การเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล

นางสาววริษฐา กณิกนันต์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต

สาขาวิชาสถิติ ภาควิชาสถิติ

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2555

COMPARISON OF THE ESTIMATION METHODS FOR THE MULTIPLE LINEAR REGRESSION MODEL WITH NONIGNORABLE – MISSING DEPENDENT AND INDEPENDENT VARIABLES

Miss Warittha Kaniknant

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	การเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล
โดย	นางสาววริษฐา กณิกนันต์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.กัลยา วานิชย์บัญชา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(อาจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี)

.....กรรมการ

(รองศาสตราจารย์ ดร.ธีระพร วีระถาวร)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.ธิดาพร ศุภภากร)

วริษฐา กณิกนันต์ : การเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล. (COMPARISON OF THE ESTIMATION METHODS FOR THE MULTIPLE LINEAR REGRESSION MODEL WITH NONIGNORABLE – MISSING DEPENDENT AND INDEPENDENT VARIABLES) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : อ.ดร.อนุภาพ สมบูรณ์สวัสดี, 129 หน้า.

การศึกษาในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณตัวแบบการถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบนอนอิกนอร์เรเบิล วิธีการประมาณที่ใช้ในการศึกษาครั้งนี้คือ วิธี EM Algorithm (EM) วิธี K-Nearest Neighbor (KNN) และวิธี Predictive Mean Matching (PMM)

ข้อมูลที่ใช้ในการศึกษาได้จากการจำลองโดยมีสัดส่วนของการสูญหายของข้อมูล 3 ระดับ คือ 10% 20% และ 30% และมีระดับการสูญหายแบบนอนอิกนอร์เรเบิล 3 ระดับคือ ไม่มี ปานกลาง และสูง ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error : AMSE) ของแต่ละวิธีการจะใช้เปรียบเทียบโดยวิธีการที่ดีที่สุดจะมีค่า AMSE น้อยที่สุดจะเป็นวิธีการที่ดีที่สุด ผลการวิจัย พบว่า i) ในกรณีส่วนใหญ่วิธีการ KNN จะเป็นวิธีการประมาณที่ดีที่สุด โดยเฉพาะเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีขนาดปานกลางและสูง (30 และ 90) ii) สัดส่วนการสูญหายและระดับการสูญหายแบบนอนอิกนอร์เรเบิลที่สูงมีผลทำให้วิธีการ EM เป็นวิธีการประมาณค่าที่ดีที่สุดในบางกรณี โดยเฉพาะเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีขนาดเล็ก(10) iii) วิธีการประมาณทุกวิธีจะมีประสิทธิภาพน้อยลงเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน สัดส่วนของการสูญหาย และระดับการสูญหายแบบนอนอิกนอร์เรเบิลเพิ่มสูงขึ้น

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2555

5481675726 : MAJOR STATISTICS

KEYWORDS : MULTIPLE LINEAR REGRESSION / NONIGNORABLE – MISSING DEPENDENT AND INDEPENDENT VARIABLES

WARITTHA KANIKNANT : COMPARISON OF THE ESTIMATION METHODS FOR THE MULTIPLE LINEAR REGRESSION MODEL WITH NONIGNORABLE – MISSING DEPENDENT AND INDEPENDENT VARIABLES. ADVISOR : ANUPAP SOMBOONSAVATDEE, Ph.D., 129 pp.

The objective of this study is to compare the estimation methods for the multiple linear regression model with nonignorable-missing dependent and independent variables. The estimation methods considered in study are EM Algorithm (EM) , K-Nearest Neighbor (KNN) and Predictive Mean Matching (PMM).

Data are simulated with three levels of missing proportion of data of 10%, 20%, 30% and three levels of nonignorable missingness of none, medium, high. The average mean square errors (AMSEs) of all methods are compared with the best method will have the smallest value of AMSE. The findings are the followings : i) KNN method performs best when the standard deviation of error is medium and high (30 and 90), ii) EM method performs best especially when the standard deviation of error is small (10), iii) The performances of all estimation methods perform decrease as the standard deviation of errors, the missing proportion, or level of nonignorable missingness increase.

Department: Statistics Student's Signature .....

Field of Study: Statistics Advisor's Signature .....

Academic Year: 2012

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความเอาใจใส่ของอาจารย์ ดร.อนุภาพ สมบูรณ์สวัสดี อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่ได้ให้คำปรึกษาและแนวทางการวิจัยต่างๆ มากมาย ตลอดจนให้คำแนะนำในการแก้ไขข้อบกพร่องมาโดยตลอด จนวิทยานิพนธ์เล่มนี้เสร็จสมบูรณ์ ผู้วิจัยจึงกราบขอบพระคุณเป็นอย่างสูง

ผู้วิจัยขอกราบขอบพระคุณรองศาสตราจารย์ ดร.กัลยา วานิชย์บัญชา ประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.ธีระพร วีระถาวร และอาจารย์ ดร.ธิดาพร ศุภภากร กรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ ตรวจสอบ ตลอดจนชี้แนะแนวทางการแก้ไขวิทยานิพนธ์เล่มนี้ให้สำเร็จสมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ผู้ซึ่งคอยให้กำลังใจและเป็นแรงผลักดันให้วิทยานิพนธ์เล่มนี้เสร็จสมบูรณ์ รวมทั้งขอบคุณรุ่นพี่ และเพื่อนๆ ทุกคนที่คอยให้คำปรึกษาและเป็นกำลังใจให้มาโดยตลอด

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการวิจัยด้านต่างๆ การพยากรณ์เป็นเทคนิคหนึ่งที่มีความจำเป็นและมีส่วนสำคัญที่จะทำให้สามารถคาดคะเนเหตุการณ์ล่วงหน้าได้ ซึ่งอาจส่งผลต่อการวางแผนเพื่อให้สถานการณ์ต่างๆ เป็นไปตามเป้าหมาย เทคนิคการพยากรณ์เป็นที่นิยมใช้ในด้านต่างๆ มากมาย เช่น ทางด้านเศรษฐกิจการเงิน สังคมศาสตร์ วิทยาศาสตร์ การแพทย์ อุตสาหกรรม เป็นต้น

ในการเก็บรวบรวมข้อมูลจากแบบสำรวจของการทำวิจัยด้านต่างๆ ในปัจจุบันนี้ได้มีการพัฒนารูปแบบการเก็บข้อมูลขึ้นมามากมายเพื่อรองรับความสะดวกในการตอบแบบสอบถามสำหรับผู้ตอบและผู้สัมภาษณ์ แต่ยังมีปัญหาหลักอย่างหนึ่งที่เกิดขึ้นอยู่เสมอนั่นคือ ข้อมูลสูญหาย (Missing Data) ซึ่งอาจเกิดขึ้นได้ในบางกรณี ยกตัวอย่างเช่น ผู้ตอบแบบสอบถามอาจไม่เข้าใจคำถามในแบบสอบถาม หรือบางคำถามผู้ตอบอาจไม่สามารถให้ข้อมูลได้ ซึ่งปัญหาเหล่านี้อาจเกิดจากการออกแบบสอบถามไม่ครอบคลุม หรืออาจเกิดจากความผิดพลาดในการบันทึกข้อมูล

เทคนิคการพยากรณ์ที่นิยมใช้คือ การวิเคราะห์การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ซึ่งในการวิเคราะห์ข้อมูลจะประกอบไปด้วยตัวแปรอิสระ (Independent Variable) ตั้งแต่ 2 ตัวขึ้นไป และตัวแปรตาม (Dependent Variable) ซึ่งตัวแปรอิสระและตัวแปรตามจะมีความสัมพันธ์กันในรูปแบบเชิงเส้น ดังนั้น หากเกิดเหตุการณ์ในกรณีที่ข้อมูลสูญหายทั้งตัวแปรอิสระและตัวแปรตาม ก็ย่อมส่งผลให้เกิดปัญหาในการวิเคราะห์ขึ้นมาทันที วิธีการแก้ไขปัญหาดังกล่าว สามารถทำได้หลายวิธี ซึ่งวิธีการที่ง่ายที่สุดคือการตัดข้อมูลบางส่วนที่ไม่สมบูรณ์ทิ้งไป แต่วิธีการนี้จะทำให้สูญเสียข้อมูลที่ใช้ในการวิเคราะห์ อาจเป็นไปได้ว่าข้อมูลที่เหลืออาจไม่สามารถเป็นตัวแทนของประชากรทั้งหมดได้ จึงอาจนำไปสู่ข้อสรุปที่ผิดพลาด ดังนั้นจึงมีการคิดค้นวิธีการต่างๆ เพื่อนำมาใช้ในการประมาณค่าสูญหาย ทั้งนี้ความมีประสิทธิภาพของแต่ละวิธีการจะมากหรือน้อยก็ขึ้นอยู่กับความเหมาะสมของลักษณะข้อมูลที่จะนำไปใช้ด้วย ดังนั้นเราจึงสนใจที่จะศึกษาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายแบบต่างๆ ที่เกิดขึ้นทั้งในตัว

แปรอิสระและตัวแปรตาม ในกรณีที่มีการสูญหายแบบ Nonignorable ซึ่งเป็นการสูญหายที่ควรจะได้รับการแทนที่ มิเช่นนั้นแล้วจะส่งผลกระทบอย่างมากต่อผลการวิเคราะห์ซึ่งอาจนำไปสู่ข้อสรุปที่ผิดพลาดได้ เนื่องจากการสูญหายแบบ Nonignorable จะเกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของตัวแปรนั้นไม่มีความสัมพันธ์กับตัวแปรอื่น แต่จะมีความสัมพันธ์กับตัวมันเอง เช่น ในการสำรวจข้อมูลเกี่ยวกับรายได้นั้น พบว่า ผู้ตอบแบบสอบถามที่มีรายได้สูงส่วนใหญ่ มักไม่ตอบคำถามในเรื่องของรายได้ของตนเอง การสูญหายเช่นนี้จึงถือว่าเป็นการสูญหายแบบ Nonignorable

ในการวิจัยครั้งนี้ผู้วิจัยได้ศึกษารวบรวมงานวิจัยที่เกี่ยวข้องกับวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุ โดยงานวิจัยที่เกี่ยวข้องมีดังต่อไปนี้

วารุณี ตรีบำรุงศักดิ์ (2537) ได้ทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายเมื่อตัวแปรตามมีการสูญหายในการพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุ โดยได้ทำการประมาณค่าสูญหายด้วยวิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย วิธี EM Algorithm และวิธีฮันท์ การเปรียบเทียบกระทำภายใต้สถานการณ์ของขนาดตัวอย่าง 10 , 20 , 30 , 50 และ 70 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 5 , 10 , 15 , 20 และ 25 สัดส่วนของการสูญหายของตัวแปรตามคือ 10% , 20% , 30% , 40% , 50% , 60% และ 70% ในการเปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายจะใช้ค่า RMSE ผลการวิจัยพบว่าในกรณีที่ตัวอย่างมีขนาดเล็ก ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนไม่สูงมาก และสัดส่วนของการสูญหายมาก วิธีการของฮันท์จะมีความเหมาะสมมากที่สุด แต่ถ้าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง วิธีค่าเฉลี่ยจะดีที่สุด และในกรณีที่ตัวอย่างมีขนาดใหญ่ วิธีสูญหายจะเหมาะสมเกือบทุกกรณี

เพียงออ ยีสา (2551) ได้ทำการศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุเพื่อการพยากรณ์โดยพิจารณาข้อมูล 2 ลักษณะ คือข้อมูลภาคตัดขวาง และข้อมูลอนุกรมเวลาที่มีปัจจัยด้านแนวโน้มและปัจจัยฤดูกาล โดยทำการประมาณค่าสูญหายด้วยวิธี Regression Imputation (RI) วิธี Nearest Neighbor Imputation (NNI) วิธี Weighted Nearest Neighbor (WNR) และวิธี EM Algorithm (EM) โดยจำลองข้อมูล

ซึ่งกำหนดขนาดตัวอย่าง 50 , 100 และ 200 ส่วนเบี่ยงเบนมาตรฐาน 5 , 10 , 15 , 20 และ 25 ร้อยละของการสูญหายของตัวแปรเป็น 5 , 10 และ 20 ตามลำดับ ซึ่งกำหนดให้ตัวแปรอิสระมีการแจกแจงแบบปกติ เกณฑ์ในการเปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายจะใช้ค่า MAPE ผลการวิจัยสรุปได้ว่า สำหรับข้อมูลภาคตัดขวาง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำถึงปานกลาง วิธี RI และ EM ให้ผลดีกว่าวิธีอื่น แต่เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูง วิธี WNR ให้ผลดีกว่าอื่น ส่วนข้อมูลที่เป็นอนุกรมเวลา ถ้าข้อมูลมีอิทธิพลของฤดูกาลสูง วิธีที่ดีที่สุด คือ WNR แต่ถ้าข้อมูลมีอิทธิพลของแนวโน้มสูง วิธี RI และ EM จะเป็นวิธีที่เหมาะสมที่สุด

เนื่องจากมีงานวิจัยเป็นจำนวนมากที่ศึกษาเกี่ยวกับเรื่องการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุ ซึ่งงานวิจัยส่วนใหญ่จะศึกษาในกรณีที่ข้อมูลเกิดการสูญหายอย่างสุ่มซึ่งจะส่งผลกระทบต่อการวิเคราะห์น้อยกว่าการสูญหายแบบ Nonignorable ที่ยังไม่ค่อยได้รับความนิยมในการนำมาศึกษา แต่ก็ยังมีงานวิจัยชิ้นหนึ่งที่ศึกษาเรื่องการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบ Nonignorable ในการวิเคราะห์การถดถอยเชิงพหุ ซึ่งได้แก่งานวิจัยของ อุษณีย์ วงศ์อามาตย์ (2555) แต่ลักษณะของการสูญหายเกิดขึ้นเฉพาะในตัวแปรตามเพียงตัวแปรเดียวเท่านั้น ซึ่งในความเป็นจริงแล้วลักษณะของการสูญหายอาจเกิดขึ้นได้ทั้งในตัวแปรอิสระและตัวแปรตาม ซึ่งงานวิจัยดังกล่าวมีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธี K-Nearest Neighbor Imputation (KNN) วิธี EM Algorithm และวิธี Predictive Mean Matching Imputation (PMM) เมื่อข้อมูลตัวแปรตามมีการสูญหายแบบ Nonignorable เพียงตัวแปรเดียว โดยข้อมูลที่ใช้ในการศึกษาได้จากการจำลอง ซึ่งมีสัดส่วนของการสูญหาย 3 ระดับ คือ 10% 20% และ 30% และมีระดับการสูญหายแบบ Nonignorable 3 ระดับคือ ไม่มี ปานกลาง และสูง จากการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average Mean Square Error : AMSE) ผลสรุปเป็นดังนี้

1. วิธีการประมาณทุกวิธีสามารถประมาณได้ดีขึ้นเมื่อตัวอย่างมีขนาดใหญ่ขึ้น
2. วิธีการประมาณทุกวิธีประมาณได้แย่ลงเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของการสูญหายและระดับของการสูญหายแบบ Nonignorable มีค่าเพิ่มขึ้น

3. โดยรวมแล้ววิธี EM Algorithm ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าไม่สูง (10-30)
4. วิธี K-Nearest Neighbor Imputation (KNN) ประมาณค่าได้ดีที่สุดเมื่อส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าสูง (90)

อย่างไรก็ตาม ยังไม่มีงานวิจัยใดที่ทำการศึกษาในกรณีที่ตัวแปรอิสระและตัวแปรตามเกิดการสูญหายแบบ Nonignorable ดังนั้นในงานวิจัยนี้จึงสนใจที่จะทำการศึกษาต่อยอดโดยจะทำการศึกษาในกรณีที่ตัวแปรอิสระและตัวแปรตามเกิดการสูญหายอย่างมีความสัมพันธ์กัน นอกจากนี้ยังศึกษาเพิ่มเติมในกรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนแตกต่างกันอีกด้วย โดยจะใช้ขอบเขตที่คล้ายคลึงกัน เพื่อเปรียบเทียบวิธีการประมาณสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบ Nonignorable เพื่อให้เกิดความครอบคลุมสำหรับลักษณะการใช้งานจริง และสามารถเลือกใช้วิธีการประมาณค่าสูญหายแบบต่างๆ ในกรณีที่เกิดการสูญหายทั้งในตัวแปรอิสระและตัวแปรตามได้อย่างถูกต้องและแม่นยำ โดยในงานวิจัยชิ้นนี้ จะทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายจำนวน 3 วิธี ได้แก่

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

ส่วนเทคนิคที่นำมาใช้ในการพยากรณ์คือเทคนิคการวิเคราะห์การถดถอยเชิงเส้นพหุ และจะทำการหาค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดเพื่อนำค่าสัมประสิทธิ์การถดถอยมาใช้ในการสร้างสมการพยากรณ์ต่อไป ซึ่งวิธีการประมาณค่าสูญหายที่ดีที่สุดจะเป็นวิธีการที่ให้ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์น้อยที่สุด

1.2 **วัตถุประสงค์ของการวิจัย**

1. เพื่อศึกษาวิธีการประมาณค่าสูญหายของตัวแปรอิสระและตัวแปรตามที่มีความสัมพันธ์กันอย่างมีเงื่อนไข ในกรณีที่มีการสูญหายแบบ Nonignorable

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายเมื่อตัวแปรอิสระและตัวแปรตามเมื่อมีการสูญหายแบบ Nonignorable ทั้ง 3 วิธี ได้แก่ วิธี EM Algorithm วิธี K-Nearest Neighbor Imputation และวิธี Predictive Mean Matching Imputation (PMM)

1.3 ข้อตกลงเบื้องต้น

1. ในงานวิจัยนี้จะสนใจกรณีที่ตัวแปรอิสระ (x) และตัวแปรตาม (y) มีความสัมพันธ์กันภาย ใต้การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) ซึ่งมีรูปแบบดังนี้

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

เมื่อ y_i คือ ค่าสังเกตของตัวแปรตามของข้อมูลตัวที่ i

x_{i1} คือ ค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 1

x_{i2} คือ ค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 2

x_{i3} คือ ค่าสังเกตของข้อมูลตัวที่ i ของตัวแปรอิสระตัวที่ 3

β_p คือ สัมประสิทธิ์การถดถอยตัวที่ p เมื่อ $p = 0,1,2,3$

ε_i คือ ค่าความคลาดเคลื่อนของข้อมูลตัวที่ i

n คือ จำนวนค่าสังเกตทั้งหมด

m คือ จำนวนค่าสังเกตที่ทราบค่า

$n-m$ คือ จำนวนค่าสังเกตที่สูญหาย

2. ความคลาดเคลื่อนเป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ โดยมีค่าเฉลี่ยเท่ากับ 0 ความแปรปรวนเท่ากับ σ^2 ($\varepsilon_i \sim N(0,\sigma^2)$)
3. $\varepsilon_i, \varepsilon_k$ ไม่มีสหสัมพันธ์กัน นั่นคือ $E(\varepsilon_i, \varepsilon_k) = 0$ เมื่อ $i \neq k$
4. การสูญหายของข้อมูลเกิดขึ้นที่ตัวแปรอิสระตัวใดตัวหนึ่ง และตัวแปรตาม อย่างมีความสัมพันธ์กันด้วยความน่าจะเป็นในระดับต่างๆ ตามรูปแบบที่กำหนด โดยจะแบ่งช่วงของตัวแปรอิสระและตัวแปรตามออกเป็น 3 ช่วง และกำหนดให้ร้อยละของการสูญหายในแต่ละช่วงแตกต่างกันไป

1.4 คำจำกัดความที่ใช้ในการวิจัย

ในงานวิจัยนี้มีคำจำกัดความที่ใช้ในงานวิจัยดังนี้

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

<u>ช่วงกลาง</u> คือ ช่วงของพื้นที่ใต้โค้งปกติมาตรฐานที่อยู่ใน (z, z') เมื่อ $z, z' \sim N(0,1)$ และ $z < z'$ ดังนั้น ในช่วงนี้จะมีพื้นที่เป็น $P(z < Z < z') \times 100\%$ ของพื้นที่ทั้งหมด

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

วิธีในการแบ่งช่วงของตัวแปรอิสระคือ

ถ้า $x_{ip} \le \mu_{x_{ip}} + z_{\frac{1}{3}} \sigma_{x_{ip}}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ใน<u>ช่วงต้น</u>

$\mu_{x_{ip}} + z_{\frac{1}{3}} \sigma_{x_{ip}} < x_{ip} \le \mu_{x_{ip}} + z_{\frac{2}{3}} \sigma_{x_{ip}}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ใน<u>ช่วงกลาง</u>

และ $\mu_{x_{ip}} + z_{\frac{2}{3}} \sigma_{x_{ip}} < x_{ip}$ ตัวแปร x_{ip} นี้จะถูกจัดให้อยู่ใน<u>ช่วงปลาย</u>

ซึ่ง $\sigma_{x_{ip}}$ จะแบ่งตามกรณีที่เราจะศึกษา ตามขอบเขตของการวิจัย ซึ่งมี 2 ลักษณะคือ

- $\sigma_{x_{i1}} = \sigma_{x_{i2}} = \sigma_{x_{i3}} = \sqrt{300}$
- $\sigma_{x_{i1}} = \sqrt{100}, \sigma_{x_{i2}} = \sqrt{300}, \sigma_{x_{i3}} = \sqrt{500}$

โดยที่ $p = 1,2,3$

$i = 1,2,...,m,m+1,...,n$

$z_{\frac{1}{3}} = -0.43$

$z_{\frac{2}{3}} = +0.43$

<u>ความน่าจะเป็นของการสูญหายในแต่ละช่วง</u> คือ จำนวนตัวอย่างที่สูญหายในแต่ละช่วง / จำนวนตัวอย่างทั้งหมดที่อยู่ในช่วงนั้น

<u>สัดส่วนของการสูญหาย</u> คือ ความน่าจะเป็นของการสูญหาย $\times 100\%$

1.5 ขอบเขตของการวิจัย

1. ตัวแปรอิสระมีการแจกแจงแบบปกติ (Normal Distribution) ซึ่งมีฟังก์ชันการแจกแจงคือ

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

ซึ่งมีค่าเฉลี่ยเท่ากับ μ และความแปรปรวนเท่ากับ σ^2 ซึ่งในการศึกษาครั้งนี้เราจะศึกษาการสูญหายของข้อมูลตัวแปรอิสระตัวใดตัวหนึ่งโดยคำนึงถึงระดับของความแปรปรวน กล่าวคือจะศึกษาการสูญหายของข้อมูลในกรณีที่ความแปรปรวนเท่ากันในตัวแปรอิสระและความแปรปรวนมีขนาดเล็ก กลาง และใหญ่ ซึ่งจะแบ่งลักษณะการแจกแจงของตัวแปรอิสระออกเป็น 2 รูปแบบ ดังต่อไปนี้

<u>แบบที่ 1</u> $X_1 \sim N(0,300), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$ โดยจะศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนเท่ากัน

<u>แบบที่ 2</u> $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ โดยจะศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนขนาดเล็ก ปานกลางและใหญ่

โดยจะกำหนดให้ตัวแปรอิสระไม่มีความสัมพันธ์กันนั่นคือ มีค่าสหสัมพันธ์เท่ากับ 0

2. ค่าความคลาดเคลื่อน (ε) มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ σ^2 $(\varepsilon \sim N(0,\sigma^2))$ โดยจะกำหนดให้ $\sigma=10,30$ และ 90 เมื่อพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variation) ที่ 75% 100% และ 225% ตามลำดับ
3. ตัวแปรตาม (y) เกิดจากความสัมพันธ์ระหว่างตัวแปรอิสระ (x) ภายใต้การถดถอยเชิงเส้นพหุ คือ

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\varepsilon_i \qquad ;i=1,2,...,m,m+1,...,n$$

โดยกำหนดให้ $\beta_0=42$ และ $\beta_1=\beta_2=\beta_3=1$ เนื่องจากในการศึกษาครั้งนี้ต้องการเปรียบเทียบชุดข้อมูลของตัวแปรอิสระที่มีความแปรปรวนแตกต่างกัน ดั้งนั้นถ้าหากเปลี่ยนค่า β_1,β_2 และ β_3 จะส่งผลให้ตัวแปรตามที่ได้จากชุดข้อมูลของตัวแปรอิสระแต่ละ

ชุดมีความแปรปรวนแตกต่างกันด้วย ดังนั้นเพื่อควบคุมความแปรปรวนของตัวแปรตามให้มีขนาดเท่ากัน จึงกำหนดให้ $\beta_1 = \beta_2 = \beta_3 = 1$

4. ขนาดตัวอย่างมี 3 ขนาดคือ 50 100 และ 200
5. การสูญหายของข้อมูลที่เกิดขึ้นที่ตัวแปรตาม และตัวแปรอิสระตัวใดตัวหนึ่งเท่านั้น โดยจะทำการแบ่งตัวแปรอิสระและตัวแปรตามออกเป็น 3 ช่วงด้วยอัตราส่วนเท่าๆ กันและให้แต่ละช่วงมีสัดส่วนของการสูญหายแตกต่างกัน โดยจะสร้างตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่ ด้วยความน่าจะเป็น 0.1 0.2 และ 0.3 เพื่อให้เกิดการสูญหายเฉลี่ยในช่วงต้น ช่วงกลาง และช่วงปลาย ตามลำดับ แล้วจะจับคู่กับข้อมูลตัวแปรอิสระและตัวแปรตาม
6. พื้นที่ใต้เส้นโค้งที่มีการแจกแจงแบบปกติของข้อมูลตัวแปรอิสระและตัวแปรตามจะถูกแบ่งออกเป็น 3 ช่วง ด้วยอัตราส่วน 1 : 1 : 1 โดยจะเรียกว่า ช่วงต้น : ช่วงกลาง : ช่วงปลาย ตามลำดับ
7. การสูญหายของข้อมูลเกิดขึ้นอย่างมีความสัมพันธ์กันระหว่างตัวแปรตามและตัวแปรอิสระ และเป็นการสูญหายแบบ Nonignorable กล่าวคือจะเกิดการสูญหายในตัวแปรใดตัวแปรหนึ่งแบบ Nonignorable ถ้าความน่าจะเป็นของการสูญหายของตัวแปรนั้นไม่มีความสัมพันธ์กับค่าของตัวแปรอื่นๆ แต่จะมีความสัมพันธ์กับค่าของตัวมันเอง เช่น ข้อมูลรายได้จะถือว่ามีการสูญหายแบบ Nonignorable ถ้าครอบครัวที่มีรายได้สูงส่วนใหญ่มักจะไม่เปิดเผยรายได้ของตนเอง จึงทำให้เกิดข้อมูลสูญหาย

ซึ่งในการกำหนดสัดส่วนของการสูญหายจะแบ่งตัวแปรตามและตัวแปรอิสระออกเป็น 3 ช่วง และจะกำหนดให้สัดส่วนของการสูญหายของข้อมูลแตกต่างกันตามระดับของการสูญหายแบบ Nonignorable โดยจะกำหนดให้ช่วงของตัวแปรตามและตัวแปรอิสระที่มีค่ามากจะมีสัดส่วนของการสูญหายมากกว่าช่วงของตัวแปรตามและตัวแปรอิสระที่มีค่าน้อย ซึ่งจะส่งผลให้แต่ละช่วงมีความน่าจะเป็นในการสูญหายที่สูง-ต่ำแตกต่างกันไป

ระดับการสูญหายแบบ Nonignorable จะแบ่งออกเป็น 3 ระดับคือ ไม่มี ปานกลาง และสูง ซึ่งในแต่ละช่วงจะมีอัตราส่วนของการสูญหายดังต่อไปนี้

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

8. เพื่อศึกษาเพิ่มเติมจากงานวิจัยของ อุษณีย์ วงศ์อามาตย์ (2555) กำหนดให้การสูญหายข้อมูลเกิดขึ้นทั้งในตัวแปรตามและตัวแปรอิสระอย่างมีความสัมพันธ์กัน ภายใต้เงื่อนไขที่จะแสดงดังต่อไปนี้

จากทฤษฎีความน่าจะเป็นแบบมีเงื่อนไข $P(\delta_y=1|\delta_x=1)=\frac{P(\delta_y=1),P(\delta_x=1)}{P(\delta_x=1)}$

$$P(\delta_y=1|\delta_x=0)=\frac{P(\delta_y=1),P(\delta_x=0)}{P(\delta_x=0)}$$

กำหนดให้ $\delta_x=1$ หมายถึง เหตุการณ์ที่ตัวแปรอิสระเกิดการสูญหาย

$\delta_x=0$ หมายถึง เหตุการณ์ที่ตัวแปรอิสระไม่เกิดการสูญหาย

$\delta_y=1$ หมายถึง เหตุการณ์ที่ตัวแปรตามเกิดการสูญหาย

$P(\delta_x=1)=P(\delta_y=1)$ คือความน่าจะเป็นของการสูญหายแบบ Nonignorable ที่ระดับต่างๆ และเนื่องจากในงานวิจัยนี้ เราสนใจศึกษาแต่ความสัมพันธ์เชิงบวก ดังนั้น ระดับการสูญหายแบบ Nonignorable ระหว่าง x และ y ซึ่งจะใช้สัญลักษณ์ R จะเท่ากับ

$$\mathrm{R}=\frac{P(\delta_y=1|\delta_x=1)}{P(\delta_y=1|\delta_x=0)}=1,2,4$$

ซึ่ง R = 1 หมายถึง $P(\delta_y=1|\delta_x=1)$ มีความสัมพันธ์กับ $P(\delta_y=1|\delta_x=0)$ ในระดับไม่มี

R = 2 หมายถึง $P(\delta_y=1|\delta_x=1)$ มีความสัมพันธ์กับ $P(\delta_y=1|\delta_x=0)$ ในระดับปานกลาง

R = 4 หมายถึง $P(\delta_y=1|\delta_x=1)$ มีความสัมพันธ์กับ $P(\delta_y=1|\delta_x=0)$ ในระดับสูง

ดังนั้นความน่าจะเป็นของการสูญหายในแต่ละช่วง สามารถคำนวณได้จาก

$$\begin{aligned}P(\delta_y=1)&=[P(\delta_y=1),P(\delta_x=1)]+[P(\delta_y=1),P(\delta_x=0)]\\&=[P(\delta_y=1|\delta_x=1)P(\delta_x=1)]+[P(\delta_y=1|\delta_x=0)P(\delta_x=0)]\\&=[(\mathrm{R}\times P(\delta_y=1|\delta_x=0)P(\delta_x=1)]+[P(\delta_y=1|\delta_x=0)][1-P(\delta_x=1)]\\&=P(\delta_y=1|\delta_x=0)[\ \mathrm{R}\times P(\delta_x=1)+(1-P(\delta_x=1))]\end{aligned}$$

$$P(\delta_y = 1 | \delta_x = 0) = \frac{P(\delta_y = 1)}{\mathrm{R} \times P(\delta_x = 1) + (1 - P(\delta_x = 1))}$$

ดังนั้น

เมื่อ $P(\delta_y = 1 | \delta_x = 0) = \frac{P(\delta_y = 1)}{\mathrm{R} \times P(\delta_x = 1) + (1 - P(\delta_x = 1))}$

จะได้ว่า $P(\delta_y = 1 | \delta_x = 1) = \mathrm{R} \times P(\delta_y = 1 | \delta_x = 0)$

9. ความน่าจะเป็นของการสูญหายโดยเฉลี่ยจะกำหนดให้เท่ากับ 0.1 , 0.2 และ 0.3 โดยให้แต่ละช่วงมีการสูญหายดังต่อไปนี้

ความน่าจะเป็นของการสูญหายโดยเฉลี่ย	ระดับการสูญหายแบบ Nonignorable	ความน่าจะเป็นของการสูญหายในแต่ละช่วง		
		ช่วงต้น	ช่วงกลาง	ช่วงปลาย
0.1	ไม่มี	0.10	0.10	0.10
	ปานกลาง	0.07*	0.10	0.13
	สูง	0.04	0.10	0.16
0.2	ไม่มี	0.20	0.20	0.20
	ปานกลาง	0.14	0.20	0.26
	สูง	0.08	0.20	0.32
0.3	ไม่มี	0.30	0.30	0.30
	ปานกลาง	0.21	0.30	0.39
	สูง	0.12	0.30	0.48

เนื่องจากข้อมูลตัวแปรตามและตัวแปรอิสระมีการสูญหายอย่างมีความสัมพันธ์กันเราจึงต้องคำนวณหาความน่าจะเป็นของการสูญหายในแต่ละช่วงซึ่งจะยกตัวอย่างแสดงวิธีการคำนวณโดยใช้ขอบเขตที่เรากำหนดจากข้อ 8 ดังต่อไปนี้

อ้างอิงจาก * จะเห็นว่า $P(\delta_x = 1) = P(\delta_y = 1) = 0.07$ และสัดส่วนของการสูญหายระหว่าง x และ y เท่ากับ 2 (R = 2)

จาก $P(\delta_y = 1 | \delta_x = 0) = \frac{P(\delta_y = 1)}{[\mathrm{R} \times P(\delta_x = 1)] + [1 - P(\delta_x = 1)]}$

$$= \frac{0.07}{[2\times 0.07]+[1-0.07]} = \frac{7}{107}$$

จะได้ว่า $P(\delta_y = 1 | \delta_x = 1) = \text{R} \times P(\delta_y = 1 | \delta_x = 0) = 2 \times \frac{7}{107} = \frac{14}{107}$

ดังนั้น $P(\delta_y = 1 | \delta_x = 1) = \frac{14}{107}$ และ $P(\delta_y = 1 | \delta_x = 0) = \frac{7}{107}$

ส่วนความน่าจะเป็นในตำแหน่งอื่นๆ สามารถคำนวณหาได้ในทำนองเดียวกัน ผลที่ได้เป็นดังตารางต่อไปนี้

ความน่าจะเป็นของการสูญหายโดยเฉลี่ย	ระดับการสูญหายแบบ Nonignorable	ความน่าจะเป็นของการสูญหายในแต่ละช่วง		
		ช่วงต้น $P(\delta_y=1\|\delta_x=1), P(\delta_y=1\|\delta_x=0)$	ช่วงกลาง $P(\delta_y=1\|\delta_x=1), P(\delta_y=1\|\delta_x=0)$	ช่วงปลาย $P(\delta_y=1\|\delta_x=1), P(\delta_y=1\|\delta_x=0)$
0.1	ไม่มี	$\frac{1}{10}, \frac{1}{10}$	$\frac{1}{10}, \frac{1}{10}$	$\frac{1}{10}, \frac{1}{10}$
	ปานกลาง	$\frac{14}{107}, \frac{7}{107}$ *	$\frac{2}{11}, \frac{1}{11}$	$\frac{26}{113}, \frac{13}{113}$
	สูง	$\frac{4}{28}, \frac{1}{28}$	$\frac{4}{13}, \frac{1}{13}$	$\frac{16}{37}, \frac{4}{37}$
0.2	ไม่มี	$\frac{2}{10}, \frac{2}{10}$	$\frac{2}{10}, \frac{2}{10}$	$\frac{2}{10}, \frac{2}{10}$
	ปานกลาง	$\frac{14}{57}, \frac{7}{57}$	$\frac{2}{6}, \frac{1}{6}$	$\frac{26}{63}, \frac{13}{63}$
	สูง	$\frac{8}{31}, \frac{2}{31}$	$\frac{4}{8}, \frac{1}{8}$	$\frac{32}{49}, \frac{8}{49}$
0.3	ไม่มี	$\frac{3}{10}, \frac{3}{10}$	$\frac{3}{10}, \frac{3}{10}$	$\frac{3}{10}, \frac{3}{10}$
	ปานกลาง	$\frac{42}{121}, \frac{21}{121}$	$\frac{6}{13}, \frac{3}{13}$	$\frac{78}{139}, \frac{39}{139}$
	สูง	$\frac{12}{34}, \frac{3}{34}$	$\frac{12}{19}, \frac{3}{19}$	$\frac{48}{61}, \frac{12}{61}$

10. การศึกษาในครั้งนี้จะทำการจำลองข้อมูลภายใต้สถานการณ์ต่างๆ ที่เป็นไปตามเงื่อนไขข้างต้นที่แตกต่างกันโดยใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์เป็นจำนวน 5,000 รอบ

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีประมาณค่าสูญหายวิธีใดที่ให้ค่าประมาณใกล้เคียงกับค่าจริงมากที่สุดนั้นจะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์ (Average mean square error : AMSE) วิธีการที่ให้ค่า AMSE ต่ำสุดจะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด และจะใช้ค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiency : RE) ซึ่งเป็นอัตราส่วนระหว่างค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีการ EM กับวิธีการประมาณค่าสูญหายแบบอื่นๆ มาใช้ในการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการให้มีความชัดเจนมากยิ่งขึ้น โดยสามารถคำนวณได้จากสูตรดังต่อไปนี้

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

$$AMSE = \frac{1}{5,000}\sum_{q=1}^{5,000} MSE_q$$

$$RE = \frac{AMSE_{EM}}{AMSE_k} \qquad ;k = KNN, \text{PMM}$$

เมื่อ y_i' แทน ค่าจริงของข้อมูลตัวแปรตามตัวที่ i

$\hat{y}_{qi}$ แทน ค่าพยากรณ์ของข้อมูลตัวแปรตามตัวที่ i จากการทำซ้ำรอบที่ q

MSE_q แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำรอบที่ q

$AMSE$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำทั้งหมด 5,000 รอบ

RE แทน ค่าประสิทธิภาพสัมพัทธ์

$AMSE_{EM}$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการประมาณค่าสูญหายด้วยวิธีการ EM

$AMSE_k$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการประมาณค่าสูญหายด้วยวิธีการ KNN และ PMM

1.7 วิธีดำเนินการวิจัย

1. สร้างข้อมูลตัวแปรอิสระที่มีการแจกแจงตามที่กำหนด
2. สร้างข้อมูลความคลาดเคลื่อน (ε) ที่มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ σ^2 โดยจะกำหนดให้ σ = 10 , 30 และ 90
3. สร้างข้อมูลตัวแปรตามที่มาจากรูปแบบความสัมพันธ์ของการถดถอยเชิงเส้นพหุ $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ โดยกำหนดให้ $\beta_0 = 42$ และ $\beta_1 = \beta_2 = \beta_3 = 1$
4. ในแต่ละสถานการณ์จะสร้างข้อมูลตัวแปรอิสระที่ทำให้เกิดการสูญหาย โดยนำข้อมูลเหล่านั้นมาแบ่งเป็น 3 ช่วงโดยให้แต่ละช่วงมีสัดส่วนเท่าๆ กัน จากนั้นจะทำให้แต่ละช่วงของตัวแปรอิสระมีการสูญหาย โดยจะสร้างตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่ จำนวน 3 ช่วง ซึ่งมีขนาดเท่ากับจำนวนตัวแปรอิสระที่อยู่ในแต่ละช่วงโดยจะมีความน่าจะเป็นของการสูญหายที่แตกต่างกันไปตามที่กำหนด แล้วจะจับคู่ ข้อมูลตัวแปรอิสระกับ ข้อมูลตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่
5. ถ้าหากตัวแปรอิสระเกิดการสูญหาย $(P(\delta_x = 1))$ จะทำให้เกิดการสูญหายของตัวแปรตามด้วยความน่าจะเป็น $P(\delta_y = 1 \mid \delta_x = 1)$ หรือในทำนองเดียวกันถ้าหากตัวแปรอิสระไม่เกิดการสูญหาย $(P(\delta_x = 0))$ แล้วจะส่งผลให้ตัวแปรตามเกิดการสูญหายด้วยความน่าจะเป็น $P(\delta_y = 1 \mid \delta_x = 0)$ โดยที่ค่าความน่าจะเป็นของการสูญหายในแต่ละช่วงจะแตกต่างกันไปตามที่กำหนด

   จะสังเกตได้ว่าถ้าหากตัวแปรอิสระเกิดการสูญหายแล้ว ความน่าจะเป็นที่ตัวแปรตามเกิดการสูญหายจะมากกว่าในกรณีที่ตัวแปรอิสระไม่สูญหาย
6. ประมาณค่าข้อมูลเพื่อแทนที่ข้อมูลที่สูญหายในตัวแปรอิสระและตัวแปรตามด้วย วิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM)

7. ประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method : OLS)
8. สร้างสมการถดถอยเชิงเส้นพหุจากค่าสัมประสิทธิ์การถดถอยเพื่อใช้ในการพยากรณ์
9. คำนวณหาค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี
10. สรุปผลการวิจัยที่ได้ในแต่ละสถานการณ์

1.8 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการตัดสินใจเลือกวิธีการประมาณค่าสูญหายของตัวแปรอิสระและตัวแปรตามที่มีความสัมพันธ์กันในสมการถดถอยเชิงเส้นพหุ
2. เพื่อเป็นแนวทางในการศึกษาวิธีประมาณค่าสูญหายในรูปแบบอื่นๆ ต่อไป

