส่วนที่ 4.2 สามารถสรุปปัจจัยที่ส่งผลต่อค่า AMSE ได้ดังนี้

1. **ขนาดตัวอย่าง** เมื่อตัวอย่างมีขนาดใหญ่ขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะทำให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง
2. **สัดส่วนของการสูญหาย** เมื่อข้อมูลมีสัดส่วนของการสูญหายเพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะสูงขึ้น เพราะเมื่อข้อมูลมีการสูญหายมากขึ้น จะมีความคลาดเคลื่อนจากการประมาณค่าสูญหายและการประมาณค่าพารามิเตอร์มากขึ้นด้วย
3. **ระดับของการสูญหายแบบ Nonignorable** เมื่อระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะมีแนวโน้มสูงขึ้น เพราะวิธี KNN วิธี EM และวิธี PMM เป็นวิธีการประมาณค่าสูญหายที่ทำภายใต้เงื่อนไขของข้อมูลที่มีการสูญหายแบบสุ่ม ดังนั้นเมื่อมีการเพิ่มระดับความสัมพันธ์ของ

ข้อมูลสูญหายกับตัวแปรตาม จะส่งผลให้มีค่าความคลาดเคลื่อนจากการประมาณค่าสูญหายมากขึ้น

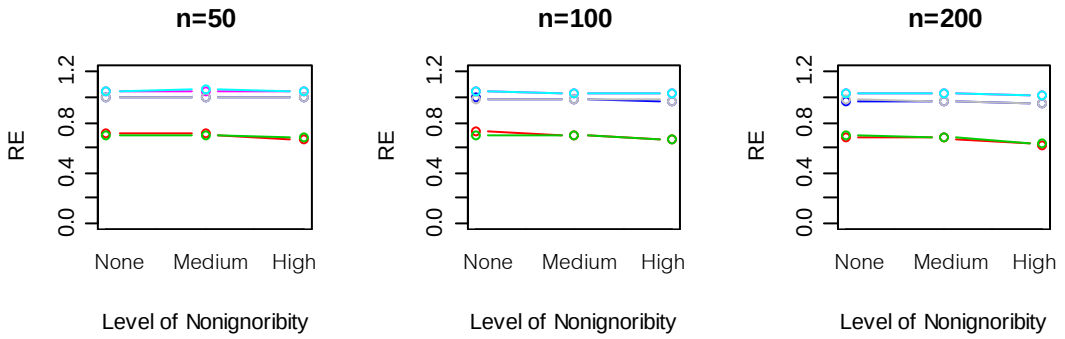
4. **ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน** เมื่อความคลาดเคลื่อนมีความแปรปรวนมากขึ้น ค่า AMSE ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีจะสูงขึ้น เพราะข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งส่งผลให้ความสามารถในการประมาณค่าสูญหายลดลง ค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการพยากรณ์เพิ่มขึ้น

4.3 การเปรียบเทียบผลการวิจัยของชุดตัวแปรอิสระแบบที่ 1 กับชุดตัวแปรอิสระแบบที่ 2

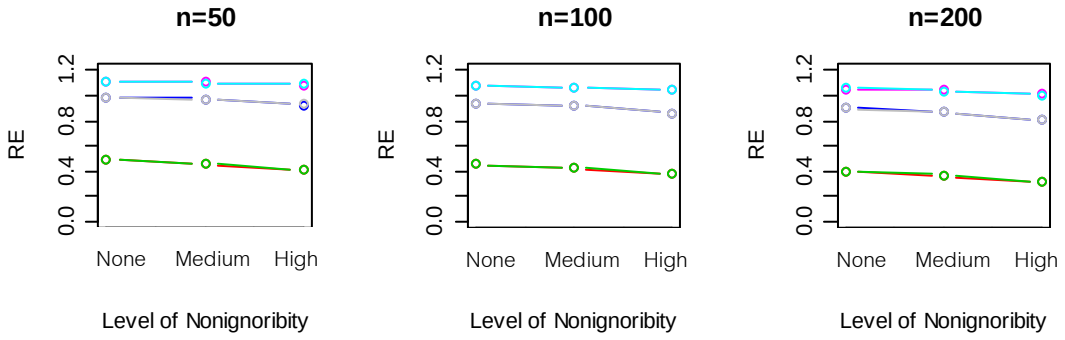
ในส่วนนี้ผู้วิจัยทำการศึกษาเปรียบเทียบกรณีที่ชุดตัวแปรอิสระเป็นแบบที่ 1 กับชุดตัวแปรอิสระเป็นแบบที่ 2 และในการเปรียบเทียบจะพิจารณาที่ระดับปัจจัยเดียวกัน โดยทำการศึกษาที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าเท่ากับ 10, 30 และ 90 ขนาดของตัวอย่างเท่ากับ 50, 100 และ 200 สัดส่วนของการสูญหายเท่ากับ 10%, 20% และ 30% ระดับของการสูญหายแบบ Nonignorable เป็น ไม่มี, ปานกลาง และสูง ซึ่งการวิจัยในส่วนนี้จะนำผลการวิจัยที่ได้จากตารางที่ 4.1.1 – 4.1.3 เทียบกับผลการวิจัยที่ได้จากตารางที่ 4.2.1 – 4.2.3

ตารางที่ 4.1.1	เปรียบเทียบกับ	ตารางที่ 4.2.1
ตารางที่ 4.1.2	เปรียบเทียบกับ	ตารางที่ 4.2.2
ตารางที่ 4.1.3	เปรียบเทียบกับ	ตารางที่ 4.2.3

