ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (สถิติ)

ปริญญา

สถิติ สถิติ

สาขา **ภาควิชา**

เรื่อง การศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุ

A Comparative Study of Missing Data Estimation Methods for Multivariate Data

นามผู้วิจัย นางสาวน้ำทิพย์ พนมไทย

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์บุญอ้อม โฉมที, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(รองศาสตราจารย์ประสิทธิ์ พยัคฆพงษ์, M.S.)

หัวหน้าภาควิชา .....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุ

A Comparative Study of Missing Data Estimation Methods for Multivariate Data

โดย

นางสาวน้ำทิพย์ พนมไทย

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)
พ.ศ. 2554

น้ำทิพย์ พนมไทย 2554: การศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหาย สำหรับข้อมูลตัวแปรพหุ ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์บุญอ้อม โฉมที, Ph.D. 95 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุ 2 วิธี คือ Expectation Maximization Algorithm (EM) และ Regularized Expectation Maximization Algorithm (REM) เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) โดยที่ถ้าวิธีการประมาณค่าสูญหายวิธีใดให้ค่า MSE ต่ำสุด แสดงว่าวิธีนั้นมีประสิทธิภาพดีกว่า ข้อมูลที่ใช้ในการศึกษาครั้งนี้ผู้วิจัยได้จำลองข้อมูล ด้วยเทคนิคมอนติคาร์โล จำแนกสถานการณ์โดยกำหนดจำนวนตัวแปรอิสระ (X) เป็น 3, 4, 5, 7 และตัวแปรตาม (Y) เป็น 2 และ 3 ขนาดตัวอย่าง (n) เป็น 50 70 และ 100 และกำหนดให้ระดับความสัมพันธ์ระหว่างตัวแปรอิสระระดับสูง ($\rho = 0.7 - 0.9$) และระดับต่ำ ($\rho = 0.1 - 0.3$) ตัวแปรตามมีข้อมูลสูญหายแบบสุ่ม (Missing at Random, MAR) ที่ระดับการสูญหาย 10% , 20% และ 30% ดังนั้นข้อมูลที่จำลองตามสถานการณ์ต่าง ๆ มีจำนวนทั้งสิ้น 144 สถานการณ์ โดยทำซ้ำแต่ละสถานการณ์จำนวน 1,000 ครั้ง และศึกษาวิธีประมาณค่าสูญหาย กับข้อมูลจริงจำนวน 2 ชุด คือ 1) ข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ และ2) ข้อมูลกลุ่มความชื้น ของข้อมูลผลการตรวจอากาศชั้นบน ระหว่างวันที่ 1 มีนาคม 2549 - 31 ตุลาคม 2551 จากสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา สำนักฝนหลวงและการบินเกษตร

ผลการศึกษาจากการพิจาณาค่า MSE สำหรับข้อมูลจำลอง พบว่า กรณีระดับความสัมพันธ์ระหว่างตัวแปรอิสระสูง และจำนวนตัวแปรอิสระ 3 และ 4 ตัวแปร ขนาดตัวอย่าง 50 ที่ทุกระดับการสูญหาย วิธี EM เป็นวิธีที่เหมาะสม ส่วนขนาดตัวอย่าง 70 และ 100 ที่ทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม สำหรับจำนวนตัวแปรอิสระ 5 และ 7 ตัวแปร ที่ทุกขนาดตัวอย่าง และเกือบทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม กรณีระดับความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ และจำนวนตัวแปรอิสระ 3 และ 4 ตัวแปร ที่ทุกขนาดตัวอย่าง เกือบทุกระดับการสูญหาย วิธี EM เป็นวิธีที่เหมาะสม สำหรับจำนวนตัวแปรอิสระ 5 และ 7 ตัวแปร ที่ทุกขนาดตัวอย่าง เกือบทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม ส่วนผลการศึกษากับข้อมูลจริงจำนวน 2 ชุด พบว่า วิธี REM เป็นวิธีที่เหมาะสม ซึ่งผลการศึกษาส่วนใหญ่สอดคล้องกับผลการศึกษาจากข้อมูลจำลอง

_____ / _____ / _____

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Namthip Phanomthai 2011: A Comparative Study of Missing Data Estimation Methods for Multivariate Data. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Boonorm Chomtee, Ph.D. 95 pages.

The purpose of this research is to compare the missing data estimation methods for multivariate data between Expectation Maximization Algorithm (EM) and Regularized Expectation Maximization Algorithm (REM). The criterion of comparison is the mean squares error (MSE). The lower MSE indicates the higher effective estimation method. The datasets used in this study were simulated by the Monte Carlo technique. The comparisons were done under the conditions of independent variables (X) were 3, 4, 5, 7 and dependent variables (Y) were 2 and 3 ; sample sizes (n) are 50, 70 and 100 with the high levels of correlations among independent variables ($\rho = 0.7 - 0.9$) and low levels of correlations among independent variables ($\rho = 0.1 - 0.3$). The dependent variables were assigned to be missing at random (MAR) by 10%, 20% and 30% of missing data rates. These gave rise to a total of 144 possible situations with repeated 1,000 times under each situation. Also, the two real datasets; 1) Stability and forecasting indices group and 2) Moisture group. In upper air observation data during March 1, 2006 - October 31, 2008 from Pimai Radar Station of Bureau of Royal Rainmaking and Agricutural Aviation were used to compare the 2 missing data estimation methods.

The results based on MSE for simulation data showed that in cases of high levels of correlations among independent variables, 3 and 4 of independent variables, sample size 50, at all levels of missing data, EM is a suitable method. The sample sizes 70 and 100 at all levels of missing data, REM is a suitable method. For 5 and 7 of independent variables, at all sample sizes and almost level of missing data, REM is a suitable method. In cases of low levels of correlations among independent variables, 3 and 4 of independent variable, at all sample sizes and almost all level of missing data, EM methods is a suitable method. For 5 and 7 of independent variables, at all sample sizes and almost all levels of missing data, REM is a suitable method. In addition, the result for the two real datasets indicated that REM is a suitable method, which was consistent with that of simulation data.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยได้รับคำปรึกษาแนะนำ การสนับสนุน และตรวจแก้ไข ปรับปรุงข้อบกพร่องต่างๆ จาก ผู้ช่วยศาสตราจารย์ ดร.บุญอ้อม โฉมที อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รองศาสตราจารย์ประสิทธิ์ พยัคฆพงษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม และกราบขอบพระคุณ รองศาสตราจารย์ ดร. อภิญญา หิรัญวงษ์ ประธานกรรมการสอบ และ ผู้ช่วยศาสตราจารย์ ดร. กุศยา ปลั่งพงษ์พันธ์ ผู้ทรงคุณวุฒิภายนอก และท่านอาจารย์ทุกท่านที่ประสิทธิ์ประสาทความรู้ให้แก่ข้าพเจ้าตลอดมา

ขอขอบพระคุณวารสารวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น ที่ให้ความอนุเคราะห์ในการตีพิมพ์บทความ

ขอขอบคุณ พี่ ๆ เพื่อน ๆ และน้อง ๆ ทุกคนที่ให้ความช่วยเหลือและเป็นกำลังใจขณะทำวิทยานิพนธ์

สุดท้ายนี้ คุณค่าและประโยชน์อันจะพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ครอบครัว บูรพาจารย์ และผู้มีพระคุณทุกท่าน

น้ำทิพย์ พนมไทย

เมษายน 2554

สารบัญ

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

คำอธิบายสัญลักษณ์และคำย่อ

EM Algorithm	=	การคำนวณค่าข้อมูลสูญหายด้วยวิธี Expectation Maximization
REM Algorithm	=	การคำนวณค่าข้อมูลสูญหายด้วยวิธี Regularized Expectation Maximization
MCAR	=	การสูญหายแบบสุ่มอย่างสมบูรณ์
MAR	=	การสูญหายแบบสุ่ม
NMAR	=	การสูญหายแบบไม่สุ่ม
n	=	ขนาดตัวอย่าง
p	=	จำนวนตัวแปร
ρ	=	ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ
σ	=	ค่าเบี่ยงเบนมาตรฐานของข้อมูลประชากร
Σ	=	เมทริกซ์ความแปรปรวนร่วม
$\mathbf{X}$	=	เมทริกซ์ของข้อมูล
$\mathbf{x}$	=	เวกเตอร์ของข้อมูล
X	=	ตัวแปรที่สนใจศึกษา
x	=	ค่าของตัวแปร

การศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุ

A Comparative Study of Missing Data Estimation Methods for Multivariate Data

คำนำ

การวิเคราะห์ข้อมูลในทางสถิติ อาจจำแนกวิธีการวิเคราะห์ตามจำนวนตัวแปรที่ศึกษาเป็น 2 ประเภทใหญ่ๆ ด้วยกัน คือ การวิเคราะห์ทางสถิติระดับตัวแปรเดียว (Univariate Statistical Analysis) และการวิเคราะห์ทางสถิติระดับตัวแปรพหุ (Multivariate Statistical Analysis) ซึ่งการวิเคราะห์ระดับตัวแปรพหุนั้น เป็นการวิเคราะห์สำหรับตัวแปรที่เก็บรวบรวมข้อมูลที่วัดจากหน่วยตัวอย่างเดียวกัน ซึ่งในทางปฏิบัติตัวแปรเหล่านั้นอาจมีความสัมพันธ์กันมากกว่า 2 ตัวแปร เช่น การศึกษาพฤติกรรมในหลายๆด้านกับข้อมูลส่วนบุคคล เช่น เพศ รายได้ อายุ และทัศนคติบางเรื่องที่ส่งผลต่อพฤติกรรม หรือด้านการแพทย์ เช่น การศึกษาโรคหัวใจกับอายุ น้ำหนัก ปริมาณคอเรสเตอรอล การเป็นเบาหวาน ความดัน และกรรมพันธุ์ จะพบว่า ตัวแปรเหล่านี้ทุกตัวอาจมีความสัมพันธ์กัน

เทคนิคที่ใช้วิเคราะห์ตัวแปรพหุมีหลายเทคนิค เช่น การวิเคราะห์จำแนกประเภท (Discriminant Analysis) การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) การวิเคราะห์ปัจจัย (Factor Analysis) การวิเคราะห์ความแปรปรวนพหุ (Multivariate Analysis of Variance: MANOVA) และการวิเคราะห์การถดถอยโลจิสติค (Logistic Regression Analysis) เป็นต้น ซึ่งมีข้อสมมติเบื้องต้น (Assumption) ว่า ข้อมูลต้องถูกสุ่มมาจากประชากรที่มีการแจกแจงปกติพหุ (multivariate normal distribution) เมตริกซ์ความแปรปรวนร่วมของทุกกลุ่มต้องเท่ากัน และข้อมูลถูกสุ่มอย่างเป็นอิสระต่อกัน

การที่ชุดข้อมูลตัวแปรพหุมีข้อมูลหรือตัวแปรบางค่าสูญหาย จะมีผลกระทบต่อการวิเคราะห์ข้อมูล กล่าวคือ เมื่อพบว่าหน่วยตัวอย่างหนึ่งสำหรับตัวแปรใดๆมีข้อมูลสูญหายไปแม้เพียงตัวเดียว ก็จะตัดหน่วยตัวอย่างนั้นทิ้งทั้งหน่วย โดยไม่คำนึงว่าในหน่วยตัวอย่างนั้นๆ จะยังคงมีตัวแปรตัวอื่นๆ อีกมากที่มีข้อมูลครบถ้วน (Heeringa, 2000) การทำแบบนี้ทำให้กลุ่มตัวอย่างมีขนาดลดน้อยลง ซึ่งมีผลทำให้อำนาจการทดสอบต่ำและการประมาณค่าพารามิเตอร์มีความเอนเอียง (Roth, 1994) และยังส่งผลให้การสรุปผลที่ได้จากกลุ่มตัวอย่างอ้างอิงไปยังประชากรมีความคลาดเคลื่อนสูง

โดยเฉพาะอย่างยิ่งทำให้สูญเสียรายละเอียดหรือบางส่วนของข้อมูลชุดนั้นๆ ซึ่งอาจจะมีผลกระทบต่อการสรุปผลการวิเคราะห์ข้อมูล (วารุณี, 2538)

โดยทั่วไปรูปแบบการสูญหายของข้อมูล แบ่งออกเป็น 3 ประเภท คือ 1) การสูญหายแบบสุ่มสมบูรณ์ (Missing Complete At Random: MCAR) เกิดขึ้นเมื่อความน่าจะเป็นของข้อมูลสูญหายไม่มีความสัมพันธ์กับค่าของข้อมูลตัวอื่นๆ 2) การสูญหายแบบสุ่ม (Missing At Random: MAR) เกิดขึ้นเมื่อความน่าจะเป็นของข้อมูลสูญหายอาจจะขึ้นอยู่กับตัวแปรอื่นหรือสามารถทำนายจากตัวแปรอื่นได้ และ 3) การสูญหายแบบไม่สุ่ม (Not Missing At Random: NMAR) เมื่อสาเหตุของการสูญหายจะเกี่ยวข้องกับตัวแปรที่มีข้อมูลสูญหาย แต่ไม่มีความสัมพันธ์กับค่าของตัวแปรอื่น ๆ (Little and Rubin, 2002)

Little and Rubin (2002) แบ่งวิธีการจัดการข้อมูลสูญหายไว้ 4 วิธี คือ1)วิธีที่ใช้ข้อมูลสมบูรณ์ (Procedures based on completely recorded units) เช่น การตัดข้อมูลสูญหายแบบลิสท์ไวส์ (listwise deletion) การตัดข้อมูลสูญหายแบบแพร์ไวส์ (pairwise deletion) 2)วิธีการแทนค่า (Imputation based procedures) เช่น การแทนค่าแบบฮอทเดค (hot deck imputation) การแทนค่าโดยใช้ค่าเฉลี่ย (mean imputation) การแทนค่าโดยวิธีการถดถอย (regression imputation) 3)วิธีการถ่วงน้ำหนัก (Weighting procedures) และ 4)วิธีการที่ได้จากการนิยามโมเดล(Model-based procedures) เช่น วิธี Expectation Maximization (EM) วิธี Multiple Imputations (MI) เป็นต้น เพื่อให้การแทนค่าข้อมูลที่สูญหายมีความสมบูรณ์และใกล้เคียงกับความเป็นจริงก่อนที่จะนำไปวิเคราะห์

Raaijmakers (1999) กล่าวว่า วิธีการจัดการข้อมูลสูญหายมีหลายวิธี เช่น การตัดทิ้ง (Ignoring and discarding data) การแทนค่า (Imputation) การนิยามโมเดล (Model-based procedures) เป็นต้น ความแตกต่างของวิธีต่างๆ จะน้อยลงเมื่อขนาดของกลุ่มตัวอย่างมากขึ้น จำนวนข้อมูลสูญหายน้อยลง ตัวแปรที่มีข้อมูลสูญหายน้อย และความสัมพันธ์ระหว่างตัวแปรน้อยลง ในขณะที่ Schafer (1997) กล่าวถึงการตัดข้อมูลสูญหายทิ้งไป ว่า ถ้ามีจำนวนข้อมูลสูญหายน้อยประมาณ 5% การตัดหน่วยตัวอย่างทิ้งไปจะไม่มีผลเท่าไรนัก แต่ถ้าข้อมูลมีการสูญหายมากขึ้นการตัดหน่วยตัวอย่างทิ้งจะทำให้ข้อมูลนั้นไม่มีประสิทธิภาพและ ข้อมูลที่เหลืออยู่จะไม่เป็น ตัวแทนที่ดีของประชากร Roth (1994) พบว่า การเลือกวิธีการจัดการข้อมูลสูญหายจะมีความสำคัญเมื่อจำนวนข้อมูลสูญหายอยู่ระหว่าง 15-20% และจะมีความสำคัญมากที่สุดเมื่อจำนวนข้อมูลสูญหายประมาณ 30 - 40%

Schneider (2001) ศึกษาวิธีประมาณค่าสูญหายสำหรับข้อมูลเกี่ยวกับบรรยากาศ โดยใช้การประมาณค่าด้วยวิธีการวนซ้ำ (iterative method) พบว่ามีประสิทธิภาพดีกว่า วิธีแทนค่าเพียงครั้งเดียว (noniterative imputation)

จากผลการวิจัยที่เกี่ยวกับวิธีการจัดการข้อมูลสูญหายดังกล่าวข้างต้น ประกอบกับหน่วยงานที่ผู้วิจัยทำงานอยู่ในปัจจุบัน คือ สำนักฝนหลวงและการบินเกษตร ซึ่งเป็นหน่วยงานที่เก็บข้อมูลด้านลักษณะอากาศและการดัดแปรสภาพอากาศ โดยมีข้อมูลเกี่ยวกับผลการตรวจอากาศชั้นบน (upper air observation) เช่น อุณหภูมิ ความชื้น ความเร็วลม ทิศทางลม เป็นต้น ซึ่งโดยทั่วไปข้อมูลเหล่านี้มีความสัมพันธ์กันสูง และข้อมูลมีการสูญหายมาก สำหรับตัวแปรบางตัว ซึ่งการสูญหายอาจเนื่องมาจากความแปรปรวนของสภาพอากาศ หรือความผิดพลาดจากการตรวจวัด ทำให้ผู้วิจัยสนใจที่จะศึกษาและเปรียบเทียบประสิทธิภาพของวิธีประมาณค่าสูญหายในชุดข้อมูลตัวแปรพหุ 2 วิธี ได้แก่ วิธีประมาณค่าสูญหายแบบ Expectation Maximization Algorithm (EM) และแบบ Regularized Expectation Maximization Algorithm (REM) ที่ตัวแปรอิสระความสัมพันธ์กันสูง ($\rho = 0.7 - 0.9$) และต่ำ($\rho = 0.1 - 0.3$) ในสถานการณ์ต่างๆ ที่มีจำนวนตัวแปรอิสระ (X) เป็น 3, 4, 5, 7 และตัวแปรตาม (Y) เป็น 2 และ 3 ขนาดตัวอย่าง (n) เป็น 50, 70 และ 100 เมื่อตัวแปร Y มีการสูญหายแบบสุ่ม (Missing at Random, MAR) ที่ระดับการสูญหาย 10% , 20% และ 30%