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงประเภทของข้อมูลที่สูญหาย แนวทางการแก้ไขปัญหาข้อมูลสูญหายทั่วไป ตลอดจนวิธีการประมาณค่าสูญหายที่ได้รับความนิยม และวิธีการประมาณค่าสูญหายที่เลือกมาใช้ในงานวิจัยชิ้นนี้เมื่อข้อมูลตัวแปรอิสระและตัวแปรตามเกิดการสูญหาย

ลักษณะการเกิดข้อมูลสูญหายมี 3 ประเภท ดังต่อไปนี้

1. Missing Completely at Random (MCAR) คือความน่าจะเป็นที่หน่วยตัวอย่างของตัวแปรจะเกิดการสูญหาย มีค่าเท่าๆ กัน หรือลักษณะของข้อมูลสูญหายที่เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด นั่นคือข้อมูลที่สูญหายเป็นอิสระจากตัวแปรต่าง ๆ สามารถทำการตรวจสอบลักษณะของข้อมูลสูญหายกลุ่มนี้โดยการแบ่งกลุ่มของค่าสังเกตเป็นกลุ่มข้อมูลปกติและข้อมูลสูญหาย ในกรณีนี้เมื่อทำการทดสอบจะไม่พบความแตกต่างอย่างมีนัยสำคัญระหว่างทั้งสองกลุ่มสำหรับตัวแปรต่างๆในฐานข้อมูล สำหรับสาเหตุที่ทำให้ข้อมูลเกิดการสูญหายมีอยู่หลากหลายเหตุผลอาจเกิดขึ้นเนื่องจากเครื่องมือเสีย อุปกรณ์เกิดข้อบกพร่อง สภาพอากาศเลวร้าย กลุ่มเป้าหมายที่ศึกษาล้มป่วยหรือการนำเข้าข้อมูลไม่ถูกต้อง สำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นสามารถเลือกทำการวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้ เช่น สมมติให้ตัวแปรที่ศึกษามี 2 ตัวแปรคือ รายได้ และอายุงาน ในกรณีนี้ความน่าจะเป็นที่รายได้จะสูญหายมีค่าเท่าๆกันทุกหน่วยตัวอย่าง ไม่ว่าอายุงานและรายได้จะเป็นอย่างไร
2. Missing at Random (MAR) คือความน่าจะเป็นที่หน่วยตัวอย่างของตัวแปรหนึ่งจะเกิดกรสูญหายโดยขึ้นกับอีกตัวแปรหนึ่ง แต่ไม่ขึ้นอยู่กับตัวแปรที่มีค่าสูญหายนั้นๆ หรืออาจกล่าวได้ว่าเป็นลักษณะของข้อมูลสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่มจากค่าสังเกตทั้งหมด แต่เกิดขึ้นอย่างสุ่มภายในบางส่วนหรือบางกลุ่มของค่าสังเกต นั่นคือ

ค่าของข้อมูลสูญหายขึ้นอยู่กับตัวแปรตัวอื่น ๆ ในฐานข้อมูลซึ่งไม่ได้เป็นตัวแปรที่เกิดข้อมูลสูญหาย ยกตัวอย่างเช่น หากพบว่าเฉพาะกลุ่มผู้ได้รับการศึกษาน้อยที่ไม่ให้ความร่วมมือในการตอบข้อคำถามเกี่ยวกับทัศนคติในการเสพยาเสพติด ในลักษณะนี้สามารถกล่าวได้ว่าข้อมูลทัศนคติในการเสพยาเสพติดมีค่าสูญหายแบบ MAR ทั้งนี้เนื่องจากเป็นค่าสูญหายที่เกิดขึ้นเฉพาะในบางส่วนของตัวแปรระดับการศึกษา สำหรับข้อมูลสูญหายประเภทนี้ยังไม่ส่งผลกระทบรุนแรงเท่ากับข้อมูลสูญหายในประเภท NMAR

3. Not missing at random (NMAR) คือความน่าจะเป็นที่หน่วยตัวอย่างของตัวแปรหนึ่งจะเกิดการสูญหายขึ้นอยู่กับตัวแปรนั้นๆ ที่มีลักษณะของหน่วยตัวอย่างเหมือนกัน หรืออาจกล่าวได้ว่าเป็นการสูญหายซึ่งไม่ได้เกิดขึ้นอย่างสุ่ม โดยค่าของข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลสมบูรณ์ในตัวแปรเดียวกัน รวมถึงตัวแปรตัวอื่นด้วย เช่น หากข้อมูลสูญหายของระดับรายได้ขึ้นอยู่ระดับรายได้ในแต่ละช่วงอายุ ข้อมูลสูญหายที่เกิดขึ้นจัดอยู่ในประเภท NMAR หรือในบางกรณีค่าของข้อมูลสูญหายอาจไม่ขึ้นอยู่กับตัวแปรใด ๆ ในฐานข้อมูลเลย แต่ขึ้นอยู่กับตัวแปรอื่นที่ไม่ได้ถูกเก็บรวบรวมไว้ในการศึกษาครั้งนั้น เช่น ค่าน้ำหนักตัวที่ลดลงขึ้นอยู่กับน้ำหนักตัวตอนเริ่มต้น แต่เนื่องจากตัวแปรน้ำหนักตอนเริ่มต้นไม่ได้ถูกรวบรวมไว้ในฐานข้อมูล ดังนั้นค่าสูญหายของน้ำหนักตัวที่ลดลงจึงขึ้นอยู่กับตัวแปรภายนอกฐานข้อมูล

ลักษณะข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลสูญหายที่สามารถส่งผลกระทบอย่างรุนแรงในการวิเคราะห์ข้อมูล ซึ่งการสูญหายแบบ Nonignorable ที่เลือกมาศึกษาในงานวิจัยชิ้นนี้ก็ถูกจัดอยู่ในประเภท NMAR

ในทางปฏิบัติ ลักษณะของข้อมูลสูญหายประเภท MCAR มักไม่พบบ่อยนัก ที่พบบ่อยครั้งมักเป็นข้อมูลสูญหายประเภท MAR ดังนั้นวิธีการทางสถิติต่าง ๆ ที่พัฒนาขึ้นมาเพื่อแก้ปัญหาข้อมูลสูญหาย มักดำเนินการภายใต้ข้อสมมติของ MAR เป็นส่วนใหญ่ ดังนั้นในงานวิจัยนี้จึงได้ทำการประยุกต์ใช้วิธีการต่างๆ เพื่อให้สอดคล้องกับการประมาณค่าสูญหายแบบ NMAR

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี (2550) ได้เสนอแนวทางการแก้ไขปัญหาข้อมูลสูญหายทั่วๆไปว่าการจัดการกับข้อมูลสูญหายมีหลายวิธีการให้เลือกใช้ การพิจารณาเลือกใช้วิธีการใดขึ้นอยู่กับลักษณะของข้อมูลสูญหายที่เกิดขึ้น หากเลือกวิธีการที่ไม่เหมาะสมมาใช้อาจเป็นการเพิ่มค่าความคลาดเคลื่อนและทำลายผลลัพธ์ที่ควรจะได้ สำหรับวิธีการจัดการกับข้อมูลสูญหายที่มักถูกเลือกนำมาใช้มีดังนี้

- Listwise data deletion: เป็นวิธีการจัดการกับข้อมูลสูญหายที่ง่ายมาก นั่นคือไม่สนใจข้อมูลสูญหายที่เกิดขึ้น โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลเฉพาะส่วนที่สมบูรณ์ แนวทางนี้จะมีความเหมาะสมในกรณีที่ข้อมูลสูญหายมีจำนวนน้อยมาก และ/หรือผลจากการวิเคราะห์มีความชัดเจนมาก ซึ่งวิธีการนี้มักถูกกำหนดให้ใช้เป็นหลัก (by default) สำหรับจัดการกับข้อมูลสูญหายในโปรแกรมคอมพิวเตอร์ทางสถิติทั่วๆไป หากไม่เจาะจงเลือกใช้วิธีการอื่นในการจัดการกับข้อมูลสูญหาย

- Pairwise data deletion: เป็นวิธีการจัดการกับข้อมูลสูญหายสำหรับกรณีที่ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคู่ โดยจะทำการวิเคราะห์ข้อมูลจากข้อมูลส่วนที่มีค่าสมบูรณ์ทั้งสองตัวแปร

- Mean substitution: เป็นวิธีการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่าในแต่ละกลุ่มย่อยของตัวแปรอื่น ซึ่งเป็นวิธีที่พัฒนามาจากการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่า ทั้งนี้เนื่องจากข้อสมมติที่ว่าค่าของข้อมูลสูญหายควรจะต้องขึ้นอยู่กับลักษณะของหน่วยตัวอย่าง โดยลักษณะของหน่วยตัวอย่างที่ใกล้เคียงกันควรจะมีค่าข้อมูลที่สนใจคล้ายคลึงกัน

- Regression method: ทำการสร้างสมการถดถอยระหว่างตัวแปรใด ๆ ที่ต้องการจากข้อมูลที่สมบูรณ์โดยกำหนดให้ตัวแปรตามเป็นตัวแปรที่มีข้อมูลไม่สมบูรณ์ จากนั้นใช้สมการถดถอยที่ได้ทำการประมาณค่าของข้อมูลที่ไม่สมบูรณ์

- Hot deck imputation: เป็นวิธีการพิจารณาเลือกหน่วยตัวอย่างที่มีลักษณะคล้ายคลึงกันมากที่สุดกับหน่วยตัวอย่างที่เกิดค่าสูญหาย จากนั้นแทนค่าที่สูญหายด้วยค่าของหน่วยตัวอย่างที่คล้ายคลึงนั้น

- Expectation Maximization (EM) approach: วิธีการนี้เป็นการอาศัยหลักของกระบวนการวนซ้ำ (iterative procedure) ระหว่าง 2 ขั้นตอน โดยขั้นตอนแรก เป็นขั้นตอนที่เรียกว่า Expectation (E) step ซึ่งจะทำการประมาณค่าคาดหวังจากฟังก์ชัน likelihood ภายใต้ข้อมูลที่สมบูรณ์ สำหรับขั้นตอนที่สอง เป็นขั้นตอนที่เรียกว่า Maximization (M) step เพื่อทำการ

แทนค่าคาดหวังของข้อมูลสูญหายด้วยค่าที่ได้จาก E step และทำการประมาณค่าคาดหวังจากฟังก์ชัน likelihood ในกรณีถ้าไม่เกิดข้อมูลสูญหาย โดยจะทำการวนซ้ำระหว่าง 2 ขั้นตอนจนกว่าจะเกิดค่าที่ลู่เข้า (convergence) หรือค่าที่มีการเปลี่ยนแปลงน้อยมาก ใช้ค่านั้นแทนค่าข้อมูลสูญหายที่เกิดขึ้น

- Raw maximum likelihood methods: เป็นวิธีการที่อาศัยข้อมูลสมบูรณ์ในการสร้างค่า maximumlikelihood ภายใต้ตัวแบบทางสถิติที่เหมาะสม ไม่ว่าจะเป็น structural equation model, regression model, ANOVA และ ANCOVA models

- Multiple imputation (MI): เป็นวิธีการที่ผสมผสานระหว่างวิธีการ EM และ Raw maximumlikelihood methods ร่วมกับความสามารถของคุณสมบัติ hot deck เพื่อทำการสร้างชุดจำลองของข้อมูลที่ได้ทำการแทนค่าข้อมูลสูญหายด้วย imputed value แล้วขึ้นมาหลาย ๆ ชุด (ประมาณ 5 ถึง 10 ชุด) จากนั้นทำการวิเคราะห์ข้อมูลจากชุดต่าง ๆ บันทึกผลการวิเคราะห์ที่ได้ โดยผลการวิเคราะห์ที่ได้เหล่านี้จะถูกรวมเข้าด้วยกันเพื่อทำการสรุปผลการศึกษา

วิธีการประมาณข้อมูลสูญหายที่ถูกนำมาใช้ในการเปรียบเทียบประสิทธิภาพในงานวิจัยชิ้นนี้ ได้แก่ วิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และ วิธี Predictive Mean Matching Imputation (PMM) ส่วนวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุ ซึ่งก็คือวิธีกำลังสองน้อยที่สุดแบบสามัญ (Ordinary Least Squares Method : OLS) และเนื่องจากงานวิจัยนี้ กำหนดให้ตัวแปรอิสระและตัวแปรตามมีการสูญหายอย่างมีความสัมพันธ์กันภายใต้ความน่าจะเป็นแบบมีเงื่อนไข ซึ่งมีรายละเอียดดังต่อไปนี้

2.1 **ความน่าจะเป็นแบบมีเงื่อนไข** (Conditional Probability)

ถ้าหากมีเหตุการณ์ 2 เหตุการณ์ซึ่งมีความสัมพันธ์กัน ความน่าจะเป็นของเหตุการณ์หนึ่งจะมากหรือน้อยจะขึ้นอยู่กับว่าเหตุการณ์อีกเหตุการณ์หนึ่งจะเกิดขึ้นหรือไม่ โดยกำหนดให้เหตุการณ์ A และ B จะใช้สัญลักษณ์ $P(A|B)$ หมายความว่า ความน่าจะเป็นของการเกิดเหตุการณ์ A เมื่อ B ได้เกิดขึ้นแล้ว สามารถคำนวณหาได้จาก

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \qquad \text{โดยที่ } P(B)\neq 0$$

แต่ถ้าเหตุการณ์ A และ B เกิดขึ้นพร้อมกันแล้วทฤษฎีการคูณ จะช่วยในการหาความน่าจะเป็นของเหตุการณ์ดังกล่าว จะได้ว่า $P(A\cap B)=P(B)\cdot P(A|B)$

ในทำนองเดียวกันถ้าหากมีเหตุการณ์ $A_1,A_2,...,A_n$ เกิดขึ้นพร้อมกัน จะได้ความน่าจะเป็นของเหตุการณ์ดังกล่าวคือ

$$P(A_1\cap A_2\cap...\cap A_n)=P(A_1)\cdot P(A_2|A_1)\cdot P(A_3|A_1,A_2)\cdot...\cdot P(A_n|A_1,A_2,...,A_{n-1})$$

ดังนั้น ถ้าหากเราต้องการหาความน่าจะเป็นที่จะเกิดเหตุการณ์หนึ่งซึ่งขึ้นอยู่กับอีก 2 เหตุการณ์ จะใช้ความน่าจะเป็นแบบมาร์จินัล (Marginal Probability) จะได้ว่า

$$P(A)=P(A\cap B_1)+P(A\cap B_2)$$
$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)$$

2.2 **วิธีกำลังสองน้อยที่สุดแบบสามัญ**(Ordinary Least Squares Method: OLS)

ในการหาสัมประสิทธิ์การถดถอยของตัวแบบความถดถอยเชิงเส้นนั้น วิธีที่ได้รับความนิยมมากที่สุดคือวิธีกำลังสองน้อยที่สุดแบบสามัญ (OLS) โดยจะหาค่าสัมประสิทธิ์ความถดถอยที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อน (SSE) มีค่าน้อยที่สุด

จากความสัมพันธ์ระหว่างตัวแปรตาม (y) และตัวแปรอิสระ (x) จะได้สมการถดถอยเชิงพหุที่แสดงความสัมพันธ์ คือ

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}+\varepsilon_i \qquad ;i=1,2,...,m,m+1,...,n$$

หรือจะเขียนในรูปเมทริกซ์ได้คือ

$$\tilde{y}=X\,\tilde{\beta}+\tilde{\varepsilon}$$

เมื่อ

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

จากวิธี OLS ในการหาค่าสัมประสิทธิ์ความถดถอยที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อน (Sum Square of Error : SSE) มีค่าน้อยที่สุด มีวิธีการดังต่อไปนี้

กำหนดตัวประมาณ $\tilde{\beta}$ คือ $\tilde{b}$

จาก $SSE = \tilde{\varepsilon}'\tilde{\varepsilon}$

$$= (\tilde{y} - X\tilde{b})'(\tilde{y} - X\tilde{b})$$
$$= \tilde{y}'\tilde{y} - \tilde{y}'X\tilde{b} - \tilde{b}'X\tilde{y} + \tilde{b}'X'X\tilde{b}$$
$$= \tilde{y}'\tilde{y} - 2\tilde{b}'X\tilde{y} + \tilde{b}'X'X\tilde{b}$$

จะหาอนุพันธ์ (Differentiate) เทียบกับ $\tilde{b}$ แล้วกำหนดให้เท่ากับ $\tilde{0}$

$$\frac{\partial}{\partial\tilde{b}}(\tilde{y}'\tilde{y} - 2\tilde{b}'X\tilde{y} + \tilde{b}'X'X\tilde{b}) \overset{set}{=} \tilde{0}$$

$$-2X\tilde{y} + 2X'X\tilde{b} = \tilde{0}$$

$$(X'X)\tilde{b} = X\tilde{y} \qquad \text{โดยที่ } (X'X) \neq 0$$

นั่นคือ $$\tilde{b} = (X'X)^{-1}X\tilde{y}$$

2.3 วิธีการประมาณค่าสูญหายโดยวิธี EM Algorithm

การประมาณค่าสูญหายด้วยวิธี EM Algorithm (Expectation Maximization) (Dempster Laird and Rubin, 1977) วิธีการนี้เป็นวิธีการที่ได้รับความนิยมในงานวิจัยเป็นจำนวนมาก ยกตัวอย่างเช่น

ในงานวิจัยของ อุษณีย์ วงศ์อามาตย์ (2555) ที่ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงพหุ ซึ่งวิธีการ EM ก็เป็นหนึ่งในวิธีการที่เลือกมาเปรียบเทียบในงานวิจัยดังกล่าว ผลการวิจัยพบว่าโดยรวมแล้ววิธี EM เป็นวิธีการประมาณค่าที่ดีที่สุดในกรณีส่วนใหญ่ โดยเฉพาะกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนไม่สูง (10-30) นอกจากนี้ยังมีงานวิจัยที่นำเสนอโดย เพียงออ ยีสา (2551) ที่ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น ซึ่งพบว่าวิธีการ EM ให้ผลที่มีประสิทธิภาพมากกว่าวิธีอื่นๆ

วิธีการ EM จะเป็นการหาค่าประมาณภาวะน่าจะเป็นสูงสุดของพารามิเตอร์โดยกระบวนการวนซ้ำ ซึ่งการประมาณค่าสูญหายด้วยวิธี EM Algorithm แบ่งออกเป็น 2 ขั้นตอน คือ

<u>ขั้นตอนที่ 1</u> E-Step (Expectation Step) เป็นขั้นตอนที่หาค่าคาดหวังของค่าที่สูญหายไปภายใต้เงื่อนไขชุดข้อมูลที่ไม่สูญหายและพารามิเตอร์ตัวปัจจุบัน ค่าที่ได้นี้จะนำไปประมาณค่าที่สูญหาย

<u>ขั้นตอนที่ 2</u> M-Step (Maximization Step) เป็นขั้นตอนที่ประมาณค่าภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ ด้วยการแทนค่าสูญหายที่ได้จากขั้นตอนที่ 1 ทำจนกระทั่งได้ตัวพารามิเตอร์ที่คงที่ นั่นคือตัวประมาณภาวะน่าจะเป็นสูงสุด

Little and Rubin (1987) ได้ประยุกต์วิธีการของ EM มาใช้ในการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุ ซึ่งลักษณะที่สนใจคือประมาณค่าสูญหายของตัวแปรอิสระและตัวแปรตามตาม โดยมีขั้นตอนดังนี้

1. สมมติว่ามีข้อมูลดังนี้

$$\begin{bmatrix} \tilde{y}_0 \\ \tilde{y}_1 \end{bmatrix} = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \tilde{\beta} + \tilde{\varepsilon}$$

เมื่อ $\tilde{y}_0$ แทน เวกเตอร์ของตัวแปรตามที่ทราบค่าขนาด $m \times 1$

$\tilde{y}_1$ แทน เวกเตอร์ของตัวแปรตามที่มีการสูญหายขนาด $(n-m) \times 1$

X_0 แทน เมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามที่ทราบค่า ขนาด $m \times (p+1)$

X_1 แทน เมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามที่มีการสูญหาย ขนาด $(n-m) \times (p+1)$

$\tilde{\beta}$ แทน เวกเตอร์ของพารามิเตอร์

2. ประมาณค่าสัมประสิทธิ์การถดถอยค่าเริ่มต้น ($\hat{\tilde{\beta}}^{(0)}$) ด้วยวิธีกำลังสองน้อยที่สุด (OLS) จากชุดข้อมูลที่สมบูรณ์ โดยจะเรียกว่าว่าสัมประสิทธิ์การถดถอยรอบที่ 0

$$\hat{\tilde{\beta}}^{(0)} = (X_0' X_0)^{-1} X_0' \tilde{y}_0$$

3. เมื่อได้ $\hat{\tilde{\beta}}^{(0)}$ แล้วจะทำให้เราสามารถหาค่าประมาณสูญหายได้จากการหาค่าคาดหวัง

ขั้นตอนนี้คือ E-Step เราจะหาค่าคาดหวังรอบที่ 1 ตามกรณีต่อไปนี้

<u>กรณีที่ 1</u> ตัวแปรอิสระทุกตัวทราบค่า แต่ตัวแปรตาม (y) เกิดการสูญหาย

$$E(y_i \mid \tilde{y}_0, X_0, \hat{\tilde{\beta}}^{(0)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \hat{\beta}_2^{(0)} x_{i2} + \hat{\beta}_3^{(0)} x_{i3} & ;i = m+1,...,n \end{cases}$$

ดังนั้นจะได้ $y_i^{(1)} = E(y_i \mid \tilde{y}_0, X_0, \hat{\tilde{\beta}}^{(0)})$

<u>กรณีที่ 2</u> ตัวแปรอิสระ $(x_{ip};\ p = 1,2,3)$ ตัวใดตัวหนึ่งเกิดการสูญหาย แต่ตัวแปรตาม (y) ทราบค่า

$$E(x_{ip}) = \overline{x}_{ip} \qquad ;i = 1,2,...,m$$

$$;p = 1,2,3$$

ดังนั้นจะได้ $x_{ip}^{(1)} = E(x_{ip}) = \overline{x}_{ip}$

<u>กรณีที่ 3</u> ตัวแปรอิสระ $(x_{ip};\ p = 1,2,3)$ ตัวใดตัวหนึ่งเกิดการสูญหาย และตัวแปรตาม (y) เกิดการสูญหาย เพื่อให้เกิดความเข้าใจยิ่งขึ้น จะแบ่งออกเป็น 3 กรณี ตามลักษณะของการสูญหายที่ตัวแปรอิสระ

- เกิดการสูญหายที่ตัวแปรอิสระ x_{1i}

$$E(y_i \mid \tilde{y}_0, \overline{x}_{i1}, x_{i2}, x_{i3}, \hat{\tilde{\beta}}^{(0)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \beta_0^{(0)} + \beta_1^{(0)} \overline{x}_{i1} + \beta_2^{(0)} x_{i2} + \beta_3^{(0)} x_{i3} & ;i = m+1,...,n \end{cases}$$

- เกิดการสูญหายที่ตัวแปรอิสระ x_{2i}

$$E(y_i \mid \tilde{y}_0, x_{i1}, \overline{x}_{i2}, x_{i3}, \hat{\tilde{\beta}}^{(0)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \beta_0^{(0)} + \beta_1^{(0)} x_{i1} + \beta_2^{(0)} \overline{x}_{i2} + \beta_3^{(0)} x_{i3} & ;i = m+1,...,n \end{cases}$$

- เกิดการสูญหายที่ตัวแปรอิสระ x_{3i}

$$E(y_i \mid \tilde{y}_0, x_{i1}, x_{i2}, \overline{x}_{i3}, \hat{\tilde{\beta}}^{(0)}) = \begin{cases} y_i & ;i = 1,2,...,m \\ \beta_0^{(0)} + \beta_1^{(0)} x_{i1} + \beta_2^{(0)} x_{i2} + \beta_3^{(0)} \overline{x}_{i3} & ;i = m+1,...,n \end{cases}$$

ดังนั้นจะได้ $y_i^{(1)} = E(y_i \mid \tilde{y}_0, \overline{x}_{i1}, x_{i2}, x_{i3}, \hat{\tilde{\beta}}^{(0)})$ และ $x_{i1}^{(1)} = \overline{x}_{i1}$

$y_i^{(1)} = E(y_i \mid \tilde{y}_0, x_{i1}, \overline{x}_{i2}, x_{i3}, \hat{\tilde{\beta}}^{(0)})$ และ $x_{i2}^{(1)} = \overline{x}_{i2}$

$y_i^{(1)} = E(y_i \mid \tilde{y}_0, x_{i1}, x_{i2}, \overline{x}_{i3}, \hat{\tilde{\beta}}^{(0)})$ และ $x_{i3}^{(1)} = \overline{x}_{i3}$ ตามลำดับ

1. เมื่อหาค่าคาดหวังสำหรับทุกกรณีได้แล้ว จะเข้าสู่ขั้นตอน M-Step ในการทำซ้ำรอบที่ 1

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

2. หาค่าสัมบูรณ์ผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ 0 กับค่าสัมประสิทธิ์การถดถอยรอบที่ 1 ของค่าสัมประสิทธิ์ทุกๆค่า $(|\hat{\tilde{\beta}}^{(0)} - \hat{\tilde{\beta}}^{(1)}|)$

3. ถ้าหาก $|\hat{\tilde{\beta}}^{(0)} - \hat{\tilde{\beta}}^{(1)}| < 0.001$ จะได้ค่าประมาณของข้อมูลที่สูญหายรอบที่ 1 แต่ถ้า $|\hat{\tilde{\beta}}^{(0)} - \hat{\tilde{\beta}}^{(1)}| > 0.001$ แสดงว่าตัวประมาณที่ได้มายังไม่ดีพอ จึงต้องทำการประมาณรอบใหม่โดยการวนซ้ำใหม่ ในขั้นตอนต่อไป

4. เข้าสู่ขั้นตอน E-Step ในการหาค่าคาดหวังรอบที่ t โดยที่ t =2,3,… ทำการหาค่าคาดหวังในแต่ละกรณีในลักษณะเช่นเดียวกับขั้นตอนที่ 3 แต่จะใช้สัมประสิทธิ์การถดถอยรอบที่ 1 ที่เราประมาณได้จากขั้นตอนที่ 4

5. เมื่อหาค่าคาดหวังได้จากขั้นตอนที่ 7 แล้ว จะเข้าสู่ขั้นตอน M-Step ในรอบที่ t โดยที่ t=2,3,…

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

9. หาค่าสัมบูรณ์ผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ t-1 กับค่าสัมประสิทธิ์การถดถอยรอบที่ t ของค่าสัมประสิทธิ์ทุกๆค่า $(|\hat{\tilde{\beta}}^{(t-1)} - \hat{\tilde{\beta}}^{(t)}|)$

10. ถ้าหาก $|\hat{\tilde{\beta}}^{(t-1)} - \hat{\tilde{\beta}}^{(t)}| < 0.001$ จะได้ค่าประมาณของข้อมูลที่สูญหายรอบที่ t โดยที่ t=2,3,… แต่ถ้า $|\hat{\tilde{\beta}}^{(t-1)} - \hat{\tilde{\beta}}^{(t)}| > 0.001$ จะต้องทำการประมาณรอบใหม่โดยทำตามขั้นตอนที่ 7-9 ไปเรื่อยๆ จนกว่าค่าสัมบูรณ์ผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยน้อยกว่า 0.001 จึงจะได้ค่าประมาณของข้อมูลสูญหายที่แท้จริง

11. นำค่าสูญหายที่ประมาณได้ แทนค่าสูญหายที่หายไป แล้วทำการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด (OLS) จะได้สมการถดถอยใหม่มาใช้ในการพยากรณ์

2.4 วิธีการประมาณค่าสูญหายแบบ K-Nearest Neighbor Imputation (KNN)

วิธี K-Nearest Neighbor (KNN) (James E.S. Macleod, 1987) เป็นเทคนิคหนึ่งที่ถูกนำมาประยุกต์ใช้โดยการหาความสัมพันธ์ระหว่างกลุ่มข้อมูลที่จะนำมาประมาณค่าที่สูญหาย ซึ่งเป็นวิธีการประมาณที่จัดอยู่ใน Hot Deck Method ที่จะแทนค่าที่สูญหายด้วยค่าสังเกตที่ทราบค่าซึ่งวิธีการนี้ ค่อนค้างที่จะมีประสิทธิภาพมากกว่า Hot Deck Method วิธีอื่นๆ วิธีการนี้จะประมาณค่าโดยใช้ค่าที่ใกล้ที่สุด โดยพิจารณาเลือกหน่วยตัวอย่าง K ชุดจากข้อมูล (x_i, y_i) ที่ทราบค่าซึ่งมีลักษณะคล้ายคลึงกับหน่วยตัวอย่างที่เกิดค่าสูญหายมากที่สุด โดยจะกำหนดให้ $K \approx \sqrt{m}$ (อ้างถึง จาก Josson และ Wohlin, 2006) ซึ่ง K จะเป็นจำนวนคี่ที่มีค่าใกล้เคียงกับ $\sqrt{m}$ มากที่สุด เมื่อ m เป็นจำนวนข้อมูลที่สมบูรณ์ จากนั้นแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของหน่วยตัวอย่างที่คล้ายกัน นอกจากนี้วิธีการ KNN ยังเป็นวิธีการที่ได้รับความนิยมในงานวิจัยต่างๆมากมาย ยกตัวอย่างเช่น

อุษณีย์ วงศ์อามาตย์ (2555) ที่ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงพหุ ซึ่งมีการใช้วิธี KNN ในการเปรียบเทียบประสิทธิภาพกับวิธีการอื่นๆ พบว่าโดยส่วนใหญ่วิธีการ KNN ประมาณค่าได้ดีที่สุดเมื่อข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของค่าความคลาดเคลื่อนมีค่าสูง (90)

ไก้รุ่ง เฮงพระพรหม ศจีมาจ ณ วิเชียร และพยุง มีสัจ (2553) ได้นำเสนองานวิจัยเรื่อง การแทนค่าข้อมูลที่ขาดหายด้วยขั้นตอนวิธีเชิงพันธุกรรมสำหรับข้อมูลมะเร็งลำไส้ใหญ่ Missing Value Imputation Using GA for Colon Cancer ซึ่งในงานวิจัยดังกล่าวได้มีการนำวิธีการ KNN มาเปรียบเทียบประสิทธิภาพในการแทนที่ข้อมูลที่สูญหาย ซึ่งกล่าวว่าการเก็บรักษาข้อมูลที่ไม่สมบูรณ์เป็นขั้นตอนสำคัญในกระบวนการเตรียมข้อมูลก่อนการวิเคราะห์ข้อมูลทางสถิติข้อมูลไมโครอาร์เรย์ปกติมักประกอบด้วยตัวอย่างการทดลองน้อยแต่มีมิติสูง และประกอบด้วยค่าข้อมูลที่ขาดหายเป็นจำนวนมาก จึงทำให้ประสิทธิภาพในการวิเคราะห์ข้อมูลลดลง โดยอัลกอริทึมในการ

วิเคราะห์ข้อมูลทางสถิติส่วนใหญ่ต้องการข้อมูลที่สมบูรณ์สำหรับใช้ในการวิเคราะห์ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการแทนค่าสำหรับข้อมูลไมโครอาร์เรย์ด้วยวิธีการเลือกคุณลักษณะสำหรับข้อมูลไมโครอาร์เรย์ด้วยวิธีการขั้นตอนวิธีพันธุกรรม (Genetic Algorithm: GAimpute) จากนั้นทำการเปรียบเทียบประสิทธิภาพกับวิธีการแถวเฉลี่ย วิธี KNN แบบปกติ วิธีการถดถอยเชิงเส้นพหุคูณ รวมถึงวิธีการแบบ KNNFS ซึ่งทำการทดลองกับข้อมูลมะเร็งลำไส้ใหญ่ ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอให้ประสิทธิภาพที่ดีกว่าในเทอมของ Normalize Root Mean Squared Error (NRMSE)

นอกจากนี้ ยังมีงานวิจัยที่นำวิธีการ KNN ไปประยุกต์เพิ่มเติมเพื่อเพิ่มประสิทธิภาพในการประมาณค่าสูญหาย เช่น งานวิจัยที่นำเสนอโดย นรุตม์ บุตรพลอย (2553) เรื่อง การประยุกต์ Soft Computing และ k-Nearest Neighbor เพื่อใช้ประมาณค่าสูญหายของข้อมูล

เนื่องจากในการศึกษาครั้งนี้ การสูญหายปรากฏขึ้นทั้งในตัวแปรอิสระ $x_{ip}; p = 1,2,3$ และตัวแปรตาม y ดังนั้น ในการประมาณค่าสูญหาย จึงต้องแยกการประมาณออกเป็น 3 กรณี ซึ่งได้แก่

<u>กรณีที่ 1</u> ตัวแปรอิสระ $(x_{ip}; p = 1,2,3)$ ทุกตัวทราบค่า แต่ตัวแปรตาม (y) เกิดการสูญหาย

<u>กรณีที่ 2</u> ตัวแปรอิสระตัวใดตัวหนึ่งเกิดการสูญหาย $(x_{ip}; p = 1,2,3)$ แต่ตัวแปรตาม (y) ทราบค่า

<u>กรณีที่ 3</u> ตัวแปรอิสระตัวใดตัวหนึ่งเกิดการสูญหาย $(x_{ip}; p = 1,2,3)$ และตัวแปรตาม (y) เกิดการสูญหาย

โดยที่ในแต่ละกรณีจะเป็นการหาความคล้ายระหว่างหน่วยตัวอย่างในชุดข้อมูลที่สมบูรณ์เสมอ เช่น กรณีที่ 1 ถ้าตัวแปรอิสระทุกตัวทราบค่า แต่ตัวแปรตาม (y) เกิดการสูญหาย เราจะพิจารณาจากระยะทางยุคลิด (Euclidean Distance) ในการประมาณค่าสูญหายของตัวแปรตามจากระยะห่างของตัวแปรอิสระในชุดข้อมูลที่สมบูรณ์เท่านั้น และในทำนองเดียวกัน ถ้าตัวแปรอิสระบางตัวเกิดการสูญหาย แต่ตัวแปรตาม (y) ทราบค่า เราก็จะพิจารณาเปรียบเทียบระยะห่างระหว่างคู่ของชุดข้อมูลที่สมบูรณ์เช่นกัน

ตัวอย่าง เช่น

$y_1 = x_{11} + x_{12} + x_{13}$ .....ชุดข้อมูลที่ 1

$\otimes = x_{21} + x_{22} + x_{23}$ .....ชุดข้อมูลที่ 2 อยู่ใน<u>กรณีที่ 1</u>

$y_3 = \otimes + x_{32} + x_{33}$ .....ชุดข้อมูลที่ 3 อยู่ใน<u>กรณีที่ 2</u>

$y_4 = x_{41} + x_{42} + x_{43}$ .....ชุดข้อมูลที่ 4

$\otimes = \otimes + x_{52} + x_{53}$ .....ชุดข้อมูลที่ 5 อยู่ในกรณีที่ 3

$y_6 = x_{61} + x_{62} + x_{63}$ .....ชุดข้อมูลที่ 6

หมายเหตุ $\otimes$ แทนข้อมูลที่สูญหาย

จะเห็นว่าจากชุดข้อมูลที่ 2 (กรณีที่ 1) ตำแหน่งตัวแปรตามเกิดการสูญหาย แต่ตัวแปรอิสระทุกตัวทราบค่า ในการพิจารณาหาความคล้ายระหว่างหน่วยตัวอย่าง จะพิจารณาจากระยะทางยุคลิด (Euclidean Distance) ซึ่งหาได้จากสมการต่อไปนี้

$$D_{ij} = \sqrt{\sum_{p=1}^{3} (x_{ip} - x_{jp})^2} \qquad ; i = 1,2,...,m \text{ และ } j = m+1,...,n$$

กำหนดให้ y_j^* แทนค่าประมาณของข้อมูลตัวแปรตามที่สูญหายด้วยวิธี KNN โดยจะมีขั้นตอนดังต่อไปนี้

1. คำนวณหาค่า D_{ij} สำหรับตัวแปรอิสระทุกคู่ที่เป็นไปได้ (ซึ่งในตัวอย่างนี้คือข้อมูลจากชุดที่ 1,4 และ 6) แล้วนับจำนวน D_{ij} ที่มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างตัวที่ j
2. คำนวณหาค่าเฉลี่ยของตัวแปรตาม ที่สอดคล้องกับค่า D_{ij} ที่ต่ำสุด K ตัวของชุดตัวอย่างตัวที่ j โดยกำหนดให้เป็น $\overline{y}_j^*$
3. จะได้ว่า $y_j^* = \overline{y}_j^*$

จากชุดข้อมูลที่ 3 (กรณีที่ 2) ตัวแปรอิสระตัวใดตัวหนึ่ง $(x_{ip}; p = 1,2,3)$ เกิดการสูญหาย แต่ตัวแปรตาม (y) ทราบค่าจะทำการประมาณค่าสูญหายจากชุดข้อมูลที่มีความสมบูรณ์ ซึ่งหาได้จากสมการต่อไปนี้

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

กำหนดให้ $x_{ip}^*; p = 1,2,3$ (แล้วแต่กรณีที่ศึกษา) แทนค่าประมาณของข้อมูลตัวแปรตามที่สูญหายด้วยวิธี KNN โดยจะมีขั้นตอนดังต่อไปนี้

1. คำนวณหาค่า D_{ij} สำหรับตัวแปรอิสระและตัวแปรตามทุกคู่ที่เป็นไปได้ แล้วนับจำนวน D_{ij} และ ที่มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างตัวที่ j

2. คำนวณหาค่าเฉลี่ยของตัวแปรอิสระ ที่สอดคล้องกับค่า D_{ij} ที่ต่ำสุด K ตัวของชุดตัวอย่างตัวที่ j โดยกำหนดให้เป็น $\overline{x}_{ip}^{*}$
3. จะได้ว่า $x_{ip}^{*} = \overline{x}_{ip}^{*}$

จากชุดข้อมูลที่ 5 (กรณีที่ 3) ทั้งตัวแปรอิสระตัวใดตัวหนึ่ง $(x_{ip}; p = 1,2,3)$ และตัวแปรตาม (y) เกิดการสูญหาย จะทำการประมาณค่าสูญหายจากชุดข้อมูลที่มีความสมบูรณ์เช่นเดียวกัน ซึ่งหาได้จากสมการต่อไปนี้

$$D_{ij} = \sqrt{\sum_{p=1}^{3}(x_{ip} - x_{jp})^2} \qquad ;i = 1,2,...,m \text{ และ } j = m+1,...,n$$

กำหนดให้ $x_{ip}^{*}; p = 1,2,3$ และ y_j^{*} แทนค่าประมาณของข้อมูลตัวแปรอิสระและตัวแปรตามที่สูญหายด้วยวิธี KNN โดยจะมีขั้นตอนดังต่อไปนี้

6. คำนวณหาค่า D_{ij} สำหรับตัวแปรอิสระทุกคู่ที่เป็นไปได้ แล้วนับจำนวน D_{ij} ที่มีค่าต่ำสุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างตัวที่ j
7. คำนวณหาค่าเฉลี่ยของตัวแปรอิสระและตัวแปรตาม ที่สอดคล้องกับค่า D_{ij} ที่ต่ำสุด K ตัวของชุดตัวอย่างตัวที่ j โดยกำหนดให้เป็น $x_{ip}^{*}; p = 1,2,3$ และ $\overline{y}_j^{*}$
8. จะได้ว่า $x_{ip}^{*} = \overline{x}_{ip}^{*}$ และ $y_j^{*} = \overline{y}_j^{*}$

2.5 **วิธีการประมาณค่าสูญหายแบบ** Predictive Mean Matching Imputation (PMM)

วิธีการประมาณค่าสูญหายแบบ Predictive Mean Matching Imputation (PMM) เป็นวิธีการที่เกิดจากการรวมสองแนวคิดเข้าด้วยกันระหว่างการหาค่าคาดหวังและการแทนที่ ซึ่งมีวิธีการคือจะหาค่าคาดหวังของข้อมูลที่สูญหายได้จากข้อมูลและพารามิเตอร์ที่ทราบค่า เมื่อได้ค่าคาดหวังแล้วจะทำการแทนที่ข้อมูลสูญหายด้วยการหาค่าสัมบูรณ์ระหว่างผลต่างที่น้อยที่สุดของค่าคาดหวังของตัวแปรที่ทราบค่ากับค่าคาดหวังของค่าที่สูญหาย