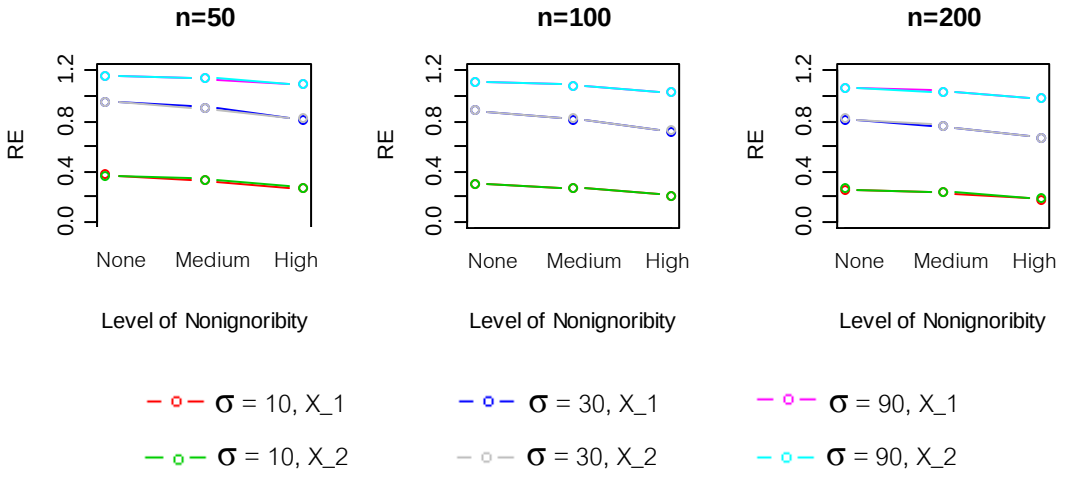
สัดส่วนของการสูญหายเท่ากับ 10%



สัดส่วนของการสูญหายเท่ากับ 20%

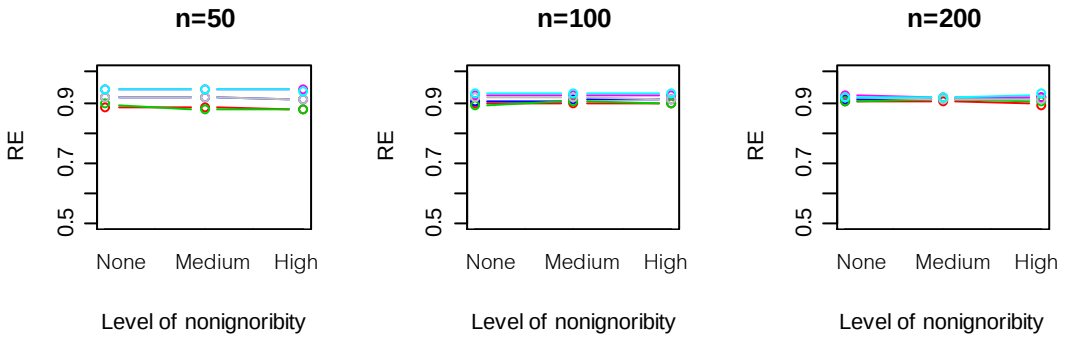


สัดส่วนของการสูญหายเท่ากับ 30%

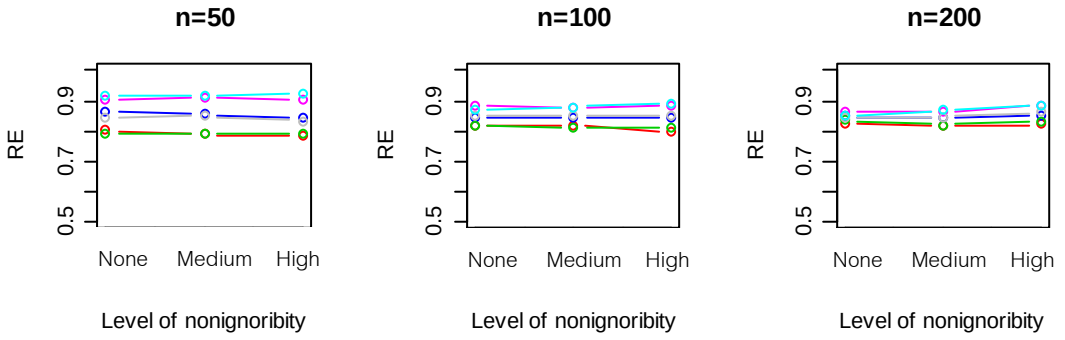


<u>ภาพที่</u> 4.3.1 แสดงการเปรียบเทียบชุดของตัวแปรอิสระแบบที่ 1 กับแบบที่ 2 ที่ระดับส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนต่างๆ ระหว่างค่า RE กับระดับของการสูญหายแบบ Nonignorable เมื่อวิธีประมาณค่าสูญหายที่ใช้เทียบอัตราส่วนกับวิธี EM คือวิธี KNN

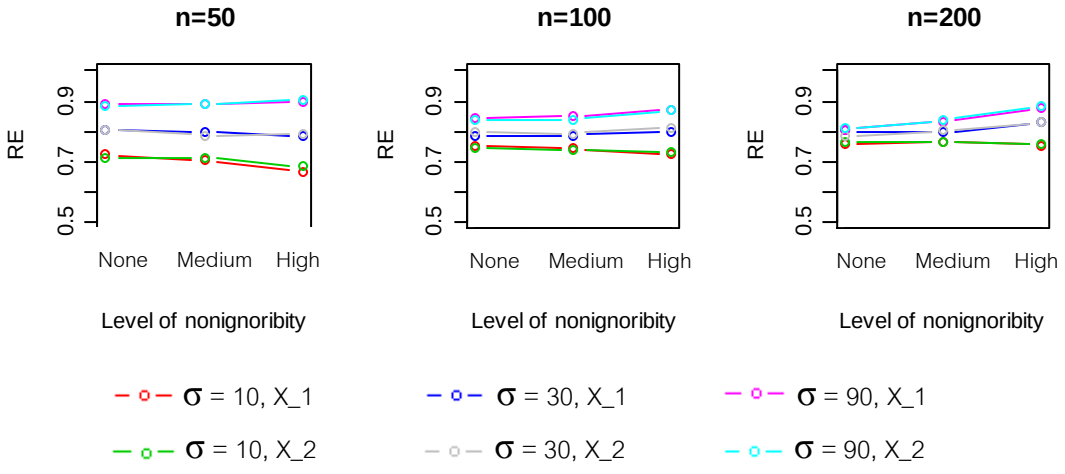
สัดส่วนของการสูญหายเท่ากับ 10%



สัดส่วนของการสูญหายเท่ากับ 20%



สัดส่วนของการสูญหายเท่ากับ 30%



<u>ภาพที่ 4.3.2</u> แสดงการเปรียบเทียบชุดของตัวแปรอิสระแบบที่ 1 กับแบบที่ 2 ที่ระดับส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนต่างๆ ระหว่างค่า RE กับระดับของการสูญหายแบบ Nonignorable เมื่อวิธีประมาณค่าสูญหายที่ใช้เทียบอัตราส่วนกับวิธี EM คือวิธี PMM

เมื่อพิจารณาผลการวิจัยที่ได้ดังตารางที่ 4.1.1-4.1.3 ซึ่งเป็นผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อชุดของตัวแปรอิสระเป็นแบบที่ 1 เทียบกับผลการวิจัยที่ได้ดังตารางที่ 4.2.1-4.2.3 ซึ่งเป็นผลการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable เมื่อชุดของตัวแปรอิสระเป็นแบบที่ 2 พบว่า ที่ระดับของขนาดตัวอย่าง สัดส่วนของการสูญหาย ระดับของการสูญหายแบบ Nonignorable และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเดียวกัน ค่า AMSE ที่ได้จากชุดตัวแปรอิสระแบบที่ 1 และแบบที่ 2 จะมีค่าใกล้เคียงกันหรือไม่แตกต่างกันมาก และจากภาพที่ 4.3.1-4.3.2 ซึ่งเป็นภาพแสดงการเปรียบเทียบลักษณะของชุดตัวแปรอิสระแบบที่ 1 กับแบบที่ 2 โดยพิจารณาจากค่า RE ของวิธี KNN และวิธี PMM จะเห็นว่า ที่ระดับของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเดียวกัน ตำแหน่งแสดงค่า RE ที่ได้จากชุดตัวแปรอิสระแบบที่ 1 และแบบที่ 2 นั้นค่อนข้างจะเป็นตำแหน่งเดียวกัน และเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูงขึ้น ก็จะทำให้ค่า RE เพิ่มขึ้นด้วย ซึ่งแสดงว่าความคลาดเคลื่อนจากการประมาณค่าสูญหายด้วยวิธี KNN และวิธี PMM จะใกล้เคียงกับความคลาดเคลื่อนจากการประมาณค่าสูญหายด้วยวิธี EM มากขึ้น นั่นคือวิธี KNN และวิธี PMM จะมีประสิทธิภาพใกล้เคียงกับวิธีEM มากขึ้น โดยที่ระดับส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 จะมีค่า RE สูงที่สุด

เมื่อพิจารณาภาพที่ 4.3.1 ซึ่งแสดงค่า RE ของวิธี KNN พบว่า ที่สัดส่วนของการสูญหายระดับเดียวกัน แต่ขนาดตัวอย่างต่างกัน ค่า RE จะไม่มีความแตกต่างกันมาก และเมื่อระดับของการสูญหายแบบ Nonignorable เพิ่มขึ้น ค่า RE จะมีแนวโน้มลดลง ซึ่งแสดงว่า ประสิทธิภาพของวิธี KNN เทียบกับวิธี EM จะลดลงเรื่อยๆ เมื่อระดับของการสูญหายแบบ Nonignorable สูงขึ้น และโดยส่วนมากที่ระดับส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 จะมีค่า RE มากกว่า 1 แสดงว่า ถ้าข้อมูลตัวแปรตามมีการกระจายมาก การประมาณค่าสูญหายด้วยวิธี KNN จะมีประสิทธิภาพมากกว่าการประมาณค่าสูญหายด้วยวิธี EM

เมื่อพิจารณาภาพที่ 4.3.2 ซึ่งแสดงค่า RE ของวิธี PMM พบว่า ที่สัดส่วนของการสูญหายระดับเดียวกัน เมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า RE ของแต่ละระดับส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน จะมีค่าใกล้เคียงกันมากขึ้น และค่า RE มีค่าน้อยกว่า 1 ซึ่งแสดงว่า สำหรับทุกระดับปัจจัยของขนาดตัวอย่าง สัดส่วนของการสูญหาย ระดับของการสูญหายแบบ Nonignorable และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน การประมาณค่าสูญหายด้วยวิธี EM จะมีประสิทธิภาพมากกว่าการประมาณค่าสูญหายด้วยวิธี PMM