วัตถุประสงค์

1. เพื่อศึกษาเปรียบเทียบวิธีประมาณค่าสูญหายแบบ EM Algorithm และแบบ REM Algorithm ในชุดข้อมูลตัวแปรพหุ ที่มีระดับความสัมพันธ์กันสูงและต่ำ

2. เพื่อเป็นแนวทางในการเลือกใช้วิธีการประมาณค่าสูญหายที่เหมาะสมกับลักษณะข้อมูล

ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลที่ได้จากการจำลองโดยเทคนิคมอนติคาร์โลตามสถานการณ์ที่กำหนด ทำการจำลองซ้ำ 1,000 ครั้ง โดยใช้โปรแกรม R (The R Project for Statistical Computing) และมีจำนวนสถานการณ์ทั้งหมด 144 สถานการณ์ โดยแบ่งเป็น สถานการณ์ที่ใช้วิธี EM จำนวน 72 สถานการณ์ และสถานการณ์ที่ใช้วิธี REM จำนวน 72 สถานการณ์ ดังภาพที่ 1

2. ข้อมูลมีการแจกแจงแบบปกติพหุ หรือ $N_p(\boldsymbol{\mu}, \Sigma)$, p = จำนวนตัวแปร มีค่าเฉลี่ย $\boldsymbol{\mu} = \mathbf{0}$ และเมทริกซ์ความแปรปรวนร่วม (covariance matrix) Σ โดยที่

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & K & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & L & \sigma_{2p} \\ M & M & O & M \\ \sigma_{p1} & \sigma_{p2} & L & \sigma_{pp} \end{bmatrix}$$

เมื่อ σ_{ij} เป็นความแปรปรวนร่วมระหว่างตัวแปร X_i และ X_j

σ_{ii} เป็นความแปรปรวนของตัวแปร X_i

โดยที่ $\sigma_{ij} = \rho_{ij}(\sigma_i \sigma_j)$; i, j = 1, 2, 3 ,….., p

และ ρ_{ij} เป็นสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X_i และ X_j

กำหนดค่า ρ_{ij} ในความสัมพันธ์ระดับสูงเป็น .70 , .80 และ .90

กำหนดค่า ρ_{ij} ในความสัมพันธ์ระดับต่ำเป็น .10 , .20 และ .30

3. ชุดข้อมูลแบ่งเป็น 4 แบบ โดยกำหนดจำนวนตัวแปร และระดับความสัมพันธ์ คือ

แบบที่ 1 ตัวแปรอิสระ 3 ตัว คือ (X_{1j} , X_{2j} และ X_{3j}) และตัวแปรตาม 2 ตัว คือ (Y_{1j} และ Y_{2j}) ในภาพที่ 1 ใช้สัญลักษณ์ 3X, 2Y

แบบที่ 2 ตัวแปรอิสระ 4 ตัว คือ (X_{1j}, X_{2j}, X_{3j} และ X_{4j}) และตัวแปรตาม 2 ตัว คือ (Y_{1j} และ Y_{2j}) ในภาพที่ 1 ใช้สัญลักษณ์ 4X, 2Y

แบบที่ 3 ตัวแปรอิสระ 5 ตัว คือ (X_{1j} , X_{2j},…., X_{5j}) และตัวแปรตาม 3 ตัว คือ (Y_{1j} , Y_{2j} และ Y_{3j}) ในภาพที่ 1 ใช้สัญลักษณ์ 5X, 3Y

แบบที่ 4 ตัวแปรอิสระ 7 ตัว คือ (X_{1j}, X_{2j},…., X_{7j}) และตัวแปรตาม 3 ตัว คือ (Y_{1j} , Y_{2j} และ Y_{3j}) ในภาพที่ 1 ใช้สัญลักษณ์ 7X, 3Y

ในความสัมพันธ์ระดับสูง กำหนดค่า $\sum$ ดังนี้

กรณีที่1 ตัวแปรอิสระ 3 ตัว $\sum = \begin{bmatrix} 1 & 0.7 & 0.8 \\ 0.7 & 1 & 0.9 \\ 0.8 & 0.9 & 1 \end{bmatrix}$

ในกรณีที่ 2 – 4 กำหนดค่า ρ_{ij} ในความสัมพันธ์ระดับสูงเป็น .70 - .80 เนื่องจากการจำลองข้อมูลที่ระดับความสัมพันธ์ .90 กรณีตัวแปรอิสระ 4 ตัวขึ้นไป จะทำให้เกิด Singular Metrix

กรณีที่ 2 ตัวแปรอิสระ 4 ตัว $\sum = \begin{bmatrix} 1 & 0.7 & 0.8 & 0.7 \\ 0.7 & 1 & 0.7 & 0.8 \\ 0.8 & 0.7 & 1 & 0.7 \\ 0.7 & 0.8 & 0.7 & 1 \end{bmatrix}$

กรณีที่ 3 ตัวแปรอิสระ 5 ตัว $\sum = \begin{bmatrix} 1 & 0.7 & 0.8 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 \\ 0.8 & 0.7 & 1 & 0.7 & 0.7 \\ 0.7 & 0.8 & 0.7 & 1 & 0.8 \\ 0.7 & 0.8 & 0.7 & 0.8 & 1 \end{bmatrix}$

กรณีที่ 4 ตัวแปรอิสระ 7 ตัว $\sum = \begin{bmatrix} 1 & 0.7 & 0.8 & 0.7 & 0.7 & 0.8 & 0.7 \\ 0.7 & 1 & 0.7 & 0.8 & 0.8 & 0.7 & 0.7 \\ 0.8 & 0.7 & 1 & 0.7 & 0.7 & 0.7 & 0.8 \\ 0.7 & 0.8 & 0.7 & 1 & 0.8 & 0.7 & 0.7 \\ 0.7 & 0.8 & 0.7 & 0.8 & 1 & 0.7 & 0.7 \\ 0.8 & 0.7 & 0.7 & 0.7 & 0.7 & 1 & 0.8 \\ 0.7 & 0.7 & 0.8 & 0.7 & 0.7 & 0.8 & 1 \end{bmatrix}$

ในความสัมพันธ์ระดับต่ำ กำหนดค่า Σ ดังนี้

กรณีที่ 1 ตัวแปรอิสระ 3 ตัว $\Sigma = \begin{bmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.1 \\ 0.2 & 0.1 & 1 \end{bmatrix}$

ในกรณีที่ 2 – 4 กำหนดค่า ρ_{ij} ในความสัมพันธ์ระดับต่ำเป็น .20 - .30 เพื่อให้สอดคล้องกับการจำลองข้อมูลที่ระดับความสัมพันธ์สูง

กรณีที่ 2 ตัวแปรอิสระ 4 ตัว $\Sigma = \begin{bmatrix} 1 & 0.3 & 0.2 & 0.3 \\ 0.3 & 1 & 0.3 & 0.2 \\ 0.2 & 0.3 & 1 & 0.3 \\ 0.3 & 0.2 & 0.3 & 1 \end{bmatrix}$

กรณีที่ 3 ตัวแปรอิสระ 5 ตัว $\Sigma = \begin{bmatrix} 1 & 0.3 & 0.2 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.2 & 0.3 & 1 & 0.2 \\ 0.3 & 0.2 & 0.3 & 0.2 & 1 \end{bmatrix}$

กรณีที่ 4 ตัวแปรอิสระ 7 ตัว $\Sigma = \begin{bmatrix} 1 & 0.3 & 0.2 & 0.3 & 0.3 & 0.2 & 0.3 \\ 0.3 & 1 & 0.3 & 0.2 & 0.2 & 0.3 & 0.3 \\ 0.2 & 0.3 & 1 & 0.3 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.3 & 1 & 0.2 & 0.3 & 0.3 \\ 0.3 & 0.2 & 0.3 & 0.2 & 1 & 0.3 & 0.3 \\ 0.2 & 0.3 & 0.3 & 0.3 & 0.3 & 1 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.3 & 0.3 & 0.2 & 1 \end{bmatrix}$

4. ค่าความคลาดเคลื่อนแบ่งเป็น 2 แบบ คือ

กรณีที่1 ตัวแปรตาม 2 ตัว

ความคลาดเคลื่อนมีการแจกแจงปกติด้วยค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 และ 5

กรณีที่ 2 ตัวแปรตาม 3 ตัว

ความคลาดเคลื่อนมีการแจกแจงปกติด้วยค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 , 5 และ 10

5. กำหนดขนาดตัวอย่าง 50 70 และ 100

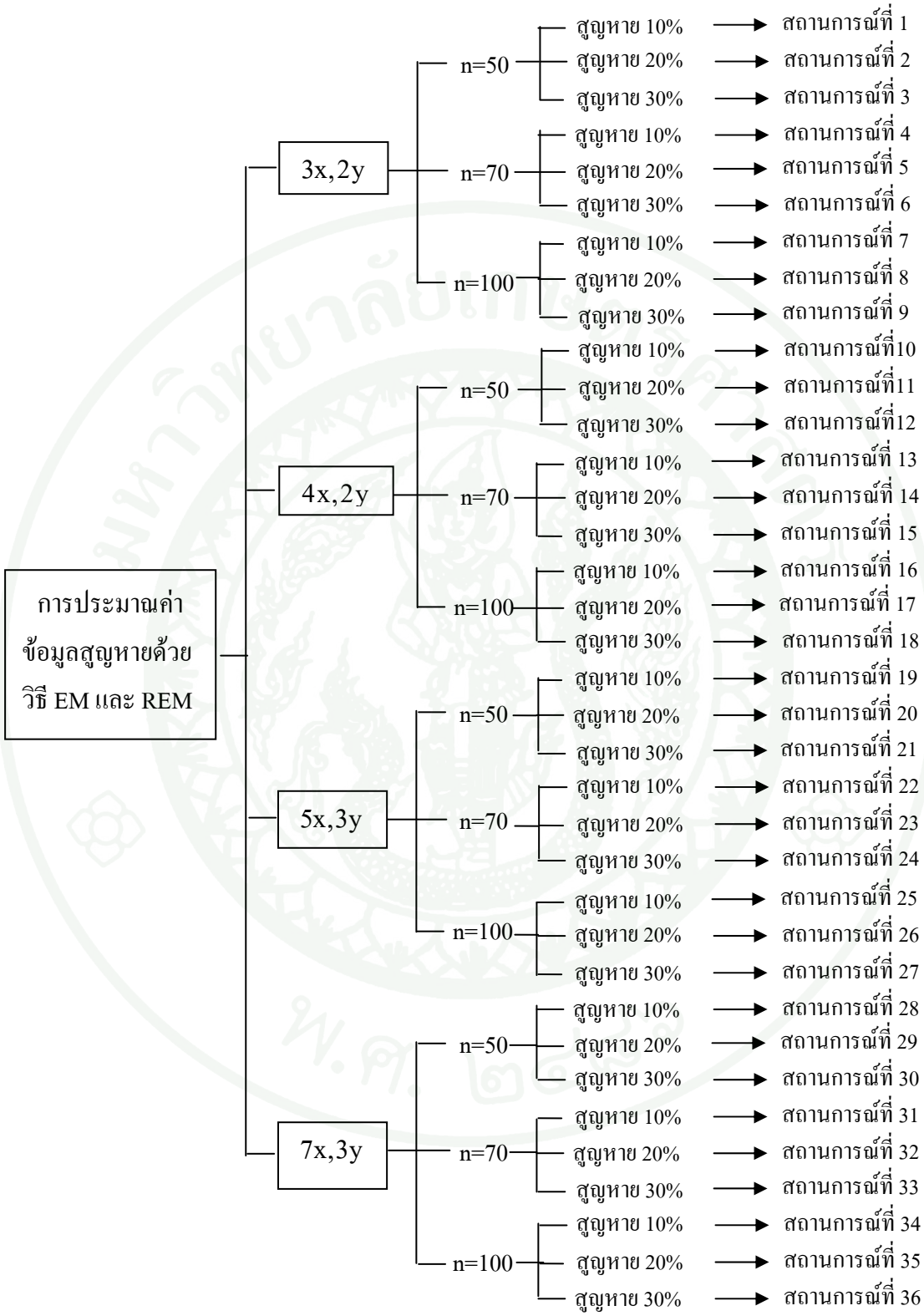
6. กำหนดการสูญหายของข้อมูลเฉพาะข้อมูลของตัวแปรตามเป็นการสูญหายแบบสุ่ม (Missing At Random: MAR) ที่ระดับ 10 20 และ 30 เปอร์เซ็นต์

7. เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squares Error: MSE) สามารถคำนวณได้ดังนี้

$$MSE = \frac{1}{1,000}\sum_{j=1}^{1,000}\left(\frac{\sum_{i=1}^{n}(Y_{ij}-\hat{Y}_{ij})^2}{n}\right)$$

8. ทดลองใช้วิธี EM และ REM กับข้อมูล 2 ชุดคือ 1) ข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ 2) ข้อมูลกลุ่มความชื้น ของข้อมูลผลการตรวจอากาศชั้นบน (upper air observation) ระหว่างวันที่ 1 มีนาคม 2549 - 31 ตุลาคม 2551 จากสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา สำนักฝนหลวงและการบินเกษตร กระทรวงเกษตรและสหกรณ์

9. กำหนดระดับนัยสำคัญในการทดสอบการแจกแจงแบบปกติพหุเป็น 0.01



ภาพที่ 1 การจำลองสถานการณ์ที่ประมาณค่าข้อมูลสูญหายด้วย EM Algorithm และ REM Algorithm 36 สถานการณ์ ในแต่ละระดับความสัมพันธ์ของตัวแปร (ระดับสูงและต่ำ) รวมจำนวน 72 สถานการณ์

ประโยชน์ที่คาดว่าจะได้รับ

1. ทราบวิธีการประมาณข้อมูลที่สูญหายที่เหมาะสมสำหรับชุดข้อมูลตัวแปรพหุที่มีระดับความสัมพันธ์กันสูงและต่ำ

2. ทราบแนวทางในการเลือกวิธีการประมาณค่าข้อมูลสูญหาย ที่เหมาะสมกับลักษณะข้อมูลที่คล้ายคลึงกับสถานการณ์ในการศึกษาครั้งนี้

การตรวจเอกสาร

การตรวจเอกสารแบ่งออกเป็น 2 ส่วน คือ ส่วนแรกกล่าวถึงผลงานวิจัยที่เกี่ยวข้องที่ได้มีผู้ศึกษาแล้วและส่วนที่สองเป็นรายละเอียดของทฤษฎีและวิธีการทางสถิติต่างๆที่ใช้ในการศึกษาวิจัย

ผลงานวิจัยที่เกี่ยวข้อง

ชะไมพร (2522) ศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายของตัวแปรอิสระในการวิเคราะห์การถดถอยวิธีต่างๆ 6 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีอันดับศูนย์หรือวิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง วิธีอันดับศูนย์ดัดแปลง วิธีถดถอยอันดับหนึ่งหรือวิธีสมการถดถอย วิธีถดถอยสองขั้น และวิธีผสมหรือวิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและสมการถดถอย โดยศึกษาจากข้อมูลทุติยภูมิ กำหนดให้ราคาข้าวเปลือกเจ้าชนิด 100% ราคาข้าวสารเจ้าชนิด 5% และปริมาณข้าวสารเจ้าส่งออก เป็นตัวแปรอิสระ ส่วนราคาข้าวสารเจ้าชนิด 100% เป็นตัวแปรตาม โดยนำข้อมูลที่สมบูรณ์มาจัดกระทำให้สูญหายแบบสุ่ม เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) ซึ่งใช้เป็นดัชนีในการตัดสินใจว่าวิธีประมาณค่าวิธีใดสามารถประมาณค่าได้ใกล้เคียงกับค่าที่สูญหายมากกว่ากัน ผลการศึกษาพบว่า วิธีที่ให้ค่า R^2 สูงกว่าวิธีอื่น ๆ มีอยู่ 3 วิธีเรียงตามลำดับจากมากไปน้อยคือ วิธีถดถอยสองขั้น วิธีถดถอยอันดับหนึ่งหรือวิธีใช้สมการถดถอย และวิธีกำลังสองน้อยที่สุด

พรศิริ (2529) ศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลที่สูญหายของตัวแปรตาม ในการวิเคราะห์ตัวแปรพหุคูณ 4 วิธี คือ วิธีค่าเฉลี่ย วิธีวิเคราะห์การถดถอยพหุคูณเชิงเส้น วิธีวิเคราะห์ความถดถอยเชิงเส้นดัดแปลง และวิธีวิเคราะห์ส่วนประกอบหลัก สำหรับกลุ่มตัวอย่างขนาด 30 50 70 100 และ 200 จำนวนตัวแปรอิสระเท่ากับ 3 5 7 และ 10 ตัวแปร ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเท่ากับ .10 .20 … .90 และกำหนดสัดส่วนข้อมูลที่สูญหายของแต่ละตัวแปรอิสระมีค่าเท่ากับ 10% ทำการศึกษาโดยใช้วิธีมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 1,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า MSE โดยวิธีใดที่ให้ค่า MSE ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่าวิธีประมาณค่าข้อมูลสูญหายในการวิเคราะห์ตัวแปรพหุคูณทั้ง 4 วิธี ให้ค่า MSE ไม่แตกต่างกันที่ระดับนัยสำคัญ .05 ไม่ว่าจะเป็นสถานการณ์ใดก็ตามที่มีข้อมูลสูญหายเกิดขึ้น