จากงานวิจัยของอุษณีย์ วงศ์อามาตย์ (2555) ที่ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงพหุ ซึ่งวิธีการ PMM เป็นหนึ่งในวิธีการที่เลือกมาเปรียบเทียบประสิทธิภาพ Van Buuren และ Groothuis-Oudshoorn (2011) ได้นำเสนอวิธี PMM โดยใช้เทคนิคการปรับค่าสัมประสิทธิ์การถดถอยใหม่จากตัวประมาณเบส์ (Bayes' Estimators) โดยใช้การแจกแจงโดยหลักเกณฑ์ที่ไม่ทราบข้อมูล (Noninformative Prior Distribution) ซึ่งค่าสัมประสิทธิ์การถดถอยใหม่จะนำไปสร้างสมการพยากรณ์ เพื่อประมาณค่าตัวแปรตามทั้งที่เกิดการสูญหายและที่มีข้อมูลสมบูรณ์ โดยในการพิจารณาว่าควรจะใช้ค่าคาดหวังของตัวแปรตามที่ทราบค่าตัวใดมาแทนที่ค่าที่สูญหายนั้นจะใช้ การหาค่าสัมบูรณ์ระหว่างค่าคาดหวังของข้อมูลตัวแปรตามที่ทราบค่า กับค่าคาดหวังของตัวแปรตามที่สูญหาย ซึ่งขั้นตอนของวิธีการดังกล่าวโดยละเอียด สามารถศึกษาเพิ่มเติมได้ในงานวิจัยของอุษณีย์ วงศ์อามาตย์ (2555)

เนื่องจากในงานวิจัยของอุษณีย์ วงศ์อามาตย์ (2555) ศึกษาเฉพาะกรณีที่ตัวแปรตามเกิดการสูญหาย ดังนั้นวิธีการ PMM ที่ใช้ในการประมาณค่าสูญหายจึงเป็นวิธีการประมาณค่าสูญหายเฉพาะตัวแปรตามเท่านั้น แต่ในงานวิจัยนี้ศึกษากรณีที่เกิดการสูญหายทั้งในตัวแปรอิสระและตัวแปรตาม ผู้วิจัยจึงได้ทำการปรับเปลี่ยนขั้นตอนบางประการ เพื่อให้สามารถประมาณค่าสูญหายในกรณีที่เกิดการสูญหายของตัวแปรอิสระและตัวแปรตามในชุดข้อมูลเดียวกันได้ แต่ยังคงไว้ซึ่งหลักการของวิธีการ PMM โดยมีขั้นตอนดังต่อไปนี้

1. ในการประมาณค่าสูญหายจากรูปแบบความสัมพันธ์ของสมการถดถอยเชิงเส้น จะเริ่มจากการประมาณค่าตัวแปรอิสระที่เกิดการสูญหายก่อน โดยการประมาณค่าสูญหายของตัวแปรอิสระนั้นจะประมาณโดยใช้วิธีการที่คล้ายคลึงกับวิธี KNN กล่าวคือ จะแทนที่ข้อมูลตัวแปรอิสระที่สูญหายด้วยค่าของตัวแปรอิสระในชุดข้อมูลที่สมบูรณ์ที่ใกล้เคียงกับชุดของตัวแปรอิสระที่เกิดการสูญหายมากที่สุด
2. เมื่อได้ค่าประมาณของตัวแปรอิสระที่สูญหายแล้ว จะทำการประมาณตัวแปรตาม ตามรูปแบบของสมการถดถอยเชิงเส้นพหุ โดยมีขั้นตอนดังต่อไปนี้

    2.1) ทำการประมาณสัมประสิทธิ์การถดถอยเพื่อสร้างสมการถดถอยจากชุดข้อมูลที่สมบูรณ์

$$\hat{\tilde{\beta}}^{(0)} = (X_0{}' X_0)^{-1} X_0{}' \tilde{y}_0$$

จะได้ $\hat{y}_i = \hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \hat{\beta}_2^{(0)} x_{i2} + \hat{\beta}_3^{(0)} x_{i3} \quad ; i = 1, 2, ..., m, m+1, ..., n$

โดยที่ $\tilde{y}_0$ แทน เวกเตอร์ของตัวแปรตามที่ทราบค่าขนาด $m \times 1$

X_0 แทน เมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามที่ทราบค่า ขนาด $m \times (p+1)$

$\hat{y}_i$ แทน ค่าพยากรณ์ของตัวแปรตามของข้อมูลตัวที่ i

2.2) ทำการประมาณค่าสูญหายของตัวแปรตามจากรูปแบบความสัมพันธ์ของสมการถดถอยเชิงเส้นพหุ จากนั้นแทนที่ข้อมูลสูญหายด้วยค่าประมาณของตัวแปรตาม ($\hat{y}_i$) ในชุดข้อมูลที่สมบูรณ์ ที่ใกล้เคียงกับค่าประมาณตัวแปรตาม ($\hat{y}_i$) ของชุดข้อมูลที่สูญหายมากที่สุด

บทที่ 3

วิธีดำเนินการวิจัย

ในงานวิจัยนี้ จะทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามและตัวแปรอิสระมีการสูญหายแบบ Nonignorable ซึ่งลักษณะของข้อมูลที่ทำการศึกษาจะเป็นข้อมูลภาคตัดขวาง(Cross-Section Data) โดยจะเริ่มจากการสร้างชุดข้อมูลของตัวแปรอิสระและชุดข้อมูลของความคลาดเคลื่อน เพื่อนำไปสู่การสร้างชุดข้อมูลของตัวแปรตามตามรูปแบบความสัมพันธ์ของสมการถดถอยเชิงเส้นพหุ จากนั้นจะสร้างข้อมูลที่มีการแจกแจงแบบเบอร์นูลลี เพื่อให้เกิดการสูญหายของข้อมูลในตัวแปรอิสระและตัวแปรตาม

เมื่อได้ชุดข้อมูลที่เกิดการสูญหายแล้ว จะทำการแทนที่ข้อมูลที่สูญหายโดยใช้วิธีการประมาณค่าสูญหาย 3 วิธีการ ซึ่งได้แก่ วิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) หลังจากนั้นจะทำการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการด้วยค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าพยากรณ์กับค่าจริง (AMSE) ซึ่งวิธีการใดที่ให้ค่า AMSE น้อยที่สุดจะเป็นวิธีการประมาณค่าสูญหายที่ดีที่สุด นอกจากนั้นเพื่อเป็นการเปรียบเทียบประสิทธิภาพที่ชัดเจนยิ่งขึ้นจะนำค่า AMSE ที่ได้จากวิธีการ EM มาเป็นค่ามาตรฐานในการเปรียบเทียบกับค่า AMSE ที่ได้จากวิธีการอื่นๆ เพื่อนำไปสู่การหาค่าประสิทธิภาพสัมพัทธ์ (RE) รายละเอียดในแต่ละขั้นตอนเป็นดังนี้

3.1 จำลองชุดข้อมูล

ในการวิจัยนี้ จะทำการศึกษาภายใต้สถานการณ์จำลองทั้งหมด 324 สถานการณ์ ที่แตกต่างกันตามลักษณะการแจกแจงของตัวแปรต่างๆ และลักษณะของการสูญหายของข้อมูล โดยในแต่ละสถานการณ์จะทำการจำลองเป็นจำนวน 5,000 รอบ มีขั้นตอนดังต่อไปนี้

1. สร้างชุดข้อมูลตัวแปรอิสระ(x_1, x_2, x_3)ซึ่งจะแบ่งออกเป็น 2 รูปแบบคือ

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

โดยที่ ศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนเท่ากัน

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

โดยที่ ศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนแตกต่างกันโดยจะกำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน

2. สร้างชุดข้อมูลความคลาดเคลื่อน (ε) มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 10 30 และ 90

3. สร้างชุดข้อมูลตัวแปรตาม ($y_i; i = 1,2,...,n$) ตามความสัมพันธ์ของสมการถดถอยเชิงเส้นพหุ $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ โดยกำหนดให้ $\beta_0 = 42$ และ $\beta_1 = \beta_2 = \beta_3 = 1$

4. ขนาดตัวอย่างเท่ากับ 50 100 และ 200

5. ในแต่ละสถานการณ์จะสร้างชุดข้อมูลตัวแปรอิสระ (x_1, x_2, x_3)ที่ทำให้เกิดการสูญหาย โดยนำข้อมูลเหล่านั้นมาแบ่งเป็น 3 ช่วงโดยให้แต่ละช่วงมีสัดส่วนเท่าๆ กัน ซึ่งวิธีในการแบ่งช่วงของตัวแปรอิสระคือ

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

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

และ $\mu_{x_{ip}} + z_{\frac{2}{3}}\sigma_{x_{ip}} < x_{ip}$ ตัวแปรตาม x_{ip} นี้จะถูกจัดให้อยู่ในช่วงปลาย

โดยที่ $p = 1,2,3$

$i = 1,2,...,m,m+1,...,n$

$z_{\frac{1}{3}} = -0.43$

$z_{\frac{2}{3}} = +0.43$

6. จากนั้นจะทำให้แต่ละช่วงของตัวแปรอิสระมีการสูญหาย โดยจะสร้างตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่ตามความน่าจะเป็นในแต่ละช่วง(จากขอบเขตของการวิจัย)ของการเกิดเหตุการณ์ที่สนใจ ซึ่งมีขนาดเท่ากับจำนวนตัวแปรอิสระ แล้วจับคู่ข้อมูลตัวแปรอิสระกับข้อมูลตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่ที่มีค่าเป็น 0 กับ 1 ถ้าเกิดเหตุการณ์ที่สนใจ(ตัวแปรที่มีการแจกแจงแบบเบอร์นูลลี่จะเท่ากับ 1) แสดงว่าจะให้ข้อมูลตัวแปรอิสระตัวนั้นเกิดการสูญหาย

7. ถ้าหากตัวแปรอิสระเกิดการสูญหาย $(P(\delta_x = 1))$ จะทำให้ตัวแปรตามเกิดการสูญหายด้วยความน่าจะเป็น $P(\delta_y = 1 | \delta_x = 1)$ หรือในทำนองเดียวกันถ้าหาก ตัวแปรอิสระไม่เกิดการสูญหาย $(P(\delta_x = 0))$ แล้วจะส่งผลให้ตัวแปรตามเกิดการสูญหายด้วยความน่าจะเป็น $P(\delta_y = 1 | \delta_x = 0)$ โดยที่ค่าความน่าจะเป็นของการสูญหายในแต่ละช่วงจะแตกต่างกันไปตามขอบเขตของการวิจัยที่กำหนด

3.2 ประมาณข้อมูลที่สูญหายด้วยวิธีการต่าง ๆ

หลังจากทำการจำลองชุดข้อมูลที่เกิดการสูญหายตามที่กำหนดแล้ว ขั้นตอนต่อไป จะเป็นการประมาณค่าสูญหายด้วยวิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) ตามลำดับ โดยวิธี EM จะเป็นวิธีการคำนวณที่ใช้พารามิเตอร์ ส่วนวิธี KNN จะเป็นวิธีที่มีรูปแบบการคำนวณโดยไม่ใช้พารามิเตอร์ และวิธี PMM จะเป็นวิธีการที่มีความคล้ายคลึงระหว่างวิธี EM และ KNN ซึ่งรายละเอียดขั้นตอนของวิธีการประมาณค่าสูญหายในแต่ละวิธีนั้นจะอยู่ในบทที่ 2

3.3 ประมาณค่าสัมประสิทธิ์การถดถอย

เมื่อได้ค่าประมาณที่สูญหายจากการประมาณทั้ง 3 วิธีแล้ว ขั้นตอนต่อไปจะทำการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดแบบสามัญ(Ordinary Least Squares Method: OLS) ตามทฤษฎีที่ได้อธิบายไว้ในบทที่ 2 จะได้ค่าสัมประสิทธิ์การถดถอยที่แตกต่างกันไปในแต่ละวิธีของการประมาณค่าสูญหาย

3.4 สร้างสมการพยากรณ์

เมื่อได้ค่าสัมประสิทธิ์การถดถอยจากข้อ 3.3 แล้ว จะทำการสร้างสมการพยากรณ์ เพื่อใช้ในการพยากรณ์ค่าของตัวแปรตาม ซึ่งการประมาณค่าสูญหายในแต่ละวิธีการจะได้สมการพยากรณ์ที่แตกต่างกัน เนื่องจากค่าสัมประสิทธิ์การถดถอยที่หาได้จากขั้นตอน 3.3 แตกต่างกัน โดยมีรูปแบบสมการดังต่อไปนี้

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

3.5 เปรียบเทียบประสิทธิภาพของวิธีการประมาณข้อมูลที่สูญหาย

ในการเปรียบเทียบประสิทธิภาพว่าวิธีการไหนจะเป็นวิธีการประมาณข้อมูลที่สูญหายได้ดีที่สุดนั้น จะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และค่าประสิทธิภาพสัมพัทธ์ (RE) โดยจะคำนวณหาค่าดังกล่าวจากแต่ละวิธีการที่มีการทำซ้ำทั้งหมด 5,000 รอบในแต่ละสถานการณ์ที่จำลอง โดยสามารถคำนวณได้จากสูตรดังต่อไปนี้

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

$$AMSE = \frac{1}{5,000} \sum_{q=1}^{5,000} MSE_q$$

$$RE = \frac{AMSE_{EM}}{AMSE_k} \quad ; k = knn, pmm$$

เมื่อ y_i' แทน ค่าจริงของข้อมูลตัวแปรตามตัวที่ i

$\hat{y}_{qi}$ แทน ค่าพยากรณ์ของข้อมูลตัวแปรตามตัวที่ i จากการทำซ้ำรอบที่ q

MSE_q แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจาก

การทำซ้ำรอบที่ q

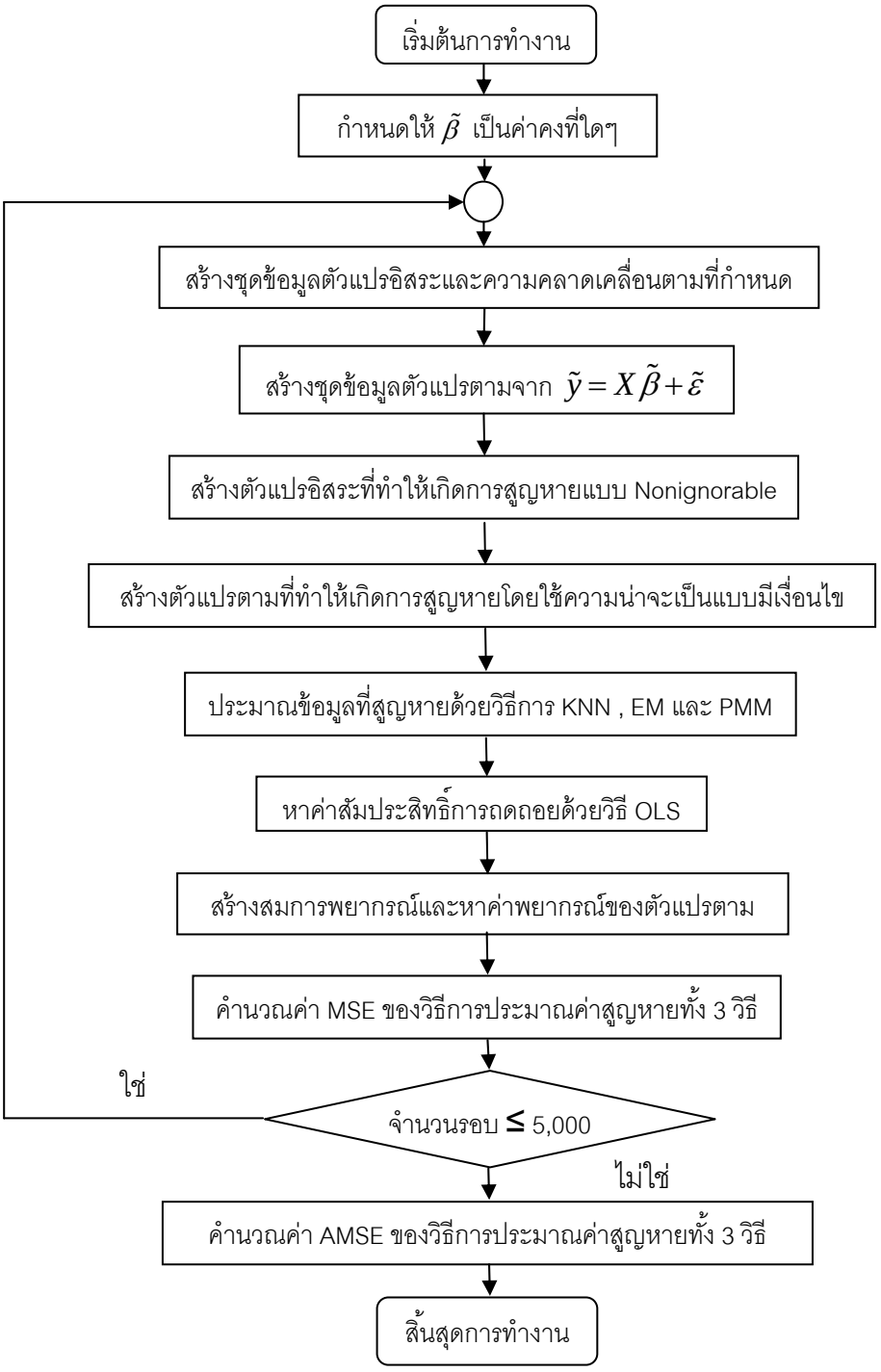
$AMSE$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการทำซ้ำทั้งหมด 5,000 รอบ

RE แทน ค่าประสิทธิภาพสัมพัทธ์

$AMSE_{EM}$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการประมาณค่าสูญหายด้วยวิธีการ EM

$AMSE_{k}$ แทน ค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ตัวแปรตามจากการประมาณค่าสูญหายด้วยวิธีการ KNN และ PMM

<u>ภาพที่ 3.1</u> แผนผังการเขียนโปรแกรม



บทที่ 4

ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ซึ่งได้แก่ วิธี K-Nearest Neighbor Imputation (KNN) วิธี EM Algorithm และวิธี Predictive Mean Matching Imputation (PMM) ซึ่งในการเปรียบเทียบประสิทธิภาพจะพิจารณาจากค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์ (AMSE) ซึ่งวิธีการที่ดีที่สุดจะเป็นวิธีการที่ให้ค่า AMSE ต่ำที่สุด ส่วนค่าประสิทธิภาพสัมพัทธ์ (RE) จะเป็นการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการ โดยใช้วิธี EM เป็นวิธีการหลักในการเปรียบเทียบเพื่อให้เห็นความแตกต่างของแต่ละวิธีการชัดเจนมากยิ่งขึ้น

ในการนำเสนอผลการวิจัยจะใช้สัญลักษณ์ต่างๆ ซึ่งมีความหมายดังต่อไปนี้

n แทน ขนาดตัวอย่าง

None แทน ไม่มีการสูญหายแบบ Nonignorable

Medium แทน มีการสูญหายแบบ Nonignorable ในระดับปานกลาง

High แทน มีการสูญหายแบบ Nonignorable ในระดับสูง

KNN แทน การประมาณค่าสูญหายด้วยวิธี KNN

EM แทน การประมาณค่าสูญหายด้วยวิธี EM

PMM แทน การประมาณค่าสูญหายด้วยวิธี PMM

AMSE แทน ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองระหว่างค่าจริงกับค่าพยากรณ์

RE แทน ค่าประสิทธิภาพสัมพัทธ์

ผลการวิจัยจะแสดงในรูปแบบของตาราง แผนภาพประกอบกับการอธิบาย ซึ่งแต่ละส่วนจะเป็นการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการด้วยค่า AMSE และ RE โดยจะแบ่งออกเป็น 4 ส่วนตามกรณีที่ศึกษาจากลักษณะการสูญหายของตัวแปรอิสระ ได้แก่

ส่วนที่ 1 ตัวแปรอิสระเป็นแบบที่ 1 : $X_1 \sim N(0,300), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$ ศึกษาในกรณีที่เกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง โดยจะกำหนดส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน ดังต่อไปนี้

1.1) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

1.2) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

1.3) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

ส่วนที่ 2 ตัวแปรอิสระแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ศึกษาในกรณีที่เกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก โดยจะกำหนดส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน ดังต่อไปนี้

2.1) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

2.2) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

2.3) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

ส่วนที่ 3 ตัวแปรอิสระแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ศึกษาในกรณีที่เกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง โดยจะกำหนดส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน ดังต่อไปนี้

3.1) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

3.2) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

3.3) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

ส่วนที่ 4 ตัวแปรอิสระแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ศึกษาในกรณีที่เกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่ โดยจะกำหนดส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน ดังต่อไปนี้

4.1) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

4.2) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

4.3) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

4.1 ส่วนที่ 1 แสดงผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อตัวแปรอิสระเป็นแบบที่ 1 (ศึกษาการสูญหายของตัวแปรอิสระตัวใดตัวหนึ่ง)

ในส่วนนี้ผู้วิจัยได้ทำการศึกษาข้อมูลสูญหายในกรณีที่เกิดการสูญหายในตัวแปรอิสระแบบที่ 1 : $X_1 \sim N(0,300), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$ โดยส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 30 และ 90 ขนาดตัวอย่างเท่ากับ 50 100 และ 200 สัดส่วนของการสูญหายเท่ากับ 10% 20% และ 30% ระดับการสูญหายแบบ Nonignorable คือ ไม่มี ปานกลาง และสูง ซึ่งผลการวิจัยจะนำเสนอโดยแสดงค่า AMSE และ RE ที่ได้จากการประมาณค่าสูญหายของแต่ละวิธีการในตารางดังต่อไปนี้

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.1.1.1	แบบที่ 1	10
4.1.2.1	แบบที่ 1	30
4.1.3.1	แบบที่ 1	90

ตารางที่ 4.1.1.1-4.1.3.1 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า AMSE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.1.1.2	แบบที่ 1	10
4.1.2.2	แบบที่ 1	30
4.1.3.2	แบบที่ 1	90

ตารางที่ 4.1.1.2-4.1.3.2 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ภาพที่	ชนิดของตัวแปรอิสระ	ขนาดตัวอย่าง (n)
4.1.4	แบบที่ 1	50
4.1.5	แบบที่ 1	100
4.1.6	แบบที่ 1	200

ภาพที่ 4.1.4-4.1.6 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 10 30 และ 90 สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable

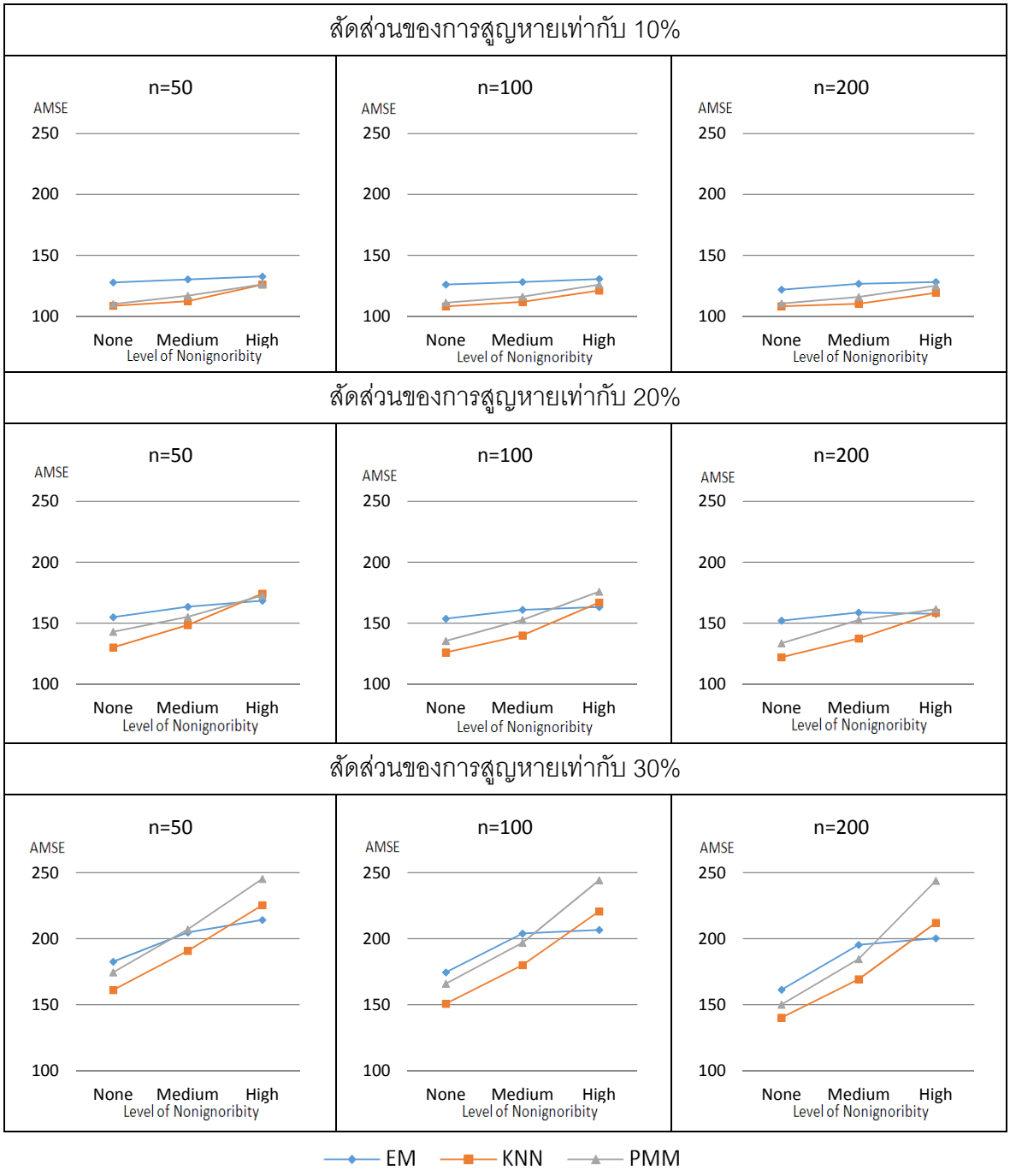
จากตารางและภาพที่ 4.1.1.1 - 4.1.3.1 พบว่า โดยส่วนใหญ่วิธีการ KNN คือวิธีการประมาณค่าสูญหายที่ให้ค่า AMSE ต่ำที่สุดเกือบทุกกรณี ยกเว้นในบางกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 และ 30 ที่ข้อมูลมีสัดส่วนของการสูญหายและระดับการสูญหายแบบ Nonignorable สูง วิธีการ EM จะเป็นวิธีการที่ดีกว่า KNN ส่วนวิธีการ PMM จะเป็นวิธีการที่มีประสิทธิภาพน้อยที่สุดในทุกกรณี

จากตารางและภาพที่ 4.1.1.2 - 4.1.3.2 เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีด้วยค่า RE โดยจะใช้วิธีการ EM เป็นวิธีการมาตรฐานในการเปรียบเทียบ โดยถ้า ค่า RE มากกว่า 1 แสดงว่า วิธีการ EM มีประสิทธิภาพในการประมาณค่าสูญหายน้อยกว่าวิธีการที่เปรียบเทียบ แต่ถ้า RE น้อยกว่า 1 แสดงว่าวิธีการ EM มีประสิทธิภาพดีกว่าวิธีการที่นำมาเปรียบเทียบ ซึ่งจากผลการวิจัยพบว่า สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable ที่สูงขึ้นจะส่งผลให้วิธีการ EM มีประสิทธิภาพดีกว่า KNN ซึ่งจะเห็นได้อย่างชัดเจนในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย(10-30) สังเกตจากตารางที่ 4.1.4-4.1.6 โดยที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง (90) ประสิทธิภาพของแต่ละวิธีการจะไม่แตกต่างกันมาก แต่อย่างไรก็ตามวิธีการ KNN ก็ยังเป็นวิธีการที่ดีที่สุดในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง

<u>ตารางที่ 4.1.1.1</u> แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	127.67	**108.45**	109.95
		Medium	AMSE	130.17	**112.30**	116.73
		High	AMSE	132.61	**125.90**	126.17
	20	None	AMSE	154.89	**130.04**	142.89
		Medium	AMSE	163.48	**148.38**	155.18
		High	AMSE	**168.43**	174.22	172.70
	30	None	AMSE	182.56	**161.09**	174.47
		Medium	AMSE	204.76	**190.76**	207.06
		High	AMSE	**214.22**	225.41	245.35
100	10	None	AMSE	125.97	**107.95**	111.08
		Medium	AMSE	128.03	**111.68**	116.02
		High	AMSE	130.52	**121.05**	125.95
	20	None	AMSE	153.63	**125.91**	135.39
		Medium	AMSE	160.94	**139.94**	152.62
		High	AMSE	**163.19**	166.91	175.82
	30	None	AMSE	174.47	**150.70**	165.92
		Medium	AMSE	203.95	**180.00**	196.94
		High	AMSE	**206.61**	220.69	244.29
200	10	None	AMSE	121.73	**108.10**	110.40
		Medium	AMSE	126.56	**110.15**	115.72
		High	AMSE	128.03	**119.15**	124.96
	20	None	AMSE	152.04	**121.96**	133.40
		Medium	AMSE	158.70	**137.38**	152.61
		High	AMSE	**157.60**	158.56	161.46
	30	None	AMSE	161.27	**140.12**	150.08
		Medium	AMSE	195.35	**169.24**	184.55
		High	AMSE	**200.29**	211.99	243.84

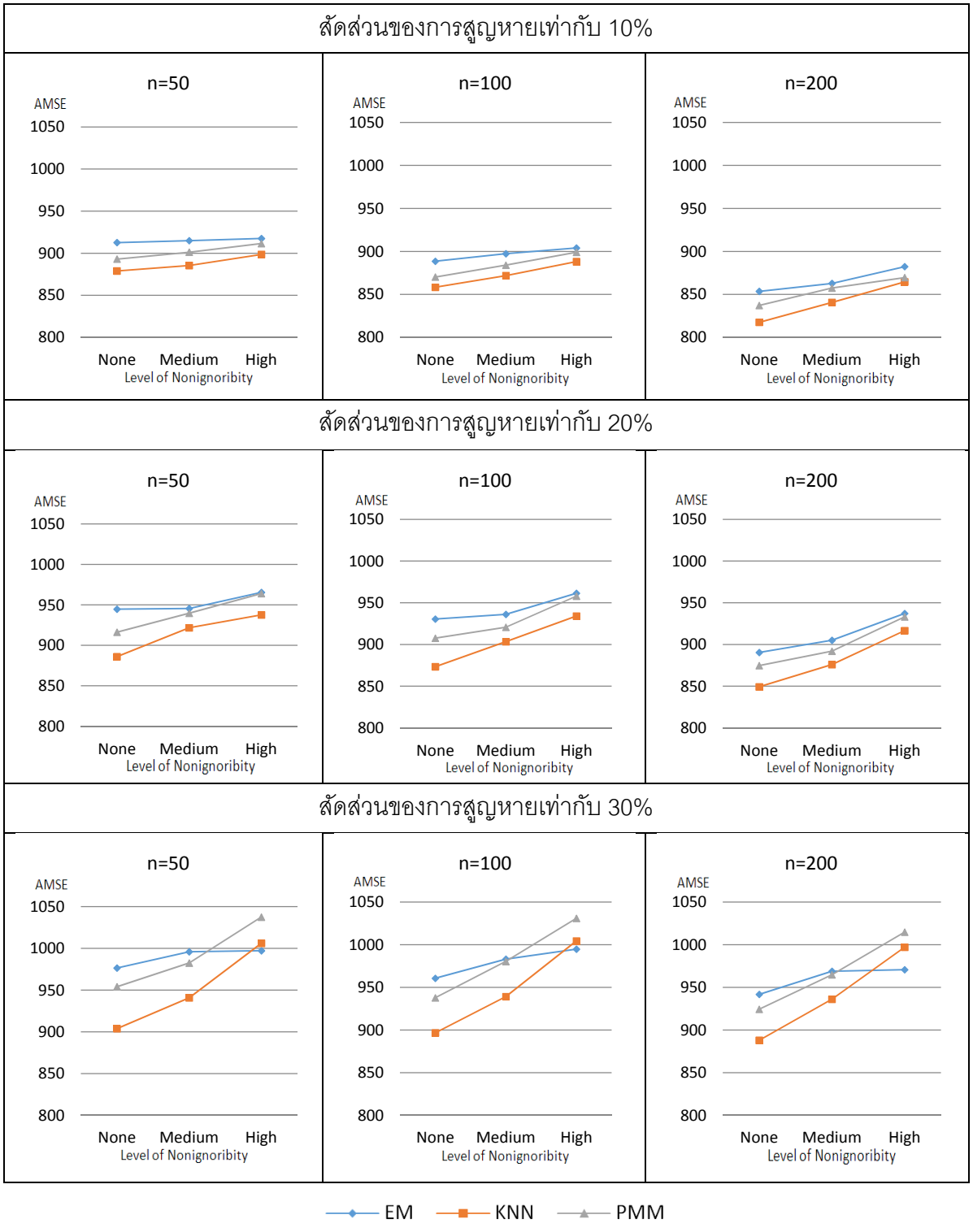
<u>ภาพที่</u> 4.1.1.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



<u>ตารางที่ 4.1.2.1</u> แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	912.68	**878.89**	893.05
		Medium	AMSE	914.90	**885.41**	901.18
		High	AMSE	917.49	**898.57**	911.51
	20	None	AMSE	944.73	**886.04**	916.22
		Medium	AMSE	945.42	**921.66**	939.64
		High	AMSE	965.43	**937.65**	963.99
	30	None	AMSE	976.52	**903.95**	954.05
		Medium	AMSE	995.92	**940.83**	982.47
		High	AMSE	**997.12**	1006.08	1037.46
100	10	None	AMSE	888.52	**858.28**	870.17
		Medium	AMSE	897.30	**871.82**	883.99
		High	AMSE	903.96	**888.09**	898.97
	20	None	AMSE	930.51	**873.33**	907.52
		Medium	AMSE	936.06	**903.32**	920.70
		High	AMSE	961.41	**934.05**	957.90
	30	None	AMSE	960.70	**896.72**	937.78
		Medium	AMSE	983.12	**939.16**	980.40
		High	AMSE	**994.80**	1004.38	1030.87
200	10	None	AMSE	853.49	**817.56**	837.16
		Medium	AMSE	862.79	**840.69**	857.37
		High	AMSE	882.31	**864.51**	869.54
	20	None	AMSE	890.39	**849.25**	874.73
		Medium	AMSE	905.01	**875.91**	891.93
		High	AMSE	937.16	**916.56**	933.11
	30	None	AMSE	941.82	**888.03**	924.37
		Medium	AMSE	968.94	**936.07**	964.83
		High	AMSE	**970.71**	997.07	1014.81

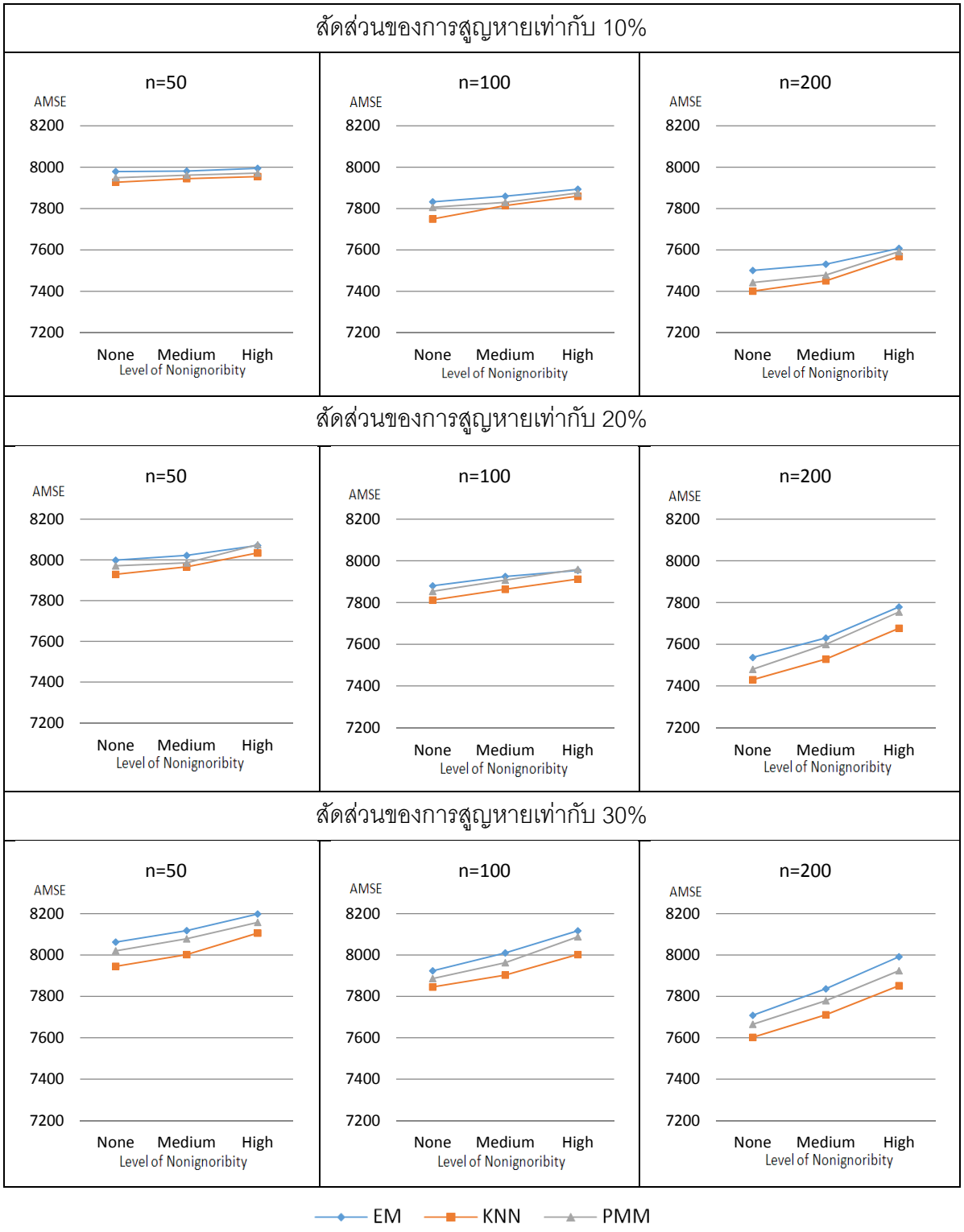
<u>ภาพที่</u> 4.1.2.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



<u>ตารางที่ 4.1.3.1</u> แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	7978.79	**7927.36**	7948.97
		Medium	AMSE	7981.14	**7944.63**	7961.51
		High	AMSE	7994.63	**7954.63**	7971.44
	20	None	AMSE	7998.50	**7929.11**	7971.09
		Medium	AMSE	8021.83	**7965.88**	7985.67
		High	AMSE	8070.63	**8034.31**	8074.53
	30	None	AMSE	8062.37	**7945.33**	8020.18
		Medium	AMSE	8117.95	**8002.12**	8078.40
		High	AMSE	8197.97	**8105.96**	8157.69
100	10	None	AMSE	7832.42	**7749.52**	7806.21
		Medium	AMSE	7859.72	**7814.17**	7829.80
		High	AMSE	7893.69	**7859.64**	7875.32
	20	None	AMSE	7880.45	**7811.86**	7853.82
		Medium	AMSE	7925.12	**7863.97**	7907.34
		High	AMSE	7954.15	**7912.21**	7959.35
	30	None	AMSE	7923.35	**7845.47**	7886.28
		Medium	AMSE	8010.14	**7902.92**	7962.74
		High	AMSE	8116.91	**8002.16**	8087.79
200	10	None	AMSE	7500.92	**7400.97**	7442.38
		Medium	AMSE	7531.13	**7450.77**	7478.38
		High	AMSE	7607.84	**7568.19**	7591.80
	20	None	AMSE	7536.61	**7430.08**	7479.65
		Medium	AMSE	7629.79	**7528.97**	7598.99
		High	AMSE	7779.14	**7676.71**	7754.83
	30	None	AMSE	7708.32	**7601.73**	7665.15
		Medium	AMSE	7836.52	**7710.76**	7779.10
		High	AMSE	7991.47	**7851.55**	7924.15