จากผลการวิจัยในส่วนที่ 4.3 **สามารถสรุป** ได้ดังนี้

ในกรณีที่ตัวแปรตามมีการสูญหายแบบ Nonignorable การใช้ชุดตัวแปรอิสระที่มีความแปรปรวนแตกต่างกัน จะไม่ทำให้ค่า AMSE เพิ่มขึ้นหรือลดลง ซึ่งก็เป็นไปตามทฤษฎีของการสูญหายแบบ Nonignorable คือค่าที่สูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับตัวแปรที่เกิดการสูญหายเท่านั้น จะไม่ขึ้นอยู่กับตัวแปรอื่นๆ นั่นคือค่าความคลาดเคลื่อนจะขึ้นอยู่กับการประมาณค่าสูญหายของตัวแปรตามเท่านั้น ลักษณะของตัวแปรอิสระจะไม่ส่งผลใดๆต่อค่าความคลาดเคลื่อน ซึ่งในกรณีนี้การที่ค่า AMSE ที่ได้จากชุดตัวแปรอิสระแบบที่ 1 และแบบที่ 2 ไม่มีความแตกต่างกัน เนื่องจากในงานวิจัยนี้ได้กำหนดให้ตัวแปรอิสระในแต่ละแบบที่ใช้ในการสร้างตัวแปรตามมีการแจกแจงปกติที่มีมีค่าเฉลี่ยเป็นศูนย์เหมือนกัน และความแปรปรวนของตัวแปรอิสระในแต่ละแบบเมื่อรวมกันแล้วจะมีค่าเท่ากัน ดังนั้นถึงแม้ว่าตัวแปรตามจะถูกสร้างจากตัวแปรอิสระต่างชุดกัน แต่ก็จะมีค่าเฉลี่ยและความแปรปรวนเท่ากัน ทำให้ที่ขนาดตัวอย่าง สัดส่วนของการสูญหาย ระดับของการสูญหายแบบ Nonignorable และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเดียวกัน ค่า AMSE ที่ได้จากชุดตัวแปรอิสระแต่ละแบบจึงไม่มีความแตกต่างกัน

จากผลการวิจัยในส่วนที่ 4.1 – 4.3 **สามารถสรุปปัจจัยต่าง ๆที่มีผลต่อค่า** AMSE **ได้ดังนี้**

1. **ขนาดตัวอย่าง** เมื่อตัวอย่างมีขนาดใหญ่ขึ้น จะทำให้ค่า AMSE ลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะช่วยให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง
2. **สัดส่วนของการสูญหาย** เมื่อข้อมูลมีสัดส่วนของการสูญหายเพิ่มขึ้น จะทำให้ค่า AMSE สูงขึ้น เพราะข้อมูลที่มีการสูญหายมาก จะทำให้มีความคลาดเคลื่อนจากการประมาณค่าสูญหายมาก ซึ่งจะส่งผลให้ได้ค่าประมาณพารามิเตอร์และค่าพยากรณ์ที่ผิดพลาดจากค่าจริงมากขึ้น
3. **ระดับของการสูญหายแบบ** Nonignorable เมื่อข้อมูลสูญหายกับตัวแปรตามมีความสัมพันธ์กันมากขึ้น จะทำให้ค่า AMSE มีแนวโน้มสูงขึ้น เพราะวิธี KNN วิธี EM และวิธี PMM เป็นวิธีการประมาณค่าสูญหายที่ถูกพัฒนาขึ้นมาภายใต้สมมติฐานของข้อมูลที่มีการสูญหายแบบสุ่ม ดังนั้นการเพิ่มระดับความสัมพันธ์ของข้อมูลสูญหายกับตัวแปรตามจึงส่งผลให้ค่าความคลาดเคลื่อนจากการประมาณค่าสูญหายเพิ่มขึ้น
4. **ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน** เมื่อความคลาดเคลื่อนมีความแปรปรวนมากขึ้น จะทำให้ค่า AMSE สูงขึ้น เพราะข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งส่งผลให้ความสามารถในการประมาณค่าสูญหายได้ใกล้เคียงค่าจริงลดลง ค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการพยากรณ์เพิ่มขึ้น
5. **ชุดของตัวแปรอิสระ** ลักษณะของชุดตัวแปรอิสระที่มีรูปแบบของความแปรปรวนแตกต่างกัน จะไม่ส่งผลต่อค่า AMSE ทั้งนี้เพราะตามทฤษฎีของการสูญหายแบบ Nonignorable ที่ค่าสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับตัวแปรที่เกิดการสูญหายเท่านั้น จะไม่ขึ้นอยู่กับตัวแปรอื่นๆ ดังนั้นไม่ว่าจะใช้ชุดของตัวแปรอิสระในรูปแบบใดก็จะไม่ส่งผลต่อการประมาณค่าสูญหาย

ผู้วิจัยสามารถสรุปผลการเปลี่ยนแปลงของค่า AMSE **ได้ดังนี้**

1. **ขนาดตัวอย่าง** : ค่า AMSE จะแปรผกผันกับขนาดตัวอย่าง คือ ค่า AMSE จะเพิ่มขึ้นเมื่อมีขนาดตัวอย่างลดลง
2. **สัดส่วนของการสูญหาย** : ค่า AMSE จะแปรผันตามสัดส่วนของการสูญหาย คือ ค่า AMSE จะเพิ่มขึ้นเมื่อสัดส่วนของการสูญหายสูงขึ้น

3. **ระดับของการสูญหายแบบ Nonignorable**: ค่า AMSE จะแปรผันตามระดับของการสูญหายแบบ Nonignorable คือ ค่า AMSE จะเพิ่มขึ้นเมื่อระดับของการสูญหายแบบ Nonignorable สูงขึ้น
4. **ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน** : ค่า AMSE จะแปรผันตามระดับความสัมพันธ์ของข้อมูลสูญหายกับตัวแปรตาม คือ ค่า AMSE จะเพิ่มขึ้นเมื่อระดับความสัมพันธ์ของข้อมูลสูญหายกับตัวแปรตามสูงขึ้น
5. **ลักษณะของตัวแปรอิสระ** : ลักษณะของชุดตัวแปรอิสระที่มีรูปแบบของความแปรปรวนเท่ากัน หรือแตกต่างกันจะไม่มีผลต่อการเพิ่มขึ้นหรือลดลงของค่า AMSE

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

ในการวิจัยครั้งนี้เป็นการวิจัยเพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามที่มีการสูญหายแบบ Nonignorable ในการวิเคราะห์การถดถอยเชิงเส้นพหุ โดยวิธีที่ใช้ในการประมาณค่าสูญหายมีทั้งหมด 3 วิธีคือ วิธีEM Algorithm (EM) วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการประมาณค่าสูญหายจะใช้ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ระหว่างค่าจริงกับค่าพยากรณ์ ซึ่งวิธีการประมาณค่าสูญหายวิธีใดที่ให้ค่า AMSE ต่ำกว่าจะเป็นวิธีที่ดีกว่า โดยมีการจำลองสถานการณ์ทั้งหมด 162 สถานการณ์ ตามเงื่อนไขดังนี้

1. ชุดตัวแปรอิสระมีการแจกแจงปกติที่มีรูปแบบแตกต่างกัน 2 ลักษณะคือ
   แบบที่ 1 : $X_1 \sim N(0,300)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$
   แบบที่ 2 : $X_1 \sim N(0,100)$, $X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$
2. ความคลาดเคลื่อนมีการแจกแจงปกติที่มีค่าเฉลี่ยเป็นศูนย์ และมีส่วนเบี่ยงเบนมาตรฐาน (σ) เป็น 10, 30 และ 90
3. ขนาดตัวอย่างมี 3 ขนาดคือ 50, 100 และ 200
4. ข้อมูลตัวแปรตามมีการสูญหายแบบ Nonignorableในลักษณะที่ข้อมูลที่เกิดการสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับค่าของตัวแปรตามเท่านั้น และมีสัดส่วนของการสูญหายเท่ากับ 10%, 20% และ 30
5. ระดับของการสูญหายแบบ Nonignorable จะแบ่งเป็น 3 ระดับคือ