ถวัลย์ (2531) ศึกษาเปรียบเทียบวิธีประมาณค่าข้อมูลสูญหายของตัวแปรตามในกลุ่มตัวอย่างขนาดเล็ก 3 วิธี คือ วิธีค่าเฉลี่ย วิธีสมการถดถอย และ วิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยและสมการถดถอย สำหรับกลุ่มตัวอย่างขนาด 5 10 และ 15 กำหนดจำนวนข้อมูลที่สูญหายครั้งละ 1 และ 2 ค่า โดยใช้ข้อมูลที่มีลักษณะการแจกแจงแบบปกติสองตัวแปรและในกรณีที่ใช้สมการถดถอยช่วยในการประมาณค่านั้นได้กำหนดสัมประสิทธิ์สหสัมพันธ์ของตัวแปรต้นกับตัวแปรตามเท่ากับ .20 .40 และ .60 ทำการศึกษาด้วยวิธีมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 4,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า MSE โดยวิธีใดที่ให้ค่า MSE ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่าวิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยและสมการถดถอยให้ผลการประมาณค่าดีที่สุดเมื่อกลุ่มตัวอย่างเป็น 5 ในทุกกรณี และเมื่อกลุ่มตัวอย่างเพิ่มขึ้นเป็น 10 วิธีนี้จะประมาณค่าได้ดีเฉพาะข้อมูลสูญหายครั้งละ 2 ค่าที่มีค่าสัมประสิทธิ์สหสัมพันธ์เป็น 0.2 หรือ 0.4 เท่านั้น วิธีสมการถดถอยให้ผลการประมาณค่าดีที่สุดแทบทุกกรณีเมื่อขนาดตัวอย่างเป็น 10 หรือ 15 โดยที่จำนวนข้อมูลที่สูญหายครั้งละ 1 ค่า แต่ในจำนวนข้อมูลที่สูญหายเป็นครั้งละ 2 ค่า วิธีนี้จะประมาณได้ดีเฉพาะที่ตัวแปรต้นและตัวแปรตามมีความสัมพันธ์กันสูงระดับ 0.6 ส่วนวิธีค่าเฉลี่ยให้ผลการประมาณค่าไม่ดีในทุกกรณี

ชุติมา (2533) ศึกษาเปรียบเทียบการประมาณข้อมูลสูญหายของตัวแปรอิสระในการวิเคราะห์การถดถอยพหุคูณด้วยวิธีกำลังสองน้อยที่สุด 4 วิธี คือ วิธีสมการถดถอย วิธีเม็กซิมั่มไลลิฮูด วิธีค่าเฉลี่ย และวิธีค่ามัธยฐาน สำหรับกลุ่มตัวอย่างขนาด 30 70 และ 100 การกระจายข้อมูล 3 ระดับ โดยใช้ Coefficient of variation (C.V.) เป็นตัวกำหนด คือ .05 .20 และ1.00 จำนวนตัวแปรอิสระ 4 ระดับ คือ 2 3 5 และ 7 ค่าเบี่ยงเบนมาตรฐาน 4 ระดับ คือ 5 10 20 และ 25 และสัดส่วนข้อมูลที่สูญหายของตัวแปรอิสระ 3 ระดับ คือ 5% 10% และ 15% ทำการศึกษาโดยใช้วิธีมอนติคาร์โล ในแต่ละสถานการณ์ การกระทำซ้ำ 700 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า MSE ของสมการถดถอยของวิธีที่ไม่มีข้อมูลสูญหาย ผลการศึกษาพบว่าวิธีการประมาณข้อมูลสูญหายในการวิเคราะห์การถดถอยพหุคูณทั้ง 4 วิธี ให้ผลต่างกันตามสถานการณ์ต่าง ๆ ซึ่งโดยส่วนใหญ่วิธีค่าเฉลี่ยให้ผลดีที่สุด ยกเว้นเมื่อมีขนาดตัวอย่างน้อยและมีจำนวนตัวแปรอิสระมาก วิธีสมการถดถอยจะให้ผลดีที่สุด แต่ถ้าตัวอย่างมีขนาดใหญ่และมีจำนวนตัวแปรอิสระน้อยการตัดชุดของข้อมูลสูญหายทิ้งจะไม่มีผลกระทบต่อการวิเคราะห์การถดถอย

วารุณี (2538) ศึกษาเปรียบเทียบการประมาณข้อมูลสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ 5 วิธี คือวิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย วิธี EM และวิธีของฮันท์ (Hunt's Method) สำหรับกลุ่มตัวอย่างขนาด 10 20 30 50 70 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเป็น 5 10 15 20 และ 25 สัดส่วนการสูญหายของตัวแปรตาม 10% 20% 30%

40% 50% 60% และ 70% ทำการศึกษาด้วยวิธีมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 200 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ รากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ (RMSE) โดยวิธีใดที่ให้ค่า RMSE ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่า วิธีการของฮันท์เป็นวิธีการที่ดี เมื่อกลุ่มตัวอย่างมีขนาดเล็ก ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนน้อย และสัดส่วนการสูญหายมาก แต่ถ้าค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง วิธีค่าเฉลี่ยจะเป็นวิธีที่ดีในทุกสัดส่วนการสูญหายของตัวแปรตาม ส่วนในสถานการณ์ที่ตัวอย่างมีขนาดใหญ่ วิธีสูญหายจะเหมาะสมเกือบทุกกรณี

เชาว์ (2547) ศึกษาการพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี(Estimated parameter and the smallest standard error:EPSSE) และตรวจสอบความแม่นยำ และอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีกับอีก 2 วิธี คือแบบ EM และแบบลิสท์ไวส์ (Listwise deletion) สำหรับขนาดตัวอย่างเท่ากับ 350 โดยการสุ่มตัวอย่างแบบแบ่งชั้นแบบกลุ่ม และแบบหลายขั้นตอน ที่ความสัมพันธ์ระหว่างตัวแปรระดับต่ำ (r = .30) ปานกลาง (r = .50) และสูง (r = .70) และจำนวนข้อมูลสูญหาย 5% 10% 20% และ 30% ข้อมูลที่ใช้ศึกษามีลักษณะการแจกแจงแบบปกติสองตัวแปร และใช้วิธีมอนติคาร์โล กระทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่าความคลาดเคลื่อนเฉลี่ยของค่าเฉลี่ยเลขคณิต ความแปรปรวน สัมประสิทธิ์สหสัมพันธ์ และอำนาจการทดสอบ ผลการศึกษาพบว่าวิธีการประมาณค่าสูญหายโดยการแทนค่าแบบอีพีเอสเอสอี มีค่าความคลาดเคลื่อนเฉลี่ยไม่แตกต่างจากวิธีEM อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และวิธีการประมาณค่าสูญหายแบบลิสท์ไวส์ มีค่าความคลาดเคลื่อนเฉลี่ยสูงและแตกต่างจากวิธีการประมาณค่าสูญหายแบบอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

จริยา (2551) ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ 4 วิธี คือ วิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย และวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า (วิธีเอ็มไอ) สำหรับขนาดตัวอย่างเท่ากับ 50 70 100 และ 200 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 5 และ 15 เปอร์เซ็นต์การสูญหายของตัวแปรตามเท่ากับ 5% 10% 20% และ 30% ตามลำดับ และตัวแปรอิสระมีการแจกแจงแบบปกติพหุ ระดับความสัมพันธ์ระหว่างตัวแปรอิสระมี 3 ระดับ คือ ระดับต่ำ (0.20) ระดับปานกลาง (0.50) และระดับสูง (0.70) ทำการจำลองด้วยวิธีมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 5,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า RMSE โดยวิธีใดที่ให้ค่า RMSE ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่า วิธีสมการถดถอยและวิธีเอ็มไอให้ค่าประมาณของ RMSE ลดลงเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีการประมาณค่าสูญหายทั้ง 4 วิธี ให้ค่าประมาณของ RMSE แตกต่างกันเพิ่มขึ้นเมื่อ

เปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีสมการถดถอยและวิธีเอ็มไอ ให้ค่าประมาณของ RMSE ใกล้เคียงกัน

ประลองพล (2551) ศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์เมื่อมีค่าสูญหายเกิดขึ้นในตัวแปรอิสระของตัวประมาณตัวแบบความถดถอยโลจิสติกแบบ 2 กลุ่ม 4 วิธี คือ วิธีMean Imputation (MEAN) วิธี Maximum Likelihood Estimation (ML) วิธี Pseudo Maximum Likelihood Estimation (PML) และวิธี The Filling Method (FILL) เมื่อมีค่าสูญหายเกิดขึ้นในกรณีที่มีตัวแปรอิสระ 2 ตัว และเกิดค่าสูญหายในตัวแปรอิสระตัวที่สองเท่านั้น กำหนดขนาดตัวอย่าง 40, 100, 200 และ 400 ร้อยละการสูญหายของตัวแปรอิสระคือ 5, 10 และ 15 และค่าสหสัมพันธ์ระหว่างตัวแปรอิสระคือ 0, 0.1 และ 0.2 จำลองด้วยเทคนิคมอนติคาโล โดยในแต่ละสถานการณ์ กระทำซ้ำ 1,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่าเฉลี่ยของความแตกต่างระหว่างค่าจริงและค่าประมาณ (Bias) และระยะทางมาหาลาโนบิสเฉลี่ย (Average Mahalanobis Distance:AMH) โดยวิธีใดที่ให้ค่า AMH ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่า การเปรียบเทียบค่า Bias และค่า AMH ของทั้ง 4 วิธี พบว่า ในกรณีขนาดตัวอย่าง 40 วิธี MEANจะให้ค่า Bias และค่า AMH น้อยที่สุด แต่ในกรณีขนาดตัวอย่าง 100, 200 และ 400 วิธี FILL จะให้ค่า Bias และค่า AMH น้อยที่สุด และมีค่าใกล้เคียงกับวิธี ML และวิธี PML โดยที่ค่า Bias และค่า AMH มีแนวโน้มลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น และมีแนวโน้มเพิ่มขึ้นเมื่อสัดส่วนข้อมูลสูญหายในตัวแปรอิสระและระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

เพียงออ (2551) ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุเพื่อพยากรณ์ 4 วิธี คือ Regression Imputation (RI) วิธี Nearest Neighbor Imputation (NNI) วิธี Weighted Nearest Neighbor and Regression Imputation (WNR) และวิธี EM โดยพิจารณาข้อมูล 2 ลักษณะ คือ ข้อมูลภาคตัดขวาง และข้อมูลอนุกรมเวลา ทำการจำลองข้อมูลซึ่งกำหนดขนาดตัวอย่าง 50 , 100 และ 200 จำนวนตัวแปรอิสระเท่ากับ 2 ส่วนเบี่ยงเบนมาตรฐาน 5, 10, 15, 20 และ 25 ร้อยละการสูญหายของตัวแปรเป็น 5, 10 และ 20 ตามลำดับ ซึ่งกำหนดให้ตัวแปรอิสระมีการแจกแจงแบบปกติ ทำการศึกษาด้วยวิธีมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 1,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า Mean Absolute Percentage Error (MAPE) ในการเปรียบเทียบวิธีการพยากรณ์ที่มีความถูกต้องมากที่สุด จะต้องเป็นวิธีที่ให้ค่า MAPE ต่ำที่สุด ผลการศึกษาพบว่า สำหรับข้อมูลภาคตัดขวาง กรณีที่ค่าสหสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระทั้ง 2 ตัวสูง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำถึงปานกลางวิธี RI และ EM มีค่า MAPE ต่ำกว่าวิธีอื่น ๆ สำหรับข้อมูลอนุกรมเวลา วิธี WNR มีค่า MAPE ต่ำกว่าวิธีการ

ประมาณค่าอื่นๆในกรณีที่ข้อมูลที่มีอิทธิพลของฤดูกาลสูง และวิธี NNI จะให้ผลดีเมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับสูง สำหรับข้อมูลมีอิทธิพลจากปัจจัยแนวโน้มสูงวิธี RI และวิธี EM เป็นวิธีที่ให้ผลดีกว่าวิธีอื่นๆที่นำมาเปรียบเทียบ กรณีที่ข้อมูลมีอิทธิพลจากปัจจัยแนวโน้มและปัจจัยฤดูกาลระดับปานกลาง เมื่อส่วนเบี่ยงเบนมาตรฐานอยู่ในระดับต่ำ วิธี RI และวิธี EM มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่นๆ เมื่อส่วนเบี่ยงเบนมาตรฐานเพิ่มสูงขึ้น วิธี WNR เป็นวิธีที่มีค่า MAPE ต่ำกว่าวิธีการประมาณค่าอื่นๆ

ดวงภรณ์ (2552) ศึกษาเปรียบวิธีการประมาณค่าสูญหายในข้อมูลที่มีการวัดซ้ำ 3 วิธี คือวิธี Markov Chain Monte Carlo (MCMC) วิธี Copulas และวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย ใช้ข้อมูลที่ได้จากการจำลองสถานการณ์ที่มีการวัดซ้ำ 3 ครั้ง ภายใต้เงื่อนไขที่กำหนดคือ1) ข้อมูลมีการแจกแจงแบบปกติหลายตัวแปร มีโครงสร้างเมทริกซ์ความสัมพันธ์ 2 แบบ คือ เมทริกซ์ความสัมพันธ์แบบ Compound Symmetry (CS) หรือ แบบ Autoregressive (AR) 2) กำหนดระดับความสัมพันธ์ระหว่างค่าวัดซ้ำมี 3 ระดับ คือ ระดับต่ำ (0.3) ระดับปานกลาง (0.5) และ ระดับสูง (0.7) 3) ขนาดตัวอย่างที่ศึกษา 3 ขนาดคือ 30, 70 และ 100 และ 4) กำหนดให้ตำแหน่งการวัดซ้ำครั้งสุดท้ายมีข้อมูลสูญหายเกิดขึ้นแบบสุ่ม โดยมีระดับการสูญหายของข้อมูล 5% , 10% , 20% และ 30% ตามลำดับ ด้วยเทคนิคมอนติคาร์โล ในแต่ละสถานการณ์ กระทำซ้ำ 1,000 ครั้ง เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่า MSE ผลการศึกษาพบว่า วิธี Copulas มีประสิทธิภาพดีกว่า วิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย และ วิธี MCMC โดยให้ค่า MSE ต่ำสุด สำหรับวิธี MCMC มีประสิทธิภาพดีกว่าวิธีแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ย เฉพาะในกรณีโครงสร้างเมทริกซ์ความสัมพันธ์แบบ Autoregressive

Yuan (2000) ศึกษาและพัฒนาวิธีการจัดการข้อมูลสูญหาย โดยใช้ Multiple Imputation สำหรับข้อมูล Fitness และแสดงผลลัพธ์การวิเคราะห์ด้วยโปรแกรม SAS โดยกำหนดให้สาเหตุการสูญหายของข้อมูลเป็นแบบสุ่ม (MAR) และ เลือกใช้วิธี Regression และ วิธี Propensity score method กับรูปแบบการสูญหายแบบคล้ายกัน (monotone) และ ใช้วิธี Markov chain Monte Carlo (MCMC) กับรูปแบบการสูญหายแบบไม่เป็นระบบ (arbitrary)

Enders (2001) ศึกษาความสามารถของการประมาณค่าแบบเอฟไอเอ็มแอล (Full information maximum likelihood: FIML) ในการวิเคราะห์การถดถอยพหุคูณเมื่อมีข้อมูลสูญหาย 4 วิธี คือ วิธี FIML วิธี Listwise deletion วิธี Pairwise deletion และวิธีค่าเฉลี่ย ตัวแปรที่ศึกษามี 4 ตัวแปรคือ วิธีการประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย ขนาดของกลุ่มตัวอย่าง และขนาดของ

ความสัมพันธ์ระหว่างตัวแปร ใช้วิธีมอนติคาร์โล จำลองข้อมูลที่มีรูปแบบของการสูญหายแตกต่างกัน 3 แบบคือ การสูญหายแบบสุ่มสมบูรณ์ การสูญหายแบบสุ่ม และการสูญหายแบบไม่สุ่ม เกณฑ์ที่ใช้เปรียบเทียบ คือ สัมประสิทธิ์การถดถอย สัมประสิทธิ์การตัดสินใจ และ regression coefficient sampling variability ผลการศึกษา พบว่าการประมาณค่าข้อมูลสูญหายแบบ FIML ดีกว่าการตัดข้อมูลสูญหายออกแบบ วิธี Listwise deletion วิธี Pairwise deletion และการแทนข้อมูลด้วยวิธีค่าเฉลี่ย การประมาณค่าข้อมูลสูญหายแบบ FIML มีความเอนเอียงน้อย และมีปฏิสัมพันธ์ระหว่างวิธีประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย และขนาดความสัมพันธ์ระหว่างตัวแปร

Schneider (2001) ศึกษาและพัฒนาวิธีการประมาณข้อมูลสูญหายในการวิเคราะห์ข้อมูลบรรยากาศ ด้วยวิธี REM เปรียบเทียบกับวิธี conventional noniterative imputation โดยตัวแปรมีความสัมพันธ์กันสูง และจำนวนตัวแปรมากกว่าขนาดตัวอย่าง ด้วยการจำลองข้อมูลอุณหภูมิผิวพื้นจำนวน 53 ชุด มีตัวแปรจำนวน 1156 ตัวแปร กำหนดเปอร์เซ็นต์การสูญหายเท่ากับ 30% จากนั้นแบ่งข้อมูลออกเป็น 9 ชุด ทำให้แต่ละชุดมีข้อมูลสูญหาย 3.3% กำหนดจำนวนรอบในการวนซ้ำไม่เกิน 30 รอบ เกณฑ์ที่ใช้เปรียบเทียบ คือ Root mean square relative imputation error (rms relative imputation error) โดยวิธีใดที่ให้ค่า rms relative imputation error ต่ำกว่าเป็นวิธีที่ดีที่สุด ผลการศึกษาพบว่า วิธี REM Algorithm สามารถประมาณค่าสูญหายในข้อมูลบรรยากาศที่ไม่สมบูรณ์ได้ดีกว่าวิธีการประมาณค่าแบบไม่ใช้การวนซ้ำ (conventional noniterative imputation)