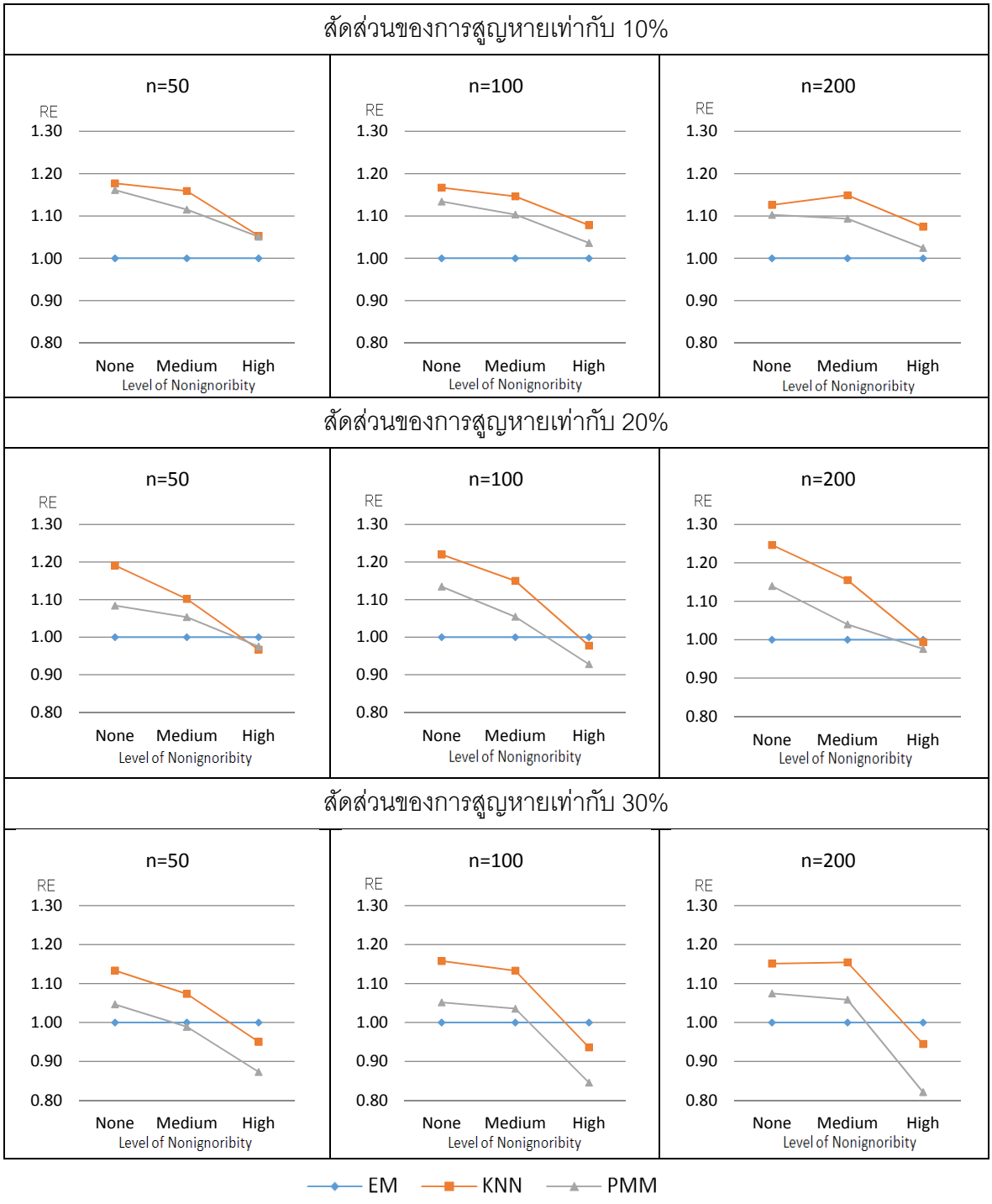
<u>ภาพที่</u> 4.1.3.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



<u>ตารางที่ 4.1.1.2</u> แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.177	1.161
		Medium	RE	1.000	1.159	1.115
		High	RE	1.000	1.053	1.051
	20	None	RE	1.000	1.191	1.084
		Medium	RE	1.000	1.102	1.053
		High	RE	1.000	0.967	0.975
	30	None	RE	1.000	1.133	1.046
		Medium	RE	1.000	1.073	0.989
		High	RE	1.000	0.950	0.873
100	10	None	RE	1.000	1.167	1.134
		Medium	RE	1.000	1.146	1.104
		High	RE	1.000	1.078	1.036
	20	None	RE	1.000	1.220	1.135
		Medium	RE	1.000	1.150	1.055
		High	RE	1.000	0.978	0.928
	30	None	RE	1.000	1.158	1.052
		Medium	RE	1.000	1.133	1.036
		High	RE	1.000	0.936	0.846
200	10	None	RE	1.000	1.126	1.103
		Medium	RE	1.000	1.149	1.094
		High	RE	1.000	1.075	1.025
	20	None	RE	1.000	1.247	1.140
		Medium	RE	1.000	1.155	1.040
		High	RE	1.000	0.994	0.976
	30	None	RE	1.000	1.151	1.075
		Medium	RE	1.000	1.154	1.059
		High	RE	1.000	0.945	0.821

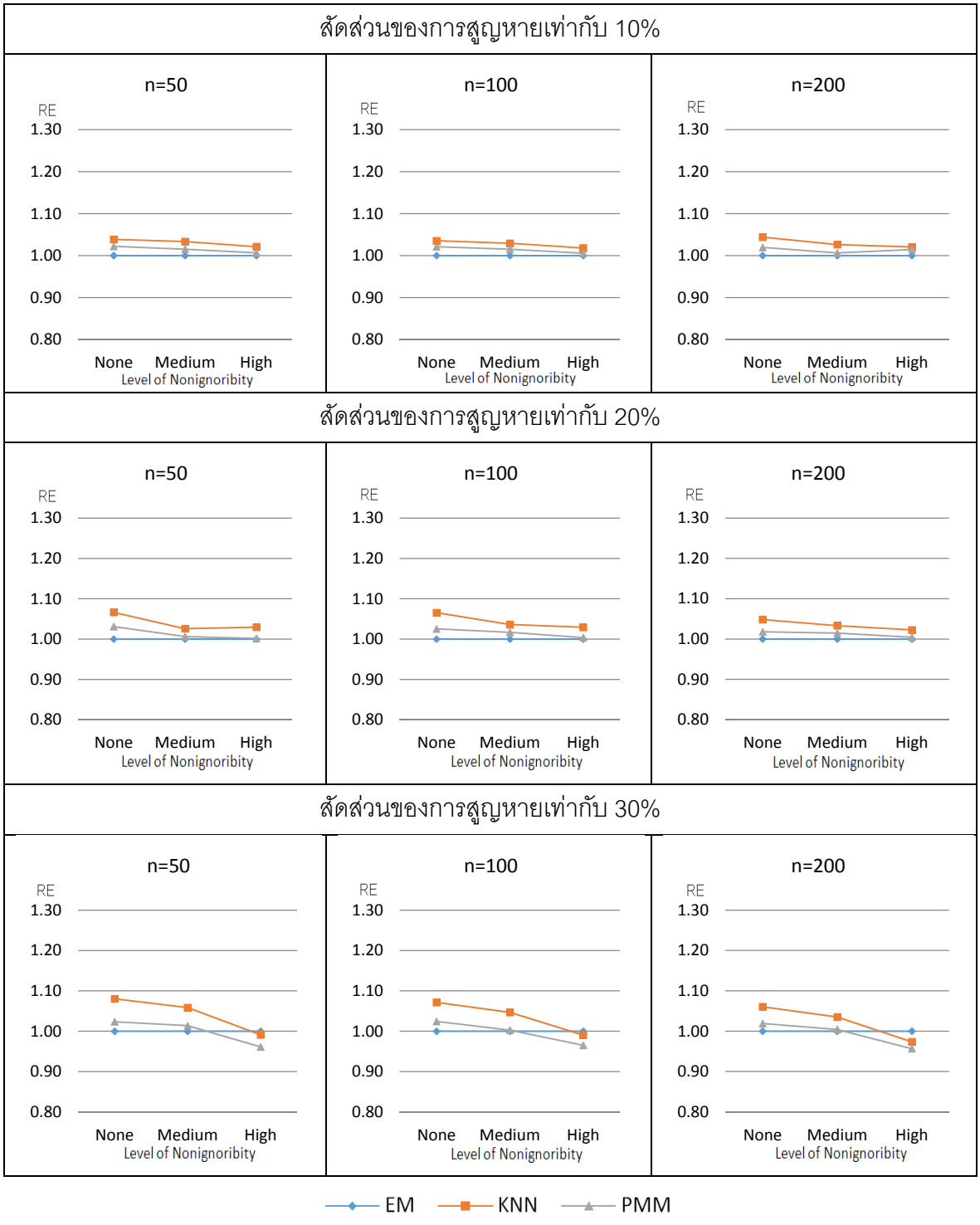
<u>ภาพที่</u> 4.1.1.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



<u>ตารางที่ 4.1.2.2</u> แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.038	1.022
		Medium	RE	1.000	1.033	1.015
		High	RE	1.000	1.021	1.007
	20	None	RE	1.000	1.066	1.031
		Medium	RE	1.000	1.026	1.006
		High	RE	1.000	1.030	1.001
	30	None	RE	1.000	1.080	1.024
		Medium	RE	1.000	1.059	1.014
		High	RE	1.000	0.991	0.961
100	10	None	RE	1.000	1.035	1.021
		Medium	RE	1.000	1.029	1.015
		High	RE	1.000	1.018	1.006
	20	None	RE	1.000	1.065	1.025
		Medium	RE	1.000	1.036	1.017
		High	RE	1.000	1.029	1.004
	30	None	RE	1.000	1.071	1.024
		Medium	RE	1.000	1.047	1.003
		High	RE	1.000	0.990	0.965
200	10	None	RE	1.000	1.044	1.020
		Medium	RE	1.000	1.026	1.006
		High	RE	1.000	1.021	1.015
	20	None	RE	1.000	1.048	1.018
		Medium	RE	1.000	1.033	1.015
		High	RE	1.000	1.022	1.004
	30	None	RE	1.000	1.061	1.019
		Medium	RE	1.000	1.035	1.004
		High	RE	1.000	0.974	0.957

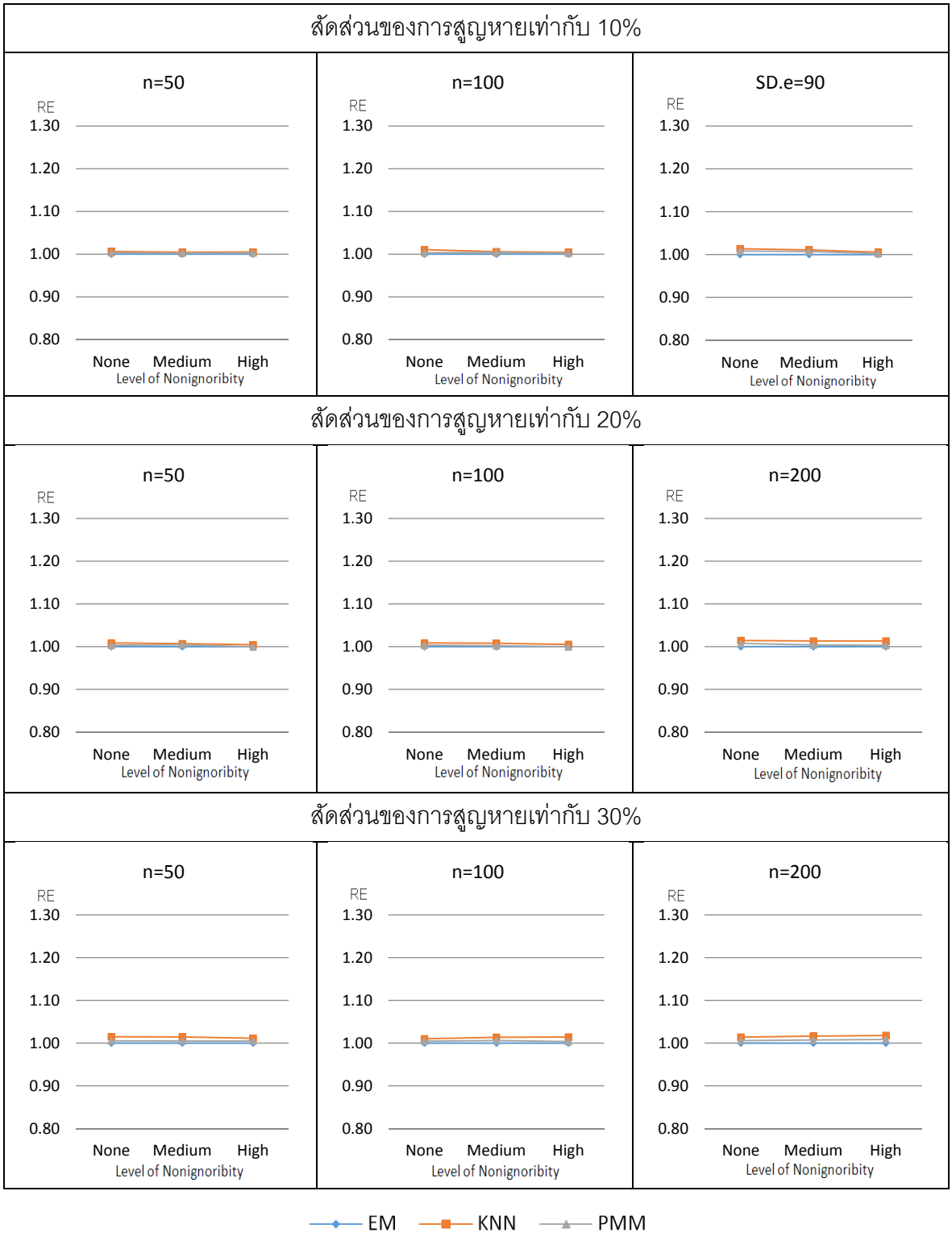
<u>ภาพที่</u> 4.1.2.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



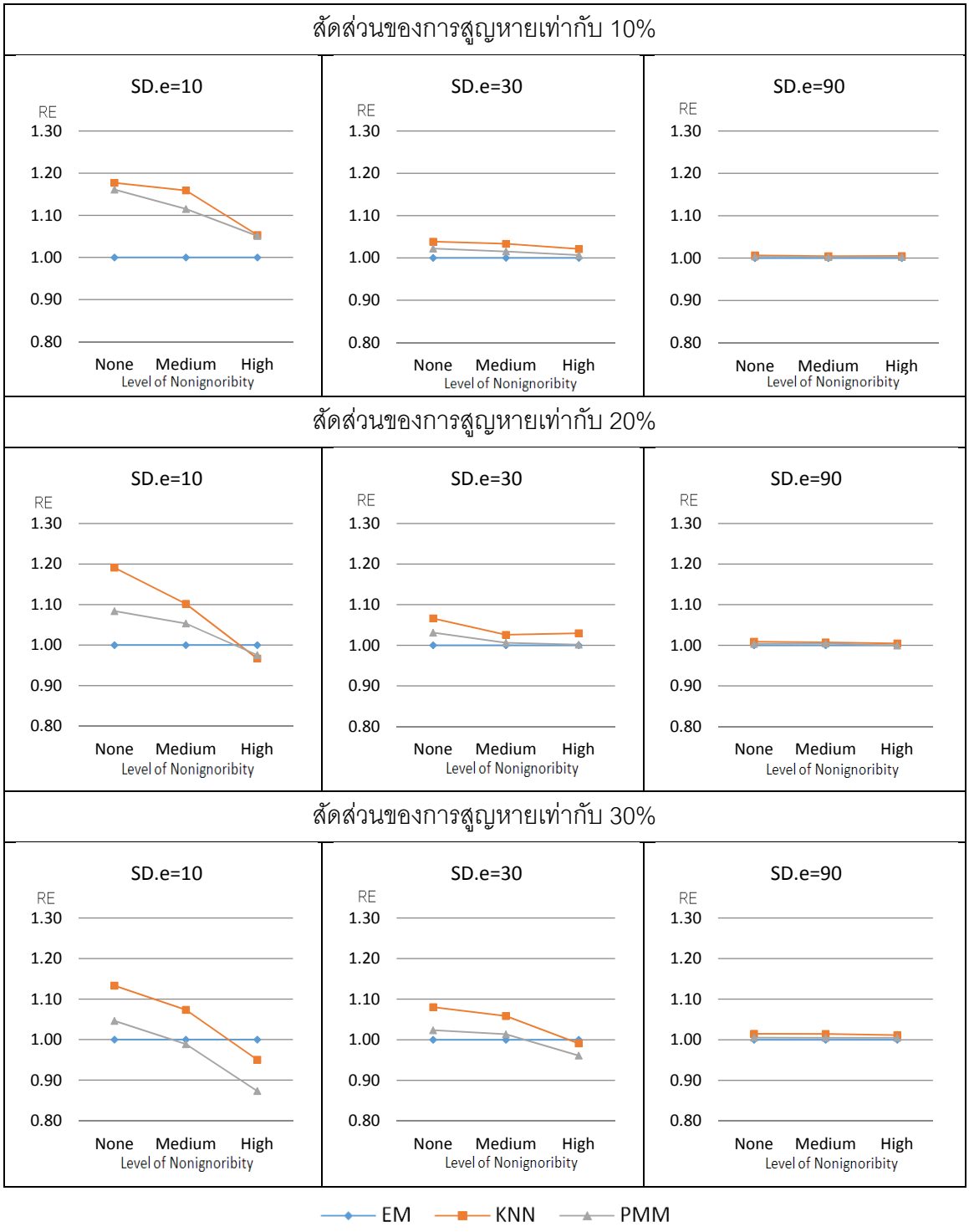
<u>ตารางที่ 4.1.3.2</u> แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.006	1.004
		Medium	RE	1.000	1.005	1.002
		High	RE	1.000	1.005	1.003
	20	None	RE	1.000	1.009	1.003
		Medium	RE	1.000	1.007	1.005
		High	RE	1.000	1.005	1.000
	30	None	RE	1.000	1.015	1.005
		Medium	RE	1.000	1.014	1.005
		High	RE	1.000	1.011	1.005
100	10	None	RE	1.000	1.011	1.003
		Medium	RE	1.000	1.006	1.004
		High	RE	1.000	1.004	1.002
	20	None	RE	1.000	1.009	1.003
		Medium	RE	1.000	1.008	1.002
		High	RE	1.000	1.005	0.999
	30	None	RE	1.000	1.010	1.005
		Medium	RE	1.000	1.014	1.006
		High	RE	1.000	1.014	1.004
200	10	None	RE	1.000	1.014	1.008
		Medium	RE	1.000	1.011	1.007
		High	RE	1.000	1.005	1.002
	20	None	RE	1.000	1.014	1.008
		Medium	RE	1.000	1.013	1.004
		High	RE	1.000	1.013	1.003
	30	None	RE	1.000	1.014	1.006
		Medium	RE	1.000	1.016	1.007
		High	RE	1.000	1.018	1.008

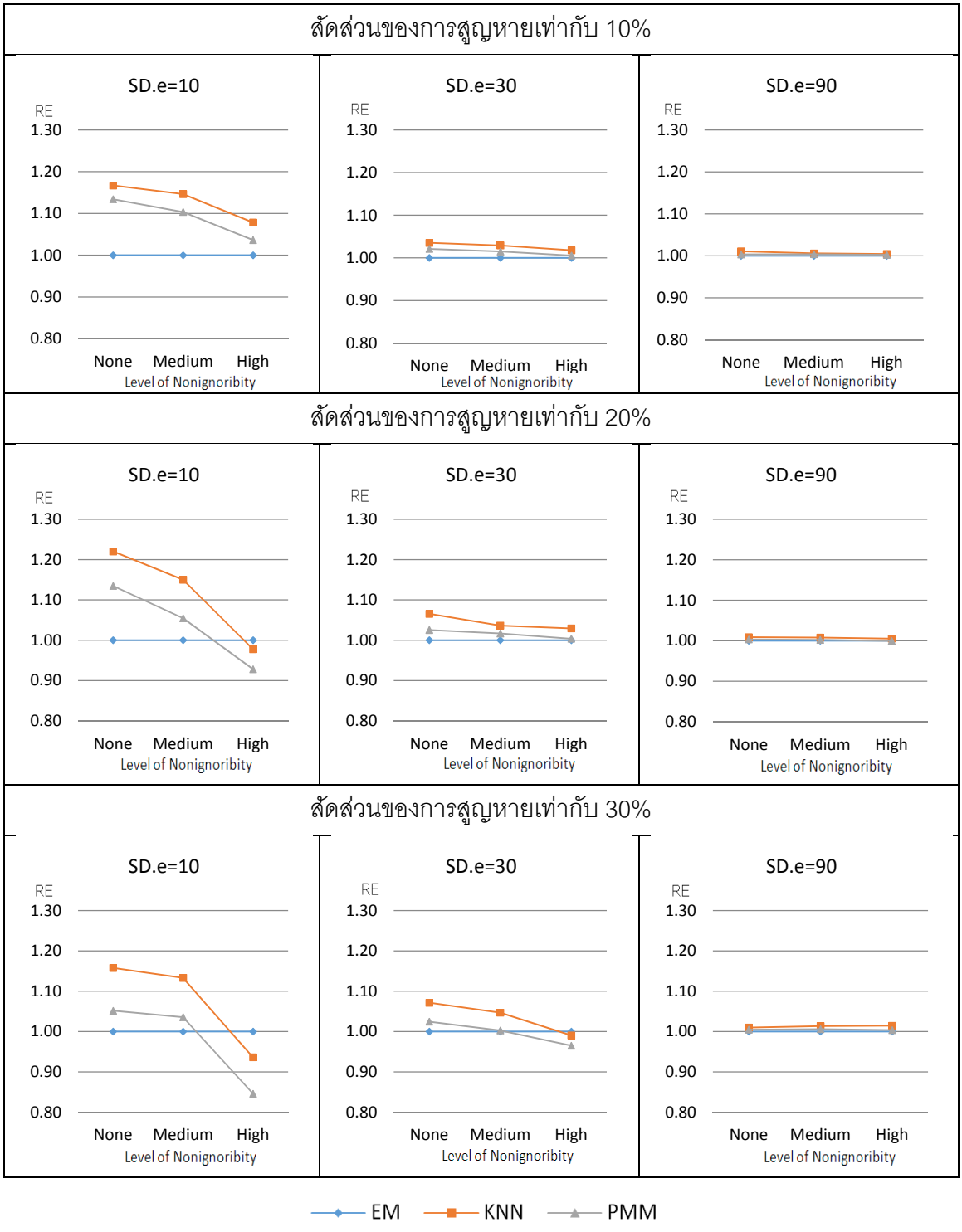
<u>ภาพที่</u> 4.1.3.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



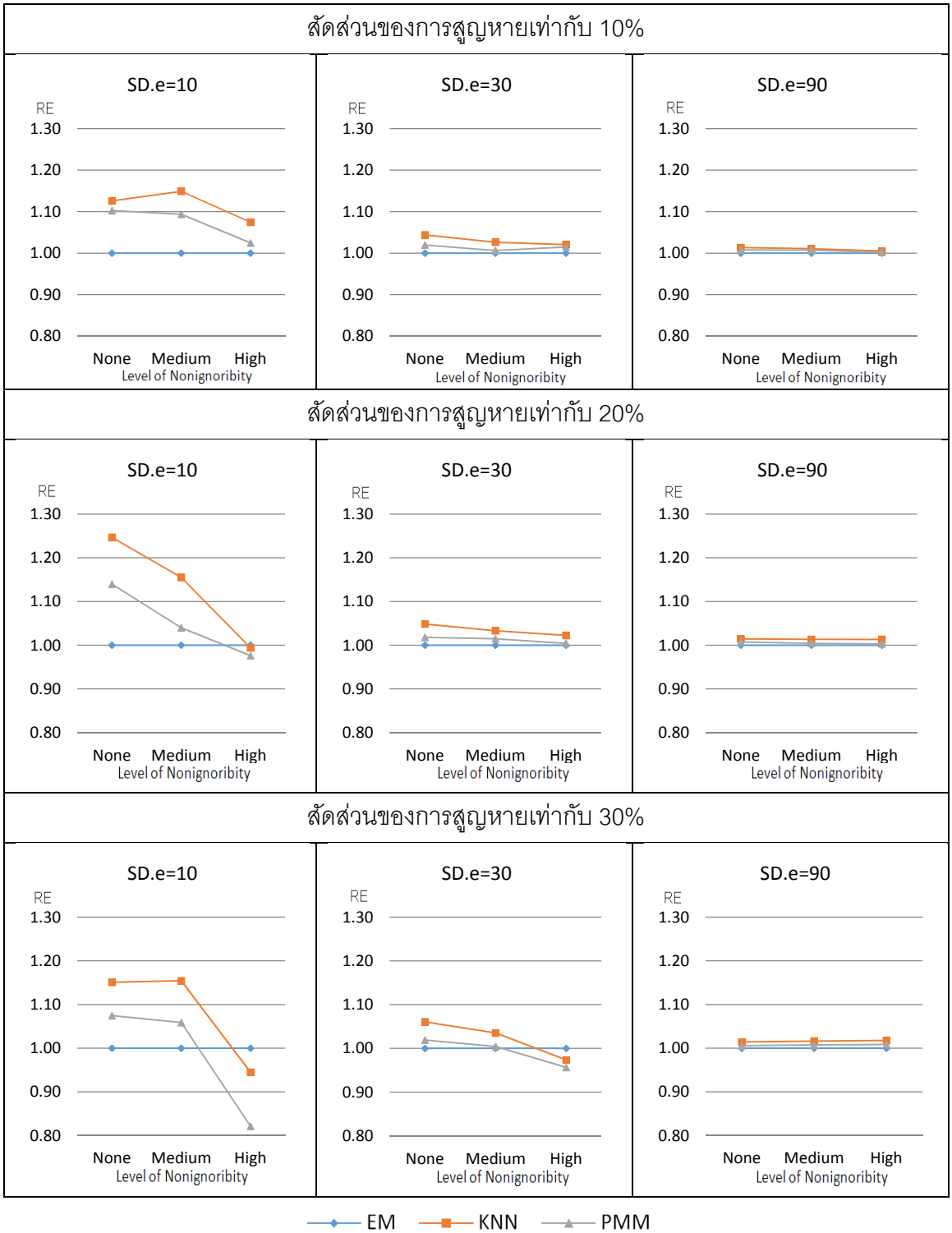
<u>ภาพที่</u> 4.1.4 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และขนาดตัวอย่างเท่ากับ 50



<u>ภาพที่ 4.1.5</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และขนาดตัวอย่างเท่ากับ 100



<u>ภาพที่ 4.1.6</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 และขนาดตัวอย่างเท่ากับ 200



4.2 ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อตัวแปรอิสระเป็นแบบที่ 2 (ศึกษาการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก)

ในส่วนนี้ผู้วิจัยได้ทำการศึกษาการสุญหายของข้อมูลในกรณีที่ตัวแปรอิสระเป็นแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ซึ่งเกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดเล็กซึ่งเท่ากับ 100 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 30 และ 90 ขนาดตัวอย่างเท่ากับ 50 100 และ 200 สัดส่วนของการสูญหายเท่ากับ 10% 20% และ 30% ระดับการสูญหายแบบ Nonignorable คือ ไม่มี ปานกลาง และสูง ซึ่งผลการวิจัยจะนำเสนอโดยแสดงค่า AMSE และ RE ที่ได้จากการประมาณค่าสูญหายของแต่ละวิธีการในตารางดังต่อไปนี้

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.2.1.1	แบบที่ 1	10
4.2.2.1	แบบที่ 1	30
4.2.3.1	แบบที่ 1	90

ตารางที่ 4.2.1.1-4.2.3.1 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า AMSE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.2.1.2	แบบที่ 1	10
4.2.2.2	แบบที่ 1	30
4.2.3.2	แบบที่ 1	90

ตารางที่ 4.2.1.2-4.2.3.2 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ภาพที่	ชนิดของตัวแปรอิสระ	ขนาดตัวอย่าง (n)
4.2.4	แบบที่ 1	50
4.2.5	แบบที่ 1	100
4.2.6	แบบที่ 1	200

ภาพที่ 4.2.4-4.2.6 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 10 30 และ 90 สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable

จากตารางและภาพที่ 4.2.1.1-4.2.3.1 ศึกษาการสูญหายของข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 30 และ 90 ตามลำดับ พบว่า ข้อมูลที่มีสัดส่วนของการสูญหาย 20%-30% และการสูญหายแบบ Nonignorable ในระดับปานกลางถึงสูง วิธีการ EM เป็นวิธีการที่มีประสิทธิภาพในการประมาณค่าสูญหายมากที่สุด เนื่องจากให้ค่า AMSE ต่ำที่สุด ซึ่งจะเห็นได้อย่างชัดเจนเมื่อข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

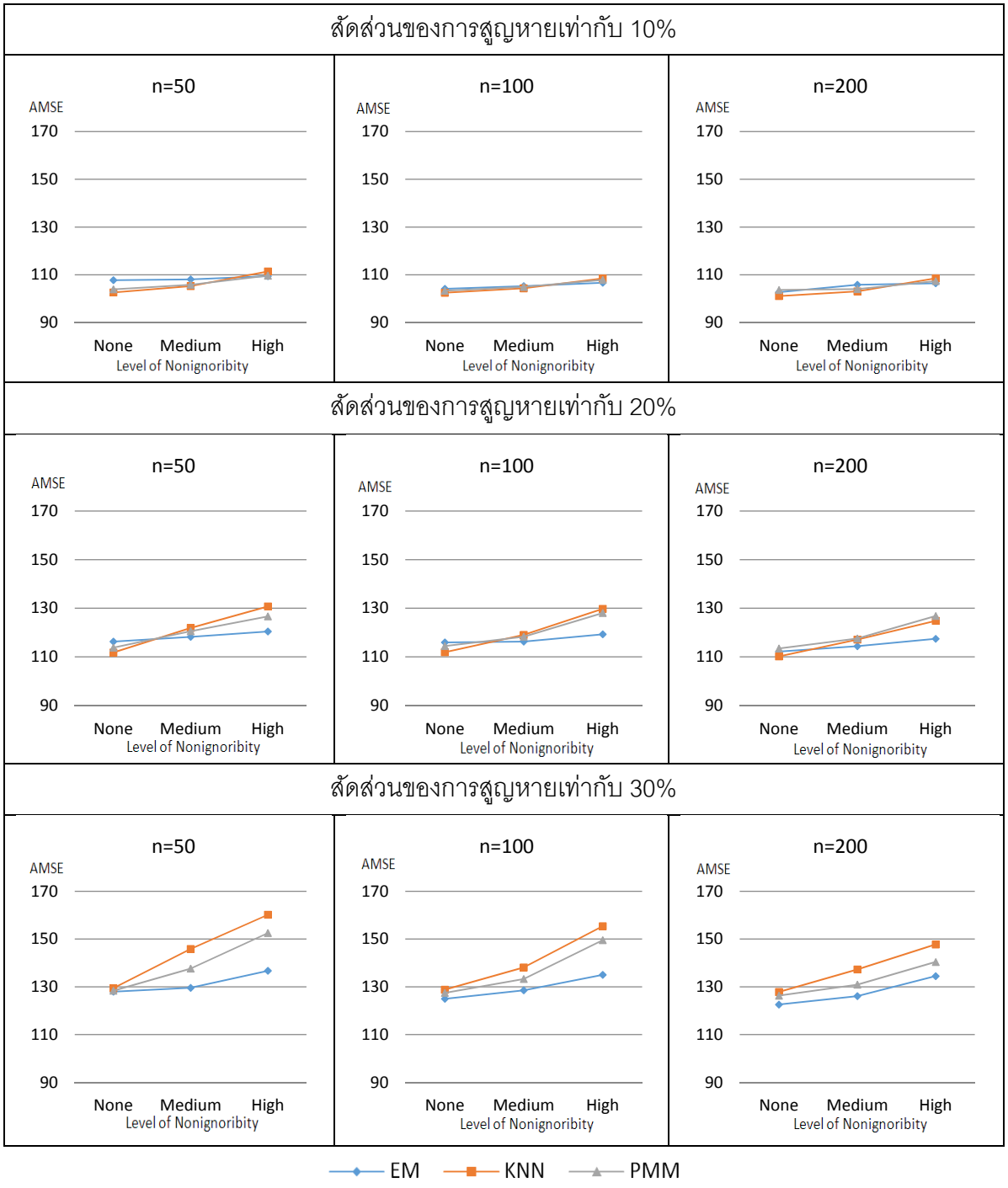
แต่ถ้าข้อมูลส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 วิธีการ KNN จะเป็นวิธีการที่มีประสิทธิภาพมากที่สุดในทุกกรณี

จากตารางและภาพที่ 4.2.1.2 - 4.2.3.2 เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีด้วยค่า RE โดยจะใช้วิธีการ EM เป็นวิธีการมาตรฐานในการเปรียบเทียบ โดยถ้า ค่า RE มากกว่า 1 แสดงว่า วิธีการ EM มีประสิทธิภาพในการประมาณค่าสูญหายน้อยกว่าวิธีการที่เปรียบเทียบ แต่ถ้า RE น้อยกว่า 1 แสดงว่าวิธีการ EM มีประสิทธิภาพดีกว่าวิธีการที่นำมาเปรียบเทียบ ซึ่งจากผลการวิจัยพบว่า สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable ที่สูงขึ้นจะส่งผลให้วิธีการ EM มีประสิทธิภาพดีกว่า KNN ซึ่งจะเห็นได้อย่างชัดเจนในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย(10)แต่ถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง(90) วิธีการประมาณแต่ละวิธีจะมีประสิทธิภาพใกล้เคียงกันมาก สังเกตได้จากตารางที่ 4.2.3.2 ที่พบว่า ค่า RE มีค่าใกล้เคียงกับ 1 ทุกกรณี ซึ่งสอดคล้องกับภาพที่ 4.2.4-4.2.6

ตารางที่ 4.2.1.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	107.78	**102.63**	103.95
		Medium	AMSE	108.12	**105.28**	105.85
		High	AMSE	**109.44**	111.44	109.66
	20	None	AMSE	116.26	**111.76**	113.78
		Medium	AMSE	**118.22**	121.89	120.55
		High	AMSE	**120.41**	130.73	126.64
	30	None	AMSE	**127.96**	129.42	128.36
		Medium	AMSE	**129.58**	145.84	137.63
		High	AMSE	**136.72**	160.19	152.50
100	10	None	AMSE	104.15	**102.50**	103.35
		Medium	AMSE	105.29	**104.37**	104.92
		High	AMSE	**106.70**	108.50	107.89
	20	None	AMSE	115.94	**111.91**	114.44
		Medium	AMSE	**116.26**	119.02	118.26
		High	AMSE	**119.26**	129.73	128.04
	30	None	AMSE	**124.98**	128.79	127.47
		Medium	AMSE	**128.54**	138.11	133.27
		High	AMSE	**135.01**	155.36	149.52
200	10	None	AMSE	102.75	**101.11**	103.68
		Medium	AMSE	105.87	**103.08**	104.06
		High	AMSE	**106.48**	108.54	107.39
	20	None	AMSE	112.15	**110.17**	113.44
		Medium	AMSE	**114.37**	117.09	117.55
		High	AMSE	**117.42**	124.82	126.80
	30	None	AMSE	**122.58**	127.83	126.32
		Medium	AMSE	**126.10**	137.28	130.92
		High	AMSE	**134.48**	147.75	140.41

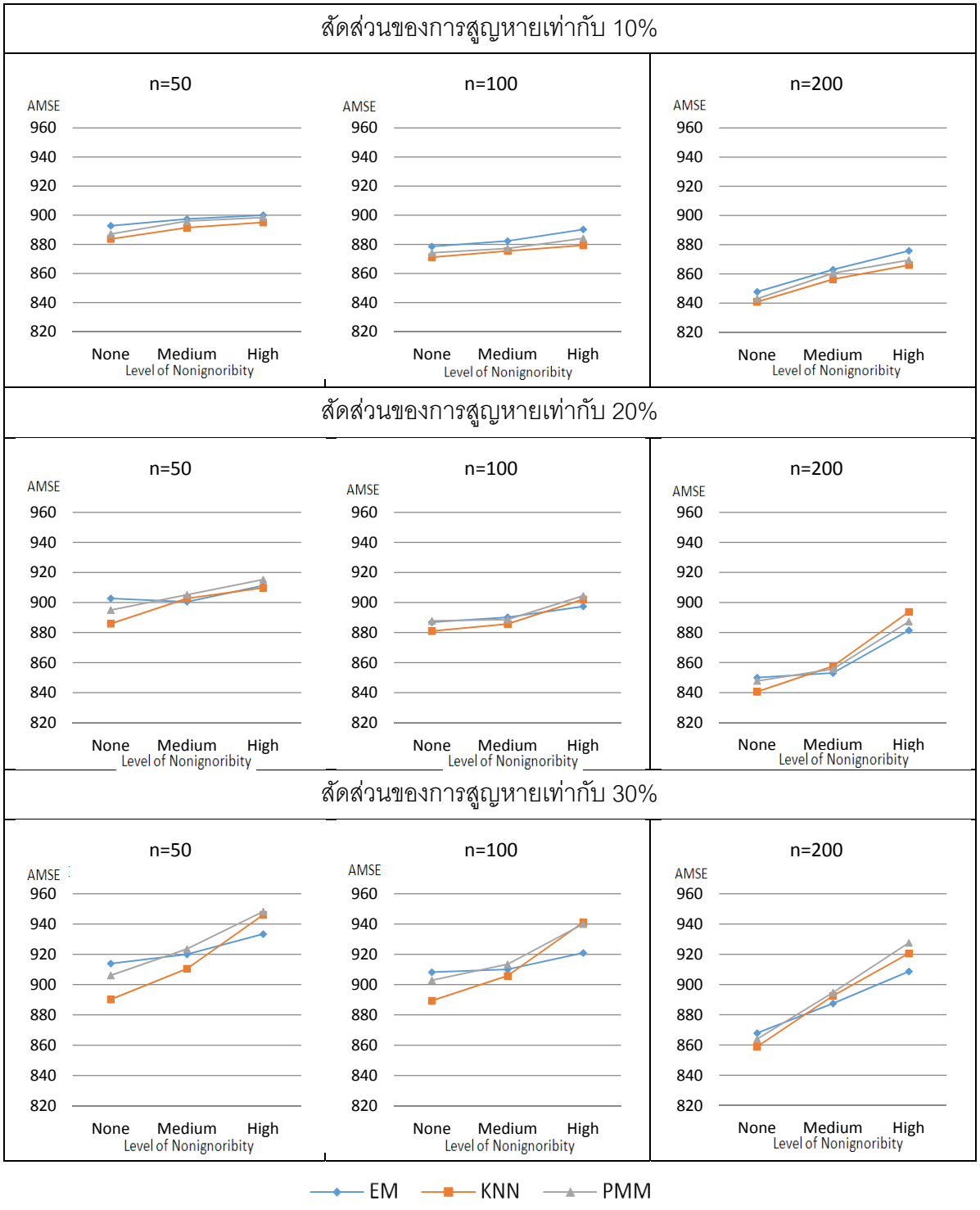
<u>ภาพที่</u> 4.2.1.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.2.2.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	892.76	**883.70**	887.09
		Medium	AMSE	897.46	**891.53**	895.96
		High	AMSE	900.07	**895.05**	898.43
	20	None	AMSE	902.71	**885.90**	894.98
		Medium	AMSE	**900.45**	902.72	905.18
		High	AMSE	911.21	**909.77**	915.21
	30	None	AMSE	913.90	**890.22**	906.04
		Medium	AMSE	919.97	**910.53**	923.53
		High	AMSE	**933.40**	946.09	948.20
100	10	None	AMSE	878.58	**871.17**	874.18
		Medium	AMSE	882.29	**875.48**	877.19
		High	AMSE	890.25	**879.30**	884.11
	20	None	AMSE	886.89	**880.99**	887.76
		Medium	AMSE	890.17	**885.70**	888.62
		High	AMSE	**897.39**	902.06	904.42
	30	None	AMSE	908.20	**889.39**	902.84
		Medium	AMSE	910.03	**905.79**	913.45
		High	AMSE	**920.92**	941.13	939.96
200	10	None	AMSE	847.71	**840.81**	842.93
		Medium	AMSE	862.92	**856.30**	860.42
		High	AMSE	875.79	**866.02**	869.39
	20	None	AMSE	849.99	**840.58**	847.74
		Medium	AMSE	**853.01**	857.57	855.64
		High	AMSE	**881.47**	893.73	887.27
	30	None	AMSE	867.92	**858.87**	863.92
		Medium	AMSE	**887.50**	892.52	894.68
		High	AMSE	**908.66**	920.57	927.55