ไม่	1	:	1	:	1
ปานกลาง	7	:	10	:	13
สูง	4	:	10	:	16

ในการจำลองสถานการณ์จะใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) โดยใช้โปรแกรม R และในแต่ละสถานการณ์จะทำการจำลองซ้ำอีกจำนวน 5,000 รอบ

5.1 ผลการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง

จากการเปรียบเทียบค่า AMSE ที่ได้จากวิธีการประมาณค่าสูญหายทั้ง 3 วิธี พบว่า สำหรับทุกลักษณะของชุดตัวแปรตาม ทุกขนาดตัวอย่าง ทุกระดับของสัดส่วนของการสูญหาย และทุกระดับของการสูญหายแบบ Nonignorable ถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 – 30 วิธีการประมาณค่าสูญหายที่ดีที่สุดคือวิธี EM ยกเว้นที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 ขนาดตัวอย่างเท่ากับ 50 สัดส่วนของการสูญหายเท่ากับ 10% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับไม่มี – ปานกลาง ที่วิธี KNN จะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด และถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 วิธีการประมาณค่าสูญหายที่ดีที่สุดคือวิธี KNN ยกเว้นที่ขนาดตัวอย่างเท่ากับ 200 สัดส่วนของการสูญหายเท่ากับ 30% และระดับของการสูญหายแบบ Nonignorable อยู่ในระดับสูง ที่วิธี EM จะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด

ในการวิเคราะห์การถดถอยเชิงเส้นพหุที่มีปัญหาข้อมูลตัวแปรตามมีการสูญหายแบบ Nonignorable ถ้าชุดข้อมูลมีขนาดเล็ก แต่มีสัดส่วนของการสูญหายสูง ระดับของการสูญหายแบบ Nonignorable สูง และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง จะทำให้ค่าพยากรณ์มีความคลาดเคลื่อนสูงขึ้น

5.2 สรุปความแตกต่างของแต่ละวิธีการประมาณค่าสูญหาย

<u>วิธี EM Algorithm</u>

ผลจากการศึกษาในครั้งนี้วิธี EM เป็นวิธีการประมาณค่าสูญหายที่มีความเหมาะสมมากกว่าวิธีการประมาณค่าสูญหายแบบอื่นๆสำหรับกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับต่ำ – ปานกลางหรือข้อมูลมีการกระจายไม่มาก ซึ่งวิธี EM จะเป็นวิธีที่ต้องใช้ค่าพารามิเตอร์ในการประมาณค่าสูญหาย โดยใช้กระบวนการวนซ้ำเพื่อหาค่าประมาณภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ และจากการศึกษาพบว่าค่าประมาณพารามิเตอร์ที่ได้จากขั้นตอนเริ่มต้นซึ่งเป็นการวิเคราะห์การถดถอย (Regression Imputation : RI) โดยพิจารณาเฉพาะชุดข้อมูลที่สมบูรณ์ กับค่าประมาณพารามิเตอร์ที่ได้จากการทำซ้ำรอบที่ 1 มีความแตกต่างกันที่ทศนิยมหลักที่ 14 ขึ้นไป ซึ่งถือว่ามีความแตกต่างกันน้อยมาก และจากการศึกษาผลการวิจัยที่ผ่านมา (วารุณี ตรีบำรุงศักดิ์, 2537; เพียงออ ยีสา, 2551) ที่มีการศึกษาในกรณีที่ตัวแปรตามมีการสูญหายแบบสุ่ม พบว่าค่าประมาณพารามิเตอร์จากวิธี EM และวิธี RI จะมีความแตกต่างกันเมื่อ

ข้อมูลมีระดับการสูญหายสูงๆคือที่ 60% - 70% ดังนั้นในทางปฏิบัติจึงสามารถใช้วิธี RI ที่มีความยุ่งยากน้อยกว่าวิธี EM มาใช้ในการประมาณค่าสูญหายได้

วิธี K-Nearest Neighbor Imputation (KNN)

วิธี KNN เป็นวิธีการประมาณค่าสูญหายที่มีความยุ่งยากซับซ้อนน้อยกว่าวิธี EM และวิธี PMM เนื่องจากในกระบวนการคำนวณไม่จำเป็นต้องหาค่าประมาณของพารามิเตอร์ ซึ่งจากผลการศึกษาในครั้งนี้พบว่า ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับสูง หรือข้อมูลมีการกระจายมาก วิธี KNN จะเป็นวิธีการประมาณค่าสูญหายที่มีความเหมาะสมมากกว่าวิธีการประมาณค่าสูญหายแบบอื่นๆ

วิธี Predictive Mean Matching Imputation (PMM)

วิธี PMM เป็นวิธีการประมาณค่าสูญหายแบบกึ่งพารามิเตอร์ คือมีทั้งส่วนที่ใช้พารามิเตอร์และไม่ใช้พารามิเตอร์ ดังนั้นจึงเป็นวิธีที่มีความยุ่งยากในการคำนวณพอสมควร แต่จากผลการศึกษาในครั้งนี้พบว่า วิธี PMM ไม่เป็นวิธีที่มีความเหมาะสมที่สุดในการประมาณค่าสูญหายไม่ว่าในกรณีใดๆ โดยเฉพาะในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับสูง จะเป็นวิธีการประมาณค่าสูญหายที่มีความคลาดเคลื่อนสูงสุด

5.3 ปัจจัยที่มีผลต่อค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของแต่ละวิธีการประมาณค่าสูญหาย

1. **ขนาดตัวอย่าง**

เมื่อตัวอย่างมีขนาดใหญ่ขึ้น จะทำให้ค่า AMSE ของทุกวิธีลดลง เนื่องจากการเพิ่มขึ้นของขนาดตัวอย่างจะช่วยให้ค่าความคลาดเคลื่อนจากการพยากรณ์ลดลง

2. **สัดส่วนของการสูญหาย**

เมื่อข้อมูลมีสัดส่วนของการสูญหายเพิ่มขึ้น จะทำให้ค่า AMSE ของทุกวิธีสูงขึ้น เพราะข้อมูลที่มีการสูญหายมาก จะทำให้มีความคลาดเคลื่อนจากการประมาณค่าสูญหายมาก ซึ่งจะส่งผลให้ได้ค่าประมาณพารามิเตอร์และค่าพยากรณ์ที่ผิดพลาดจากค่าจริงมากขึ้น

3. **ระดับของการสูญหายแบบ** Nonignorable

เมื่อข้อมูลสูญหายกับตัวแปรตามมีความสัมพันธ์กันมากขึ้น จะทำให้ค่า AMSE ของทุกวิธีมีแนวโน้มสูงขึ้น เพราะวิธี KNN วิธี EM และวิธี PMM เป็นวิธีการประมาณค่าสูญหายที่ถูก

พัฒนาขึ้นมาภายใต้สมมติฐานของข้อมูลที่มีการสูญหายแบบสุ่ม ดังนั้นการเพิ่มระดับความสัมพันธ์ของข้อมูลสูญหายกับตัวแปรตาม จึงส่งผลให้ค่าความคลาดเคลื่อนจากการประมาณค่าสูญหายเพิ่มขึ้น

4. **ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน**

เมื่อความคลาดเคลื่อนมีความแปรปรวนมากขึ้น จะทำให้ค่า AMSE ของทุกวิธีสูงขึ้น เพราะข้อมูลตัวแปรตามมีการกระจายมากขึ้น ซึ่งส่งผลให้ความสามารถในการประมาณค่าสูญหายได้ใกล้เคียงค่าจริงลดลง ค่าความคลาดเคลื่อนจากการประมาณค่าพารามิเตอร์และการพยากรณ์เพิ่มขึ้น