Truxillo (2005) ศึกษาและเปรียบเทียบวิธีการจัดการข้อมูลสูญหายในการวิเคราะห์จำแนกประเภท 7 วิธี คือ วิธี Listwise deletion วิธีค่าเฉลี่ย วิธีค่าเฉลี่ยแบบมีเงื่อนไข วิธี EM วิธี EM แบบมีเงื่อนไขสามแบบ (ใช้แถวที่เหลือค่าสังเกตน้อยที่สุด, จับคู่แถวที่ค่าสังเกตน้อยที่สุด, ใช้จำนวนแถวเฉลี่ยของค่าสังเกตแต่ละตัวแปร) สำหรับกลุ่มตัวอย่างขนาดใหญ่ (n = 100) และใช้ข้อมูล Iris ในการศึกษา กำหนดให้สาเหตุการสูญหายของข้อมูลเป็นแบบสุ่ม (MAR) และรูปแบบการสูญหายแบบไม่เป็นระบบ (arbitrary) เกณฑ์ที่ใช้เปรียบเทียบ คือ ค่าเฉลี่ยและค่าพิสัยของ First Canonical Correlation กับกรณีที่ค่าสังเกตสมบูรณ์ ผลการศึกษาพบว่า วิธี EM แบบมีเงื่อนไขโดยใช้จำนวนแถวเฉลี่ยของค่าสังเกตแต่ละตัวแปรได้ผลใกล้เคียงกับกรณีที่ค่าสังเกตสมบูรณ์ที่สุด การวิเคราะห์และแสดงผลลัพธ์การวิเคราะห์ด้วยโปรแกรม SAS

Huang and Carriere (2006) ศึกษาเปรียบเทียบวิธีการจัดการข้อมูลสูญหายในข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก 2 วิธี คือ วิธีการจัดการข้อมูลที่สูญหายระหว่างวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยใช้วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood method) กับวิธีใส่ค่าแทนข้อมูลที่สูญ

หายด้วยวิธี Multiple Imputation Procedure (MI) ใช้จำนวนค่าที่ใส่แทนข้อมูลที่สูญหาย m = 5 สำหรับข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก ข้อมูลมีการแจกแจงปกติพหุ(multivariate normal distribution) ทำการศึกษาโดยใช้สถานการณ์จำลอง เพื่อทดสอบสมมติฐานของอิทธิพลของ ทรีทเมนต์ และอิทธิพลของความคลาดเคลื่อนของทรีทเมนต์ ผลการศึกษาพบว่า วิธีการใส่ค่าแทนข้อมูลที่สูญหายด้วยวิธี MI จะให้ผลการทดสอบไม่แตกต่างจากวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยวิธีภาวะน่าจะเป็นสูงสุด

วิธีการทางสถิติ

ในส่วนนี้ผู้วิจัยจะกล่าวถึงความรู้พื้นฐานที่เกี่ยวข้องกับงานวิจัยโดยแบ่งออกเป็น 5 หัวข้อดังนี้

1. ประเภทของข้อมูลสูญหาย (Type of Missing Data)

ประสิทธิภาพของวิธีหรือเทคนิคที่นำมาใช้ในการประมาณค่าสูญหายของข้อมูลนั้นจะดีหรือไม่ส่วนหนึ่งขึ้นอยู่กับรูปแบบการสูญหายของข้อมูล และหากทราบสาเหตุที่ทำให้เกิดข้อมูลสูญหาย ก็จะสามารถเติมเต็มหรือเดาข้อมูลส่วนนั้นได้ไม่ยาก แต่ในการทำงานจริงมักจะไม่ทราบว่าการสูญหายของข้อมูลนั้นมีสาเหตุมาจากอะไร และสูญหายในลักษณะใด ดังนั้นในการทดลองต่าง ๆ จึงมักกำหนดรูปแบบการสูญหายของค่าข้อมูลออกเป็น 3 ประเภท (Little and Rubin, 2002) คือ

1.1. การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Complete At Random: MCAR) การสูญหายของข้อมูลด้วยวิธีการสุ่มอย่างสมบูรณ์เกิดขึ้นเมื่อความน่าจะเป็นของการสูญหายของข้อมูลไม่มีความสัมพันธ์กับค่าของข้อมูลตัวอื่น ๆ ไม่ว่าจะเป็นข้อมูลที่ทราบค่า หรือข้อมูลที่เกิดการสูญหายด้วยกันก็ตาม นั่นคือความน่าจะเป็นที่จะเกิดการสูญหายของค่าข้อมูลในทุก ๆ ตำแหน่งมีค่าเท่ากัน Chantala และ Suchindran (nd) ได้ยกตัวอย่างการสูญหายแบบ MCAR ไว้เช่น ข้อมูลรายได้ของพนักงานมีรูปแบบการสูญหายแบบ MCAR ถ้าพนักงานคนใดคนหนึ่งไม่รายงานรายได้ของตนเองจะถูกสมมติว่ามีค่าเฉลี่ยรายได้เหมือนกับพนักงานคนอื่น ๆ ที่รายงานรายได้ เป็นต้น

สำหรับข้อมูลสูญหายประเภทนี้จัดเป็นข้อมูลที่ก่อให้เกิดปัญหาน้อยที่สุด เพราะว่าข้อมูลสูญหายไม่มีความเกี่ยวข้องต่อผลลัพธ์ของข้อมูล เพราะฉะนั้นสามารถเลือกทำการวิเคราะห์ข้อมูลในส่วนที่สมบูรณ์ได้ (ปิยะภรณ์ และสุคนธ์, 2551)

1.2. การสูญหายแบบสุ่ม (Missing At Random: MAR) พิจารณาตัวแปร Y_1 และ Y_2 โดยให้ Y_1 มีข้อมูลสมบูรณ์ และ Y_2 มีข้อมูลสูญหาย ถ้าข้อมูลสูญหายในY_2 เป็น MAR ความน่าจะเป็นของ Y_2 อาจจะขึ้นอยู่กับ Y_1 หรือสามารถทำนายจากตัวแปร Y_1 ได้ แต่ข้อมูลสูญหายใน Y_2 จะไม่ขึ้นอยู่กับค่าสูญหายของตัวมันเอง ตัวอย่างเช่น รายได้ของพนักงานมีรูปแบบการสูญหายแบบ MAR ถ้าความน่าจะเป็นของข้อมูลรายได้ที่สูญหาย (Y_2) ขึ้นอยู่กับสถานภาพสมรส (Y_1) เช่น โสด แต่งงาน หรือหย่าร้าง แต่ความน่าจะเป็นของข้อมูลรายได้ที่สูญหายไม่ขึ้นอยู่กับข้อมูลรายได้ (Y_2) ของตัวมันเอง

1.3. การสูญหายแบบไม่สุ่ม (Not Missing At Random: NMAR) สาเหตุของการสูญหายไม่สามารถบอกได้ และสาเหตุของการสูญหายนั้นจะเกี่ยวข้องกับตัวแปรที่มีข้อมูลสูญหาย ซึ่งจะเรียกว่าเป็น Nonignorable กล่าวคือข้อมูลสูญหายสำหรับตัวแปร Y เป็น Nonignorable ถ้าความน่าจะเป็นของค่าสูญหายของ Y ไม่มีความสัมพันธ์กับค่าของตัวแปรอื่น ๆ แต่จะมีความสัมพันธ์กับค่าของตัวมันเอง เช่น ข้อมูลรายได้ จะถือว่ามีรูปแบบการสูญหายแบบ NMAR ถ้าครอบครัวที่มีรายได้สูงส่วนใหญ่จะไม่ชอบรายงานรายได้ของตนเอง จึงทำให้ข้อมูลสูญหาย หรือ คนที่ดื่มสุรามาก ๆ จะหลีกเลี่ยงการตรวจแอลกอฮอล์มากกว่าคนที่ดื่มน้อย ทำให้เกิดการสูญหายของข้อมูลและหาสาเหตุการสูญหายของข้อมูลได้ยาก

2. รูปแบบข้อมูลสูญหาย (Patterns of Missing Data)

มีรูปแบบดังนี้ (Chantala and Suchindran , n.d.)

2.1. ข้อมูลสูญหายหนึ่งตัวแปร (Univariate nonresponse) คือ ตัวแปร 1 ตัว มีข้อมูลสูญหาย

Case	Y_1	Y_2	Y_3
A	4	7	8
B	7	5	
C	5	8	
D	6	6	8

2.2. ข้อมูลสูญหายมากกว่าหนึ่งตัวแปร (Multivariate two patterns) คือ มีข้อมูลสูญหายมากกว่าหนึ่งตัวแปรในหน่วยตัวอย่างเดียวกัน

Case	Y_1	Y_2	Y_3
A	4	7	8
B	7	5	6
C	5		
D	6		

2.3. ข้อมูลสูญหายเป็นไปในทิศทางเดียวกัน (Monotone) คือ อันดับของตัวแปรหรืออันดับของค่าสังเกตในตัวแปรมีความสำคัญ นิยามคือ ให้เซตของตัวแปรคือ Y_1 , Y_2 ,..., Y_p ถ้า Y_i มีค่าสูญหายแล้ว Y_{i+1} , Y_{i+2} ,..., Y_p จะมีค่าสูญหายด้วย ดังแสดงในตารางสำหรับ p เท่ากับ 4

Case	Y_1	Y_2	Y_3	Y_4
A	4	7	8	4
B	7	5	6	
C	5	8		
D	6	6	8	5

2.4. ข้อมูลสูญหายแบบไม่เป็นระบบ (Arbitrary) โดยข้อมูลสูญหายสามารถเกิดขึ้นตรงจุดไหนก็ได้และอันดับของตัวแปรไม่มีความสำคัญ

Case	Y_1	Y_2	Y_3	Y_4
A	4	7	8	4
B		5	6	
C	5	8		7
D	6		8	5

3. เทคนิคที่ใช้ในการประมาณค่าที่สูญหายของข้อมูล (Method for Treating Missing Data)

จากการศึกษาพบว่าเทคนิคที่ใช้ในการประมาณค่าสูญหายของข้อมูลมีหลายวิธี Little and Rubin (2002) แบ่งวิธีการจัดการข้อมูลสูญหายไว้ดังนี้

3.1. การตัดข้อมูลทิ้ง (Ignoring and discarding data)

วิธีนี้เป็นวิธีที่ง่ายและสะดวกที่สุดโดยการตัดข้อมูลที่สูญหายออกไปแล้ววิเคราะห์ข้อมูลที่สมบูรณ์เท่านั้น ซึ่งนิยมใช้ในงานด้านสถิติ ตัวอย่างการตัดข้อมูลทิ้งมี 2 แบบ คือ

3.1.1. Listwise deletion เป็นการตัดค่าสังเกตหรือแถวที่มีข้อมูลสูญหายออกไปและจะนำค่าสังเกตที่สมบูรณ์เท่านั้นไปวิเคราะห์ วิธีนี้จะใช้ได้ก็ต่อเมื่อข้อมูลมีการสูญหายแบบสุ่มอย่างสมบูรณ์ (MCAR) เท่านั้น เนื่องจากการสูญหายในลักษณะอื่น ๆ นั้นจะมีความน่าจะเป็นของการสูญหายของข้อมูลขึ้นอยู่กับค่าของข้อมูลอื่น หรือค่าของข้อมูลที่เกิดการสูญหายของด้วย ข้อเสียของวิธีการนี้คือเนื่องจากมีการตัดข้อมูลที่มีการสูญหายออกทำให้ขนาดของกลุ่มตัวอย่างลดน้อยลง ดังนั้นอำนาจการทดสอบทางสถิติจึงลดลงด้วย

3.2.2. Pairwise deletion วิธีนี้จะไม่ตัดแถวที่ขาดความสมบูรณ์ทิ้งแต่จะนำแถวเหล่านั้นมาใช้ในการประมวลผลด้วย โดยจะพิจารณาทุก ๆ แถวที่มีค่าข้อมูลในตัวแปรที่กำลังสนใจถึงแม้ว่าตัวแปรอื่นจะไม่สมบูรณ์ก็ตาม วิธีการนี้มีข้อดีในส่วนของการใช้งานข้อมูลได้เต็มที่และมีประสิทธิภาพ แต่มีขั้นตอนที่ซับซ้อนกว่า Listwise เล็กน้อยและเสียเวลามากกว่า จึงได้รับความนิยมน้อย แต่วิธีนี้จะทำให้สูญเสียอำนาจการทดสอบน้อยกว่าวิธี Listwise

3.2. การแทนค่าข้อมูลสูญหายด้วยค่าประมาณที่ได้จากวิธีการต่าง ๆ (Imputation) ดังต่อไปนี้

3.2.1. การแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยหรือค่ามัธยฐาน เป็นวิธีที่ง่าย และนิยมใช้มากพอสมควรในการจัดการข้อมูลสูญหาย ในกรณีที่ข้อมูลเป็นข้อมูลตัวเลขและมีการแจกแจงแบบปกติจะแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยของตัวแปรจากค่าสังเกตที่มีข้อมูลสมบูรณ์ แต่ถ้าข้อมูลมีการแจกแจงแบบเบ้ควรแทนค่าข้อมูลสูญหายด้วยค่ามัธยฐาน และสำหรับข้อมูลเชิงคุณภาพมักแทนค่าข้อมูลสูญหายด้วยค่าฐานนิยม

3.2.2. การแทนค่าข้อมูลสูญหายแบบ Hot Deck เป็นการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของกลุ่ม ซึ่งแบ่งเป็น 2 ขั้นตอนคือ 1) การแบ่งข้อมูลออกเป็นกลุ่มที่มีลักษณะใกล้เคียงกัน 2) หาค่าเฉลี่ยหรือค่ามัธยฐานของกลุ่มที่มีข้อมูลสมบูรณ์มาแทนที่ในตำแหน่งที่สูญหายไปของแถวข้อมูลที่ไม่สมบูรณ์ซึ่งข้อมูลแต่ละแถวจะเป็นสมาชิกในกลุ่มใดกลุ่มหนึ่งที่ถูกแบ่งมาจากขั้นตอนแรก

3.2.3. การแทนค่าข้อมูลสูญหายด้วยวิธีการถดถอย (Regression imputation) วิธีนี้ใช้การวิเคราะห์การถดถอยเพื่อสร้างสมการทำนายข้อมูลสูญหายจากข้อมูลสมบูรณ์ที่มีอยู่ โดยกำหนดให้ตัวแปรอิสระ (x) มีข้อมูลสมบูรณ์และมีความสัมพันธ์กับตัวแปรตาม (y) ซึ่งเป็นตัวแปรที่มีข้อมูลสูญหาย แต่ถ้าตัวแปรอิสระกับตัวแปรตาม ไม่มีความสัมพันธ์กันอาจจะทำให้ตัวแบบการ

ประมาณค่าที่ได้ไม่ถูกต้องเท่าที่ควร เนื่องจากการสร้างตัวแบบการประมาณค่าในวิธีการนี้จะต้องอาศัยความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเป็นหลัก

3.3. วิธีที่อาศัยตัวแบบ (Model-based procedures) วิธีนี้ใช้หลักการเกี่ยวกับ Likelihood การประมาณค่าพารามิเตอร์ ที่เรียกว่า Maximum likelihood โดยมีการวนซ้ำ เช่น

3.3.1. วิธี Multiple Imputations (MI) เป็นเทคนิคในการแทนค่าสูญหายแต่ละค่าโดยใช้ค่าที่นำไปแทนตั้งแต่สองค่าขึ้นไป และใส่ค่าสูญหายซ้ำหลายครั้งซึ่งสุ่มค่าสูญหายจากการแจกแจงภายหลังของค่าพารามิเตอร์

3.3.2.วิธี Expectation Maximization (EM) ซึ่งเป็นวิธีหนึ่งที่ผู้วิจัยสนใจศึกษาเปรียบเทียบในงานวิจัยนี้ จะกล่าวรายละเอียดในหัวข้อที่ 4

4. ทฤษฎีที่เกี่ยวข้องกับ Expectation Maximization Algorithm

การแจกแจงแบบปกติพหุ (Multivariate Normal Distribution)

ให้เวกเตอร์ของตัวแปร $\mathbf{x}$ หรือ $\mathbf{x}' = (X_1, X_2, ..., X_p)$ มีการแจกแจงแบบปกติพหุ ที่มีเวกเตอร์ค่าเฉลี่ยเป็น $\boldsymbol{\mu}$ และ เมทริกซ์ความแปรปรวนร่วม (Covariance matrix) คือ $\boldsymbol{\Sigma}$ โดยที่ฟังก์ชันความน่าจะเป็นของเวกเตอร์ตัวแปร $\mathbf{x}$ คือ

$$\mathbf{f}(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^{\mathbf{P}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \mathbf{e}^{\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

โดยที่ p เป็นจำนวนตัวแปร หรือเขียนย่อๆ ว่า $\mathbf{x} \sim \mathrm{N}_{\mathrm{P}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

คุณสมบัติของเวกเตอร์ตัวแปรที่มีการแจกแจงแบบปกติพหุ

ถ้าเวกเตอร์ $\mathbf{x}' = (X_1, X_2, ..., X_p)$ มีการแจกแจงแบบปกติ นั่นคือ $\mathbf{x} \sim \mathrm{N}_{\mathrm{P}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ คุณสมบัติของเวกเตอร์ $\mathbf{x}$ เป็นดังนี้ (กัลยา, 2552)

1. ฟังก์ชันเชิงเส้นของตัวแปรสุ่ม ($\mathbf{x}$) ที่มีการแจกแจงแบบปกติจะมีการแจกแจงแบบปกติด้วย

1.1. ถ้า $\mathbf{a} = (a_1, ..., a_p)'$ เป็นเวกเตอร์ของค่าคงที่ $\mathbf{a'x} = a_1X_1 + a_2X_2 + ... + a_pX_p$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\mathbf{a'\mu}$ และความแปรปรวน $\mathbf{a'\sum a}$ หรือกล่าวได้ว่า

ถ้า $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sum)$ แล้วจะได้ว่า $\mathbf{a'x} \sim N_p(\mathbf{a'\mu}, \mathbf{a'\sum a})$

โดยที่ $E(\mathbf{a'X}) = \mathbf{a'}E(\mathbf{X}) = \mathbf{a'\mu}$ และ $Var(\mathbf{a'x}) = \mathbf{a'\sum a} = \boldsymbol{\sigma}^2_{\mathbf{a'x}}$

1.2. ถ้า $\mathbf{A}$ เป็นเมทริกซ์ของค่าคงที่ ซึ่งมีขนาด $q \times p$ ซึ่งมีลำดับที่ (rank) $= q$ โดยที่ $q \le p$ จะได้ว่า $\mathbf{Ax}$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\mathbf{A\mu}$ และความแปรปรวน $\mathbf{A\sum A'}$ เขียนแทนด้วย

$$\mathbf{Ax} \sim N_q(\mathbf{A\mu}, \mathbf{A\sum A'})$$

โดยที่ $E(\mathbf{Ax}) = \mathbf{A\mu}$ และ $Cov(\mathbf{Ax}) = \mathbf{A\sum A'}$