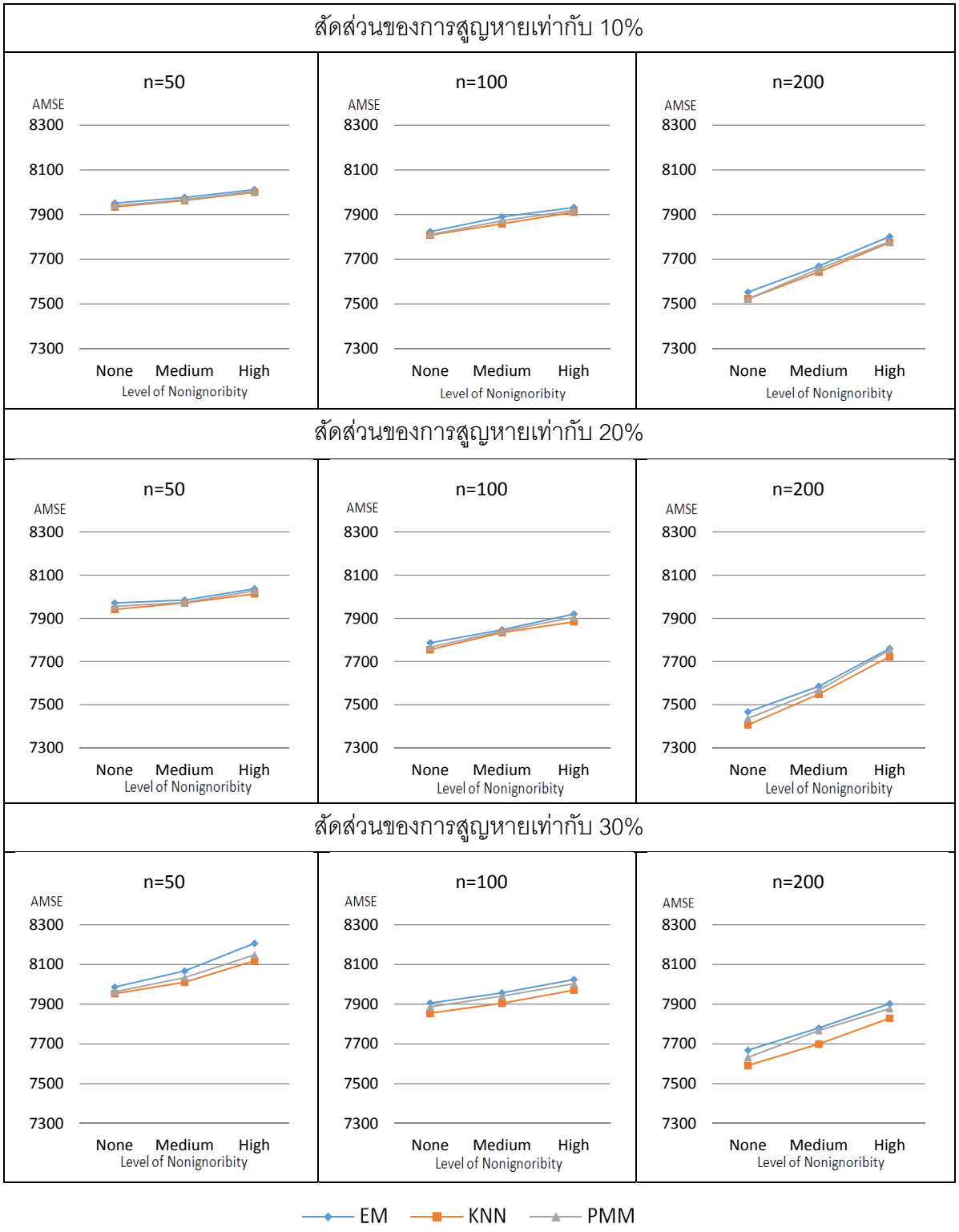
<u>ภาพที่</u> 4.2.2.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



ตารางที่ 4.2.3.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	7951.33	**7933.95**	7939.79
		Medium	AMSE	7976.76	**7963.44**	7967.67
		High	AMSE	8011.63	**7999.49**	8004.04
	20	None	AMSE	7970.75	**7941.42**	7955.67
		Medium	AMSE	7984.78	**7971.05**	7973.13
		High	AMSE	8036.81	**8013.25**	8027.57
	30	None	AMSE	7985.93	**7953.00**	7961.81
		Medium	AMSE	8067.27	**8010.32**	8033.61
		High	AMSE	8206.10	**8118.46**	8148.08
100	10	None	AMSE	7823.90	**7807.89**	7810.62
		Medium	AMSE	7889.94	**7858.67**	7871.98
		High	AMSE	7931.97	**7910.88**	7918.92
	20	None	AMSE	7786.30	**7754.63**	7765.94
		Medium	AMSE	7847.05	**7834.52**	7840.23
		High	AMSE	7919.85	**7883.90**	7904.25
	30	None	AMSE	7905.48	**7854.71**	7886.51
		Medium	AMSE	7957.08	**7904.65**	7940.74
		High	AMSE	8023.90	**7970.73**	8003.05
200	10	None	AMSE	7552.61	**7522.01**	7522.00
		Medium	AMSE	7669.37	**7642.74**	7655.93
		High	AMSE	7801.08	**7775.10**	7778.66
	20	None	AMSE	7467.03	**7406.61**	7436.64
		Medium	AMSE	7585.26	**7548.04**	7567.29
		High	AMSE	7760.31	**7722.68**	7753.28
	30	None	AMSE	7667.95	**7591.60**	7632.21
		Medium	AMSE	7779.80	**7699.36**	7767.06
		High	AMSE	7902.15	**7829.20**	7877.27

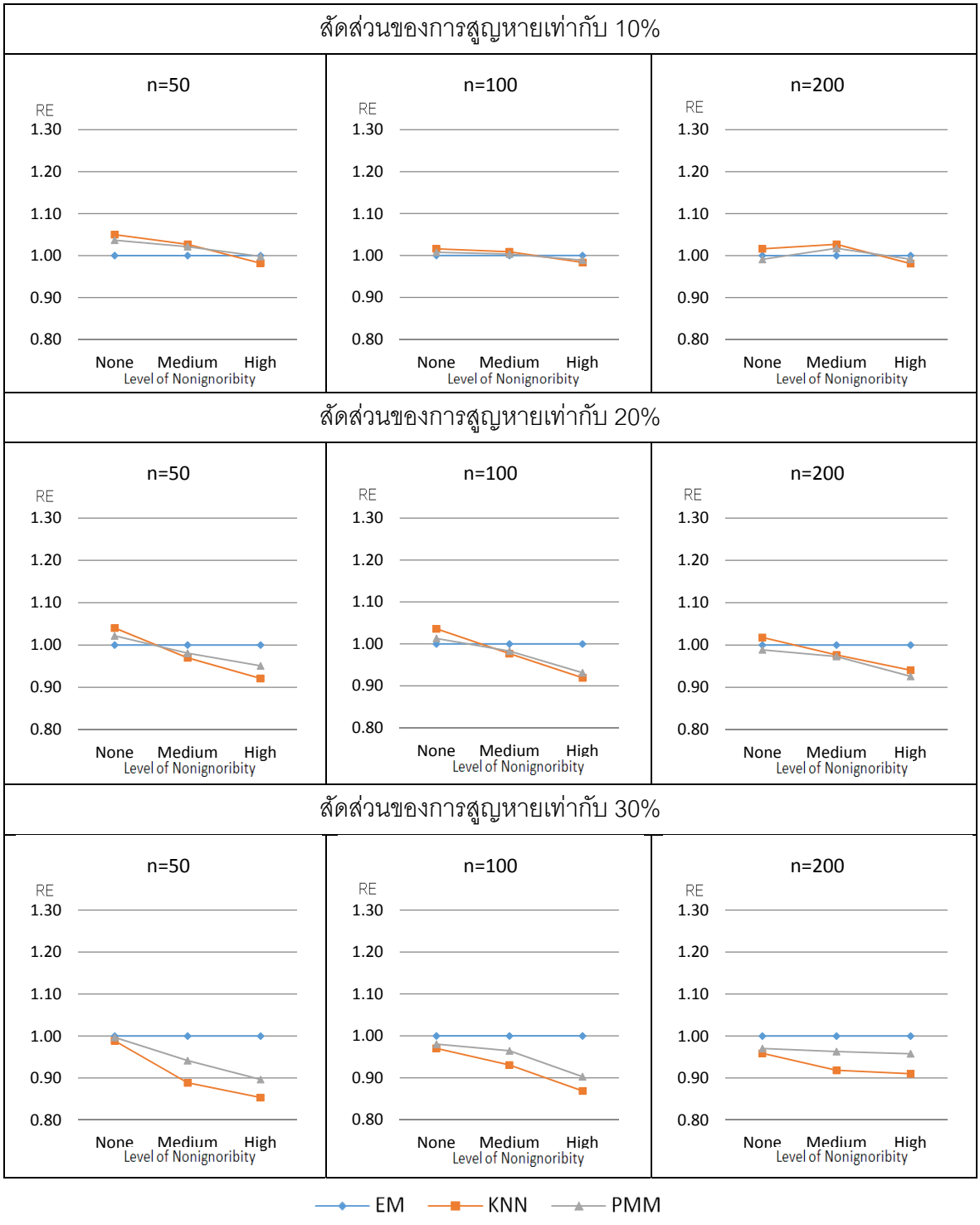
<u>ภาพที่</u> 4.2.3.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



ตารางที่ 4.2.1.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.050	1.037
		Medium	RE	1.000	1.027	1.021
		High	RE	1.000	0.982	0.998
	20	None	RE	1.000	1.040	1.022
		Medium	RE	1.000	0.970	0.981
		High	RE	1.000	0.921	0.951
	30	None	RE	1.000	0.989	0.997
		Medium	RE	1.000	0.889	0.942
		High	RE	1.000	0.853	0.896
100	10	None	RE	1.000	1.016	1.008
		Medium	RE	1.000	1.009	1.004
		High	RE	1.000	0.983	0.989
	20	None	RE	1.000	1.036	1.013
		Medium	RE	1.000	0.977	0.983
		High	RE	1.000	0.919	0.931
	30	None	RE	1.000	0.970	0.980
		Medium	RE	1.000	0.931	0.965
		High	RE	1.000	0.869	0.903
200	10	None	RE	1.000	1.016	0.991
		Medium	RE	1.000	1.027	1.017
		High	RE	1.000	0.981	0.992
	20	None	RE	1.000	1.018	0.989
		Medium	RE	1.000	0.977	0.973
		High	RE	1.000	0.941	0.926
	30	None	RE	1.000	0.959	0.970
		Medium	RE	1.000	0.919	0.963
		High	RE	1.000	0.910	0.958

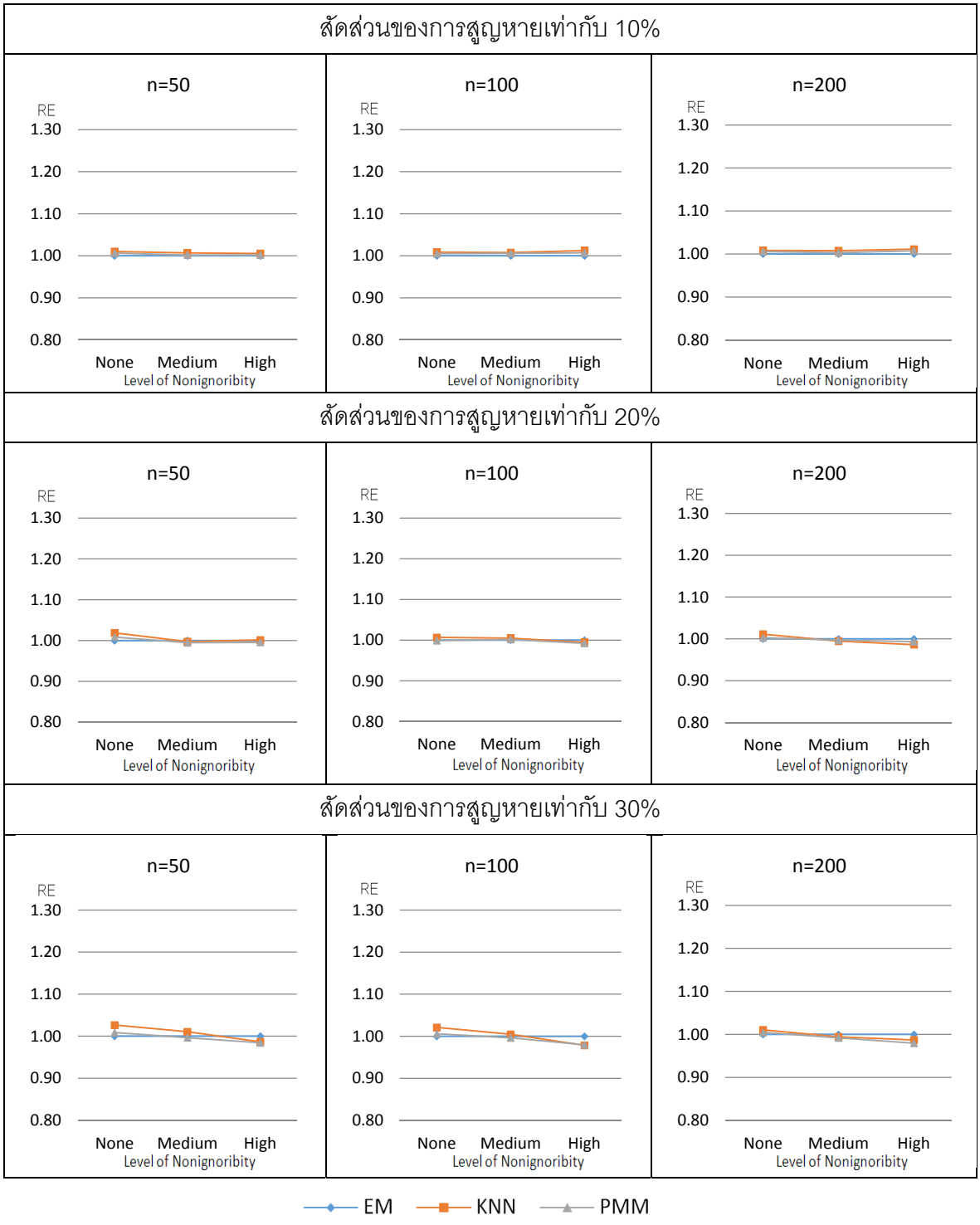
<u>ภาพที่</u> 4.2.1.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.2.2.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.010	1.006
		Medium	RE	1.000	1.007	1.002
		High	RE	1.000	1.006	1.002
	20	None	RE	1.000	1.019	1.009
		Medium	RE	1.000	0.997	0.995
		High	RE	1.000	1.002	0.996
	30	None	RE	1.000	1.027	1.009
		Medium	RE	1.000	1.010	0.996
		High	RE	1.000	0.987	0.984
100	10	None	RE	1.000	1.008	1.005
		Medium	RE	1.000	1.008	1.006
		High	RE	1.000	1.012	1.007
	20	None	RE	1.000	1.007	0.999
		Medium	RE	1.000	1.005	1.002
		High	RE	1.000	0.995	0.992
	30	None	RE	1.000	1.021	1.006
		Medium	RE	1.000	1.005	0.996
		High	RE	1.000	0.979	0.980
200	10	None	RE	1.000	1.008	1.006
		Medium	RE	1.000	1.008	1.003
		High	RE	1.000	1.011	1.007
	20	None	RE	1.000	1.011	1.003
		Medium	RE	1.000	0.995	0.997
		High	RE	1.000	0.986	0.993
	30	None	RE	1.000	1.011	1.005
		Medium	RE	1.000	0.994	0.992
		High	RE	1.000	0.987	0.980

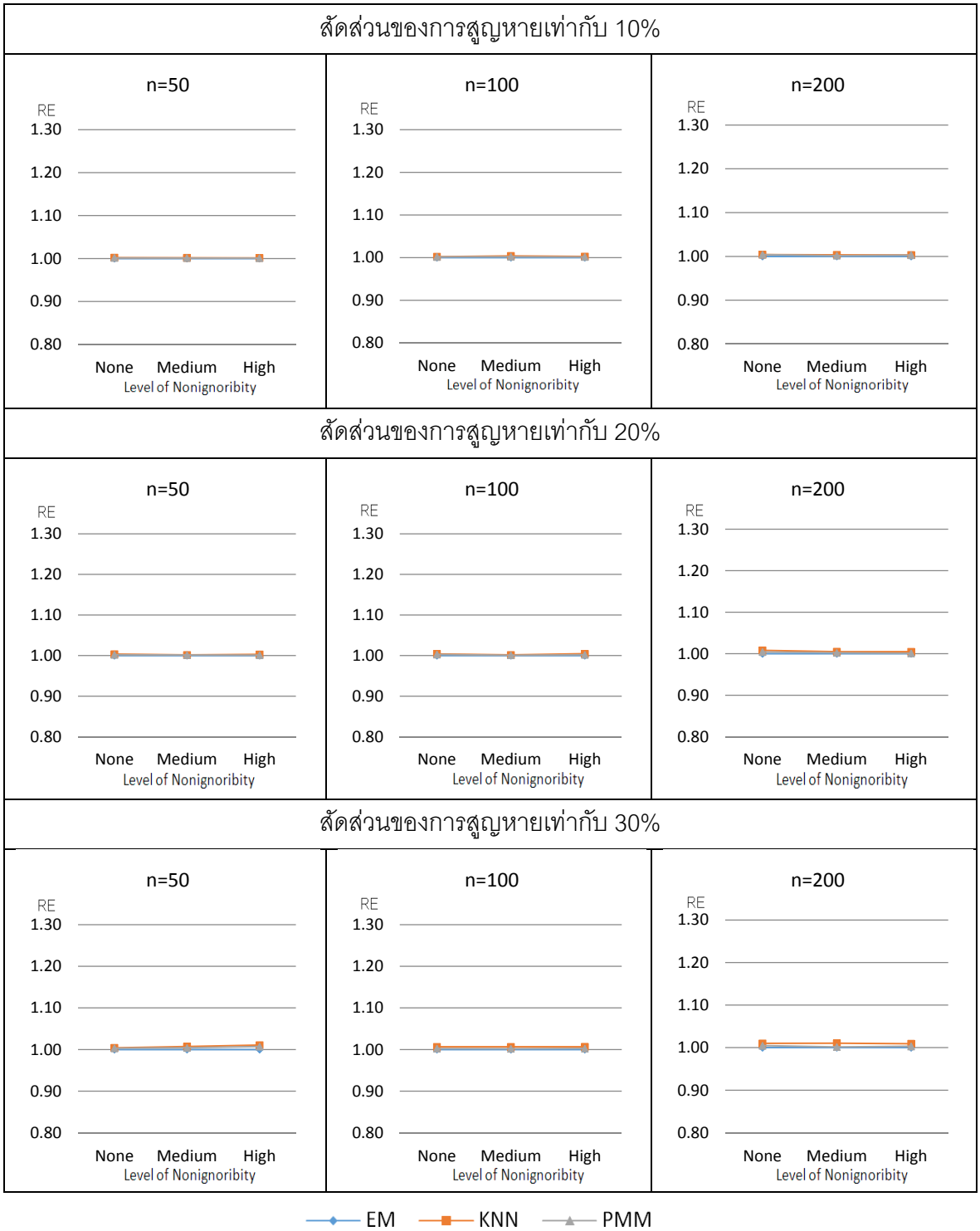
<u>ภาพที่ 4.2.2.2</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



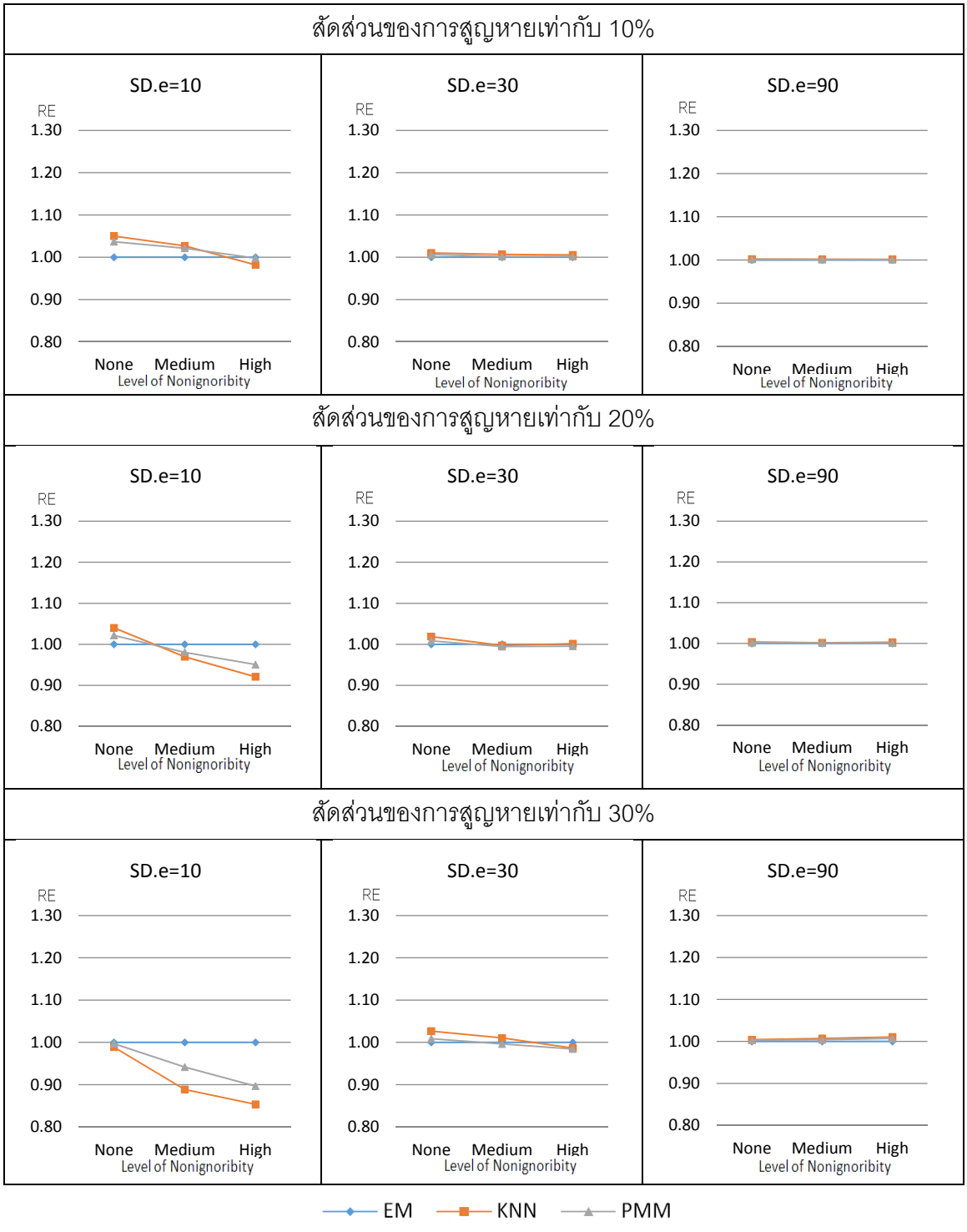
ตารางที่ 4.2.3.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.002	1.001
		Medium	RE	1.000	1.002	1.001
		High	RE	1.000	1.002	1.001
	20	None	RE	1.000	1.004	1.002
		Medium	RE	1.000	1.002	1.001
		High	RE	1.000	1.003	1.001
	30	None	RE	1.000	1.004	1.003
		Medium	RE	1.000	1.007	1.004
		High	RE	1.000	1.011	1.007
100	10	None	RE	1.000	1.002	1.002
		Medium	RE	1.000	1.004	1.002
		High	RE	1.000	1.003	1.002
	20	None	RE	1.000	1.004	1.003
		Medium	RE	1.000	1.002	1.001
		High	RE	1.000	1.005	1.002
	30	None	RE	1.000	1.006	1.002
		Medium	RE	1.000	1.007	1.002
		High	RE	1.000	1.007	1.003
200	10	None	RE	1.000	1.004	1.004
		Medium	RE	1.000	1.003	1.002
		High	RE	1.000	1.003	1.003
	20	None	RE	1.000	1.008	1.004
		Medium	RE	1.000	1.005	1.002
		High	RE	1.000	1.005	1.001
	30	None	RE	1.000	1.010	1.005
		Medium	RE	1.000	1.010	1.002
		High	RE	1.000	1.009	1.003

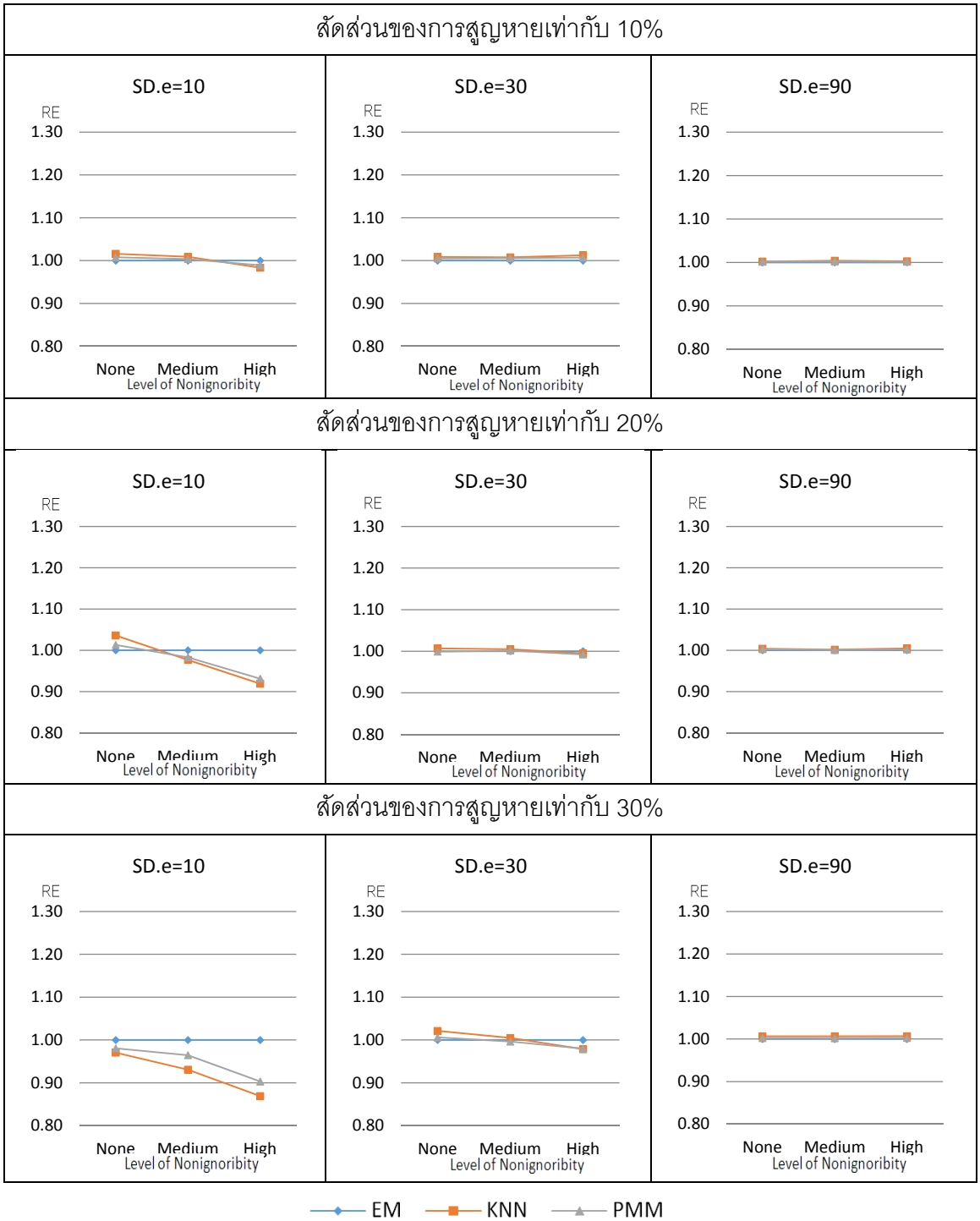
<u>ภาพที่</u> 4.2.3.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



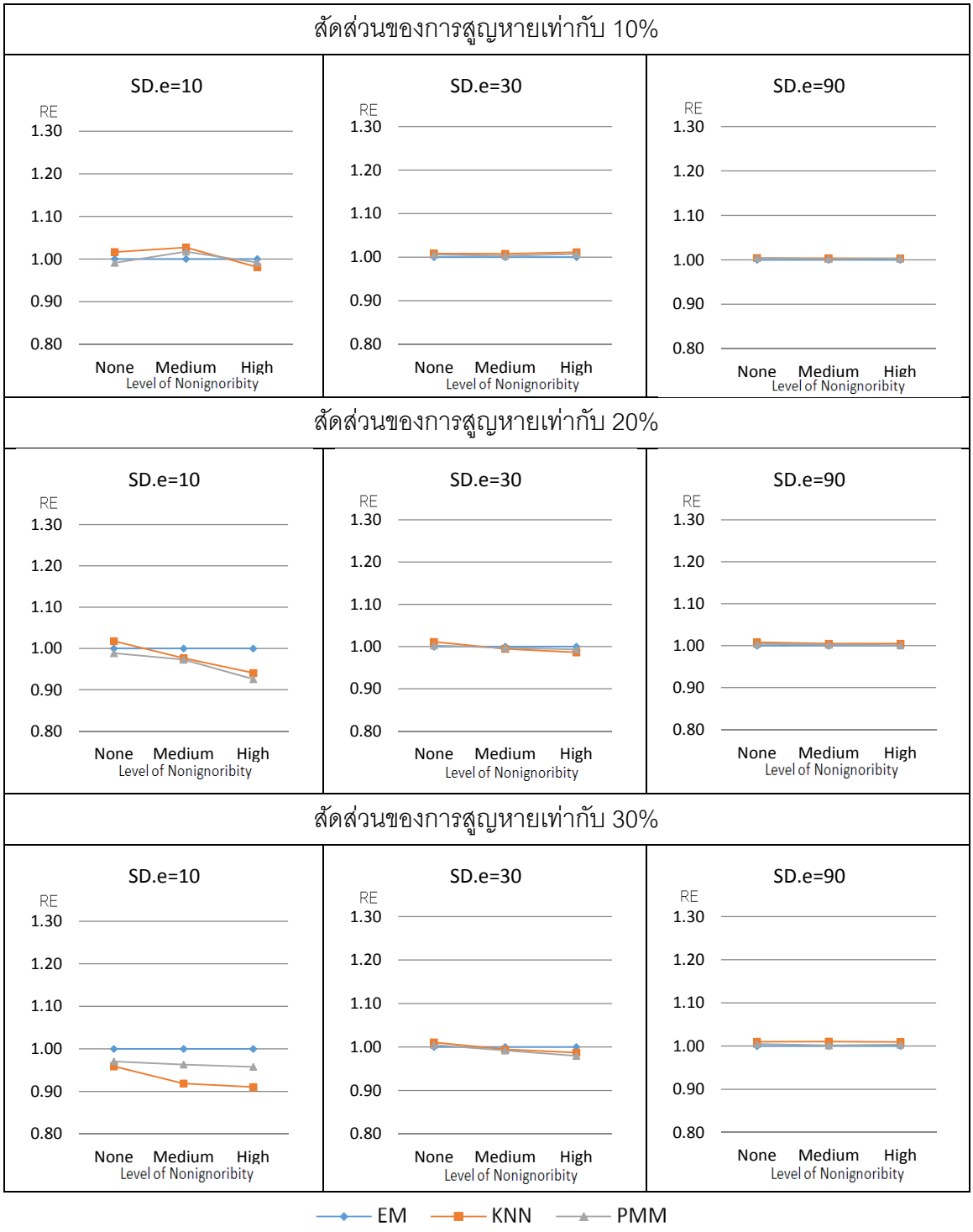
<u>ภาพที่ 4.2.4</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และขนาดตัวอย่างเท่ากับ 50



<u>ภาพที่ 4.2.5</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และขนาดตัวอย่างเท่ากับ 100



<u>ภาพที่ 4.2.6</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 1 เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>เล็ก</u>) และขนาดตัวอย่างเท่ากับ 200



4.3 ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อตัวแปรอิสระเป็นแบบที่ 2 (ศึกษาการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดกลาง)

ในส่วนนี้ผู้วิจัยได้ทำการศึกษาการสุญหายของข้อมูลในกรณีที่ตัวแปรอิสระเป็นแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ซึ่งเกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง ซึ่งเท่ากับ 300 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 30 และ 90 ขนาดตัวอย่างเท่ากับ 50 100 และ 200 สัดส่วนของการสูญหายเท่ากับ 10% 20% และ 30% ระดับการสูญหายแบบ Nonignorable คือ ไม่มี ปานกลาง และสูง ซึ่งผลการวิจัยจะนำเสนอโดยแสดงค่า AMSE และ RE ที่ได้จากการประมาณค่าสูญหายของแต่ละวิธีการในตารางดังต่อไปนี้

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.3.1.1	แบบที่ 1	10
4.3.2.1	แบบที่ 1	30
4.3.3.1	แบบที่ 1	90

ตารางที่ 4.3.1.1-4.3.3.1 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า AMSE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.3.1.2	แบบที่ 1	10
4.3.2.2	แบบที่ 1	30
4.3.3.2	แบบที่ 1	90

ตารางที่ 4.3.1.2-4.3.3.2 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ภาพที่	ชนิดของตัวแปรอิสระ	ขนาดตัวอย่าง (n)
4.3.4	แบบที่ 1	50
4.3.5	แบบที่ 1	100
4.3.6	แบบที่ 1	200

ภาพที่ 4.3.4-4.3.6 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 10 30 และ 90 สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable

จากตารางและภาพที่ 4.3.1.1-4.3.3.1 พบว่า โดยส่วนใหญ่วิธีการ KNN คือวิธีการที่ให้ค่า AMSE ต่ำที่สุดจึงมีประสิทธิภาพมากที่สุด แต่ในกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย(10) ที่สัดส่วนของการสูญหายสูง(30%) และการสูญหายแบบ Nonignorable ในระดับปานกลางถึงสูง วิธีการ EM เป็นวิธีการที่มีประสิทธิภาพในการประมาณค่าสูญหายมากที่สุด เนื่องจากให้ค่า AMSE ต่ำที่สุด

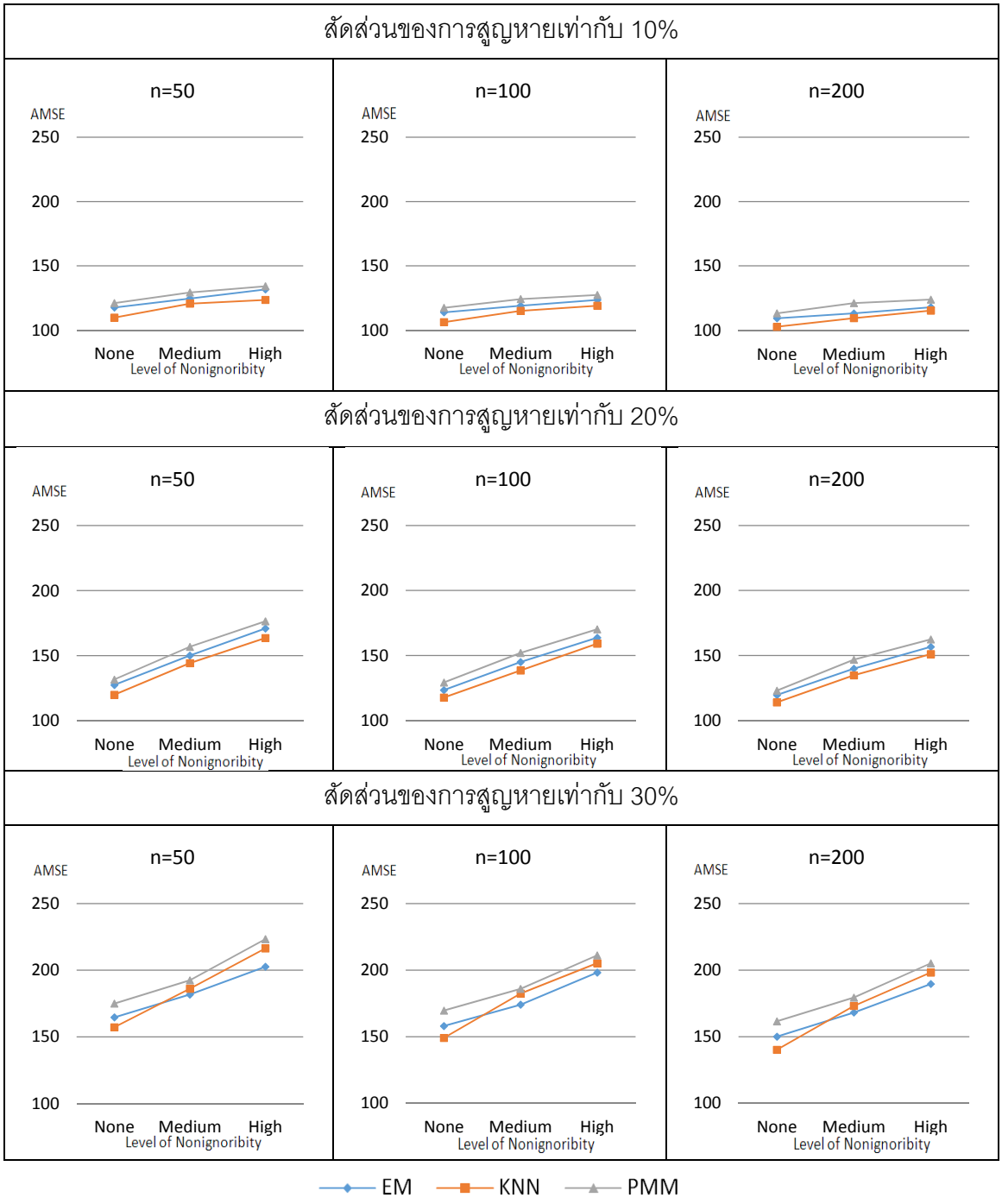
แต่ถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 และ 90 วิธีการ KNN จะเป็นวิธีการที่มีประสิทธิภาพมากที่สุดในทุกกรณี

จากตารางและภาพที่ 4.3.1.2 - 4.3.3.2 เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีด้วยค่า RE โดยจะใช้วิธีการ EM เป็นวิธีการมาตรฐานในการเปรียบเทียบ โดยถ้า ค่า RE มากกว่า 1 แสดงว่า วิธีการ EM มีประสิทธิภาพในการประมาณค่าสูญหายน้อยกว่าวิธีการที่นำมาเปรียบเทียบ แต่ถ้า RE น้อยกว่า 1 แสดงว่าวิธีการ EM มีประสิทธิภาพมากกว่าวิธีการที่นำมาเปรียบเทียบ ซึ่งจากผลการวิจัยพบว่า ที่สัดส่วนของการสูญหาย 30% และระดับการสูญหายแบบ Nonignorable ปานกลางและสูง จะส่งผลให้วิธีการ EM มีประสิทธิภาพดีกว่า KNN ซึ่งจะเห็นได้อย่างชัดเจนในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย (10) แต่ถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง(90) วิธีการประมาณแต่ละวิธีจะมีประสิทธิภาพใกล้เคียงกันมาก แต่อย่างไรก็ตามวิธีการ KNN ก็ยังเป็นวิธีการที่มีประสิทธิภาพมากที่สุด

ตารางที่ 4.3.1.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	117.59	**109.75**	121.04
		Medium	AMSE	124.57	**120.62**	129.35
		High	AMSE	131.68	**123.54**	134.17
	20	None	AMSE	127.56	**120.04**	131.65
		Medium	AMSE	150.19	**144.32**	156.84
		High	AMSE	170.90	**163.62**	176.47
	30	None	AMSE	164.60	**157.20**	174.98
		Medium	AMSE	**181.78**	186.15	192.43
		High	AMSE	**202.63**	216.39	223.34
100	10	None	AMSE	113.87	**106.27**	117.43
		Medium	AMSE	119.03	**115.01**	124.20
		High	AMSE	123.55	**119.00**	127.40
	20	None	AMSE	123.40	**117.63**	129.19
		Medium	AMSE	144.99	**138.49**	151.97
		High	AMSE	163.65	**159.13**	170.17
	30	None	AMSE	158.06	**149.10**	169.65
		Medium	AMSE	**174.06**	182.42	185.96
		High	AMSE	**198.22**	205.22	211.11
200	10	None	AMSE	109.25	**102.64**	113.15
		Medium	AMSE	113.07	**109.37**	121.09
		High	AMSE	117.82	**115.24**	123.80
	20	None	AMSE	119.52	**113.89**	122.92
		Medium	AMSE	139.91	**134.72**	146.75
		High	AMSE	156.61	**151.05**	162.36
	30	None	AMSE	150.06	**140.20**	161.57
		Medium	AMSE	**168.17**	173.15	179.41
		High	AMSE	**189.62**	198.23	205.17