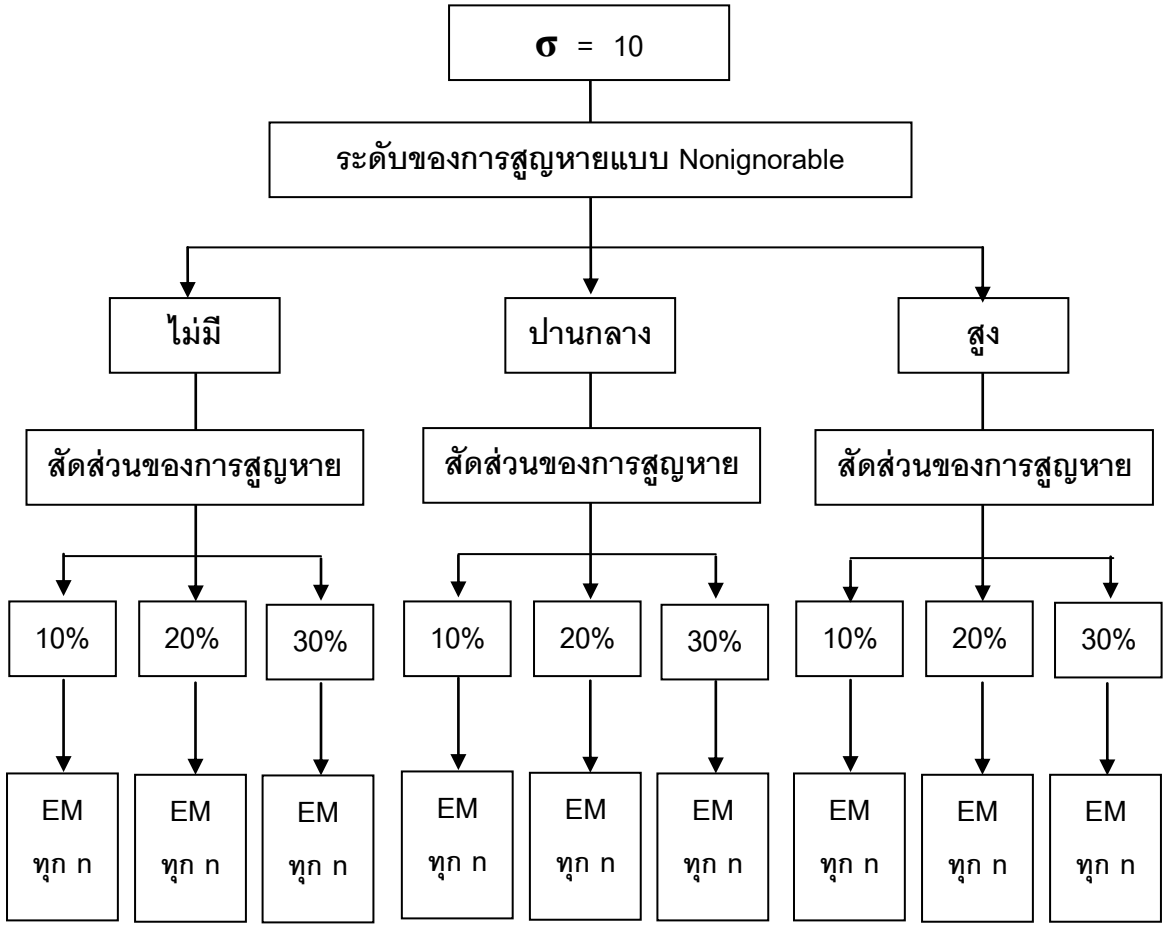
5. **ชุดของตัวแปรอิสระ**

ลักษณะของชุดตัวแปรอิสระที่มีรูปแบบของความแปรปรวนแตกต่างกัน จะไม่ส่งผลต่อค่า AMSE ของทุกวิธีการประมาณค่าสูญหาย ทั้งนี้เพราะตามทฤษฎีของการสูญหายแบบ Nonignorable ที่ค่าสูญหายจะมีความสัมพันธ์หรือขึ้นอยู่กับตัวแปรที่เกิดการสูญหายเท่านั้น จะไม่ขึ้นอยู่กับตัวแปรอื่นๆ ดังนั้นไม่ว่าจะใช้ชุดของตัวแปรอิสระในรูปแบบใดก็จะไม่ส่งผลต่อการประมาณค่าสูญหาย

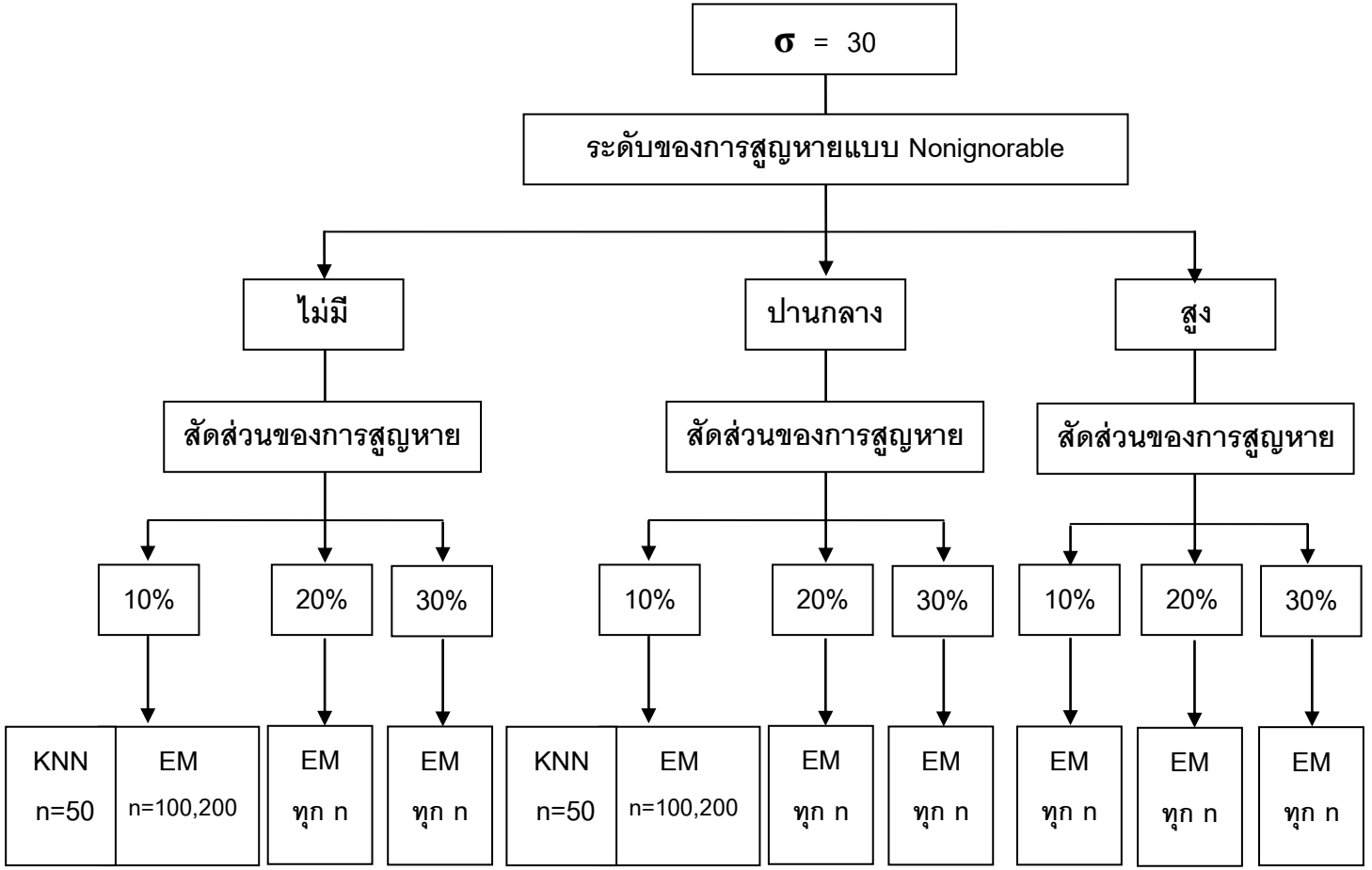
5.4 ผลสรุปการเลือกใช้วิธีการประมาณค่าสูญหายเมื่อข้อมูลตัวแปรตามมีการสูญหายแบบ Nonignorable