2. ตัวแปรปกติพหุมาตรฐาน

กำหนดให้ $\mathbf{z} = \sum^{\frac{-1}{2}}(\mathbf{x} - \boldsymbol{\mu})$ โดยที่ $\sum^{\frac{1}{2}}$ เป็นค่ารากที่สองของเมทริกซ์ความแปรปรวนร่วมที่สมมาตร (Symmetric square root matrix) ที่ทำให้ $\sum = \sum^{\frac{1}{2}}\sum^{\frac{1}{2}}$ จะได้ว่า เวกเตอร์ตัวแปรปกติพหุมาตรฐาน $\mathbf{z}$ จะมีการแจกแจงปกติพหุที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์ และเมทริกซ์ความแปรปรวนเป็น $\mathbf{I}$ และความแปรปรวนร่วมของตัวแปรทุกคู่เท่ากับศูนย์ นั่นคือ $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$ โดยที่ $\mathbf{I}$ เป็นเมทริกซ์เอกลักษณ์ (Identity matrix)

ถ้า Z_i เป็นตัวแปรที่ i ของเวกเตอร์ตัวแปรปกติพหุมาตรฐาน $\mathbf{z}$ จะได้ว่าตัวแปร Z_i เป็นตัวแปรที่มีการแจกแจงแบบปกติมาตรฐานที่มีค่าเฉลี่ยเป็นศูนย์ $(\boldsymbol{\mu}_{Z_i} = 0)$ และความแปรปรวนเป็นหนึ่ง $(\boldsymbol{\sigma}^2_{Z_i} = 1)$ นั่นคือ $Z_i \sim N(0,1)$

3. การแจกแจงแบบไคกำลังสอง

ถ้า $\mathbf{z}$ เป็นเวกเตอร์ตัวแปรปกติมาตรฐานที่เป็นอิสระกัน จะได้ว่า

$\sum_{i=1}^{p} \mathbf{z}_i^2 = \mathbf{z'z} = (\mathbf{x} - \boldsymbol{\mu})'\sum^{-1}(\mathbf{x} - \boldsymbol{\mu})$ มีการแจกแจงแบบไคกำลังสองที่มีองศาอิสระ p ($\boldsymbol{\chi}^2_p$) นั่นคือ

ถ้า $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sum)$ จะได้ว่า $(\mathbf{x} - \boldsymbol{\mu})'\sum^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \boldsymbol{\chi}^2_p$

4. การแจกแจงแบบปกติของเซตย่อยสำหรับตัวแปรปกติพหุ

4.1. ถ้า $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ จะได้ว่าเซตย่อยใดๆ ของเวกเตอร์ $\mathbf{x}$ จะมีการแจกแจงแบบปกติด้วย เช่น ถ้ากำหนดให้ $\mathbf{x}_1' = (X_1, X_2, ..., X_r)$ เป็นตัวแปร r ตัวแรกของเวกเตอร์ $\mathbf{x}$ และ $\mathbf{x}_2' = (X_{r+1}, X_{r+2}, ..., X_p)$ เป็นเซตของตัวแปรตัวที่ $r+1$ ถึงตัวแปรตัวที่ p ของเวกเตอร์ $\mathbf{x}$

ดังนั้น $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ และ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

โดยที่ $\mathbf{x}_1$ และ $\boldsymbol{\mu}_1$ เป็นเวกเตอร์ขนาด $r \times 1$ ส่วน Σ_{11} เป็นเมทริกซ์ขนาด $r \times r$ จะได้ว่า $\mathbf{x}_1$ มีการแจกแจงแบบ $N_r(\boldsymbol{\mu}_1, \Sigma_{11})$ และเมื่อ $\mathbf{x}_2$ และ $\boldsymbol{\mu}_2$ เป็นเวกเตอร์ขนาด $(p-r) \times 1$ และ Σ_{22} เป็นเมทริกซ์ขนาด $(p-r) \times (p-r)$ จะได้ว่า $\mathbf{x}_2$ มีการแจกแจงแบบ $N_{p-r}(\boldsymbol{\mu}_2, \Sigma_{22})$

ตัวอย่างเช่น กำหนดเวกเตอร์ $\mathbf{x}_3' = (X_1, X_4, X_9, X_{10})$ เป็นเซตย่อยของเวกเตอร์ $\mathbf{x}$ โดย $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ จะได้ว่า $\mathbf{x}_3$ จะมีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ย $\boldsymbol{\mu}_3$ และความแปรปรวนร่วม Σ_{33} หรือเขียนแทนด้วย $\mathbf{x}_3 \sim N_4(\boldsymbol{\mu}_3, \Sigma_{33})$

4.2. ถ้า X_i เป็นตัวแปรที่ i ในเวกเตอร์ $\mathbf{x}$ นั่นคือ X_i เป็นตัวแปรเดี่ยวที่มีการแจกแจงแบบปกติมีค่าเฉลี่ย m_i ความแปรปรวน s_i^2 หรือเขียนแทนด้วย $X_i \sim N(\mu_i, \sigma_i^2)$ แล้วไม่จำเป็นที่เวกเตอร์ $\mathbf{x}$ จะต้องมีการแจกแจงแบบปกติ

5. ความเป็นอิสระระหว่างเวกเตอร์ของตัวแปรสุ่ม

กำหนดให้ $\mathbf{x}$ และ $\mathbf{y}$ เป็นเวกเตอร์ย่อย 2 เวกเตอร์ โดยที่

$$E\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix} \text{ และ } \text{Cov}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

จะได้ว่า

5.1. $\mathbf{x}$ และ $\mathbf{y}$ จะเป็นอิสระกัน ถ้า $\Sigma_{xy} = \mathbf{0}$

5.2. ตัวแปร X_i และ X_j จะเป็นอิสระกัน ถ้า $\sigma_{ij} = 0$

6. การแจกแจงแบบมีเงื่อนไข

ถ้า $\mathbf{x}$ และ $\mathbf{y}$ ไม่เป็นอิสระกัน จะทำให้ $\Sigma_{xy} \neq \mathbf{0}$ การแจกแจงแบบมีเงื่อนไขของ x โดยกำหนด $\mathbf{y}$ คือ $\mathbf{f}(\mathbf{x}|\mathbf{y})$ จะมีการแจกแจงแบบปกติพหุที่มีค่าคาดหวังและค่าแปรปรวนร่วม ดังนี้

$$\mathbf{E}(\mathbf{x}|\mathbf{y}) = \boldsymbol{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) \text{ และ } \mathbf{Cov}(\mathbf{x}|\mathbf{y}) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$$

โดยค่า $\mathbf{E}(\mathbf{x}|\mathbf{y})$ เป็นฟังก์ชันเชิงเส้นของเวกเตอร์ $\mathbf{y}$ ในขณะที่ $\mathbf{Cov}(\mathbf{x}|\mathbf{y})$ ไม่ขึ้นกับ $\mathbf{y}$

7. การแจกแจงของผลบวกของเวกเตอร์ 2 เวกเตอร์

ถ้า $\mathbf{x}$ และ $\mathbf{y}$ เป็นเวกเตอร์ของตัวแปรที่มีขนาดเท่ากันคือ $p\times 1$ และเป็นอิสระต่อกันจะได้ว่า

$$\mathbf{x}+\mathbf{y} \sim N_p(\boldsymbol{\mu}_x+\boldsymbol{\mu}_y, \Sigma_{xx}+\Sigma_{yy})$$
$$\mathbf{x}-\mathbf{y} \sim N_p(\boldsymbol{\mu}_x-\boldsymbol{\mu}_y, \Sigma_{xx}+\Sigma_{yy})$$

ถ้าเวกเตอร์ $\mathbf{x}$ และ $\mathbf{y}$ มีการแจกแจงแบบปกติพหุ จะทำให้เวกเตอร์ $\mathbf{x}\pm\mathbf{y}$ จะมีการแจกแจงแบบปกติพหุด้วย

การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method: OLS Method)

วิธีการหาสัมประสิทธิ์การถดถอยโดยวิธี OLS คือ หาค่าประมาณของพารามิเตอร์ที่ทำให้ผลบวกกำลังสองของผลต่างระหว่างค่าสังเกตกับค่าคาดหวังของตัวแปรมีค่าต่ำที่สุด

จากสมการความสัมพันธ์ระหว่างตัวแปรตาม $\mathbf{y}$ และตัวแปรอิสระ $\mathbf{X}$ คือ

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad \text{เมื่อ} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

เมื่อ $\mathbf{y}$ คือ เวกเตอร์ของตัวแปรตามขนาด $n\times1$

$\mathbf{X}$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $n\times(p+1)$

$\boldsymbol{\beta}$ คือ เวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $(p+1)\times1$

n คือ ขนาดตัวอย่าง

p คือ จำนวนตัวแปรอิสระ

และ $\boldsymbol{\varepsilon}$ คือ เวกเตอร์ของความคลาดเคลื่อนขนาด $n\times1$

โดยทั่วไปเมื่อมีข้อมูลครบสมบูรณ์ วิธีกำลังสองน้อยที่สุดในการประมาณสัมประสิทธิ์ของการถดถอย โดยทำให้ผลบวกกำลังสองของความคลาดเคลื่อน(Sum Square of Error : SSE) มีค่าน้อยที่สุด

$$\begin{aligned} SSE &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y}-2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

การหาค่ากำลังสองน้อยที่สุดของผลบวกกำลังสองของความคลาดเคลื่อน ทำได้โดยหาอนุพันธ์ของ SSE เทียบกับ $\hat{\boldsymbol{\beta}}$ แล้วกำหนดให้เท่ากับศูนย์

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{y}'\mathbf{y}-2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})=\mathbf{0}$$

$$-2\mathbf{X}'\mathbf{y}+2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{0}$$

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y}$$

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

ดังนั้นสมการถดถอยที่ใช้พยากรณ์คือ

$$\hat{\mathbf{y}}=\mathbf{X}\hat{\boldsymbol{\beta}}$$

โดยที่ $E\left(\hat{\boldsymbol{\beta}}\right)=\boldsymbol{\beta}$ และ $V\left(\hat{\boldsymbol{\beta}}\right)=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2$

ซึ่งจะได้ตัวประมาณของ $\boldsymbol{\beta}$ หรือ $\hat{\boldsymbol{\beta}}$ ที่มีคุณลักษณะ 3 ประการ คือ มีความเป็นเส้นตรง (linear) เป็นตัวประมาณที่ไม่เอนเอียง (Unbiased estimator) และมีความแปรปรวนต่ำสุด (minimum variance) ในบรรดาตัวประมาณค่าที่ไม่เอนเอียงอื่นๆ ดังนั้น จึงเป็นเป็นตัวประมาณค่าที่มีประสิทธิภาพ ซึ่งเรียกว่า BLUE (Best linear unbiased estimator)

Expectation Maximization Algorithm

Dempster et al. (1977) ได้แนวความคิดในการหาตัวประมาณภาวะน่าจะเป็นสูงสุด เมื่อมีข้อมูลบางส่วนสูญหายสำหรับการแจกแจงแบบ Multinomial , Normal , Multivariate Normal โดยอาศัยวิธีการวนซ้ำ (Iterative Method) จนกระทั่งได้ตัวประมาณภาวะน่าจะเป็นสูงสุด ใช้ชื่อว่าวิธี EM ซึ่งมีชื่อเต็มว่า Expectation Maximization Algorithm

วิธีการนี้เป็นการหาค่าประมาณของค่าเฉลี่ยและความแปรปรวน โดยอาศัยหลักของกระบวนการวนซ้ำ ระหว่าง 2 ขั้นตอน คือ M-Step (Maximization) และ E-Step (Expectation) โดยมีข้อสมมติเบื้องต้น ดังนี้

1. ข้อมูลที่ใช้ต้องมีการแจกแจงแบบปกติพหุ (Multivariate Normal Distribution)
2. การสูญหายของข้อมูลเป็นแบบสุ่ม (Schafer, 1997)

สมมติให้ $\mathbf{Y}$ คือเมทริกซ์ขนาด $n\times p$ ที่มีขนาดตัวอย่างเท่ากับ n ด้วยเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}$ และเมทริกซ์ความแปรปรวนร่วม $\boldsymbol{\Sigma}$

ขั้นตอนที่ 1 :M-Step

เริ่มต้นด้วยค่าประมาณของ $\boldsymbol{\mu}$ และ $\boldsymbol{\Sigma}$ จากตัวอย่าง คือ $\overline{\mathbf{Y}}$ และ $\mathbf{S}$ ตามลำดับ และข้อมูลต้องไม่มีค่าสูญหาย ถ้าแต่ละแถวของชุดข้อมูลมีค่าสูญหายให้เริ่มต้นค่า $\boldsymbol{\mu}=\mathbf{0}$ และ $\boldsymbol{\Sigma}=\mathbf{I}$

ขั้นตอนที่ 2 : E-Step

ให้ $Y_{i(miss)}$ คือตัวแปรที่มีค่าสูญหาย และ $Y_{i(obs)}$ คือตัวแปรที่มีค่าสังเกต

หา $Y_{i(miss)}$ จากการคำนวณค่า $E(Y_{i(miss)} \mid Y_{i(obs)}; \hat{\mu}, \hat{\Sigma})$ และคำนวณค่า $COV(Y_{i(miss)} \mid Y_{i(obs)}; \hat{\mu}, \hat{\Sigma})$, i= 1,2,…,N โดย $\hat{\mu}, \hat{\Sigma}$ คือค่าประมาณจาก M-Step

วนซ้ำขั้นตอน M-Step และ E-Step จนกว่าค่า $(\hat{\mu}_{k+1}, \hat{\Sigma}_{k+1})$ และ $(\hat{\mu}_k, \hat{\Sigma}_k)$ มีค่าแตกต่างกันไม่เกินค่าขอบเขตที่ผู้วิจัยกำหนด

Little and Rubin (2002) ได้ประยุกต์วิธี EMโดยการประมาณค่าสูญหายด้วยการวิเคราะห์การถดถอยเชิงเส้นพหุ มีขั้นตอนดังนี้

1. จัดข้อมูลให้อยู่ในรูป

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

เมื่อ $\mathbf{y}_1$ คือ เวกเตอร์ของตัวแปรตามที่ไม่สูญหายขนาด $r \times 1$

$\mathbf{y}_2$ คือ เวกเตอร์ของตัวแปรตามที่สูญหายขนาด $(n-r) \times 1$

$\mathbf{X}_1$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $r \times (p+1)$ เมื่อชุดข้อมูลของตัวแปรตามไม่สูญหาย

$\mathbf{X}_2$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $(n-r) \times (p+1)$ เมื่อชุดข้อมูลของตัวแปรตามสูญหาย

2. ประมาณค่าสัมประสิทธิ์การถดถอยเริ่มต้น $\hat{\boldsymbol{\beta}}^0$ โดยวิธี Ordinary Least Squares (OLS) จากชุดข้อมูลที่ตัวแปรตามไม่สูญหาย ดังนี้

$$\hat{\boldsymbol{\beta}}^0 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$$

3. คำนวณ Expectation Step (E- Step)ในกระบวนการวนซ้ำรอบที่ 1 เพื่อประมาณค่าที่สูญหาย ดังนี้

$$\mathbf{E}(\mathbf{y}_i \mid \mathbf{X}, \mathbf{y}_1, \hat{\boldsymbol{\beta}}^0) = \begin{cases} \mathbf{y}_i & ; i = 1,\ 2, \ldots\ldots,\ r \\ \mathbf{X}_2\hat{\boldsymbol{\beta}}^0 & ; i = r+1,\ r+2\ , \ldots\ldots,\ n \end{cases}$$

4. แทนค่าข้อมูลสูญหายจากค่าที่ประมาณได้ใน E- Step แล้วคำนวณหาค่าประมาณของสัมประสิทธิ์การถดถอยใหม่ ($\hat{\boldsymbol{\beta}}^1$) โดยวิธี OLS ในขั้นนี้เรียกว่า Maximization step (M- Step) ในกระบวนการวนซ้ำรอบที่ 1

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

เมื่อ $\mathbf{y}$ คือ เวกเตอร์รวมของตัวแปรตามที่ไม่สูญหายขนาด $r \times 1$ และค่าประมาณของตัวแปรตามขนาด $(n-r)\times 1$ ที่ได้จากการประมาณใน E- Step

5. คำนวณค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยเริ่มต้น($\hat{\boldsymbol{\beta}}^0$) กับค่าสัมประสิทธิ์การถดถอยรอบที่ 1 ($\hat{\boldsymbol{\beta}}^1$) เขียนแทนด้วย $\left|\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}^1\right|$

6. จากข้อ 5 ค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยเริ่มต้น ($\hat{\boldsymbol{\beta}}^0$) กับค่าสัมประสิทธิ์การถดถอยรอบที่ 1 ($\hat{\boldsymbol{\beta}}^1$) ทุกค่า ต้องมีค่ามากกว่าหรือเท่ากับ δ เขียนแทนด้วย $\left|\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}^1\right| \geq \delta$ เมื่อ δ คือ เลขจำนวนจริงบวกที่ใช้ในการกำหนดความละเอียดของการคำนวณที่ผู้ศึกษาสามารถยอมรับได้โดยถ้าเงื่อนไข $\left|\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}^1\right| \geq \delta$ เป็นจริงให้ทำขั้นตอนต่อไป ในทางตรงกันข้ามจะได้ค่าประมาณค่าสูญหายด้วย E-Step รอบที่ 1

หมายเหตุ สำหรับค่า δ นั้น ผู้ใช้ควรกำหนดให้เหมาะสมกับมาตรวัดของข้อมูลนั้นๆ เช่น ถ้าข้อมูลที่มีระดับของมาตรวัดเป็นกิโลเมตร อาจจะกำหนดค่า δ เท่ากับ 0.05 กิโลเมตรหรือ 0.01 กิโลเมตร ตามความเหมาะสมของค่าประมาณที่ยอมรับได้ หรือถ้าข้อมูลมีมาตรวัดเป็นเซนติเมตร อาจกำหนดค่า δ ให้ละเอียดเพิ่มขึ้น เช่น อาจเท่ากับ 0.0001 เซนติเมตร เป็นต้น