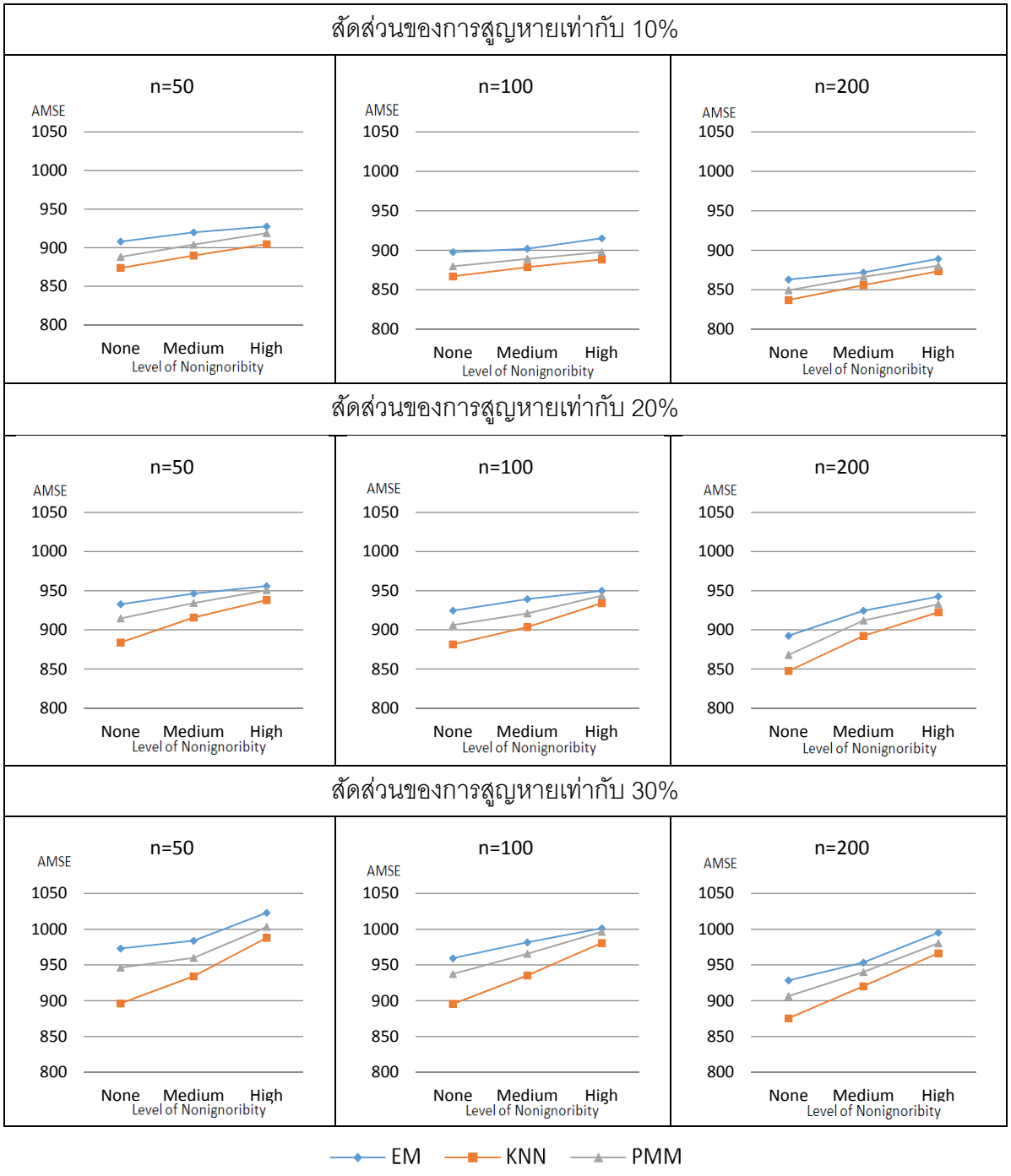
<u>ภาพที่</u> 4.3.1.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.3.2.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	907.87	**873.73**	888.03
		Medium	AMSE	919.87	**889.79**	903.92
		High	AMSE	927.43	**904.62**	918.83
	20	None	AMSE	932.73	**883.89**	914.63
		Medium	AMSE	946.32	**915.72**	934.30
		High	AMSE	955.88	**937.94**	950.65
	30	None	AMSE	973.08	**896.21**	946.23
		Medium	AMSE	983.79	**934.25**	959.91
		High	AMSE	1022.99	**988.03**	1003.21
100	10	None	AMSE	897.52	**866.79**	879.58
		Medium	AMSE	901.82	**878.57**	888.94
		High	AMSE	915.09	**888.26**	897.70
	20	None	AMSE	924.51	**881.45**	906.11
		Medium	AMSE	939.30	**903.54**	920.96
		High	AMSE	949.97	**934.07**	943.69
	30	None	AMSE	959.52	**895.59**	937.43
		Medium	AMSE	981.63	**935.37**	965.59
		High	AMSE	1001.17	**980.72**	996.32
200	10	None	AMSE	862.88	**836.85**	849.23
		Medium	AMSE	871.80	**855.72**	866.17
		High	AMSE	889.10	**873.46**	880.36
	20	None	AMSE	892.39	**847.45**	867.94
		Medium	AMSE	924.37	**892.21**	911.92
		High	AMSE	942.46	**922.53**	932.70
	30	None	AMSE	928.43	**875.42**	906.38
		Medium	AMSE	953.49	**920.20**	940.02
		High	AMSE	995.09	**966.49**	980.21

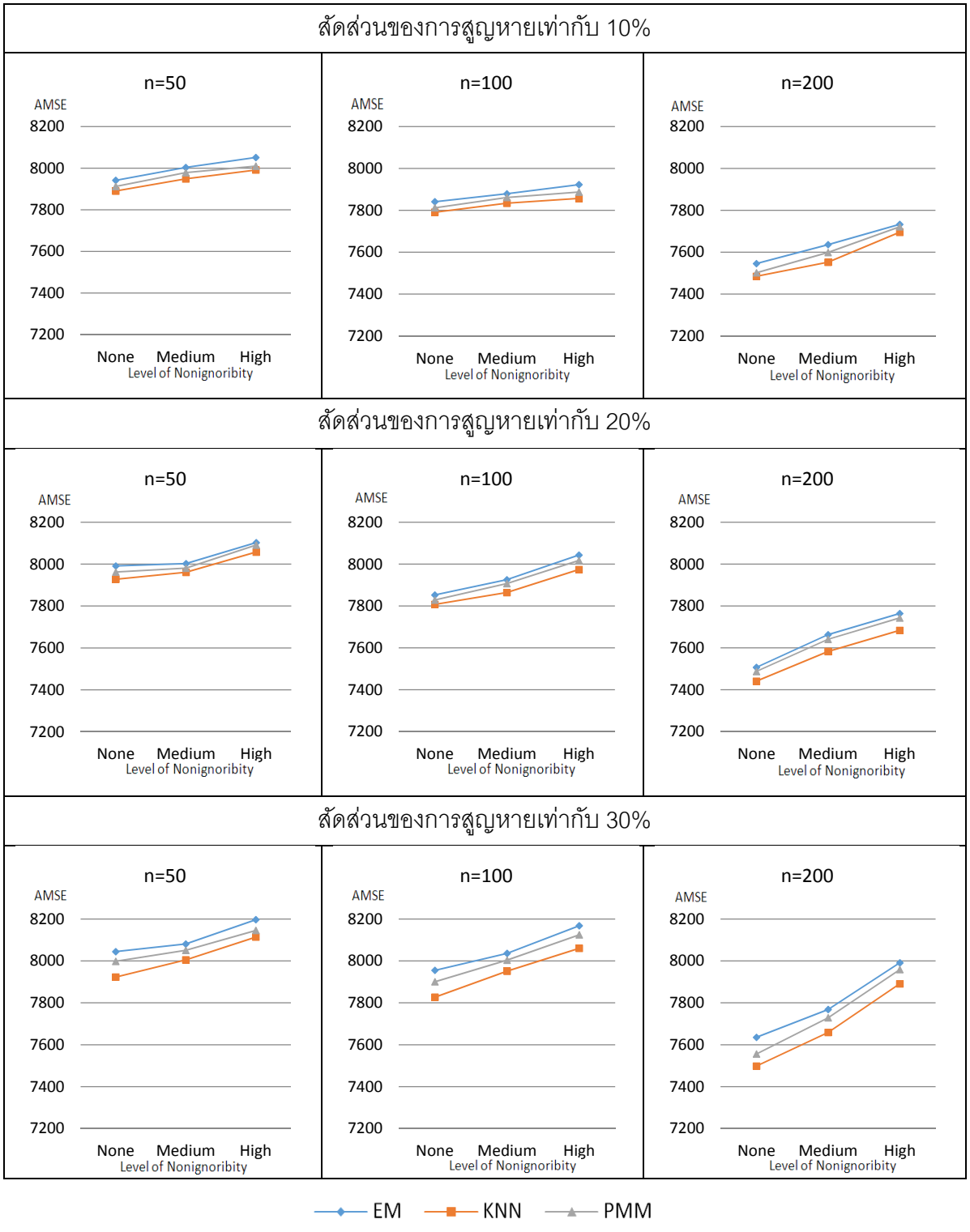
<u>ภาพที่</u> 4.3.2.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



ตารางที่ 4.3.3.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	7941.66	**7890.06**	7912.32
		Medium	AMSE	8003.43	**7948.68**	7978.56
		High	AMSE	8051.44	**7991.29**	8009.81
	20	None	AMSE	7991.50	**7927.97**	7962.40
		Medium	AMSE	8002.63	**7961.33**	7980.87
		High	AMSE	8102.83	**8057.35**	8090.91
	30	None	AMSE	8044.38	**7922.68**	7998.08
		Medium	AMSE	8080.91	**8004.89**	8050.51
		High	AMSE	8196.82	**8113.88**	8145.69
100	10	None	AMSE	7840.79	**7790.65**	7811.74
		Medium	AMSE	7878.92	**7833.85**	7861.31
		High	AMSE	7922.43	**7856.63**	7887.03
	20	None	AMSE	7852.51	**7807.00**	7828.35
		Medium	AMSE	7925.45	**7864.16**	7906.69
		High	AMSE	8043.04	**7973.86**	8017.37
	30	None	AMSE	7955.15	**7826.71**	7900.46
		Medium	AMSE	8035.87	**7951.17**	8003.63
		High	AMSE	8167.76	**8060.38**	8124.44
200	10	None	AMSE	7545.25	**7484.38**	7501.63
		Medium	AMSE	7635.99	**7551.97**	7598.57
		High	AMSE	7732.79	**7695.00**	7721.25
	20	None	AMSE	7506.61	**7440.48**	7486.98
		Medium	AMSE	7663.09	**7582.64**	7640.75
		High	AMSE	7764.12	**7683.18**	7742.59
	30	None	AMSE	7635.37	**7498.05**	7556.09
		Medium	AMSE	7768.35	**7658.84**	7728.63
		High	AMSE	7991.17	**7890.94**	7958.71

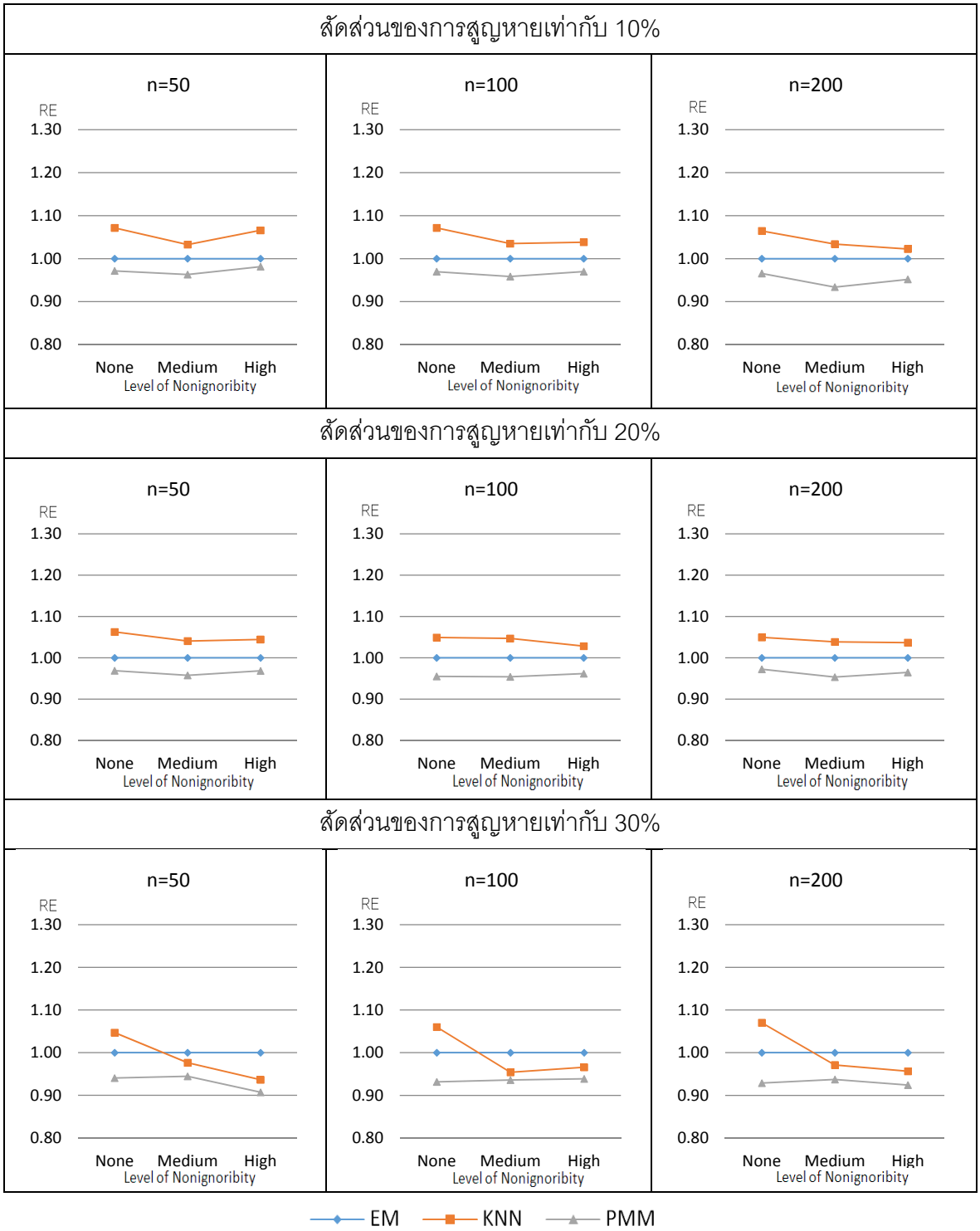
<u>ภาพที่</u> 4.3.3.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



ตารางที่ 4.3.1.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.071	0.972
		Medium	RE	1.000	1.033	0.963
		High	RE	1.000	1.066	0.981
	20	None	RE	1.000	1.063	0.969
		Medium	RE	1.000	1.041	0.958
		High	RE	1.000	1.044	0.968
	30	None	RE	1.000	1.047	0.941
		Medium	RE	1.000	0.976	0.945
		High	RE	1.000	0.936	0.907
100	10	None	RE	1.000	1.072	0.970
		Medium	RE	1.000	1.035	0.958
		High	RE	1.000	1.038	0.970
	20	None	RE	1.000	1.049	0.955
		Medium	RE	1.000	1.047	0.954
		High	RE	1.000	1.028	0.962
	30	None	RE	1.000	1.060	0.932
		Medium	RE	1.000	0.954	0.936
		High	RE	1.000	0.966	0.939
200	10	None	RE	1.000	1.064	0.966
		Medium	RE	1.000	1.034	0.934
		High	RE	1.000	1.022	0.952
	20	None	RE	1.000	1.049	0.972
		Medium	RE	1.000	1.039	0.953
		High	RE	1.000	1.037	0.965
	30	None	RE	1.000	1.070	0.929
		Medium	RE	1.000	0.971	0.937
		High	RE	1.000	0.957	0.924

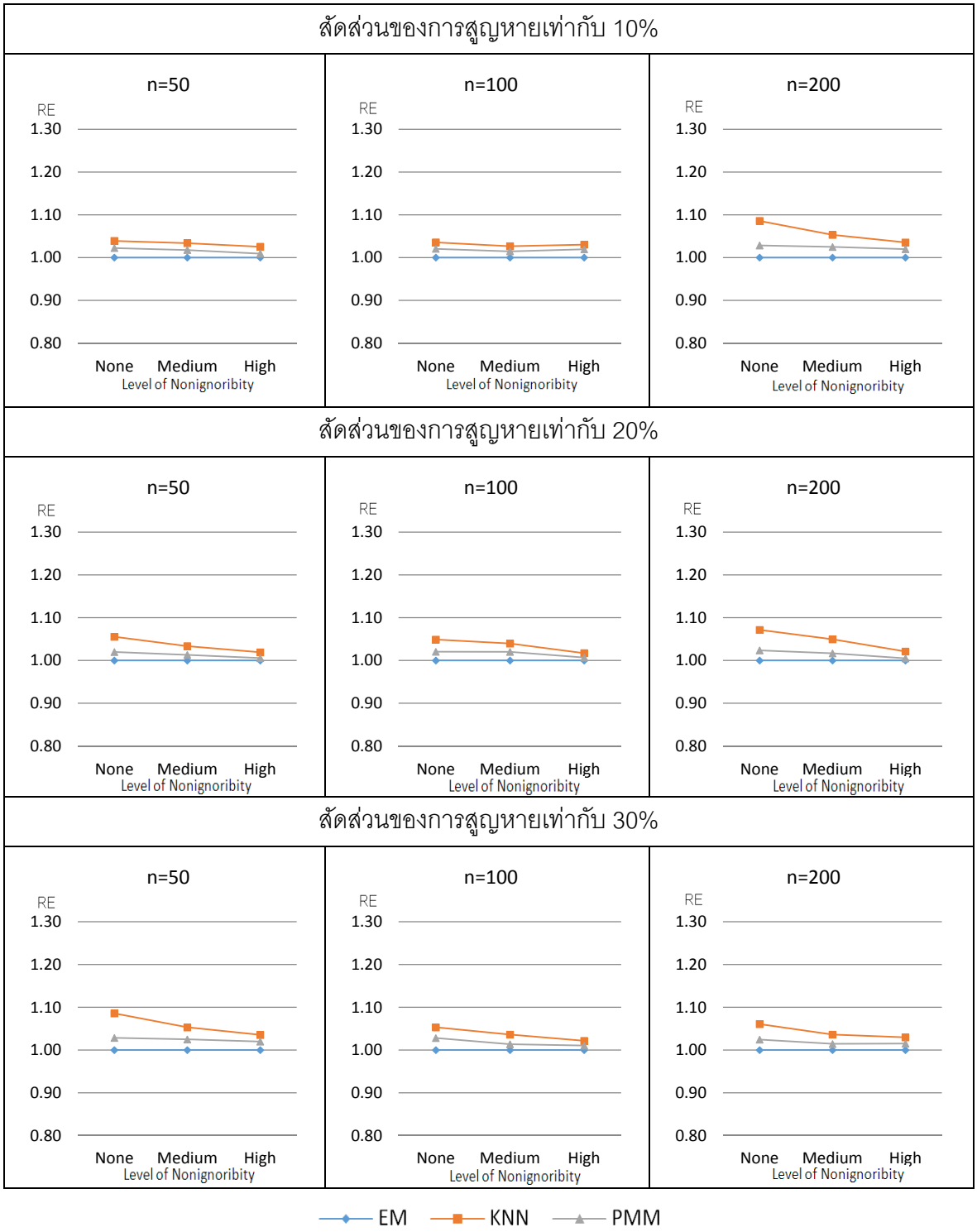
<u>ภาพที่</u> 4.3.1.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.3.2.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.039	1.022
		Medium	RE	1.000	1.034	1.018
		High	RE	1.000	1.025	1.009
	20	None	RE	1.000	1.055	1.020
		Medium	RE	1.000	1.033	1.013
		High	RE	1.000	1.019	1.006
	30	None	RE	1.000	1.086	1.028
		Medium	RE	1.000	1.053	1.025
		High	RE	1.000	1.035	1.020
100	10	None	RE	1.000	1.035	1.020
		Medium	RE	1.000	1.026	1.014
		High	RE	1.000	1.030	1.019
	20	None	RE	1.000	1.049	1.020
		Medium	RE	1.000	1.040	1.020
		High	RE	1.000	1.017	1.007
	30	None	RE	1.000	1.071	1.024
		Medium	RE	1.000	1.049	1.017
		High	RE	1.000	1.021	1.005
200	10	None	RE	1.000	1.031	1.016
		Medium	RE	1.000	1.019	1.007
		High	RE	1.000	1.018	1.010
	20	None	RE	1.000	1.053	1.028
		Medium	RE	1.000	1.036	1.014
		High	RE	1.000	1.022	1.010
	30	None	RE	1.000	1.061	1.024
		Medium	RE	1.000	1.036	1.014
		High	RE	1.000	1.030	1.015

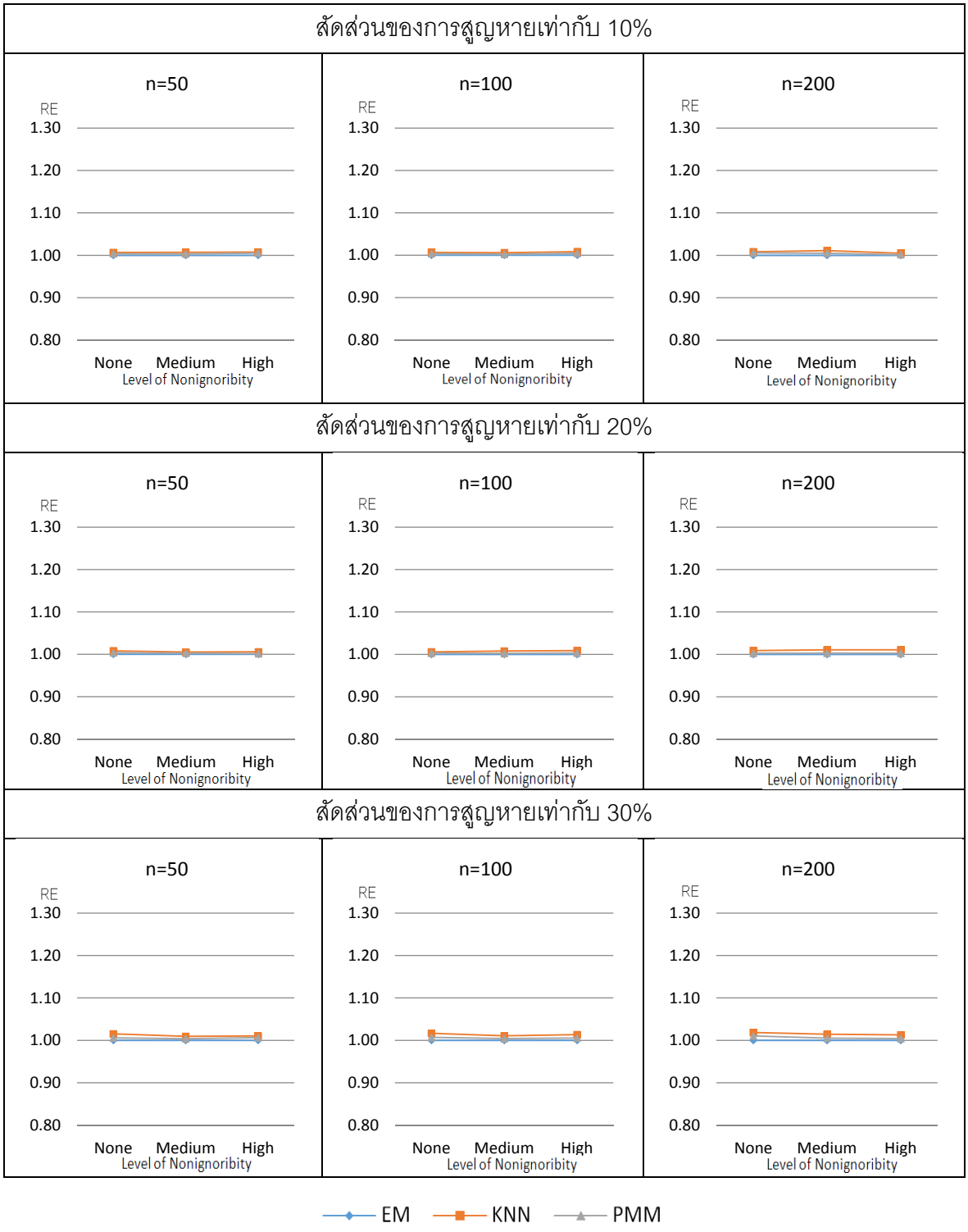
<u>ภาพที่</u> 4.3.2.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



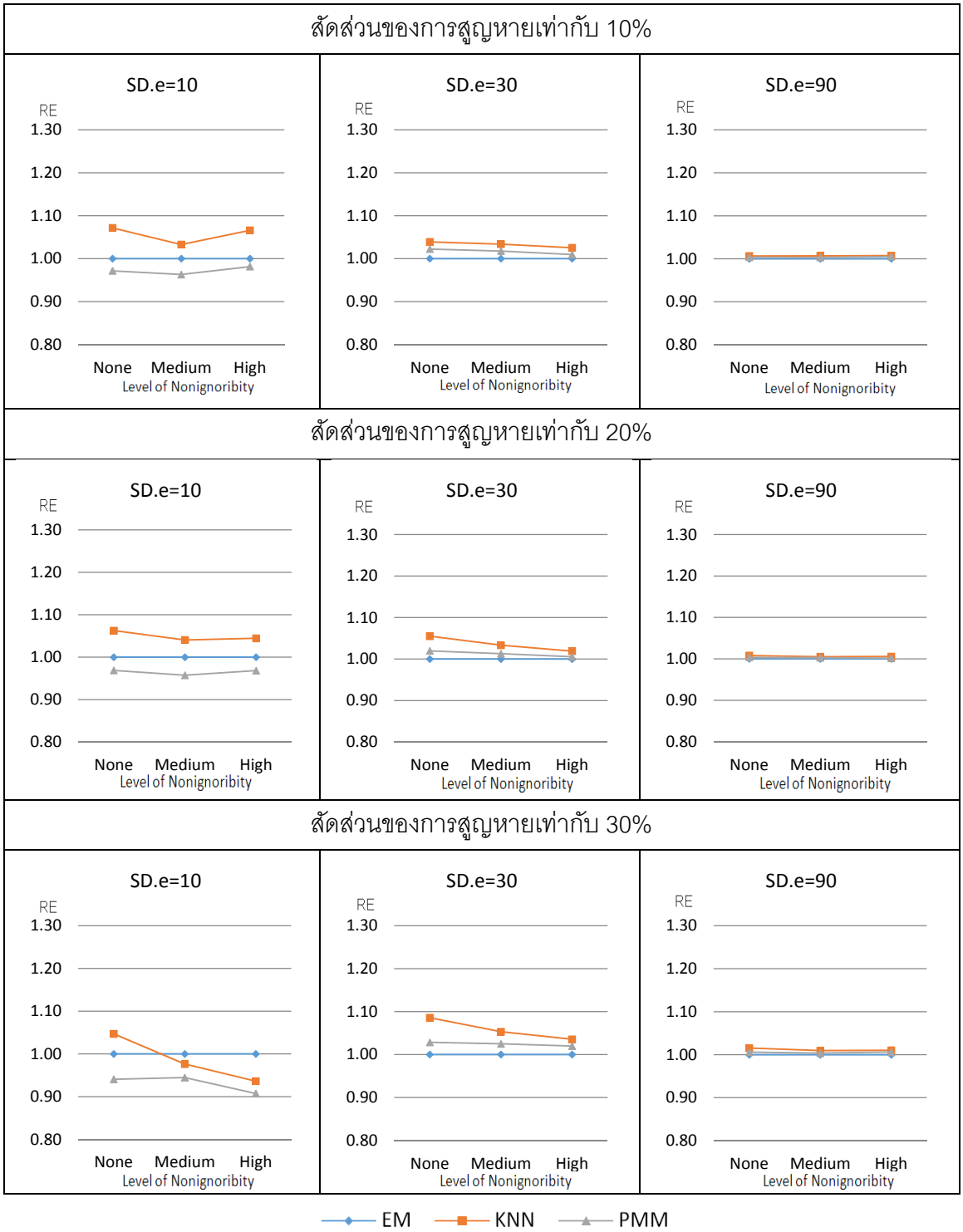
ตารางที่ 4.3.3.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.007	1.004
		Medium	RE	1.000	1.007	1.003
		High	RE	1.000	1.008	1.005
	20	None	RE	1.000	1.008	1.004
		Medium	RE	1.000	1.005	1.003
		High	RE	1.000	1.006	1.001
	30	None	RE	1.000	1.015	1.006
		Medium	RE	1.000	1.009	1.004
		High	RE	1.000	1.010	1.006
100	10	None	RE	1.000	1.006	1.004
		Medium	RE	1.000	1.006	1.002
		High	RE	1.000	1.008	1.004
	20	None	RE	1.000	1.006	1.003
		Medium	RE	1.000	1.008	1.002
		High	RE	1.000	1.009	1.003
	30	None	RE	1.000	1.016	1.007
		Medium	RE	1.000	1.011	1.004
		High	RE	1.000	1.013	1.005
200	10	None	RE	1.000	1.008	1.006
		Medium	RE	1.000	1.011	1.005
		High	RE	1.000	1.005	1.001
	20	None	RE	1.000	1.009	1.003
		Medium	RE	1.000	1.011	1.003
		High	RE	1.000	1.011	1.003
	30	None	RE	1.000	1.018	1.010
		Medium	RE	1.000	1.014	1.005
		High	RE	1.000	1.013	1.004

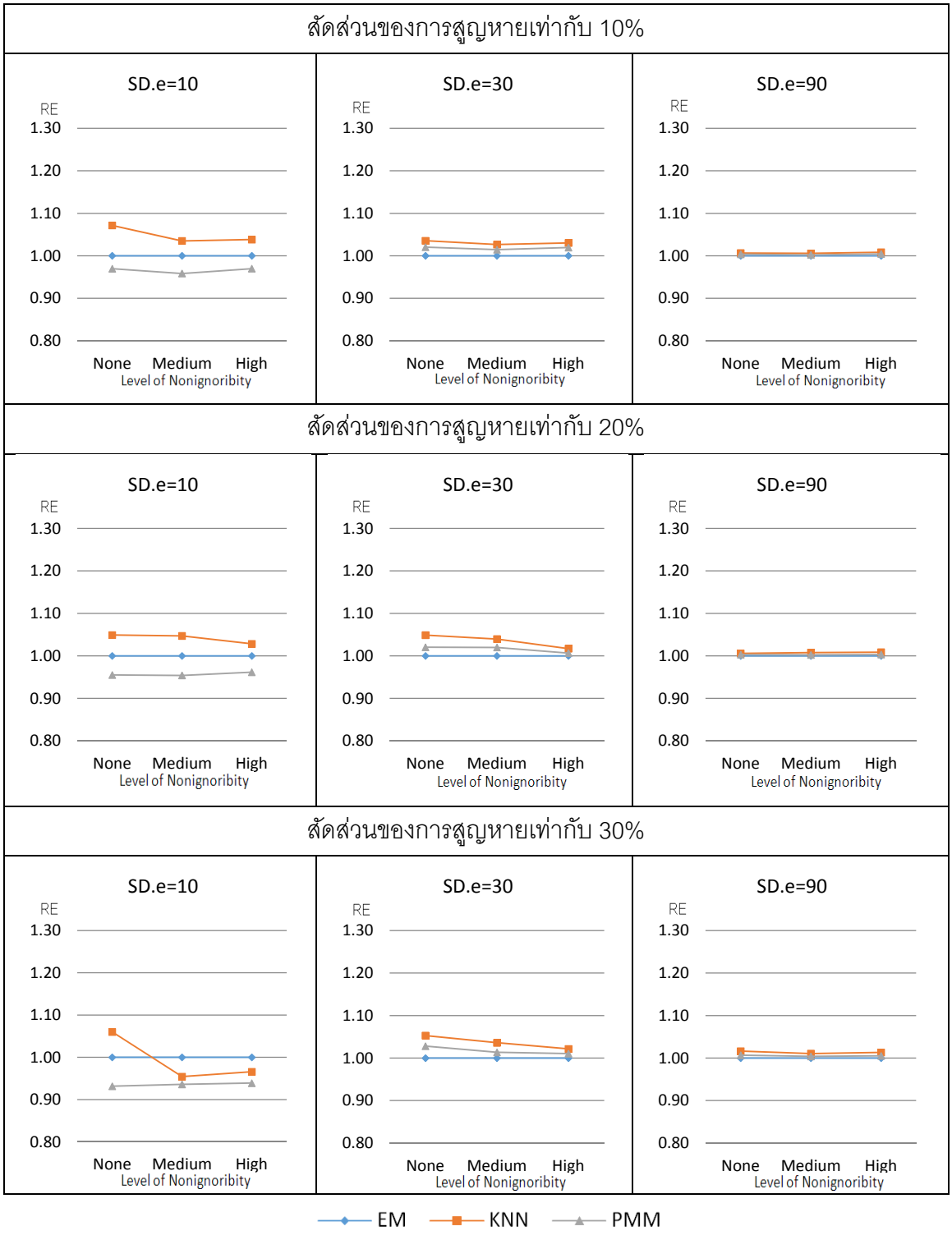
<u>ภาพที่</u> 4.3.3.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



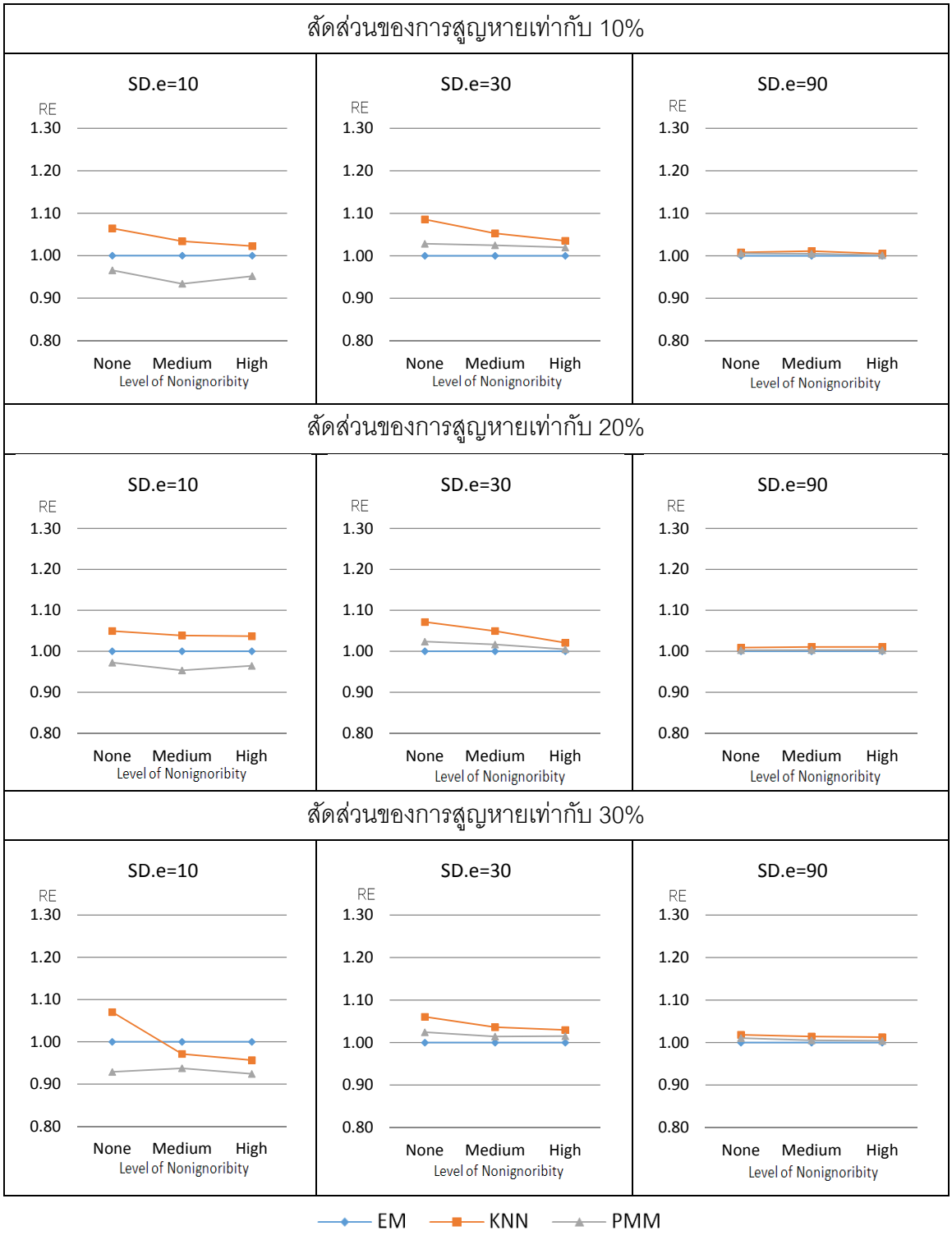
<u>ภาพที่ 4.3.4</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และขนาดตัวอย่างเท่ากับ 50



<u>ภาพที่ 4.3.5</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และขนาดตัวอย่างเท่ากับ 100



<u>ภาพที่ 4.3.6</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ปานกลาง</u>) และขนาดตัวอย่างเท่ากับ 200



4.4 ผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย เมื่อตัวแปรอิสระเป็นแบบที่ 2 (ศึกษาการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่)

ในส่วนนี้ผู้วิจัยได้ทำการศึกษาการสูญหายของข้อมูลในกรณีที่ตัวแปรอิสระเป็นแบบที่ 2 : $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$ ซึ่งเกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่ ซึ่งเท่ากับ 500 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 30 และ 90 ขนาดตัวอย่างเท่ากับ 50 100 และ 200 สัดส่วนของการสูญหายเท่ากับ 10% 20% และ 30% ระดับการสูญหายแบบ Nonignorable คือ ไม่มี ปานกลาง และสูง ซึ่งผลการวิจัยจะนำเสนอโดยแสดงค่า AMSE และ RE ที่ได้จากการประมาณค่าสูญหายของแต่ละวิธีการในตารางดังต่อไปนี้

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.4.1.1	แบบที่ 1	10
4.4.2.1	แบบที่ 1	30
4.4.3.1	แบบที่ 1	90

ตารางที่ 4.4.1.1-4.4.3.1 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า AMSE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ตารางและภาพที่	ชนิดของตัวแปรอิสระ	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
4.4.1.2	แบบที่ 1	10
4.4.2.2	แบบที่ 1	30
4.4.3.2	แบบที่ 1	90

ตารางที่ 4.4.1.2-4.4.3.2 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากขนาดตัวอย่าง สัดส่วนการสูญหายและระดับการสูญหายแบบ Nonignorable

ภาพที่	ชนิดของตัวแปรอิสระ	ขนาดตัวอย่าง (n)
4.4.4	แบบที่ 1	50
4.4.5	แบบที่ 1	100
4.4.6	แบบที่ 1	200

ภาพที่ 4.4.4-4.4.6 แสดงการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ด้วยค่า RE โดยในแต่ละตารางจะเปรียบเทียบจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 10 30 และ 90 สัดส่วนของการสูญหาย และระดับการสูญหายแบบ Nonignorable

จากตารางและภาพที่ 4.4.1.1-4.4.3.1 พบว่า โดยส่วนใหญ่วิธีการ KNN คือวิธีการที่ให้ค่า AMSE ต่ำที่สุดจึงเป็นวิธีการที่มีประสิทธิภาพมากที่สุด แต่ในกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย(10) ที่สัดส่วนของการสูญหายสูง(30%) และการสูญหายแบบ Nonignorable ในระดับสูง วิธีการ EM เป็นวิธีการที่มีประสิทธิภาพในการประมาณค่าสูญหายมากที่สุด เนื่องจากให้ค่า AMSE ต่ำที่สุด