ในการวิจัยครั้งนี้พบว่าการประมาณค่าสูญหายของข้อมูลตัวแปรตามด้วยวิธี EM จะให้ผลดีในกรณีที่ข้อมูลมีการกระจายไม่มาก เพราะการที่ข้อมูลมีการกระจายน้อย จะทำให้สามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริง และในกรณีที่ข้อมูลมีการกระจายมาก วิธี KNN จะเป็นวิธีการประมาณค่าสูญหายที่ให้ผลดีกว่า เพราะการที่ข้อมูลมีการกระจายมาก จะทำให้การประมาณค่าพารามิเตอร์มีความผิดพลาดสูง ดังนั้นการประมาณค่าที่สูญหายจากชุดข้อมูลที่มีลักษณะคล้ายคลึงกันน่าจะมีความเหมาะสมกว่า

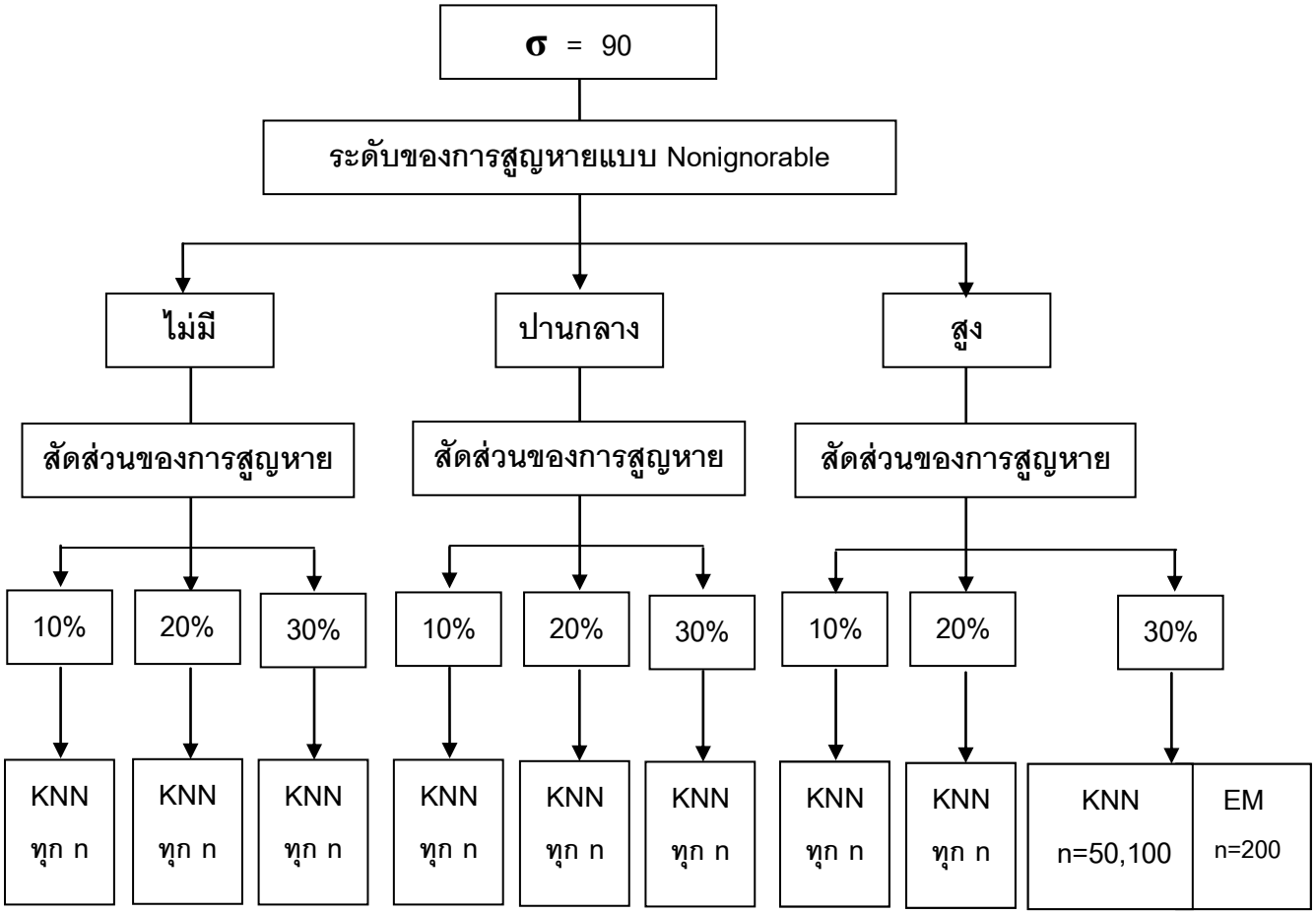
ภาพที่ 5.1 แผนผังสรุปวิธีการประมาณค่าสูญหายเมื่อตัวแปรตามมีการสูญหายแบบ Nonignorable และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



<u>ภาพที่ 5.2</u> แผนผังสรุปวิธีการประมาณค่าสูญหายเมื่อตัวแปรตามมีการสูญหายแบบ Nonignorable และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



ภาพที่ 5.3 แผนผังสรุปวิธีการประมาณค่าสูญหายเมื่อตัวแปรตามมีการสูญหายแบบ Nonignorable และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



5.5 ข้อเสนอแนะ

ผลการวิจัยครั้งนี้จะมีข้อเสนอแนะเป็น 2 ด้าน คือ

5.5.1 ด้านการนำไปใช้ประโยชน์

เพื่อเป็นแนวทางในการเลือกใช้วิธีการประมาณค่าสูญหาย เมื่อตัวแปรตามที่มีการแจกแจงปกติมีการสูญหายแบบ Nonignorable ได้อย่างเหมาะสมกับแต่ละสถานการณ์

ในทางปฏิบัติแล้วเมื่อเกิดปัญหาข้อมูลสูญหายก็มักจะตัดข้อมูลที่สูญหายทิ้ง และพิจารณาเฉพาะข้อมูลที่สมบูรณ์ แต่ในกรณีที่ข้อมูลมีการสูญหายแบบ Nonignorable นั้น การตัดชุดข้อมูลที่สูญหายทิ้งจะส่งผลกระทบต่อผลการวิเคราะห์ข้อมูลและมีความเสี่ยงในการที่จะได้ผลสรุปที่ผิดพลาดมากกว่าการสูญหายแบบสุ่ม ดังนั้นจึงต้องทำการประมาณค่าสูญหายด้วยวิธี KNN หรือวิธี EM ตามแต่ความเหมาะสมของสถานการณ์ สำหรับวิธี KNN นั้นโดยส่วนมากเพื่อความสะดวกในการใช้งานก็มักจะนิยมกำหนดให้ K = 1 (วิธี Nearest Neighbor Imputation : NNI) แต่ผลลัพธ์ที่ได้ก็จะไม่ดีเท่ากับการใช้ $K \approx \sqrt{m}$[1] เพราะวิธี NNI จะเป็นการแทนที่ข้อมูลที่สูญหายด้วยข้อมูลที่ทราบค่าที่ใกล้ที่สุดเพียงชุดเดียว ซึ่งในกรณีที่ข้อมูลมีการกระจายมาก การแทนที่ด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าจำนวน K ชุดจะมีความคลาดเคลื่อนน้อยกว่า และสำหรับวิธี EM นั้นในทางปฏิบัติสามารถนำวิธี RI มาใช้ในการประมาณค่าสูญหายแทนวิธี EM ได้ เพราะมีความยุ่งยากในการคำนวณน้อยกว่าวิธี EM แต่ก็ให้ผลการประมาณค่าสูญหายที่ไม่แตกต่างกัน

5.5.2 ด้านการศึกษาวิจัย

เพื่อเป็นแนวทางให้ผู้ที่สนใจได้ศึกษาเพิ่มเติม ซึ่งในการศึกษาครั้งต่อๆไปอาจทำการศึกษาในกรณีต่างๆดังต่อไปนี้

1. ในการศึกษาครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะกรณีที่ตัวแปรตามมีการแจกแจงปกติ และมีการสูญหายแบบ Nonignorable สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ข้อมูลมีรูปแบบการแจกแจงแบบอื่นๆ ซึ่งการที่ตัวแปรตามมีการแจกแจกเปลี่ยนไป อาจจะทำให้ได้ผลการวิจัยที่ไม่เหมือนเดิม

---

[1] ดูเพิ่มเติมที่ภาคผนวก ข., หน้า 91.