7. คำนวณ E- Step ในกระบวนการวนซ้ำรอบที่ k ; k = 2, 3,……

$$\mathbf{E}(\mathbf{y}_i \mid \mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\beta}}^{k-1}) = \begin{cases} \mathbf{y}_i & ; i = 1,\ 2, \ldots\ldots,\ r \\ \mathbf{X}_2\hat{\boldsymbol{\beta}}^{k-1} & ; i = r+1,\ r+2, \ldots\ldots,\ n \end{cases}$$

8. คำนวณ M- Step ในกระบวนการวนซ้ำรอบที่ k ; k=2, 3,……

$$\hat{\boldsymbol{\beta}}^{k} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{k}$$

9. คำนวณค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ k-1 ($\hat{\boldsymbol{\beta}}^{k-1}$) กับค่าสัมประสิทธิ์การถดถอยรอบที่ k ($\hat{\boldsymbol{\beta}}^{k}$) หรือ $\left|\hat{\boldsymbol{\beta}}^{k-1} - \hat{\boldsymbol{\beta}}^{k}\right|$

10. ถ้าค่าสัมประสิทธิ์การถดถอยทุกค่าในขั้นตอนที่ 9 มากกว่า δ ให้กลับไปทำซ้ำขั้นที่ 7 ถึง 9 แต่ถ้าน้อยกว่า δ จะได้ค่าประมาณค่าที่สูญหายด้วย E –Step รอบสุดท้ายคือ $\mathbf{E}(\mathbf{y}_i|\mathbf{X},\ \mathbf{y},\ \hat{\boldsymbol{\beta}}^{*})$ ซึ่งจะเป็นตัวประมาณภาวะน่าจะเป็นสูงสุด

5. ทฤษฎีที่เกี่ยวข้องกับ Regularized Expectation Maximization Algorithm

การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีริดจ์รีเกรสชั่น (Ridge Regression Method:RR)

Hoerl and Kennard (1970) ได้ทำการศึกษาและพัฒนาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณที่ช่วยในการแก้ปัญหากรณีที่ตัวแปรอิสระเกิดปัญหาภาวะร่วมเส้นตรง (Multicollinearity) โดยไม่ต้องตัดตัวแปรอิสระออกจากตัวแบบ วิธีการนี้จะให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่าวิธี OLS โดยมีหลักการคือ พยายามลดค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการประมาณ $\boldsymbol{\beta}$ ให้น้อยลงโดยวิธีริดจ์รีเกรสชั่น (Ridge Regression, RR) ดังนี้

$$\hat{\boldsymbol{\beta}}_{RR} = (\mathbf{X}'\mathbf{X} + \mathbf{cI})^{-1}\mathbf{X}'\mathbf{y}$$

โดยที่ c เป็นค่าคงที่ใดๆ ที่ $c \geq 0$, **I** เป็น identity matrix

ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์รีเกรสชั่นจะต้องเลือกค่า c ที่เหมาะสม ซึ่งจะให้ค่า MSE ต่ำเมื่อ $c_m = \dfrac{\sigma^2}{\boldsymbol{\beta}_m'\boldsymbol{\beta}_m}$ ซึ่งค่า c เกิดจากวิธีการคำนวณที่วนซ้ำรอบที่ m, เมื่อ m =1, 2, 3,… เมื่อต้องการค่า c เพียงค่าเดียว จึงหาค่า c จากค่าเฉลี่ยฮาร์โมนิค (harmonic mean) ของ c โดยที่

$$\frac{1}{c} = \frac{\sum(1/c_m)}{p} = \frac{\sum(\hat{\boldsymbol{\beta}}_m^2/\hat{\sigma}^2)}{p} = \frac{\sum\hat{\boldsymbol{\beta}}_m^2}{p\hat{\sigma}^2}$$

จะได้ว่า $c = \dfrac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}$

โดยที่ p คือ จำนวนตัวแปรอิสระ

$\hat{\beta}$ คือ เวกเตอร์ค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณโดยวิธี OLS

$\hat{\sigma}^2$ คือ ค่าประมาณของความแปรปรวนโดยวิธี OLS

ซึ่งได้มีการแสดงให้เห็นว่าค่าMSE ของวิธีริดจ์รีเกรสชั่นลดลงจากวิธี OLS อย่างมีนัยสำคัญ

Regularized Expectation Maximization Algorithm

Schneider (2001) ได้ศึกษาวิธี REM Algorithm ซึ่งพัฒนาจาก EM Algorithm ประกอบด้วยขั้นตอนเหมือนกับ EM Algorithm ยกเว้นว่าในแต่ละรอบที่มีการวนซ้ำเพื่อหา ค่าประมาณพารามิเตอร์ $\hat{\boldsymbol{\beta}}^{\mathbf{k}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{\mathbf{k}}$ จะถูกประมาณด้วย $\hat{\boldsymbol{\beta}}^{\mathbf{k}}_{\mathbf{RR}} = (\mathbf{X}'\mathbf{X} + \mathbf{cI})^{-1}\mathbf{X}'\mathbf{y}^{\mathbf{k}}$

เนื่องจาก Schneider พบว่า การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS ด้วย $\hat{\boldsymbol{\beta}}^{\mathbf{k}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{\mathbf{k}}$ สำหรับข้อมูลที่มีปัญหาภาวะร่วมเชิงเส้นตรง จะทำให้ค่าความคลาดเคลื่อนสูงขึ้น ดังนั้นจึงเลือกใช้การประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธีริดจ์รีเกรสชั่น นั่นคือ $\hat{\boldsymbol{\beta}}^{\mathbf{k}}_{\mathbf{RR}} = (\mathbf{X}'\mathbf{X} + \mathbf{cI})^{-1}\mathbf{X}'\mathbf{y}^{\mathbf{k}}$ ซึ่งทำให้มีความคลาดเคลื่อนน้อยกว่า

เกณฑ์การเปรียบเทียบ

งานวิจัยนี้ได้ทำการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายโดยพิจารณาจากค่า MSE ซึ่งวิธีการประมาณค่าสูญหายวิธีใดให้ค่า MSE ต่ำสุด แสดงว่าวิธีนั้นมีประสิทธิภาพดีกว่า สามารถคำนวณได้ดังนี้

$$MSE = \frac{1}{1,000}\sum_{j=1}^{1,000}\left(\frac{\sum_{i=1}^{n}(Y_{ij}-\hat{Y}_{ij})^2}{n}\right)$$

เมื่อ n คือ จำนวนข้อมูลที่มีการสูญหาย

$\hat{Y}_{ij}$ คือ ค่าประมาณของข้อมูลสูญหาย

Y_{ij} คือ ค่าจริงของข้อมูล

j คือ จำนวนรอบของการทำซ้ำ j=1, 2,..., 1,000

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องไมโครคอมพิวเตอร์ที่มีหน่วยความจำขนาด 40 GB มีความเร็วในการประมวลผล 256 MHz ที่ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

2. โปรแกรม R (The R Project for Statistical Computing) และ REvolution R (Revolution Analytics)

3. ข้อมูลผลการตรวจอากาศชั้นบน (upper air observation) ของสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา

วิธีการ

ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลที่ได้จากการจำลองสถานการณ์ที่กำหนดโดยใช้โปรแกรม R (The R Project for Statistical Computing) มีขั้นตอนดังนี้

1. จำลองข้อมูลตัวแปรอิสระที่มีการแจกแจงแบบปกติพหุ ตามขนาดตัวอย่าง และระดับความสัมพันธ์ตามที่กำหนดไว้ในขอบเขตการวิจัย

2. จำลองข้อมูลค่าความคลาดเคลื่อนที่มีการแจกแจงแบบปกติ ตามที่กำหนดไว้ในขอบเขตการวิจัย

3. การกำหนดค่า β เพื่อให้เกิดค่าความคลาดเคลื่อนของตัวประมาณน้อยที่สุดค่า β ควรมีค่าระหว่าง -1 ถึง 1 (Bolch and Huang, 1974) จึงกำหนด $\beta_1 = \beta_2 = \ldots = \beta_k = 1$ โดยที่ $\beta_0 = 0$

กรณีที่1 ตัวแปรอิสระ 3 ตัว และตัวแปรตาม 2 ตัว $\beta=\begin{bmatrix}0 & 0\\ 1 & 1\\ 1 & 1\\ 1 & 1\end{bmatrix}$

กรณีที่ 2 ตัวแปรอิสระ 4 ตัว และตัวแปรตาม 2 ตัว $\beta=\begin{bmatrix}0 & 0\\ 1 & 1\\ 1 & 1\\ 1 & 1\\ 1 & 1\end{bmatrix}$

กรณีที่ 3 ตัวแปรอิสระ 5 ตัว และตัวแปรตาม 3 ตัว $\beta=\begin{bmatrix}0 & 0 & 0\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$

กรณีที่ 4 ตัวแปรอิสระ 7 ตัว และตัวแปรตาม 3 ตัว $\beta=\begin{bmatrix}0 & 0 & 0\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$

4. สร้างตัวแปรตามที่มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ ค่าสัมประสิทธิ์การถดถอย และค่าความคลาดเคลื่อน โดยใช้รูปแบบความสัมพันธ์เชิงเส้น

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \qquad \text{เมื่อ} \quad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0},\sigma^2\mathbf{I}_n\right)$$

5. คำนวณหาจำนวนชุดข้อมูลที่สูญหายและสุ่มจำนวนชุดข้อมูลสูญหาย

5.1 คำนวณหาจำนวนชุดข้อมูลที่สูญหาย จาก

จำนวนชุดข้อมูลที่สูญหาย = $\frac{\text{ขนาดตัวอย่าง X เปอร์เซ็นต์การสูญหาย}}{100}$

ยกตัวอย่างกรณี n = 50 เปอร์เซ็นต์การสูญหาย = 10%

จำนวนชุดข้อมูลที่สูญหาย = $\frac{50 \times 10}{100} = 5$ ชุด เป็นต้น

5.2 ทำการสุ่มชุดข้อมูลที่สูญหาย กำหนดการสูญหายของชุดข้อมูลตัวแปรตามเป็นการสูญหายแบบสุ่ม (Missing At Random: MAR) และเปอร์เซ็นต์การสูญหายของชุดข้อมูลเป็น 10 20 และ 30 เปอร์เซ็นต์

6. กำหนดการประมาณค่าสูญหายด้วยวิธี EM

7. กำหนดการประมาณค่าสูญหายด้วยวิธี REM

8. ในกระบวนการวนซ้ำเพื่อหาค่าประมาณพารามิเตอร์ สำหรับวิธีประมาณข้อ 6 และ 7 นั้น กำหนดให้มีการวนซ้ำสูงสุดจำนวน m = 30 รอบ

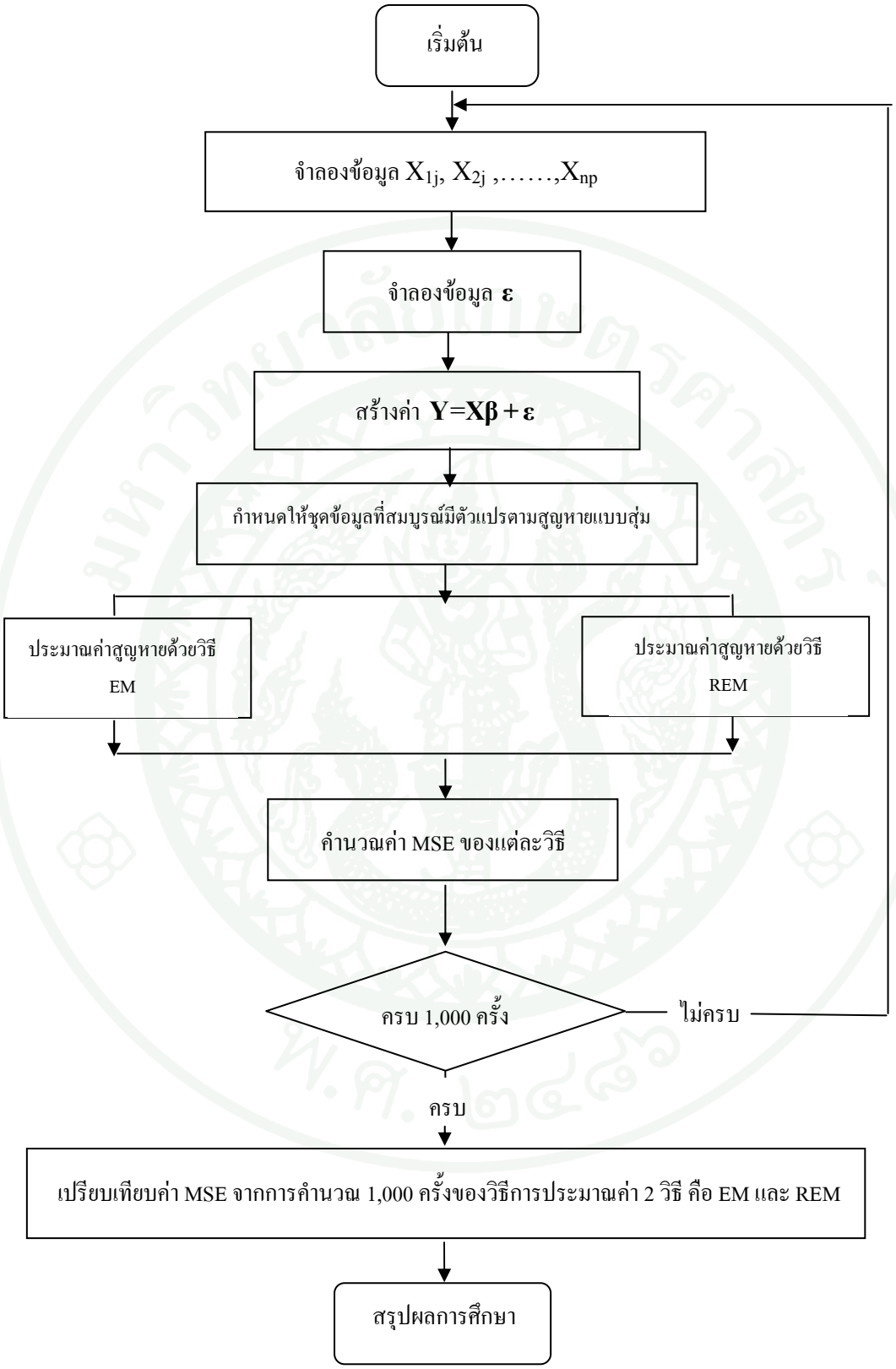
9. เปรียบเทียบค่าประมาณที่ได้จากวิธี EM และ วิธี REM กับค่าในข้อมูลสมบูรณ์ด้วยค่า MSE

10. กระทำซ้ำ 1,000 ครั้งในแต่ละสถานการณ์

11. สรุปผลในแต่ละสถานการณ์

12. นำวิธีการประมาณค่าสูญหายทั้ง 2 วิธี มาทดลองใช้กับข้อมูล 2 ชุดคือ 1) ข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ 2) ข้อมูลกลุ่มความชื้น ของข้อมูลผลการตรวจอากาศชั้นบน (upper air observation) ระหว่างวันที่ 1 มีนาคม 2549 - 31 ตุลาคม 2551 จากสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา สำนักฝนหลวงและการบินเกษตร กระทรวงเกษตรและสหกรณ์ (รายละเอียดดังภาคผนวก)

13. สรุปผลการศึกษา



ภาพที่ 2 ผังงานของขั้นตอนการดำเนินงาน

ผลและวิจารณ์

ผล

การศึกษาวิธีประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุ 2 วิธี คือ วิธี EM และวิธี REM ได้ทำการศึกษากับข้อมูลที่ได้จากการจำลองสถานการณ์โดยเทคนิคมอนติคาร์โล ด้วยโปรแกรม R ทำการจำลองซ้ำ 1,000 ครั้งในแต่ละสถานการณ์ จำนวนทั้งหมด 144 สถานการณ์ และได้นำวิธีการประมาณค่าสูญหายทั้ง 2 วิธีมาทดลองกับข้อมูลจริง โดยผลการเปรียบเทียบประสิทธิภาพของทั้ง 2 วิธี แบ่งเป็น 2 ส่วนคือ ส่วนที่เป็นข้อมูลจำลองโดยเทคนิคมอนติคาร์โล และส่วนที่เป็นข้อมูลจริง

ส่วนที่ 1 ผลการศึกษาจากข้อมูลจำลอง

ข้อมูลที่ได้จากการจำลองโดยเทคนิคมอนติคาร์โล กำหนดให้ตัวแปรอิสระมีการแจกแจงแบบปกติพหุ และตัวแปรตามมีข้อมูลสูญหายแบบสุ่มที่ระดับการสูญหาย 10 %, 20 % และ 30% ของขนาดตัวอย่าง 50 , 70 และ 100 จากการจำลอง 1,000 ครั้ง ในแต่ละสถานการณ์ และเนื่องจากลักษณะของชุดข้อมูล มี 4 แบบ ที่แต่ละระดับความสัมพันธ์ (ระดับสูงและต่ำ) โดยจำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย ดังนั้น ในการนำเสนอผลการศึกษาเปรียบเทียบประสิทธิภาพด้วยค่า MSE จะแบ่งเป็น 8 กรณี คือ สำหรับความสัมพันธ์ระหว่างตัวแปรอิสระสูง (0.70 – 0.90) 4 กรณี สำหรับความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ (0.10 – 0.30) 4 กรณี ดังต่อไปนี้

1. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 3 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

2. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 4 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

3. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 5 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

4. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 7 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

5. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 3 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

6. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 4 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

7. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 5 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

8. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 7 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

พิจารณาเปรียบเทียบค่า MSE ของแต่ละวิธี ในการจำลองสถานการณ์ โดยกระทำซ้ำ1,000 ครั้ง ในแต่ละสถานการณ์ ถ้าวิธีการประมาณค่าสูญหายวิธีใดให้ค่า MSE ต่ำสุด แสดงว่าวิธีนั้นมีประสิทธิภาพดีกว่า

1. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 3 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