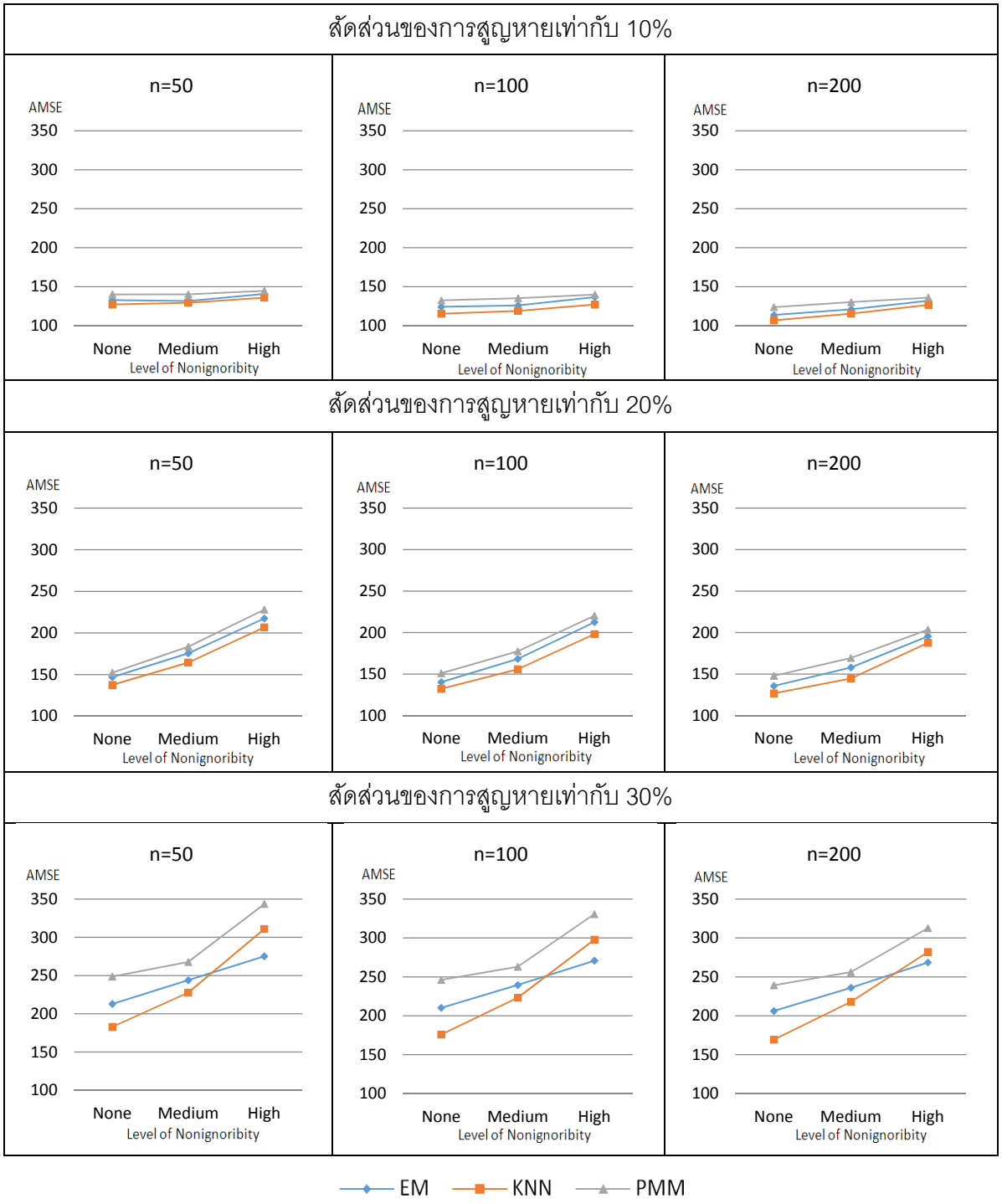
แต่ถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 และ 90 วิธีการ KNN จะเป็นวิธีการที่มีประสิทธิภาพมากที่สุดในทุกกรณี

จากตารางและภาพที่ 4.4.1.2 - 4.4.3.2 เมื่อพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีด้วยค่า RE โดยจะใช้วิธีการ EM เป็นวิธีการมาตรฐานในการเปรียบเทียบ โดยถ้า ค่า RE มากกว่า 1 แสดงว่า วิธีการ EM มีประสิทธิภาพในการประมาณค่าสูญหายน้อยกว่าวิธีการที่นำมาเปรียบเทียบ แต่ถ้า RE น้อยกว่า 1 แสดงว่าวิธีการ EM มีประสิทธิภาพมากกว่าวิธีการที่นำมาเปรียบเทียบ ซึ่งจากผลการวิจัยพบว่า เมื่อข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย(10) ที่สัดส่วนของการสูญหาย 30% และระดับการสูญหายแบบ Nonignorable สูง วิธีการ EM มีประสิทธิภาพดีกว่า KNN แต่ถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับปานกลางถึงสูง(30-90) วิธีการ KNN จะเป็นวิธีการที่มีประสิทธิภาพมากที่สุดในทุกกรณี

ตารางที่ 4.4.1.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	132.79	**127.04**	139.97
		Medium	AMSE	131.56	**129.33**	140.12
		High	AMSE	140.60	**135.91**	144.54
	20	None	AMSE	146.86	**137.55**	152.13
		Medium	AMSE	175.52	**164.44**	183.39
		High	AMSE	217.43	**206.85**	228.10
	30	None	AMSE	212.84	**182.78**	248.60
		Medium	AMSE	243.88	**227.69**	267.72
		High	AMSE	**275.20**	310.93	343.53
100	10	None	AMSE	124.10	**115.28**	132.38
		Medium	AMSE	125.85	**118.84**	135.19
		High	AMSE	136.60	**127.06**	139.88
	20	None	AMSE	140.34	**132.15**	150.97
		Medium	AMSE	168.30	**155.78**	177.53
		High	AMSE	212.57	**198.18**	220.23
	30	None	AMSE	210.21	**175.93**	246.10
		Medium	AMSE	239.53	**223.30**	262.97
		High	AMSE	**270.80**	297.73	330.67
200	10	None	AMSE	113.60	**106.84**	123.64
		Medium	AMSE	120.88	**115.47**	130.13
		High	AMSE	131.81	**126.62**	135.86
	20	None	AMSE	135.70	**126.73**	147.87
		Medium	AMSE	157.88	**144.65**	169.39
		High	AMSE	195.60	**187.68**	203.54
	30	None	AMSE	206.11	**169.37**	239.20
		Medium	AMSE	236.04	**217.91**	256.06
		High	AMSE	**268.48**	281.92	312.67

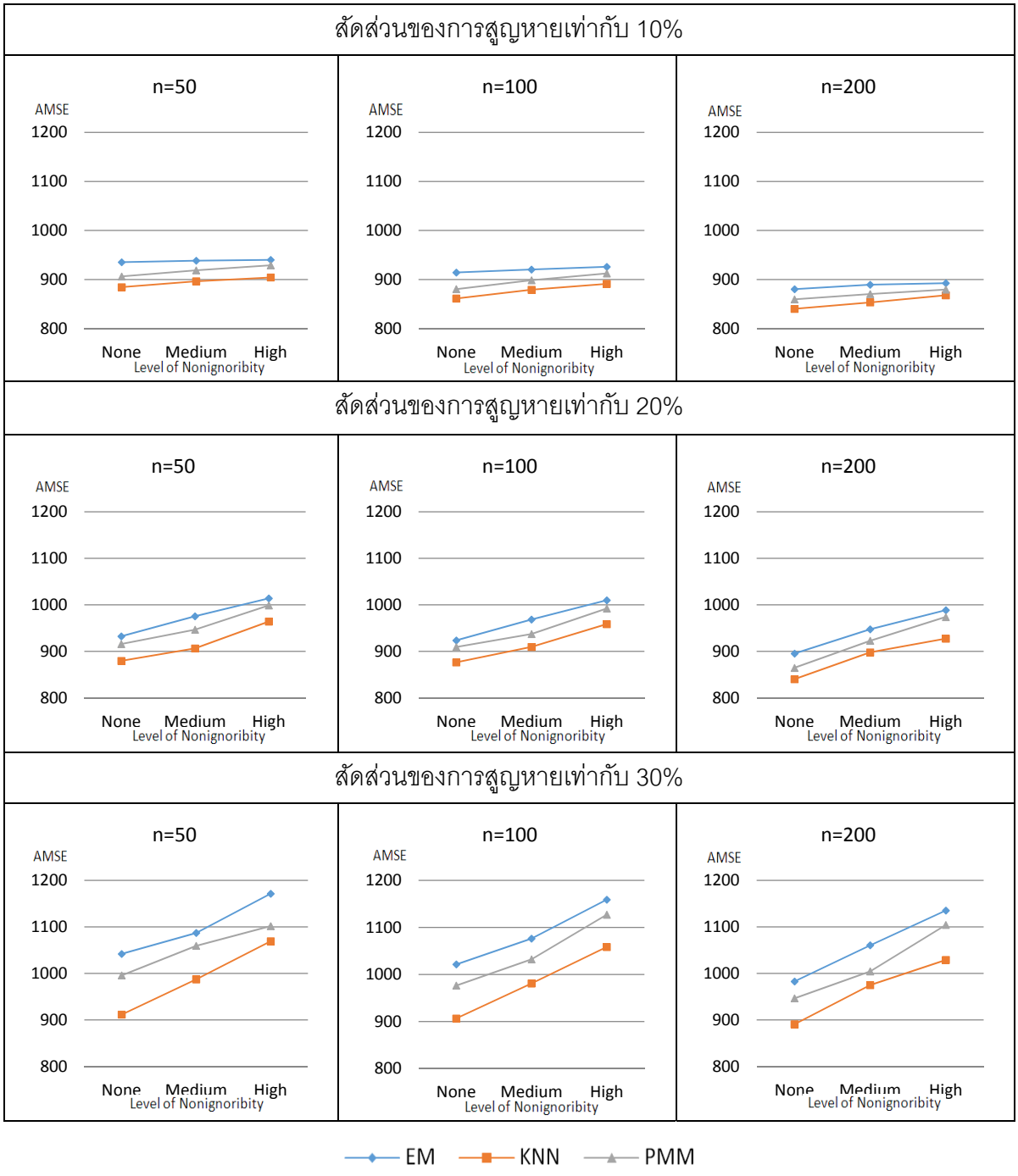
<u>ภาพที่</u> 4.4.1.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ใหญ่</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.4.2.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	935.39	**884.45**	906.28
		Medium	AMSE	938.52	**896.50**	918.83
		High	AMSE	940.17	**904.30**	929.00
	20	None	AMSE	932.27	**879.58**	915.64
		Medium	AMSE	975.59	**906.39**	946.72
		High	AMSE	1013.92	**964.26**	998.73
	30	None	AMSE	1041.91	**911.43**	995.67
		Medium	AMSE	1086.86	**987.19**	1059.26
		High	AMSE	1171.26	**1069.01**	1101.58
100	10	None	AMSE	914.35	**861.36**	880.56
		Medium	AMSE	920.47	**879.20**	898.72
		High	AMSE	925.94	**891.27**	912.57
	20	None	AMSE	923.55	**876.42**	909.39
		Medium	AMSE	968.51	**909.56**	937.30
		High	AMSE	1009.80	**958.65**	992.16
	30	None	AMSE	1021.48	**906.62**	976.59
		Medium	AMSE	1076.36	**981.07**	1032.36
		High	AMSE	1158.98	**1058.46**	1127.07
200	10	None	AMSE	880.41	**840.32**	859.71
		Medium	AMSE	889.46	**853.44**	870.46
		High	AMSE	892.62	**867.90**	879.68
	20	None	AMSE	895.18	**840.19**	864.59
		Medium	AMSE	947.40	**897.98**	922.82
		High	AMSE	988.63	**927.29**	973.91
	30	None	AMSE	982.70	**890.61**	946.44
		Medium	AMSE	1060.45	**974.86**	1004.37
		High	AMSE	1135.18	**1028.93**	1104.14

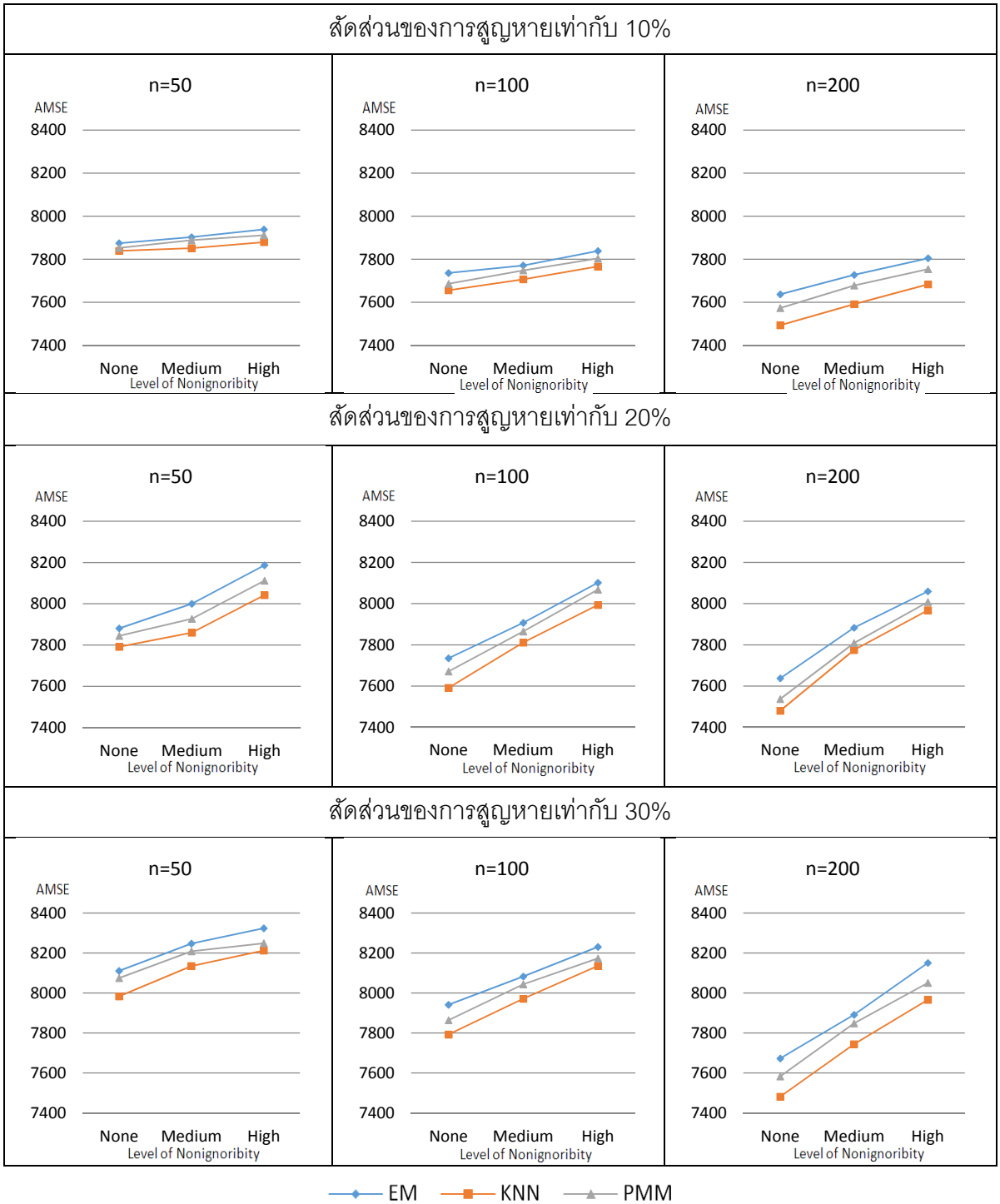
<u>ภาพที่</u> 4.4.2.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ใหญ่</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



ตารางที่ 4.4.3.1 แสดงค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	AMSE	7874.56	**7838.86**	7852.81
		Medium	AMSE	7902.75	**7851.62**	7888.82
		High	AMSE	7938.64	**7879.65**	7911.40
	20	None	AMSE	7880.89	**7791.44**	7844.75
		Medium	AMSE	7999.90	**7859.80**	7926.85
		High	AMSE	8186.51	**8041.87**	8111.83
	30	None	AMSE	8110.55	**7983.34**	8074.57
		Medium	AMSE	8246.97	**8134.93**	8208.68
		High	AMSE	8324.06	**8212.38**	8248.83
100	10	None	AMSE	7736.40	**7656.23**	7686.34
		Medium	AMSE	7771.27	**7706.91**	7748.40
		High	AMSE	7838.63	**7766.60**	7803.65
	20	None	AMSE	7735.27	**7591.37**	7671.06
		Medium	AMSE	7906.88	**7811.55**	7865.09
		High	AMSE	8101.81	**7993.71**	8068.21
	30	None	AMSE	7941.13	**7792.34**	7863.99
		Medium	AMSE	8082.40	**7971.28**	8043.19
		High	AMSE	8230.72	**8135.16**	8173.03
200	10	None	AMSE	7637.38	**7493.90**	7573.61
		Medium	AMSE	7727.63	**7591.51**	7678.38
		High	AMSE	7804.73	**7683.61**	7754.24
	20	None	AMSE	7637.73	**7480.34**	7537.35
		Medium	AMSE	7883.28	**7775.53**	7808.82
		High	AMSE	8059.51	**7968.11**	8008.31
	30	None	AMSE	7672.66	**7481.33**	7582.77
		Medium	AMSE	7891.44	**7743.83**	7848.21
		High	AMSE	8150.57	**7966.12**	8050.90

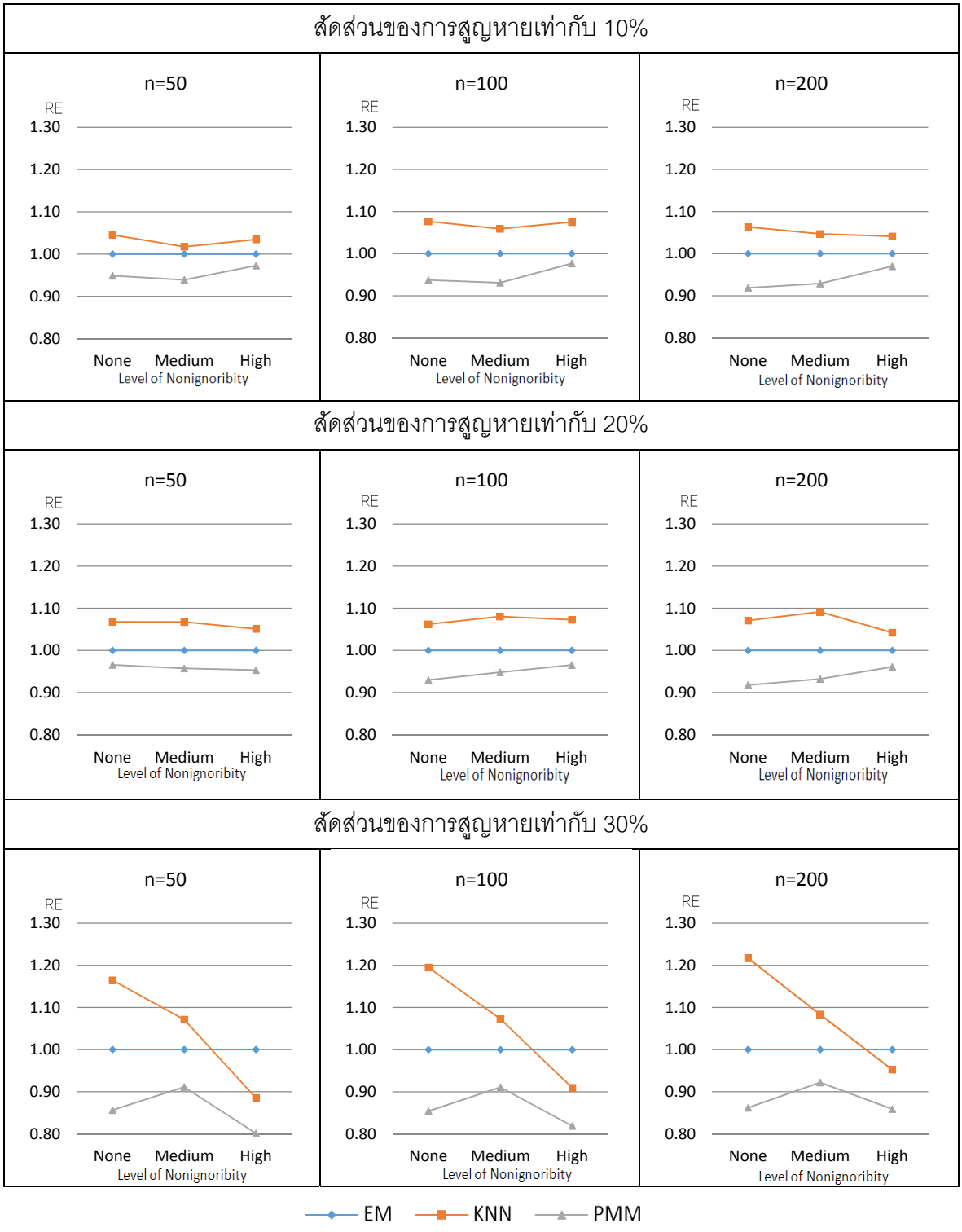
<u>ภาพที่</u> 4.4.3.1 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า AMSE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ใหญ่</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



<u>ตารางที่ 4.4.1.2</u> แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาด<u>ใหญ่</u>) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.045	0.949
		Medium	RE	1.000	1.017	0.939
		High	RE	1.000	1.035	0.973
	20	None	RE	1.000	1.068	0.965
		Medium	RE	1.000	1.067	0.957
		High	RE	1.000	1.051	0.953
	30	None	RE	1.000	1.164	0.856
		Medium	RE	1.000	1.071	0.911
		High	RE	1.000	0.885	0.801
100	10	None	RE	1.000	1.077	0.938
		Medium	RE	1.000	1.059	0.931
		High	RE	1.000	1.075	0.977
	20	None	RE	1.000	1.062	0.930
		Medium	RE	1.000	1.080	0.948
		High	RE	1.000	1.073	0.965
	30	None	RE	1.000	1.195	0.854
		Medium	RE	1.000	1.073	0.911
		High	RE	1.000	0.910	0.819
200	10	None	RE	1.000	1.063	0.919
		Medium	RE	1.000	1.047	0.929
		High	RE	1.000	1.041	0.970
	20	None	RE	1.000	1.071	0.918
		Medium	RE	1.000	1.092	0.932
		High	RE	1.000	1.042	0.961
	30	None	RE	1.000	1.217	0.862
		Medium	RE	1.000	1.083	0.922
		High	RE	1.000	0.952	0.859

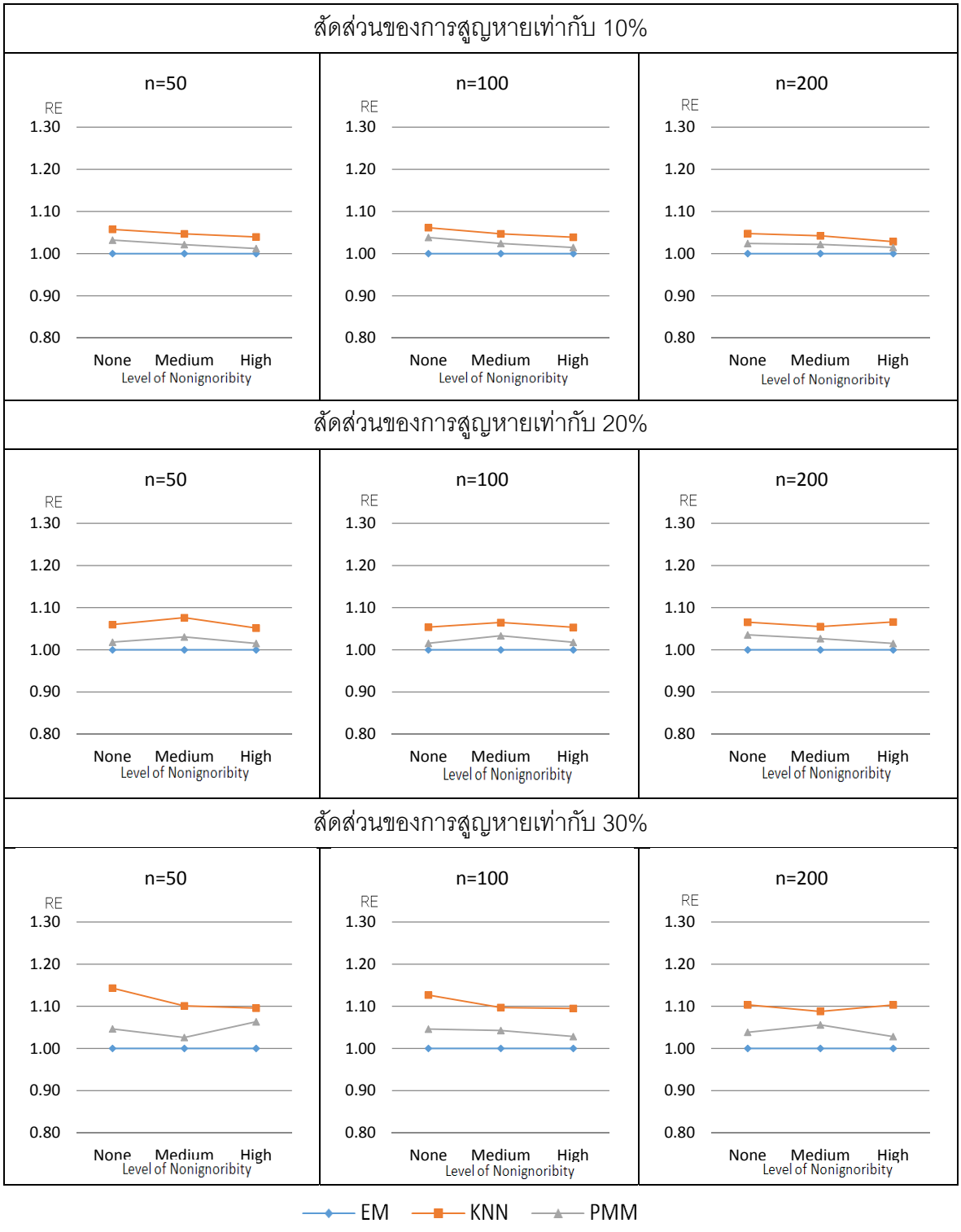
<u>ภาพที่</u> 4.4.1.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10



ตารางที่ 4.4.2.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.058	1.032
		Medium	RE	1.000	1.047	1.021
		High	RE	1.000	1.040	1.012
	20	None	RE	1.000	1.060	1.018
		Medium	RE	1.000	1.076	1.030
		High	RE	1.000	1.052	1.015
	30	None	RE	1.000	1.143	1.046
		Medium	RE	1.000	1.101	1.026
		High	RE	1.000	1.096	1.063
100	10	None	RE	1.000	1.062	1.038
		Medium	RE	1.000	1.047	1.024
		High	RE	1.000	1.039	1.015
	20	None	RE	1.000	1.054	1.016
		Medium	RE	1.000	1.065	1.033
		High	RE	1.000	1.053	1.018
	30	None	RE	1.000	1.127	1.046
		Medium	RE	1.000	1.097	1.043
		High	RE	1.000	1.095	1.028
200	10	None	RE	1.000	1.048	1.024
		Medium	RE	1.000	1.042	1.022
		High	RE	1.000	1.028	1.015
	20	None	RE	1.000	1.065	1.035
		Medium	RE	1.000	1.055	1.027
		High	RE	1.000	1.066	1.015
	30	None	RE	1.000	1.103	1.038
		Medium	RE	1.000	1.088	1.056
		High	RE	1.000	1.103	1.028

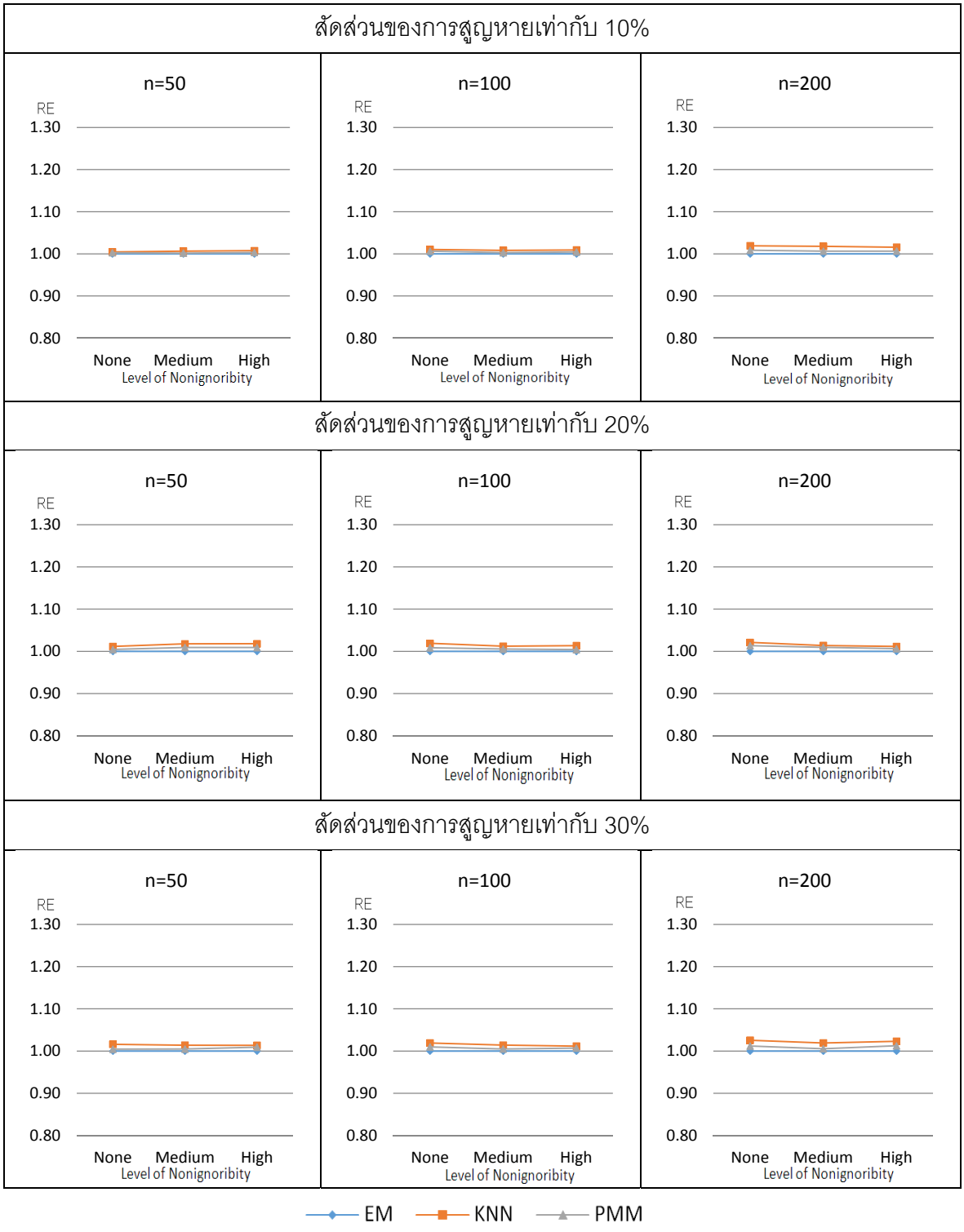
<u>ภาพที่</u> 4.4.2.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ KNN EM และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30



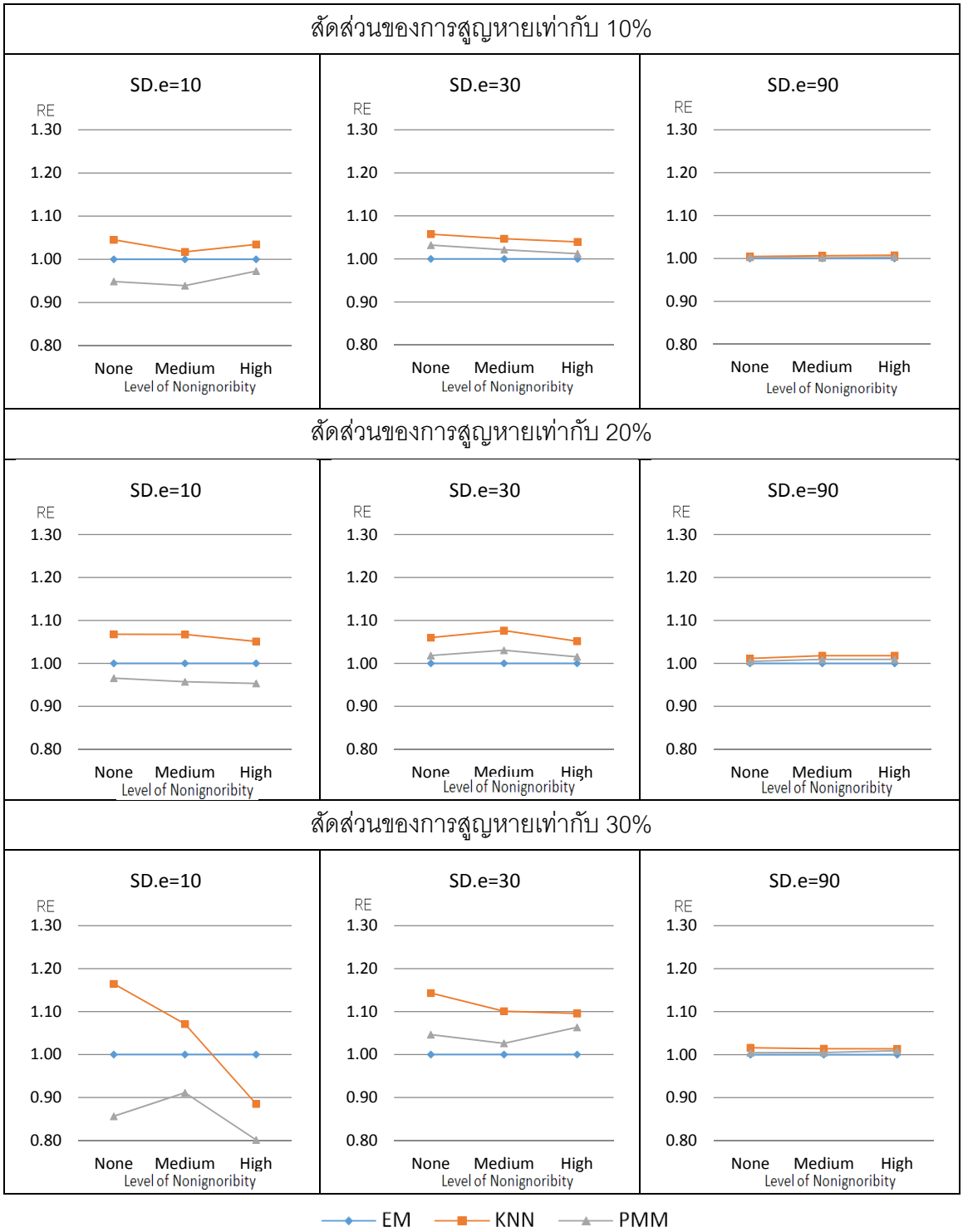
ตารางที่ 4.4.3.2 แสดงค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

n	%missing	Level of Nonignoribity		EM	KNN	PMM
50	10	None	RE	1.000	1.005	1.003
		Medium	RE	1.000	1.007	1.002
		High	RE	1.000	1.007	1.003
	20	None	RE	1.000	1.011	1.005
		Medium	RE	1.000	1.018	1.009
		High	RE	1.000	1.018	1.009
	30	None	RE	1.000	1.016	1.004
		Medium	RE	1.000	1.014	1.005
		High	RE	1.000	1.014	1.009
100	10	None	RE	1.000	1.010	1.007
		Medium	RE	1.000	1.008	1.003
		High	RE	1.000	1.009	1.004
	20	None	RE	1.000	1.019	1.008
		Medium	RE	1.000	1.012	1.005
		High	RE	1.000	1.014	1.004
	30	None	RE	1.000	1.019	1.010
		Medium	RE	1.000	1.014	1.005
		High	RE	1.000	1.012	1.007
200	10	None	RE	1.000	1.019	1.008
		Medium	RE	1.000	1.018	1.006
		High	RE	1.000	1.016	1.007
	20	None	RE	1.000	1.021	1.013
		Medium	RE	1.000	1.014	1.010
		High	RE	1.000	1.011	1.006
	30	None	RE	1.000	1.026	1.012
		Medium	RE	1.000	1.019	1.006
		High	RE	1.000	1.023	1.012

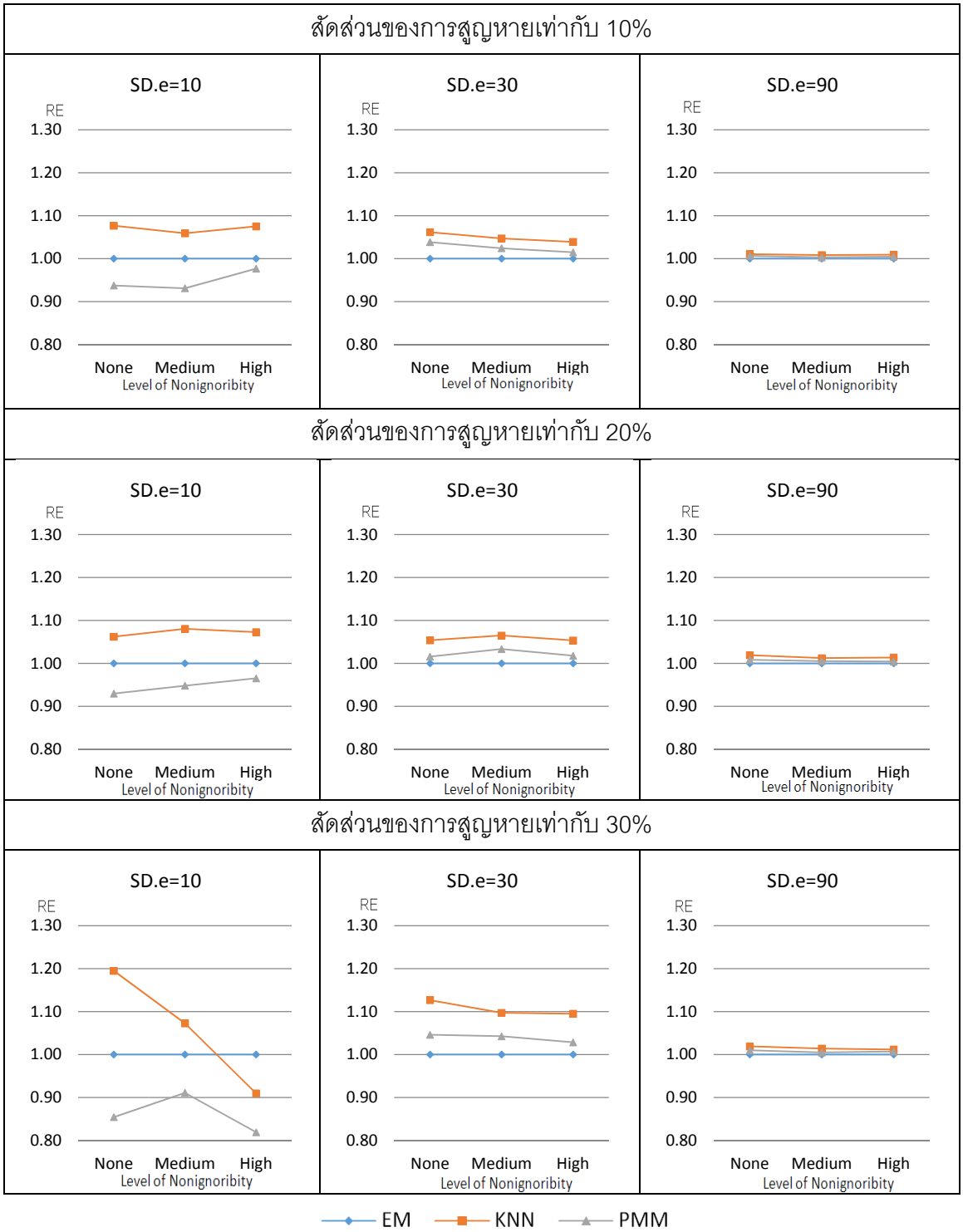
<u>ภาพที่</u> 4.4.3.2 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90