2. ชุดข้อมูลที่ใช้ในการศึกษาครั้งนี้เป็นข้อมูลเชิงปริมาณทั้งหมด จึงควรศึกษาเพิ่มเติมในกรณีที่ข้อมูลมีทั้งเชิงปริมาณและเชิงคุณภาพ เพราะการที่มีข้อมูลเชิงคุณภาพเข้ามาเกี่ยวข้องอาจจะทำให้ได้ผลการวิจัยที่แตกต่างออกไป

3. ควรทำการศึกษาเพิ่มเติมในกรณีที่ข้อมูลเป็นข้อมูลอนุกรมเวลา เพื่อศึกษาว่าการเปลี่ยนแปลงของเวลาจะส่งผลกระทบต่อค่าสูญหายและวิธีการประมาณค่าสูญหายอย่างไร

4. ศึกษาเพิ่มเติมในกรณีที่เกิดการสูญหายทั้งในตัวแปรอิสระและตัวแปรตาม โดยที่มีรูปแบบของการสูญหายที่แตกต่างกัน เพื่อศึกษาผลกระทบของทั้งตัวแปรอิสระและตัวแปรตามที่มีทำต่อวิธีการประมาณค่าสูญหาย

5. ทำการศึกษาและพัฒนาวิธี EM เพื่อใช้สำหรับประมาณค่าสูญหายในกรณีที่มีการสูญหายแบบ Nonignorable โดยในขั้น E-Step ซึ่งเป็นการหาค่าคาดหวัง อาจเพิ่มข้อมูลในส่วนที่มีการสูญหายเข้าไปในเงื่อนไขด้วย

6. ในการวิจัยครั้งนี้วิธี PMM เป็นวิธีการประมาณค่าสูญหายที่ไม่มีความเหมาะสมในทุกกรณี ทั้งนี้ก็อาจเป็นผลมาจากการเลือกใช้การแจกแจงโดยหลักเกณฑ์ (Prior Distribution) ที่ไม่มีความเหมาะสม ดังนั้นหากทราบข้อมูลเพิ่มขึ้นหรือเลือกใช้การแจกแจงโดยหลักเกณฑ์ในรูปแบบอื่นๆก็อาจจะทำให้ได้ผลการประมาณค่าสูญหายที่มีความเหมาะสมก็เป็นไปได้

รายการอ้างอิง

ภาษาไทย

ธีระพร วีระถาวร. ตัวแบบเชิงเส้น ทฤษฎีและการประยุกต์. กรุงเทพมหานคร: วิทยพัฒน์, 2541.

เพียงออ ยีสา. การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ คณะพาณิชย์ศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย. 2551.

วารุณี ตรีบำรุงศักดิ์. การพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามมีค่าสูญหาย. วิทยานิพนธ์ปริญญามหาบัณฑิต, ภาควิชาสถิติ คณะพาณิชย์ศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย. 2537.

ภาษาอังกฤษ

Jönsson, P., and Wohlin, C. Benchmarking k-nearest neighbour imputation with homogeneous likert data. Empirical Software Engineering 11, 3 (2006): 463-489.

Little, R. J. A., and Rubin, D. B. Statistical analysis with missing data. New York: John Wiley & Sons, 1987.

van Buuren, S., and Groothuis-Oudshoorn, K. Multivariate imputation by chained equations in r. Journal of Statistical Software 45 (December 2011): 1-67.

บรรณานุกรม

ภาษาไทย

กัลยา วานิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร. พิมพ์ครั้งที่ 4. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2552.

ภาษาอังกฤษ

Wang, J. X., and Miao, Y. Note on the em algorithm in linear regression model. International Mathematical Forum 4, 38 (2009): 1883-1889.

Pastor, J. B. N. Methods for the analysis of explanatory linear regression models with missing data not at random. Quanlity & Quantity 37 (November 2003): 363-376.

Ding, Y., and Ross, A. A comparison of imputation methods for handling missing scores in biometric fusion. Parrern Recognition 45 (2012): 919-933.