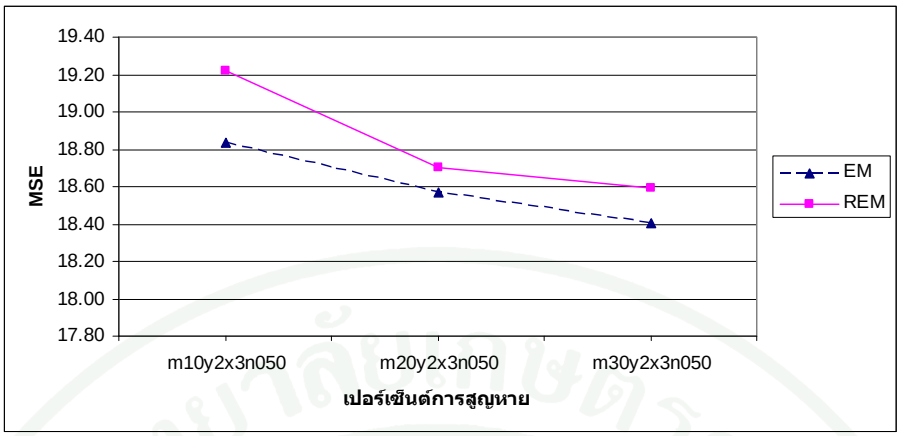
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 1 และภาพที่ 3 สำหรับขนาดตัวอย่างเท่ากับ 50 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM สำหรับขนาดตัวอย่างเท่ากับ 70 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 100 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 1 และภาพที่ 4 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 ที่ขนาดตัวอย่าง 50 และ 70 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อขนาดตัวอย่าง 100 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 20 และ 30 ที่ขนาดตัวอย่างเท่ากับ 50 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM ที่ขนาดตัวอย่าง 70 และ 100 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สรุปโดยรวมว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลง ทั้ง 2 วิธี

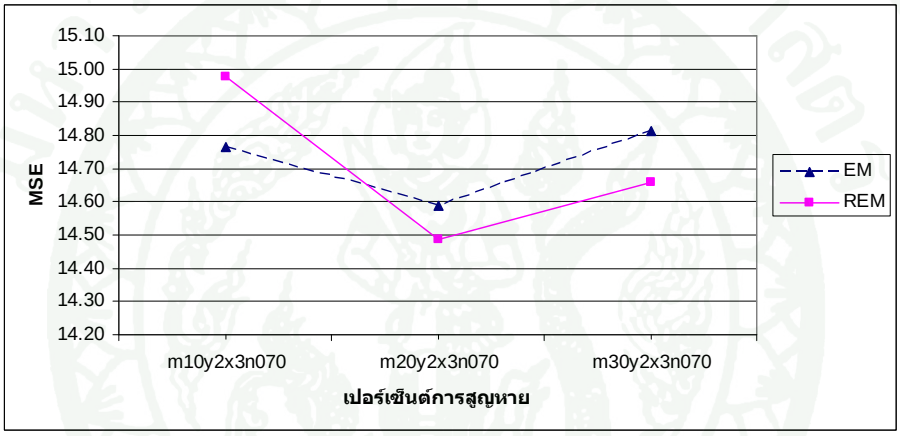
ตารางที่ 1 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 3, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	18.8324*	19.2234
	20	18.5694*	18.7039
	30	18.4017*	18.5894
70	10	14.7663*	14.9760
	20	14.5887	14.4874*
	30	14.8147	14.6596*
100	10	13.8281	13.7582*
	20	14.5771	14.4136*
	30	14.6683	14.4644*

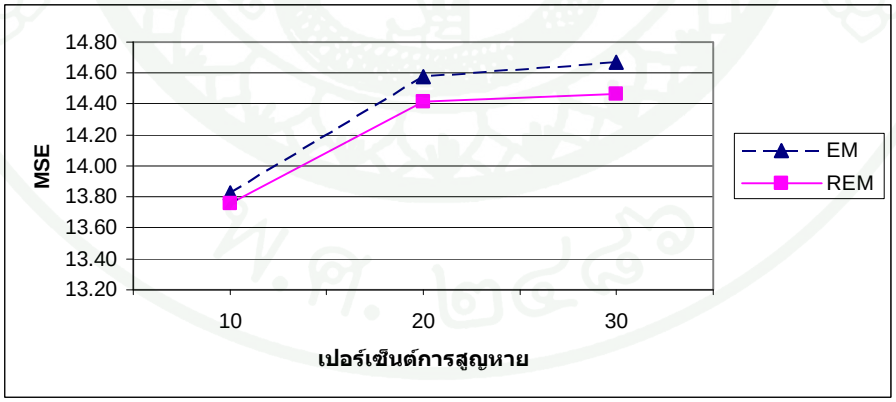
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

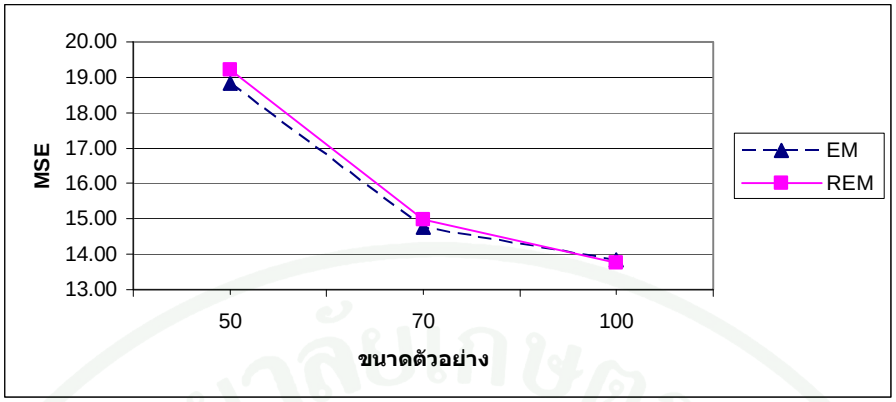


(ข) ขนาดตัวอย่างเท่ากับ 70

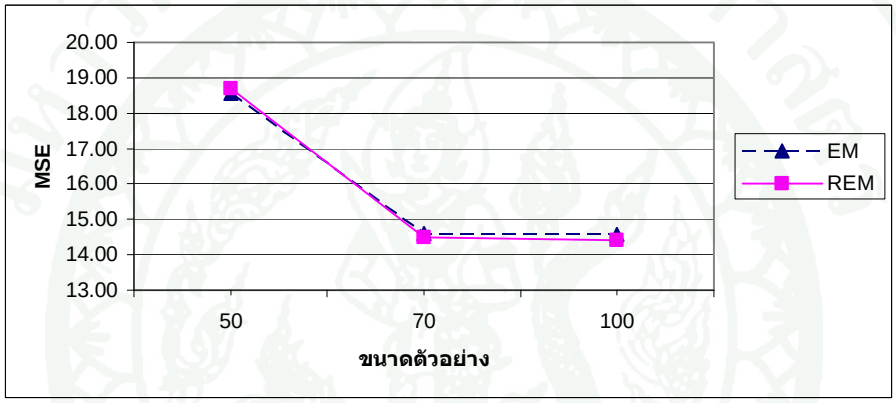


(ค) ขนาดตัวอย่างเท่ากับ 100

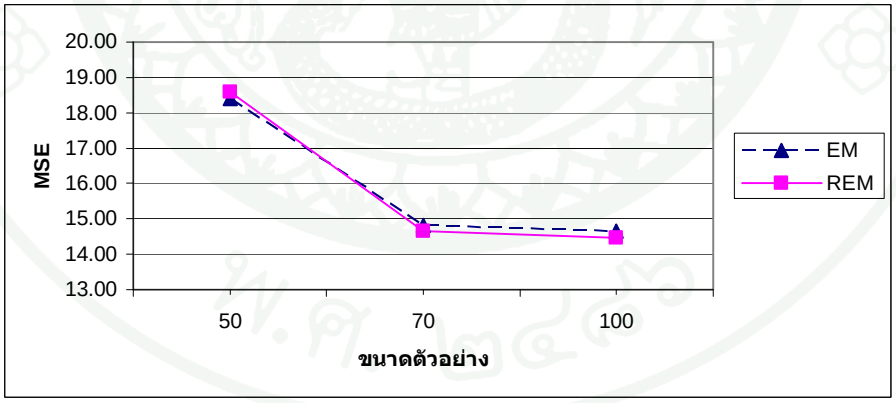
ภาพที่ 3 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 3, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 4 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, X = 3, Y= 2 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

2. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 4 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

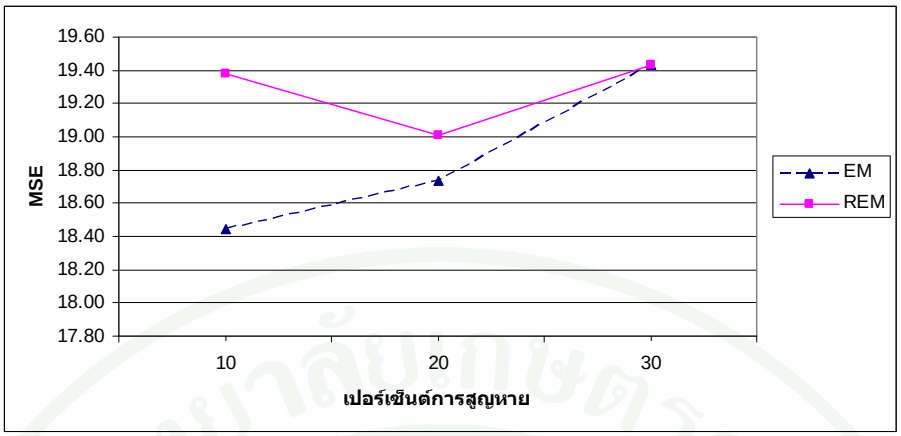
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 2 และภาพที่ 5 สำหรับขนาดตัวอย่างเท่ากับ 50 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM สำหรับขนาดตัวอย่างเท่ากับ 70 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น ค่า MSE ของทั้ง 2 วิธีมีแนวโน้มเพิ่มขึ้น

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 2 และภาพที่ 6 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20 และ 30 ที่ขนาดตัวอย่างเท่ากับ 50 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM ที่ขนาดตัวอย่าง 70 และ 100 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM สรุปโดยรวมว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลง ทั้ง 2 วิธี

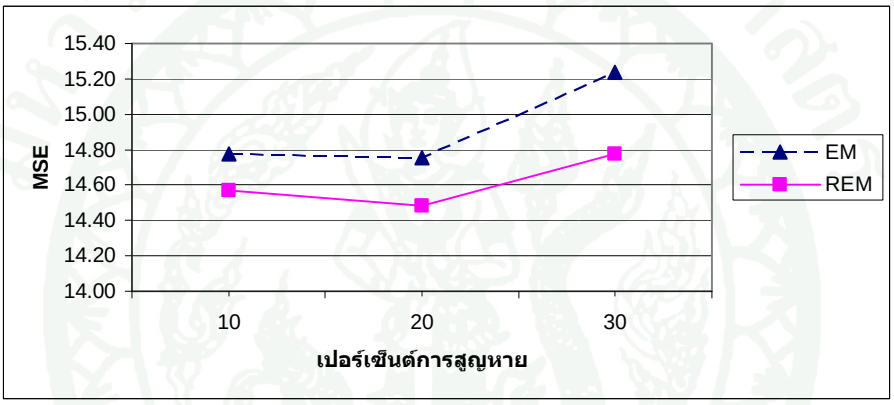
ตารางที่ 2 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 4, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	18.4424*	19.3831
	20	18.7327*	19.0086
	30	19.4335*	19.4341
70	10	14.7785	14.5717*
	20	14.7525	14.4832*
	30	15.2380	14.7736*
100	10	14.2687	13.9948*
	20	14.5933	14.2700*
	30	14.7682	14.4047*

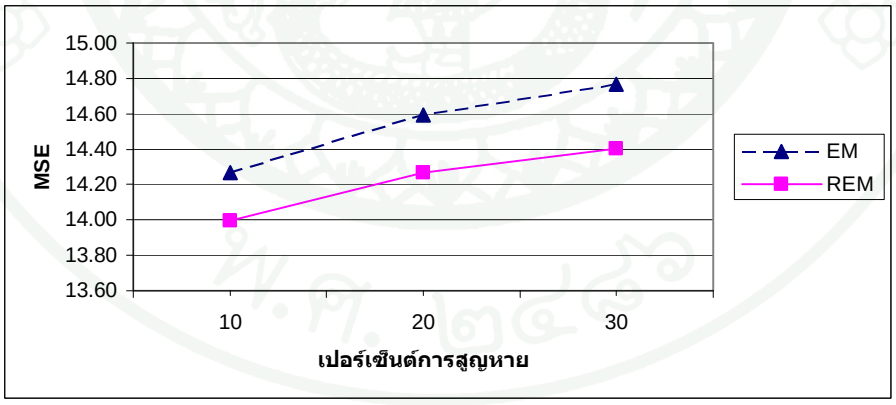
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

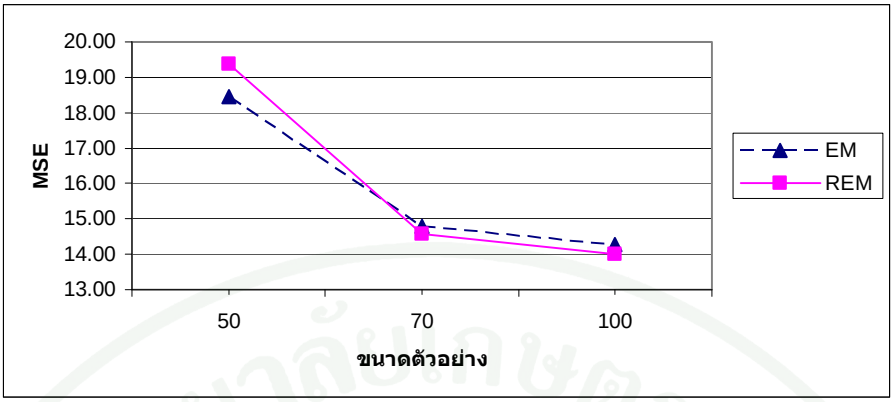


(ข) ขนาดตัวอย่างเท่ากับ 70

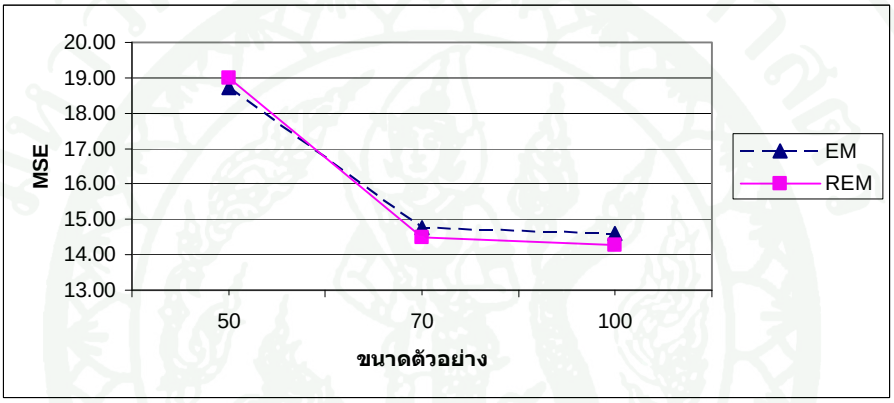


(ค) ขนาดตัวอย่างเท่ากับ 100

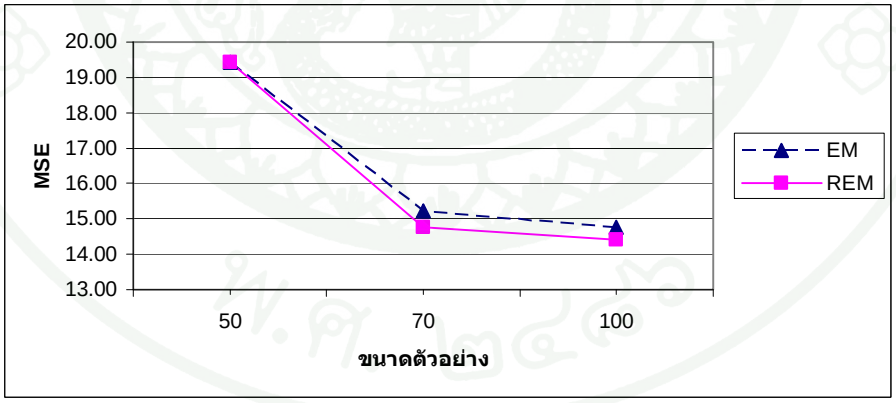
ภาพที่ 5 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 4, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 6 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, X = 4, Y= 2 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

3. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 5 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

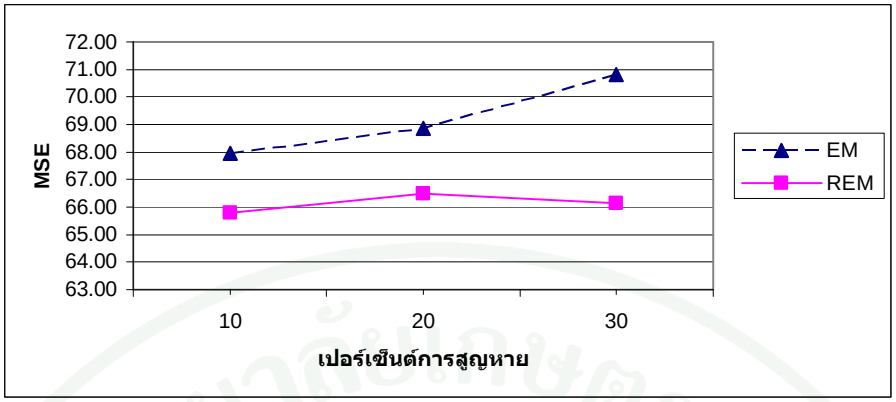
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 3 และภาพที่ 7 สำหรับขนาดตัวอย่างเท่ากับ 50 ทุกระดับการสูญหายของข้อมูล พบว่าวิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 70 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 20 และ 30 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การ สูญหายเท่ากับ 10 และ 20 พบว่าวิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 3 และภาพที่ 8 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 ที่ขนาดตัวอย่างเท่ากับ 50 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อขนาดตัวอย่าง 70 และ 100 ค่า MSE ของวิธี EM ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 20 ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อขนาดตัวอย่างเท่ากับ 100 ค่า MSE ของวิธี EM ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 30 ทุกขนาดตัวอย่าง วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM