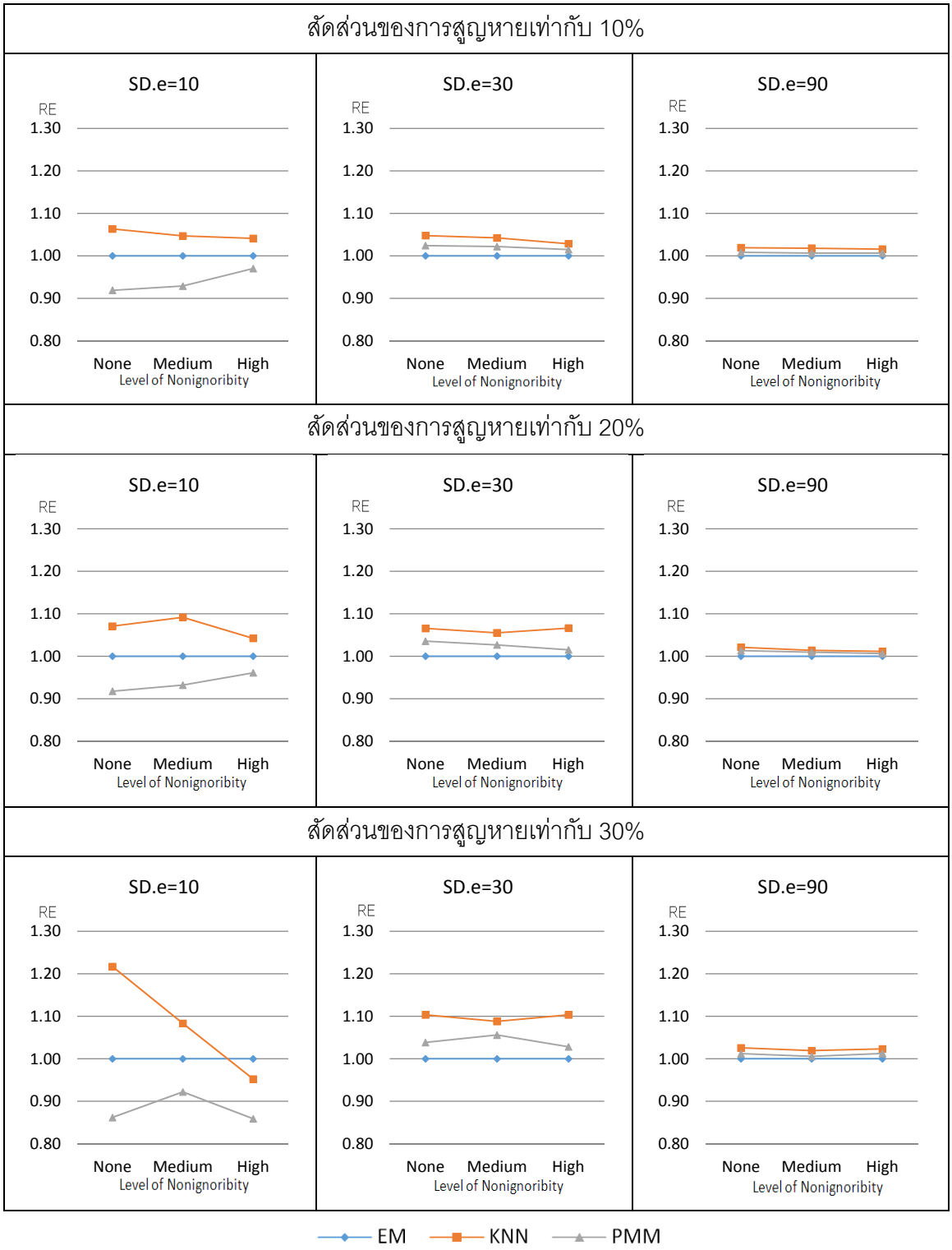
<u>ภาพที่</u> 4.4.4 แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และขนาดตัวอย่างเท่ากับ 50



<u>ภาพที่ 4.4.5</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และขนาดตัวอย่างเท่ากับ 100



<u>ภาพที่ 4.4.6</u> แสดงการเปรียบเทียบประสิทธิภาพวิธีการ EM KNN และ PMM ด้วยค่า RE เมื่อตัวแปรอิสระเป็นแบบที่ 2 (เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่) และขนาดตัวอย่างเท่ากับ 200



จากผลการวิจัยในแต่ละส่วนข้างต้นทำให้ได้ข้อสรุปว่า การที่จะเลือกใช้วิธีการประมาณค่าสูญหายในแต่ละวิธีการ จะขึ้นอยู่กับลักษณะของข้อมูลที่จะนำมาวิเคราะห์ กล่าวคือ โดยส่วนใหญ่แล้ว วิธีการ KNN จะเป็นวิธีการประมาณค่าสูญหายที่มีประสิทธิภาพมากที่สุด ยกเว้นในบางกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนต่ำ(10-30) และสัดส่วนของการสูญหาย 20% และ 30% และมีการสูญหายแบบ Nonignorable อยู่ในระดับสูง วิธีการ EM จะเป็นวิธีการที่ดีกว่าวิธี KNN ในบางกรณี นอกจากนี้ หากสังเกตค่า AMSE เมื่อส่วนเบี่ยงเบนมาตรฐานของความคาดเคลื่อนแตกต่างกัน (10 30 และ 90) จะพบว่า เมื่อข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 10 ค่า AMSE จะประมาณเท่ากับ 100 ส่วนถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความเคลื่อน 30 ค่า AMSE จะประมาณเท่ากับ 900 และถ้าข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 ค่า AMSE จะประมาณเท่ากับ 8,100 ทั้งนี้สาเหตุที่ทำให้ค่า AMSE เป็นเช่นนี้ เป็นผลมาจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนนั่นเอง

นอกจากนี้ลักษณะการสูญหายของตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของการสูญหายและระดับของการสูญหายแบบ Nonignorable ก็ยังมีอิทธิพลต่อประสิทธิภาพของการประมาณค่าสูญหายอีกด้วย ซึ่งสามารถสรุปได้ดังต่อไปนี้

1) ลักษณะของตัวแปรอิสระที่เกิดการสูญหายมีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหายในแต่วิธี โดยเมื่อศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนขนาดปานกลางและใหญ่ (300 และ 500) พบว่าโดยส่วนใหญ่วิธีการ KNN จะเป็นวิธีการที่ดีที่สุด ยกเว้นในบางกรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก (100) ซึ่งพบว่าจะมีบางกรณีที่วิธีการ EM เป็นวิธีการที่ดีที่สุด
2) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหายในแต่วิธี โดยที่ ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 วิธีการ KNN จะเป็นวิธีการที่ดีที่สุดในทุกกรณี ซึ่งแตกต่างจาก ข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 และ 30 ที่พบว่าในบางกรณีวิธีการ EM จะเป็นวิธีการที่ดีกว่าวิธี KNN โดยจะสังเกตได้ว่ายิ่งข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อยลง วิธีการ EM จะเป็นวิธีที่มีประสิทธิภาพมากยิ่งขึ้น

3) การเพิ่มหรือลดขนาดของตัวอย่างไม่ส่งผลทำให้ผลการวิจัยเปลี่ยนแปลงไป แต่ถ้าขนาดตัวอย่างใหญ่ขึ้นจะทำให้ผลการวิจัยมีความแม่นยำมากยิ่งขึ้น กล่าวคือเมื่อขนาดตัวอย่างใหญ่ขึ้น ค่า AMSE จะลดน้อยลง ซึ่งสอดคล้องกับหลักการที่ว่าเมื่อขนาดตัวอย่างใหญ่ขึ้นจะทำให้การประมาณค่าทางสถิติเกิดความคลาดเคลื่อนน้อยลง
4) สัดส่วนของการสูญหายในแต่ละระดับมีผลต่อประสิทธิภาพของความแตกต่างในแต่ละวิธีการประค่าสูญหายเล็กน้อย โดยที่วิธีการ EM มักมีประสิทธิภาพดีในบางกรณีที่สัดส่วนของการสูญหายเท่ากับ 20% และ 30% และถ้าสัดส่วนของการสูญหายสูงขึ้นค่า AMSE จะยิ่งสูงขึ้นด้วย กล่าวคือถ้าข้อมูลมีสัดส่วนของการสูญหายมาก ย่อมทำให้ค่าประมาณที่ได้แตกต่างจากค่าจริงมากยิ่งขึ้น จึงส่งผลให้ความคลาดเคลื่อนสูงขึ้น
5) ระดับของการสูญหายแบบ Nonignorable มีผลต่อความแตกต่างของค่า AMSE ในแต่ละวิธีการเช่นกัน โดยถ้าข้อมูลมีการสูญหายแบบ Nonignorable ในระดับไม่มี โดยส่วนใหญ่การประมาณค่าสูญหายด้วยวิธี KNN จะดีที่สุด แต่ถ้าข้อมูลเกิดการสูญหายแบบ Nonignorable ในระดับปานกลางและสูง วิธีการ EM มักเป็นวิธีที่มีประสิทธิภาพมากกว่าในบางกรณี

นอกจากนี้ ถ้าข้อมูลมีการสูญหายแบบ Nonignorable ในระดับไม่มี จะส่งผลให้ความคลาดเคลื่อนต่ำกว่าในกรณีที่ข้อมูลเกิดการสูญหายแบบ Nonignorable ในระดับสูง เนื่องจากถ้าเกิดการสูญหายแบบ Nonignorable (การสูญหายของข้อมูลที่ขึ้นกับตัวแปรนั้นๆเอง) ยิ่งจะเป็นการยากที่จะประมาณค่าสูญหายให้ใกล้เคียงกับค่าจริง เพราะการสูญหายแบบ Nonignorable เป็นการสูญหายที่ไม่ได้เกิดขึ้นอย่างสุ่ม

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

ในการวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย 3 วิธี ซึ่งได้แก่ วิธี EM Algorithm วิธี K-Nearest Neighbor Imputation (KNN) และวิธี Predictive Mean Matching Imputation (PMM) โดยแต่ละวิธีการจะมีขั้นตอนในการประมาณค่าสูญหายที่แตกต่างกัน จึงส่งผลให้ประสิทธิภาพของแต่ละวิธีการมีความแตกต่างกันด้วย ทั้งนี้ประสิทธิภาพของแต่ละวิธีการอาจขึ้นอยู่กับปัจจัยหลายด้าน โดยเฉพาะลักษณะของข้อมูลที่นำมาใช้ในการวิเคราะห์ ซึ่งขอบเขตที่ใช้ในการจำลองข้อมูลจากการวิจัยครั้งนี้มีกรณีศึกษาย่อยทั้งหมด 324 กรณี โดยแต่ละกรณีจะมาจากการกำหนดเงื่อนไขต่างๆ ดังต่อไปนี้

1. ลักษณะของชุดตัวแปรอิสระ แบ่งออกเป็น 2 รูปแบบ คือ

   <u>แบบที่ 1</u> $X_1 \sim N(0,300), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,300)$

   โดยจะศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนเท่ากัน

   <u>แบบที่ 2</u> $X_1 \sim N(0,100), X_2 \sim N(0,300)$ และ $X_3 \sim N(0,500)$

   โดยจะศึกษาการสูญหายในกรณีที่ตัวแปรอิสระมีความแปรปรวนขนาดเล็ก ปานกลาง และใหญ่

2. ค่าความคลาดเคลื่อน (ε) มีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน แบ่งออกเป็น 3 รูปแบบ คือ 10 30 และ 90
3. ขนาดตัวอย่าง แบ่งออกเป็น 3 ขนาดคือ 50 100 และ 200
4. ลักษณะการสูญหายของข้อมูลมี 9 แบบ ได้แก่

   4.1) สัดส่วนของการสูญหายแบ่งออกเป็น 3 รูปแบบคือ 10% 20% และ 30%

   4.2) ระดับการสูญหายแบบ Nonignorable 3 ระดับแบ่งออกเป็น สูง ปานกลาง และไม่มี

ในการจำลองข้อมูลแต่ละสถานการณ์จะใช้โปรแกรม R i386 2.15.3 ซึ่งจะทำการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) และในแต่ละสถานการณ์จะทำซ้ำทั้งหมด 5,000 รอบ

5.1 สรุปความแตกต่างของแต่ละวิธีการประมาณค่าสูญหาย

วิธีการ EM Algorithm

การประมาณค่าสูญหายด้วยวิธี EM เป็นวิธีการที่มีประสิทธิภาพและเหมาะสมที่สุดในกรณีที่เกิดการสูญหายของตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก(100) และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 หรือมีการกระจายน้อย รวมทั้งวิธีการ EM จะมีประสิทธิภาพดีในบางกรณีที่ข้อมูลมีสัดส่วนของการสูญหาย 20% และ 30% และระดับของการสูญหายแบบ Nonignorable สูง

ในขั้นตอนของวิธีการ EM นั้นจะทำการแทนที่ข้อมูลที่สูญหายของตัวแปรตามโดยต้องใช้ค่าพารามิเตอร์ในการประมาณข้อมูลที่สูญหาย ซึ่งจะใช้กระบวนการวนซ้ำเพื่อหาค่าประมาณภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ แต่ในการแทนที่ข้อมูลสูญหายในตัวแปรอิสระ จะแทนด้วยค่าเฉลี่ยของชุดตัวแปรอิสระที่เกิดการสูญหาย ซึ่งในขั้นตอนนี้อาจจะทำให้ความแม่นยำในการประมาณค่าสูญหายโดยรวมลดลง จึงอาจเป็นสาเหตุที่ทำให้วิธีการ EM ไม่มีประสิทธิภาพเท่าที่ควร เมื่อเปรียบเทียบกับผลการวิจัยของอุษณีย์ วงศ์อามาตย์ (2555) ที่ผลการวิจัยพบว่า ในกรณีส่วนใหญ่วิธีการ EM คือวิธีการที่มีประสิทธิภาพมากที่สุดเมื่อข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับ 10 และ 30

วิธีการ K-Nearest Neighbor Imputation (KNN)

การประมาณค่าสูญหายด้วยวิธีการ KNN จะมีประสิทธิภาพและเหมาะสมในทุกกรณีสำหรับชุดข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 และโดยส่วนใหญ่วิธีการ KNN ยังมีประสิทธิภาพมากที่สุดในกรณีที่ข้อมูลมีสัดส่วนของการสูญหาย 10% และ 20% และการสูญหายแบบ Nonignorable ในระดับไม่มี

สาเหตุที่วิธีการ KNN มีประสิทธิภาพมากที่สุดในกรณีดังกล่าว อาจเนื่องมาจากวิธีการ KNN เป็นวิธีการที่ไม่ต้องหาค่าประมาณของพารามิเตอร์ แต่เกิดจากการหาค่าเฉลี่ยจากชุดข้อมูลที่สมบูรณ์ที่มีความใกล้เคียงกับชุดข้อมูลที่สูญหายมากที่สุดแล้วจึงนำค่าเฉลี่ยนั้นมาแทนที่ข้อมูลที่สูญหาย จึงน่าจะประมาณค่าสูญหายได้ดีในกรณีที่ข้อมูลมีการกระจายมาก

วิธีการ Predictive Mean Matching Imputation (PMM)

วิธีการประมาณค่าสูญหาย PMM เป็นวิธีการที่เป็นกลางระหว่างวิธี KNN และวิธี EM ซึ่งในการแทนที่ข้อมูลของตัวแปรอิสระที่สูญหายนั้น จะแทนที่โดยใช้วิธีการที่คล้ายคลึงกับวิธี KNN ส่วนในการแทนที่ข้อมูลของตัวแปรตามที่เกิดการสูญหายก็จะแทนที่โดยใช้วิธีการที่คล้ายคลึงกับวิธี EM ซึ่งผลการวิจัยพบว่า ไม่มีกรณีใดที่วิธีการ PMM จะให้ประสิทธิภาพดีที่สุด อาจเนื่องมาจากค่าประมาณที่เอาไปแทนในตำแหน่งของข้อมูลที่สูญหายเป็นค่าเดียวที่ใกล้เคียงกับข้อมูลที่ทราบค่ามากที่สุด ซึ่งแตกต่างจากวิธีการ KNN ที่นำข้อมูลมาเฉลี่ย K ตัวแล้วค่อยนำมาแทนที่ข้อมูลที่สูญหาย หรือวิธีการ EM ที่มีการทำซ้ำหลายรอบเพื่อให้เกิดความแม่นยำในการประมาณค่าสูญหายมากที่สุด

5.2 ผลการเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) ของแต่ละวิธีการประมาณค่าสูญหาย

จากงานวิจัยนี้ เป็นการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายแบบต่างๆ ด้วยค่า AMSE และ RE เพื่อหาวิธีการที่มีความเหมาะสมที่สุดสำหรับการนำไปใช้ในแต่ละกรณี ซึ่งวิธีการที่ดีที่สุดจะเป็นวิธีการที่ให้ค่าความคลาดเคลื่อนระหว่างค่าจริงกับค่าพยากรณ์น้อยที่สุด ซึ่งจากผลการวิจัยพบว่า วิธีการ KNN จะเป็นวิธีการที่ให้ค่า AMSE น้อยที่สุดเกือบทุกกรณี ยกเว้นในบางกรณีที่ข้อมูลเกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก โดยส่วนใหญ่วิธีการ EM จะเป็นวิธีการที่ให้ค่า AMSE น้อยที่สุด แต่อย่างไรก็ตามการเปลี่ยนแปลงของค่า AMSE ในแต่ละวิธีการก็ยังขึ้นอยู่กับปัจจัยหลายด้าน เช่น สัดส่วนของการสูญหาย ระดับการสูญหายแบบ Nonignorable และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

นอกจากนั้น หากพิจารณาค่า AMSE ที่ได้จากกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนแตกต่างกัน (10 30 และ 90) จะพบว่า ค่า AMSE ที่ได้จากข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 จะมีค่าประมาณ 100 ส่วนค่า AMSE ที่ได้จากข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30 จะมีค่าประมาณ 900 และค่า AMSE ที่ได้จากข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90 จะมีค่าประมาณ 8,100 สาเหตุเนื่องมาจาก ค่า AMSE ในแต่ละกรณีดังกล่าวล้วนได้รับอิทธิพลมาจากส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

ผลการเปรียบเทียบประสิทธิภาพของแต่ละวิธีการจะทำการสรุปดังตารางต่อไปนี้

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

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

ส่วนที่ 1 ตัวแปรอิสระแบบที่ 1 เกิดการสูญหายในตัวแปรอิสระตัวใดตัวหนึ่ง

ส่วนที่ 2 ตัวแปรอิสระแบบที่ 2 เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก

ส่วนที่ 3 ตัวแปรอิสระแบบที่ 2 เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง

ส่วนที่ 4 ตัวแปรอิสระแบบที่ 2 เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่

<u>ตารางที่ 5.2.1</u> สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10

ชนิดของตัวแปรอิสระ	สัดส่วนของการสูญหาย	วิธีการประมาณค่าสูญหายที่ดีที่สุด		
		ระดับการสูญหายแบบNonignorable		
		ไม่มี	ปานกลาง	สูง
<u>ส่วนที่ 1</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	**EM ทุก n**
	30%	KNN ทุก n	KNN ทุก n	**EM ทุก n**
<u>ส่วนที่ 2</u>	10%	KNN ทุก n	KNN ทุก n	**EM ทุก n**
	20%	KNN ทุก n	**EM ทุก n**	**EM ทุก n**
	30%	**EM ทุก n**	**EM ทุก n**	**EM ทุก n**
<u>ส่วนที่ 3</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	**EM ทุก n**	**EM ทุก n**
<u>ส่วนที่ 4</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	**EM ทุก n**

<u>ตารางที่ 5.2.2</u> สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 30

ชนิดของตัวแปรอิสระ	สัดส่วนของการสูญหาย	วิธีการประมาณค่าสูญหายที่ดีที่สุด		
		ระดับการสูญหายแบบNonignorable		
		ไม่มี	ปานกลาง	สูง
<u>ส่วนที่ 1</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	**EM ทุก n**
<u>ส่วนที่ 2</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN(50,100),**EM(200)**	KNN(50),**EM(100,200)**
	30%	KNN ทุก n	KNN(50,100),**EM(200)**	**EM ทุก n**
<u>ส่วนที่ 3</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n
<u>ส่วนที่ 4</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n

<u>ตารางที่ 5.2.3</u> สรุปผลการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย ในกรณีที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 90

ชนิดของตัวแปรอิสระ	สัดส่วนของการสูญหาย	วิธีการประมาณค่าสูญหายที่ดีที่สุด		
		ระดับการสูญหายแบบNonignorable		
		ไม่มี	ปานกลาง	สูง
<u>ส่วนที่ 1</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n
<u>ส่วนที่ 2</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n
<u>ส่วนที่ 3</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n
<u>ส่วนที่ 4</u>	10%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	20%	KNN ทุก n	KNN ทุก n	KNN ทุก n
	30%	KNN ทุก n	KNN ทุก n	KNN ทุก n

5.3 ปัจจัยอื่น ๆ ที่มีผลต่อค่า AMSE ของแต่ละวิธีการประมาณค่าสูญหาย

5.3.1) สัดส่วนของการสูญหาย และระดับของการสูญหายแบบ Nonignorable ที่เพิ่มสูงขึ้นมีส่วนโดยตรงที่ทำให้ค่า AMSE ของแต่ละวิธีการเพิ่มขึ้น เนื่องจากถ้าข้อมูลมีระดับการสูญหายมาก ย่อมส่งผลให้ค่าพยากรณ์มีความคลาดเคลื่อนกับค่าจริงสูงยิ่งขึ้น

5.3.2) ขนาดตัวอย่างที่เพิ่มขึ้นจะทำให้ค่า AMSE ของทุกวิธีการประมาณค่าสูญหายลดลง เนื่องจากตัวอย่างใหญ่ขึ้นจะทำให้การประมาณค่าทางสถิติเกิดความคลาดเคลื่อนน้อยลง

5.3.3) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมากขึ้นจะทำให้ค่า AMSE ของแต่ละวิธีการสูงขึ้น เพราะว่าถ้าข้อมูลมีการกระจายมาก ย่อมส่งผลให้การประมาณค่าสูญหายมีความคลาดเคลื่อนกับค่าจริงสูงยิ่งขึ้น จึงทำให้ประสิทธิภาพการพยากรณ์ลดลง

5.3.4) ค่า AMSE ที่ได้จากแต่ละวิธีการประมาณค่าจะแปรผันตามค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

5.3.5) ลักษณะของชุดตัวแปรอิสระที่มีรูปแบบของความแปรปรวนแตกต่างกันทั้ง 2 ลักษณะ 4 กรณีศึกษาซึ่งได้แก่

1) ศึกษาการสูญหายของตัวแปรอิสระที่มีความแปรปรวนเท่ากัน

2) ศึกษาการสูญหายของตัวแปรอิสระที่มีความแปรปรวนแตกต่างกัน

2.1) เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก

2.2) เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดปานกลาง

2.3) เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่

จากผลการวิจัย พบว่า ค่า AMSE ของแต่ละวิธีการจะเพิ่มขึ้นตามลักษณะของชุดตัวแปรอิสระที่เกิดการสูญหายในกรณีที่มีความแปรปรวนเพิ่มขึ้น กล่าวคือ ถ้าเกิดการสูญหายของชุดข้อมูลตัวแปรอิสระที่มีความแปรปรวนขนาดเล็ก(2.1) ค่า AMSE ของทุกวิธีการจะน้อยกว่าค่า AMSE ที่ได้จากการประมาณค่าสูญหายของชุดตัวแปรอิสระที่มีความแปรปรวนขนาดใหญ่(2.3)

5.4 การเปรียบเทียบผลการวิจัยในงานวิจัยที่เกี่ยวข้อง

จากงานวิจัยของ อุษณีย์ วงศ์อามาตย์ (2555) ที่ศึกษาเกี่ยวกับการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบ Nonignorable ในการวิเคราะห์การถดถอยเชิงพหุ ในกรณีที่ตัวแปรตามเกิดการสูญหาย เนื่องจากในงานวิจัยดังกล่าว กรณีที่ศึกษายังไม่ครอบคลุมกับสถานการณ์ที่เกิดขึ้นจริง ดังนั้นในงานวิจัยนี้จึงทำการศึกษาเพิ่มเติมในกรณีที่เกิดการสูญหายในตัวแปรอิสระที่มีความแปรปรวนแตกต่างกันและยังส่งผลให้ตัวแปรตามเกิดการสูญหายตามมาด้วย เพื่อให้เกิดความครอบคลุมสำหรับการนำมาประยุกต์ใช้งานต่อไป

ความแตกต่างของผลการวิจัยชิ้นนี้กับงานวิจัยดังกล่าวพบว่า ในงานวิจัยของอุษณีย์ วงศ์อามาตย์ (2555) วิธีการประมาณค่าสูญหายที่มีประสิทธิภาพมากที่สุดในกรณีส่วนใหญ่คือ วิธีการ EM ซึ่งจะเป็นวิธีการที่ดีที่สุดในกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนอยู่ในระดับต่ำถึงปานกลาง ส่วนข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง วิธีการ KNN จะเป็นวิธีการที่ดีที่สุด ซึ่งมีความแตกต่างจากงานวิจัยชิ้นนี้ ที่พบว่าโดยส่วนใหญ่วิธีการ KNN จะเป็นวิธีการที่ดีที่สุดในทุกกรณีโดยเฉพาะข้อมูลที่มีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง แต่ยังมีบางกรณีที่วิธีการ EM มีประสิทธิภาพมากกว่าวิธี KNN คือกรณีที่ข้อมูลมีส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนต่ำ สัดส่วนและระดับการสูญหายแบบ Nonignorable สูง

สิ่งที่เหมือนกันระหว่างงานวิจัยชิ้นนี้กับงานวิจัยของอุษณีย์ วงศ์อามาตย์ (2555) คือไม่มีกรณีใดที่วิธีการ PMM ให้ประสิทธิภาพดีที่สุด ถึงแม้ว่าในงานวิจัยชิ้นนี้ ผู้วิจัยจะมีการปรับเปลี่ยนขั้นตอนย่อยภายในวิธีการ PMM ให้แตกต่างจากงานวิจัยของ อุษณีย์ วงศ์อามาตย์ (2555) แล้วก็ตาม แต่วิธีการ PMM ก็ยังไม่มีความเหมาะสมที่สุดที่จะนำไปใช้ประมาณค่าสูญหายให้กับข้อมูลที่มีลักษณะต่างๆ ได้เลย

5.5 แนวทางในการนำวิธีการประค่าสูญหายจากงานวิจัยนี้ไปประยุกต์ใช้

ในการสำรวจเพื่อเก็บรวบรวมข้อมูลจากงานวิจัยแขนงต่างๆ หากเกิดปัญหาข้อมูลสูญหาย นักวิจัยจำเป็นต้องทำการตัดสินใจดำเนินการระหว่างการพิจารณาไม่เลือกใช้ข้อมูลในรายที่เกิดปัญหานี้ หรือควรที่จะแทนค่าที่สูญหาย

ลำดับแรกอาจพิจารณาจากลักษณะการสูญหายว่าเกิดขึ้นอย่างสุ่มหรือไม่ ซึ่งสามารถตรวจสอบได้โดยการแบ่งข้อมูลออกเป็น 2 กลุ่ม คือข้อมูลชุดที่สมบูรณ์และข้อมูลชุดที่สูญหาย แล้วทำการทดสอบสมมติฐานทางสถิติ เพื่อหานัยสำคัญทางสถิติระหว่างข้อมูลทั้ง 2 กลุ่ม ถ้าพบว่าข้อมูลทั้ง 2 กลุ่มไม่มีนัยสำคัญทางสถิติ แสดงว่าลักษณะการสูญหายเป็นแบบสุ่ม แต่ถ้าหากข้อมูลทั้ง 2 กลุ่มมีนัยสำคัญทางสถิติ แสดงว่าอาจเกิดการสูญหาย แบบ Nonignorable จึงควรทำการทดสอบต่อด้วยการเขียนกราฟเพื่อหาแนวโน้มการกระจายของข้อมูล แล้วพิจารณาว่าลักษณะของ

การสูญหายของตัวแปรนั้นๆ มีความสัมพันธ์กับตัวมันเองหรือไม่ หากมีความสัมพันธ์กับตัวมันเองแสดงว่าเป็นการสูญหายแบบ Nonignorable

ลำดับต่อมา ควรทำการตรวจสอบสัดส่วนของการสูญหาย อาจตรวจสอบโดยการใช้โปรแกรมสำเร็จรูปต่างๆ เช่น excel spss เพื่อเป็นแนวทางในการพิจารณาเลือกวิธีการประมาณค่าสูญหายที่เหมาะสม เช่น ถ้าข้อมูลมีสัดส่วนของการสูญหาย 30% และระดับของการสูญหายแบบ Nonignorable สูง จะทำการประมาณค่าสูญหายด้วยวิธีการ EM เป็นต้น นอกจากนี้ถ้าหากพิจารณาที่ความแปรปรวนของตัวแปรอิสระที่เกิดการสูญหายด้วยแล้ว อาจยิ่งส่งผลให้วิธีการที่เลือกมาประมาณค่าสูญหายมีประสิทธิภาพมากยิ่งขึ้น

5.6 **ข้อเสนอแนะ**

จากงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษากรณีเฉพาะบางกรณีเท่านั้น ซึ่งในความเป็นจริงแล้วปัญหาที่พบอาจจะอยู่นอกเหนือจากขอบเขตและข้อสรุปของงานวิจัยชิ้นนี้ เช่น จากงานวิจัยชิ้นนี้ในแต่ละสถานการณ์ ถูกกำหนดให้ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กันและเกิดการสูญหายที่ตัวแปรอิสระเพียงตัวใดตัวหนึ่งเท่านั้น แต่ในความเป็นจริงแล้วตัวแปรอิสระในชุดข้อมูลเดียวกันอาจเกิดการสูญหายพร้อมๆ กันมากกว่าหนึ่งตัวก็เป็นได้ หรือตัวแปรอิสระแต่ละตัวอาจมีความสัมพันธ์กัน ซึ่งสามารถพบได้ในข้อมูลทั่วๆไป รวมทั้งยังควรศึกษาเพิ่มเติมในกรณีที่เกิดการสูญหายของชุดข้อมูลที่มีทั้งข้อมูลเชิงคุณภาพและข้อมูลเชิงปริมาณด้วย เพราะในงานวิจัยชิ้นนี้ศึกษาแต่การสูญหายของข้อมูลเชิงปริมาณเท่านั้น นอกจากนี้ยังควรศึกษาเพิ่มเติมในกรณีที่ข้อมูลเป็นข้อมูลอนุกรมเวลา เพื่อศึกษาว่าการเปลี่ยนแปลงของเวลาจะส่งผลกระทบต่อค่าสูญหายและวิธีการประมาณค่าสูญหายหรือไม่ ส่วนวิธีที่ใช้ในการประมาณค่าสูญหาย อาจมีการปรับเปลี่ยนตามความเหมาะสมของชุดข้อมูลที่มีลักษณะแตกต่างกันไป โดยเฉพาะวิธีการ PMM ที่ผู้วิจัยได้มีการปรับเปลี่ยนขั้นตอนการทำแล้ว ก็ยังไม่ทำให้ประสิทธิภาพของการประมาณค่าสูญหายดียิ่งขึ้น ดังนั้น หากจะทำการศึกษาต่อไปในอนาคต จึงควรปรับเปลี่ยนวิธีการประมาณค่าให้เป็นแบบอื่นเพื่อเพิ่มตัวเลือกในการพิจารณาหาวิธีการประมาณค่าสูญหายที่มีความเหมาะสมกับข้อมูลลักษณะต่างๆ มากที่สุด

รายการอ้างอิง

ภาษาไทย

ธีระพร วีระถาวร. ตัวแบบเชิงเส้น : ทฤษฎีและการประยุกต์. กรุงเทพมหานคร : พิทักษ์การพิมพ์, 2531.

นรุตม์ บุตรพลอย. การประยุกต์ Soft Computing และ k-Nearest Neighbor. The 3rd National Conference on Information Technology: "IT Innovation for Global Awareness" (NCIT 2010) : 25-29

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

วารุณี ตรีบำรุงศักดิ์. การพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามมีค่าสูญหาย. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2538.

อุษณีย์ วงศ์อามาตย์. การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอิกนอร์เรเบิล ในการวิเคราะห์การถดถอยเชิงพหุ. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2555.

ภาษาอังกฤษ

James E.S. Macleod. An alternative nearest neighbour classification scheme. Pattern Recognition Letters, 4:375–381.

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

K. Hengpraprohm, S. Na Wichian and P. Meesad. Missing Value Imputation Using GA for Colon Cancer. The 6TH National Conference on Computing and Information Technology (2010) : 156-160.

Piyaporn Parsitwattanasaree and Sukorn Parsitwattanasaree. Missing data and Management. Vol.4 No.3 Data Management & Biostatistics Journal (2006): 52-61

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. การวิเคราะห์สถิติ : สถิติสำหรับการบริหารและงานวิจัย. กรุงเทพมหานคร : โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2553.

ธีระพร วีระถาวร. ความน่าจะเป็นกับการประยุกต์. กรุงเทพมหานคร : นำอักษรการพิมพ์, 2537.

ศุภลักษณ์ กรรณิกา. การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวางแผนการทดลองแบบจัตุรัสละติน. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2549.

ภาษาอังกฤษ

Roderick J.A. Little and Donald B. Rubin. Statistical Analysis with Missing Data. Wiley,1987.

ภาคผนวก

รายละเอียดของโปรแกรมที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้ทำการจำลองข้อมูลเพื่อเปรียบประสิทธิภาพของวีการประมาณค่าสูญหายแบบต่างๆ โดยใช้โปรแกรม R i386 2.15.3 ซึ่งมีคำสั่งดังต่อไปนี้

```
estimatemiss<-function(dataset,k=NA,showmiss=F,sortindex=T){
        Dataset<-dataset
        if(sortindex) Dataset<-Dataset[order(as.numeric(rownames(Dataset))),]
        Dataset$e<-NULL
        splitmiss<-split(Dataset,complete.cases(Dataset))
        if(length(splitmiss)==2){
                index.x1miss<-which(is.na(Dataset$x1))
                index.ymiss<-which(is.na(Dataset$y))
                rowmiss<-union(index.x1miss,index.ymiss)
                index.ymiss.only<-setdiff(index.ymiss,index.x1miss)
                compltdata<-splitmiss[[2]]
                ymiss.only.data<-Dataset[index.ymiss.only,]
                x1notmiss.data<-rbind(compltdata,ymiss.only.data)
rname<-rownames(Dataset)
rownames(Dataset)[index.x1miss]<-paste0(rownames(Dataset)[index.x1miss],"*")
rownames(Dataset)[index.ymiss]<-paste0(rownames(Dataset)[index.ymiss],"#")
                result.knn<-result.pmm<-result.em<-Dataset
                if(is.na(k)) k<-sqrt(nrow(compltdata))
                for(i in rowmiss){
                        target<-as.numeric(Dataset[i,])
                        indexmiss<-which(is.na(target))
                        d<-t(t(compltdata[,-indexmiss])-target[-indexmiss])
                        d2<-rowSums(d^2)
```

```
                rankd<-rank(d2,ties.method="max")
                slt<-rankd<=min(rankd[rankd>=k])
                if(length(indexmiss)>1){
                result.knn[i,indexmiss]<-colMeans(compltdata[slt,indexmiss])
                } else result.knn[i,indexmiss]<-mean(compltdata[slt,indexmiss])
                if(i %in% index.x1miss){
                        if(i %in% index.ymiss & length(index.ymiss.only)>0){
                        d.<-t(t(ymiss.only.data[,-indexmiss])-target[-indexmiss])
                        d2<-c(d2,rowSums(d.^2))
                result.pmm$x1[i]<mean(x1notmiss.data$x1[d2==min(d2)])
                } else result.pmm$x1[i]<-mean(compltdata$x1[d2==min(d2)])
        }
}
        lm_begin<-lm(y~.,compltdata)
        yhat<-predict(lm_begin,result.pmm)
        y.com<-yhat[-index.ymiss]
        for(i in index.ymiss){
                dij<-abs(y.com-yhat[i])
                result.pmm$y[i]<-mean(y.com[dij==min(dij)])
        }
        result.em$x1[index.x1miss]<-mean(x1notmiss.data$x1)
        bee0<-lm_begin$coef
        bet<-c(bee0,max_diff=NA)
        yhat<-predict(lm_begin,result.em[index.ymiss,])
        repeat{
                result.em$y[index.ymiss]<-yhat
```

```
                    lm_1<-lm(y~.,result.em)
                    bee<-lm_1$coef
                    dif<-max(abs(bee-bee0))
                    bet<-rbind(bet,c(bee,max_diff=dif))
                    if(all(bee==bee0)){
                    if(all.equal(bee,bee0)==T){
                    if(dif<0.001){ break
                    } else if(nrow(bet)>10000){ break
                    bet<-rbind(bet,"over flow!!!")
                    } else {
                        bee0<-bee
                        yhat<-lm_1$fitted[index.ymiss]
                    }
                }
                if(showmiss==F){
                    result<-list(KNN=result.knn,PMM=result.pmm,
                    EM=result.em,EM.beta=bet)
                } else result<-list(missing=splitmiss[[1]],KNN=result.knn,
                PMM=result.pmm,EM=result.em,EM.beta=bet)
        } else result<-"no missing data!!!"
        return(result)
}
```

```
## Simulate data##
n
sd.e<-c(10,30,90)
N
r<-sd.e/sd.e[1]
n.e<-length(sd.e)
DATA <- list()
DATA.complete <- list()
mse.knn <- mse.em <- mse.pmm <- matrix(NA,N,n.e)
temp<-list()
length(temp)<-N
KNN <- list()
for(g in 1:n.e) KNN[[g]]<- temp
names(KNN)<- paste("Var", sd.e^2,sep="")
EM <- PMM <- KNN
BETA.KNN <- list()
length(BETA.KNN)<-n.e
names(BETA.KNN)<- names(KNN)
BETA.EM <- BETA.PMM <- BETA.KNN
for(t in 1 : N){
## Generate x1,x2,x3 (equal-var) ##
x1<-rnorm(n,0,sqrt(100))
x2<-rnorm(n,0,sqrt(300))
x3<-rnorm(n,0,sqrt(500))
e0<-rnorm(n,0,sd.e[1])
```

```
per_mis1<-0.07

per_mis2<-0.10

per_mis3<-0.13

## cut x1 ##

mean_x1<-mean(x1)

sd_x1<-sd(x1)

c1_x1<-mean_x1+(-0.43)*(sd_x1)

c2_x1<-mean_x1+(0.43)*(sd_x1)

## divide x1 3part ##

x1[x1<=c1_x1]->x1_part1

x1[x1>c1_x1&x1<=c2_x1]->x1_part2

x1[x1>c2_x1]->x1_part3

## Generate x_miss binomial(0,1) ##

rbinom(length(x1_part1),1,per_mis1)->a1

rbinom(length(x1_part2),1,per_mis2)->b1

rbinom(length(x1_part3),1,per_mis3)->c1

miss_x1<-c(a1,b1,c1)

## Generate y_miss condition of Prob. ##

rbinom(length(x1_part1),1,14/107)->a2

miss_y_part1<-ifelse(a2==0,0,ifelse(a1==0,rbinom(length(x1_part1),1,0.5),1))

rbinom(length(x1_part2),1,2/11)->b2

miss_y_part2<-ifelse(b2==0,0,ifelse(b1==0,rbinom(length(x1_part2),1,0.5),1))

rbinom(length(x1_part3),1,26/113)->c2

miss_y_part3<-ifelse(c2==0,0,ifelse(c1==0,rbinom(length(x1_part3),1,0.5),1))

miss_y<-c(miss_y_part1,miss_y_part2,miss_y_part3)
```

```
E<-c()
for(g in 1:n.e) E<-cbind(E,e0*r[g])
Y<-42+x1+x2+x3+E
colnames(E)<-paste("e",1:n.e,sep="")
colnames(Y)<-paste("y",1:n.e,sep="")
data.complete<-data.frame(x1,x2,x3,E,Y)
data_sort<-data.complete[order(data.complete$x1),]
data_sort$x1[miss_x1==1]<-NA
data_sort[miss_y==1,ncol(data_sort)-n.e+1:n.e]<-NA
Y.data_sort<-data_sort[,ncol(data_sort)-n.e+1:n.e]
DATA[[t]] <- data_sort
DATA.complete[[t]] <- data.complete
for(g in 1:n.e){
data_sort<-data.frame(data_sort[,c("x1","x2","x3")],y=Y.data_sort[,g])
estimate<-estimatemiss(data_sort)
es.knn<-estimate["KNN"]
es.em<-estimate["EM"]
es.pmm<-estimate["PMM"]
KNN[[g]][[t]] <- es.knn
EM[[g]][[t]] <- es.em
PMM[[g]][[t]] <- es.pmm
## Find Regression model of KNN ##
fit.knn<-lm(y ~ x1 + x2 + x3, es.knn$KNN)
BETA.KNN[[g]] <- rbind(BETA.KNN[[g]],fit.knn$coef)
yhat.knn <- fit.knn$fitted
mse.knn[t,g] <- mean((Y[,g]-yhat.knn)^2)
```

```
## Find Regression model of EM ##
fit.em<-lm(y ~ x1 + x2 + x3, es.em$EM)
BETA.EM[[g]] <- rbind(BETA.EM[[g]],fit.em$coef)
yhat.em <- fit.em$fitted
mse.em[t,g] <- mean((Y[,g]-yhat.em)^2)
## Find Regression model of PMM ##
fit.pmm<-lm(y ~ x1 + x2 + x3, es.pmm$PMM)
BETA.PMM[[g]] <- rbind(BETA.PMM[[g]],fit.pmm$coef)
yhat.pmm <- fit.pmm$fitted
mse.pmm[t,g] <- mean((Y[,g]-yhat.pmm)^2)
}
pie(c(t,N-t),c(t,N-t), radius=1,clockwise=T)
}
# Find AMSE #
amse.knn <- colMeans(mse.knn)
amse.em <- colMeans(mse.em)
amse.pmm <- colMeans(mse.pmm)
amse.knn
amse.em
amse.pmm
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาววริษฐา กณิกนันต์ เกิดวันพฤหัสบดีที่ 4 พฤษภาคม พ.ศ.2532 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร ในปีการศึกษา 2553 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2554