ภาคผนวก

ภาคผนวก ก

รายละเอียดของโปรแกรมที่ใช้ในงานวิจัย

ในการวิจัยครั้งนี้ได้ใช้โปรแกรม R เวอร์ชัน 2.13.0 ในการจำลองข้อมูลและการประมาณค่าสูญหายของแต่ละวิธี ซึ่งมีคำสั่งดังต่อไปนี้

```
library(mice)

# Impute MSE #
impute.mse <- function(Y,X,Y_true)
{
        lm_imp <- lm(Y ~ X[ ,1] + X[ ,2] + X[ ,3])
        Y_fit <- fitted(lm_imp)
        E <- Y_fit - Y_true
        SSE <- E^2
        MSE <- mean(SSE)
        return(MSE)
}
# Standardize X #
std <- function(X)
{       X1bar <- mean(X[ ,1])
        X2bar <- mean(X[ ,2])
        X3bar <- mean(X[ ,3])
        sd_X1 <- sd(X[ ,1])
        sd_X2 <- sd(X[ ,2])
        sd_X3 <- sd(X[ ,3])
        new_X1 <- (X[ ,1]-X1bar)/sd_X1
        new_X2 <- (X[ ,2]-X2bar)/sd_X2
        new_X3 <- (X[ ,3]-X3bar)/sd_X3
        new_X <- data.frame(I.x1=new_X1,I.x2=new_X2,I.x3=new_X3)
        return(new_X) }
```

```
# Find K for KNN Method #
impute.K <- function(Nobs)
{       k <- round(sqrt(Nobs))is.wholenumber <- function(k, tol =
        .Machine$double.eps^0.5)  abs(k - round(k)) < tol
        is.wholenumber(k/2) -> chk
        if(chk=="FALSE"){K<-k}
        if(chk=="TRUE"){
        lw <- abs(((k-1)^2)-Nobs)
        up <- abs(((k+1)^2)-Nobs)
        min(lw,up) -> min
        if(min==lw&lw!=up){K<-k-1}
        if(min==up&lw!=up){K<-k+1}
        if(min==lw&lw==up){K<-k-1}}
        return(K)}
# IMPUTE KNN #
impute.knn <- function(Y,X,k,Nobs,Nmis)
{       out <- c()
        for(i in Nobs+1:Nmis){
        Ynonmis <- Y[1:Nobs]
        Xnonmis <- X[1:Nobs,1:3]
        Xmisi <- X[i,1:3]
        x_1 <- rep(Xmisi[1,1],Nobs)
        x_2 <- rep(Xmisi[1,2],Nobs)
        x_3 <- rep(Xmisi[1,3],Nobs)
        Xmis <- cbind(x_1,x_2,x_3)
        D   <-   ((Xnonmis[,1]-Xmis[,1])^2)+((Xnonmis[,2]-Xmis[,2])^2)+((Xnonmis[,3]-
        Xmis[,3])^2)
        ix <- sort.int(D,index.return=TRUE)$ix
        Y_ix <- Ynonmis[ix]
```

```
        Ybar <- sum(Y_ix[1:k])/k
        out <- c(out,Ybar) }
    return(out)}
# IMPUTE EM #
impute.em <- function(Y,X,Nobs,Nmis)
{       lm_begin <- lm(Y ~ X[ ,1] + X[ ,2] + X[ ,3])
        bee <- as.matrix(coef(lm_begin))
        bee <- matrix(bee,1,4)
        dif<-c()
        bet<-c()
        repeat{
        #--- E-Step ----#
        out <- c()
        out <- c(out,Y[1:Nobs])
        Xmis <- X[Nobs+1:Nmis, ]
        beta <- bee
        E_yhat <- beta[1,1]+beta[1,2]*Xmis[ ,1]+beta[1,3]*Xmis[ ,2]+beta[1,4]*Xmis[ ,3]
        out <- c(out,E_yhat)
        #--- M-Step ---#
        lm_1 <- lm(out ~ X[ ,1] + X[ ,2] + X[ ,3])
        bee <- as.matrix(coef(lm_1))
        bee <- matrix(bee,1,4)
        diff <- abs(bee-beta)
        dif <- rbind(dif,diff)
        bet <- rbind(bet,beta)
        if(diff[1,1]<0.001 & diff[1,2]<0.001 & diff[1,3]<0.001 & diff[1,4]<0.001)
        {break}
}       return(out)}
```

```
# IMPUTE PMM #
impute.pmm<-function(Y, RY, X)
{       X <- cbind(1, as.matrix(X))
        parm <- .norm.draw(Y, RY, X)
        Yhat <- X %*% parm$beta
        return(apply(as.array(Yhat[!RY]),1,pmm.match,Yhat=Yhat[RY],Y=Y[RY]))  }
#.PMM.MATCH #
pmm.match<-function(z, Yhat=Yhat, Y=Y)
{       D <- abs(Yhat-z)
        Yest <- Y[D==min(D)]
        if (length(Yest)>1) Yest <- sample(Yest,1)
        return(Yest)        }
## Main Menu##
missing1 <- c(10,7,4,20,14,8,30,21,12)
missing2 <- c(10,10,10,20,20,20,30,30,30)
missing3 <- c(10,13,16,20,26,32,30,39,48)
AMSE <- list()
for(j in 1:9){
percent1 <- missing1[j]/100
percent2 <- missing2[j]/100
percent3 <- missing3[j]/100
N
DATA <- list()
nonmis <- c()
mis <- c()
KK <- c()
Output <- list()
MSE_COM <- c()
MSE_KNN <- c()
```

```
MSE_NNI <- c()
MSE_EM <- c()
MSE_PMM <- c()
for(t in 1:N){
n
SD_e
SD_x1
SD_x2
SD_x3
per_mis1 = percent1
per_mis2 = percent2
per_mis3 = percent3
#  Generate e,x1,x2,x3,y #
e <- rnorm(n,0,SD_e)
x1 <- rnorm(n,0,SD_x1)
x2 <- rnorm(n,0,SD_x2)
x3 <- rnorm(n,0,SD_x3)
y <- 42+x1+x2+x3+e
fulldata <- data.frame(x1=x1,x2=x2,x3=x3,e=e,y=y)
#  Find cut point #
cut1 <- (qnorm(1/3)*sd(y))+mean(y)
cut2 <- (qnorm(2/3)*sd(y))+mean(y)
# Divided y into 3 parts #
y[y<=cut1] -> part1
y[y>cut1&y<=cut2] -> part2
y[y>cut2] -> part3
rbinom(length(part1),1,per_mis1) -> a
rbinom(length(part2),1,per_mis2) -> b
rbinom(length(part3),1,per_mis3) -> c
```

```
cbind(part1,a) -> p1
cbind(part2,b) -> p2
cbind(part3,c) -> p3
rbind(p1,p2,p3) -> miss
colnames(miss)[c(1,2)] <-c ("y","R")
merge(fulldata,miss,by='y',all.y=TRUE) -> missdata
missdata$y[which(missdata$R==1)]=NA
merge(fulldata,missdata,by=c("x1","x2","x3","e")) -> data
colnames(data)[c(5,6)] <- c("y.full","y.miss")
data[with(data,order(data$y.miss,na.last=TRUE)),] -> data
y.true <- data$y.full - data$e
data <- cbind(data,y.true)
DATA[[t]] <- data
m <- sum(data$R==0)
nn <- sum(data$R==1)
nonmis <- c(nonmis,m)
mis <- c(mis,nn)
dat_Y <- data$y.miss
dat_X <- data[,1:3]
#  Estimated regression coefficient + MSE for complete data #
mse.com <- impute.mse(data$y.full,dat_X,data$y.true)
MSE_COM <- c(MSE_COM,mse.com)
##  Complete Case  ##
if(m==n){
Output[[t]] <- data
mse.knn <- mse.com
mse.nni <- mse.com
mse.em  <- mse.com
mse.pmm <- mse.com  }
```

```
##  Missing Case  ##
if(m!=n){
##  NNI + KNN Method  ##
K <- impute.K(m)
KK <- c(KK,K)
std_x<-std(dat_X)
# NNI #
out_nni <- c()
out_nni <- c(out_nni, dat_Y[1:m])
out_nni <- c(out_nni,impute.knn(dat_Y,std_x,1,m,nn))
## KNN ##
out_knn <- c()
out_knn <- c(out_knn, dat_Y[1:m])
out_knn <- c(out_knn,impute.knn(dat_Y,std_x,K,m,nn))
#  Estimated regression coefficient + MSE for NNI + KNN Method #
mse.nni <- impute.mse(out_nni,dat_X,data$y.true)
mse.knn <- impute.mse(out_knn,dat_X,data$y.true)
#  EM Method #
out_em <- impute.em(dat_Y,dat_X,m,nn)
#  Estimated regression coefficient + MSE for EM Method  #
mse.em <- impute.mse(out_em,dat_X,data$y.true)
# PMM Method #
RRY <- rep(T, n)
RRY[m+1:nn] <- F
out_pmm <- c()
out_pmm <- c(out_pmm, dat_Y[1:m])
out_pmm <- c(out_pmm, impute.pmm(dat_Y,RRY,dat_X))
# Estimated regression coefficient + MSE for PMM Method #
mse.pmm <- impute.mse(out_pmm,dat_X,data$y.true)
```

```
Output[[t]] <- cbind(data[,c(1:6,8)],out_nni,out_knn,out_em,out_pmm)
}
MSE_NNI <- c(MSE_NNI,mse.nni)
MSE_KNN <- c(MSE_KNN,mse.knn)
MSE_EM <- c(MSE_EM,mse.em)
MSE_PMM <- c(MSE_PMM,mse.pmm)
}
## IMPUTE  AMSE ##
AMSE[[j]]<-cbind(mean(MSE_COM),mean(MSE_NNI),mean(MSE_KNN),
mean(MSE_EM),mean(MSE_PMM))
pie(c(j,9-j),radius=1,clockwise=T)
}
```

ภาคผนวก ข

<u>ตารางที่ 1</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ของทั้ง 4 วิธี เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย	EM	NNI	KNN	PMM
50	10	None	723.141	763.1815	685.5151	768.9203
		Medium	736.183	775.3753	699.249	778.5785
		High	737.8014	773.8669	702.7598	781.902
	20	None	834.9311	897.8029	747.3564	922.3917
		Medium	857.3068	919.9935	775.7086	940.4895
		High	909.5812	986.3753	837.0822	1004.505
	30	None	980.3351	1055.121	843.1639	1098.473
		Medium	1011.39	1103.429	888.2242	1137.105
		High	1204.29	1292.677	1099.207	1341.93
100	10	None	367.3659	388.5528	352.5124	397.0479
		Medium	362.6913	386.0516	349.1303	392.0174
		High	380.9417	406.9966	368.0581	411.4513
	20	None	423.5191	469.446	392.5287	477.4507
		Medium	433.1827	481.6733	406.0306	493.3501
		High	488.102	538.4953	467.474	553.4344
	30	None	474.3056	542.132	425.7903	563.5465
		Medium	538.1941	613.9251	496.3531	634.31
		High	723.2525	812.3707	704.7652	827.824

<u>ตารางที่ 1(ต่อ)</u> แสดงค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ของทั้ง 4 วิธี เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%การสูญหาย	ระดับของการสูญหาย	EM	NNI	KNN	PMM
200	10	None	181.0283	193.3641	174.9084	196.6516
		Medium	182.2302	195.3486	177.2552	198.473
		High	196.6496	211.8122	192.9311	214.7278
	20	None	201.8056	230.9516	191.3193	233.0671
		Medium	225.8031	254.4321	216.785	261.3875
		High	291.3785	320.4179	287.7619	329.4321
	30	None	229.794	273.2226	214.4574	285.5349
		Medium	301.1511	342.7986	289.7244	361.726
		High	484.6579	536.813	493.1248	551.6971

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอุษณีย์ วงศ์อามาตย์ เกิดวันศุกร์ที่ 20 พฤศจิกายน พ.ศ. 2530 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล ในปีการศึกษา 2552 และเข้าศึกษาต่อในหลักสูตรสถิติศาสตรมหาบัณฑิต (สต.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553