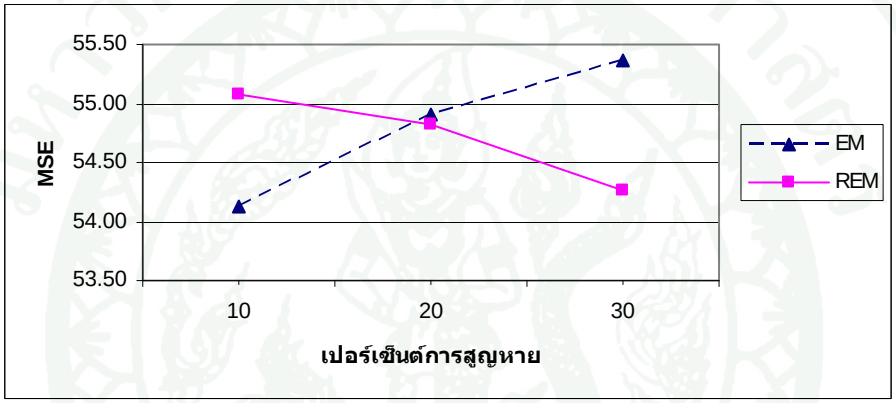
ตารางที่ 3 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, X = 5, Y= 3 , และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	67.9531	65.8019*
	20	68.8602	66.4719*
	30	70.8009	66.1381*
70	10	54.1340*	55.0776
	20	54.9033	54.8231*
	30	55.3576	54.2544*
100	10	42.8445*	43.1814
	20	42.3615*	42.6551
	30	42.7057	42.1616*

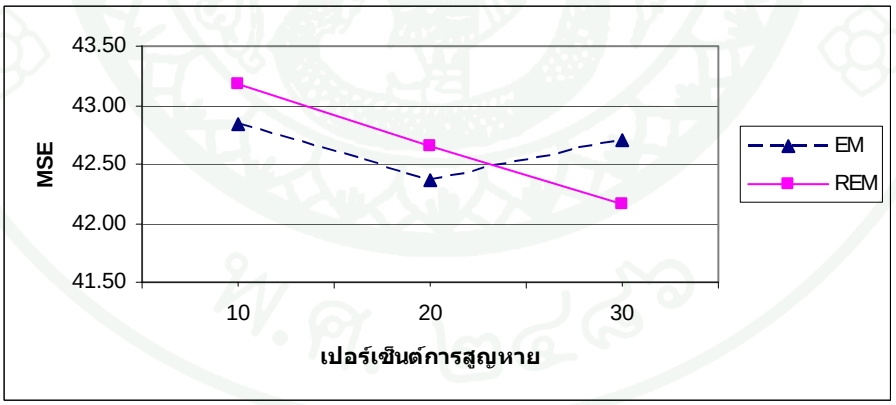
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

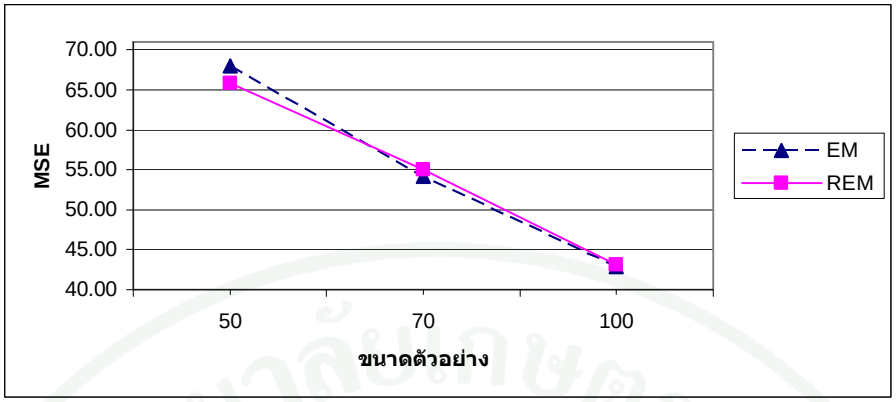


(ข) ขนาดตัวอย่างเท่ากับ 70

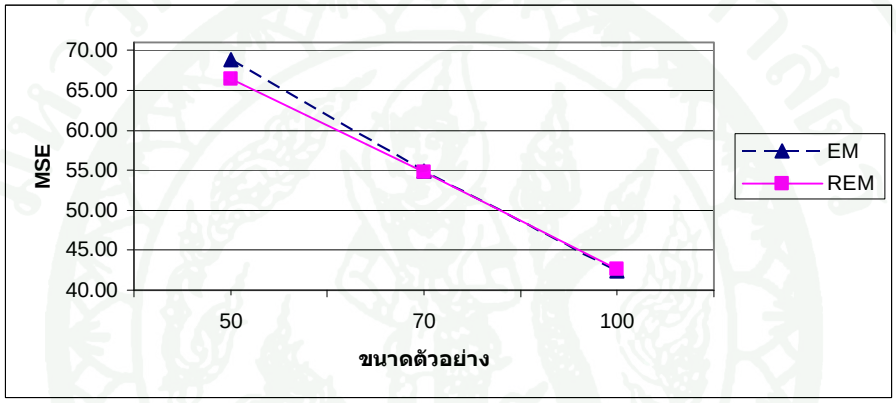


(ค) ขนาดตัวอย่างเท่ากับ 100

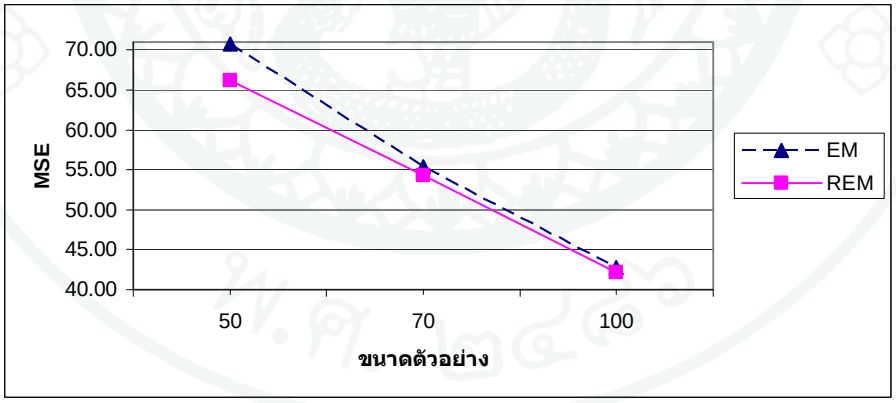
ภาพที่ 7 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 5, Y = 3$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 8 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 5, Y = 3$, และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

4. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระสูง เมื่อมีตัวแปรอิสระ (X) 7 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

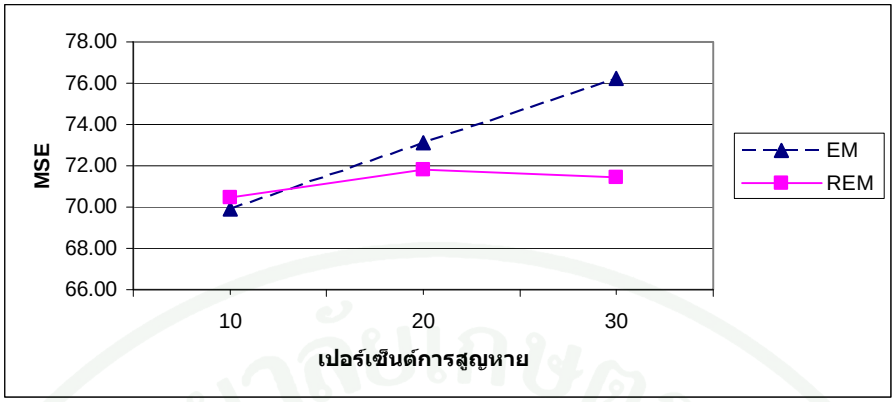
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 4 และภาพที่ 9 สำหรับขนาดตัวอย่างเท่ากับ 50 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 70 และ 100 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 70 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหาย 30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 100 ทุกระดับการสูญหายของข้อมูล วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 4 และภาพที่ 10 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อขนาดตัวอย่างเท่ากับ 100 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 20 ที่ขนาดตัวอย่างเท่ากับ 50 และ 100 วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อขนาดตัวอย่างเท่ากับ 70 ค่า MSE ของวิธี EM ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 30 ทุกขนาดตัวอย่าง วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM

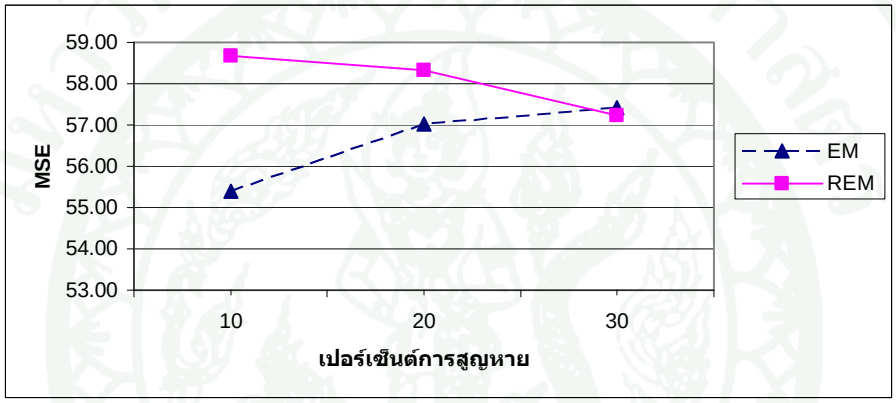
ตารางที่ 4 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 7, Y = 3$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	69.9140*	70.4695
	20	73.0969	71.8094*
	30	76.2210	71.4269*
70	10	55.3842*	58.6791
	20	57.0171*	58.3250
	30	57.4077	57.2401*
100	10	42.9180	41.8997*
	20	43.8865	42.7422*
	30	44.4505	43.0289*

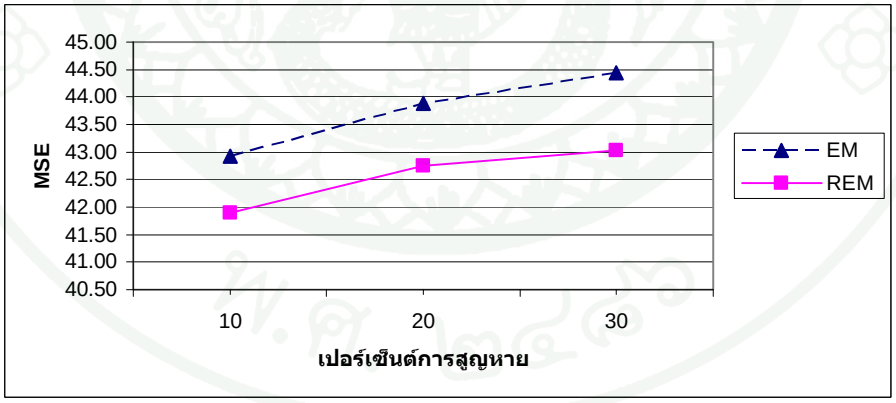
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

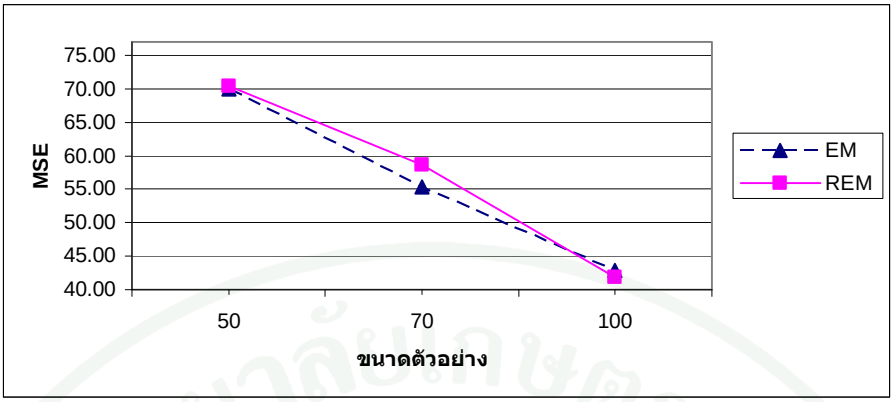


(ข) ขนาดตัวอย่างเท่ากับ 70

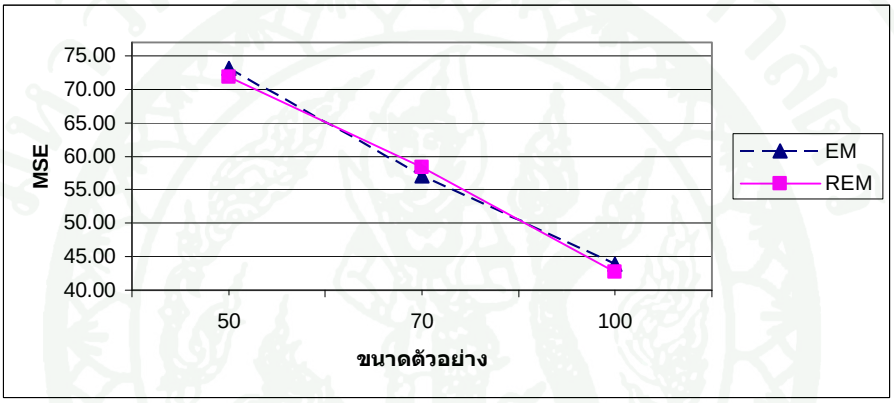


(ค) ขนาดตัวอย่างเท่ากับ 100

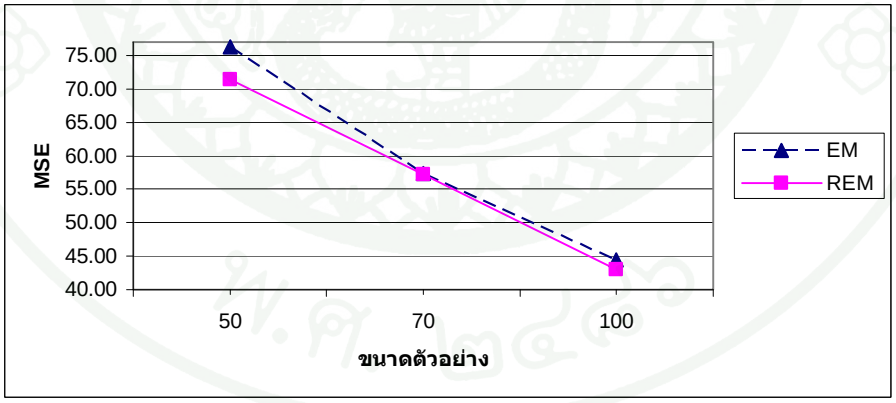
ภาพที่ 9 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, $X = 7, Y = 3$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 10 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.7 - 0.9$, X = 7, Y= 3 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

5. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 3 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

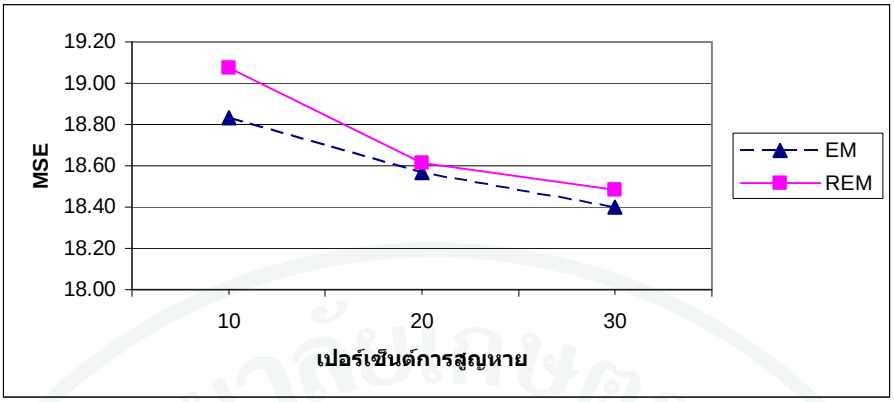
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 5 และภาพที่ 11 สำหรับขนาดตัวอย่างเท่ากับ 50 และ 100 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM สำหรับขนาดตัวอย่างเท่ากับ 70 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM ในขณะที่เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 5 และภาพที่ 12 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 ทุกขนาดตัวอย่าง พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 30 ที่ขนาดตัวอย่างเท่ากับ 50 และ 100 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อขนาดตัวอย่างเท่ากับ 70 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM สรุปโดยรวมได้ว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลงทั้ง 2 วิธี

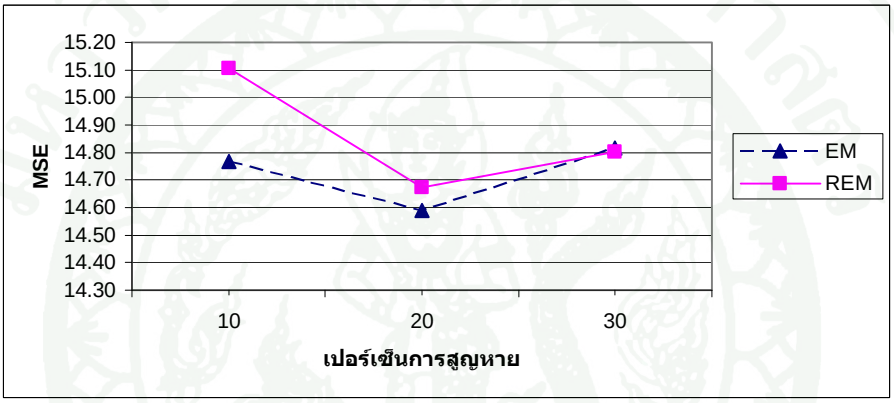
ตารางที่ 5 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 3, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	18.8324*	19.0752
	20	18.5694*	18.6134
	30	18.4017*	18.4816
70	10	14.7663*	15.1051
	20	14.5887*	14.6737
	30	14.8147	14.8014*
100	10	13.8281*	14.0539
	20	14.5771*	14.6951
	30	14.6683*	14.6938

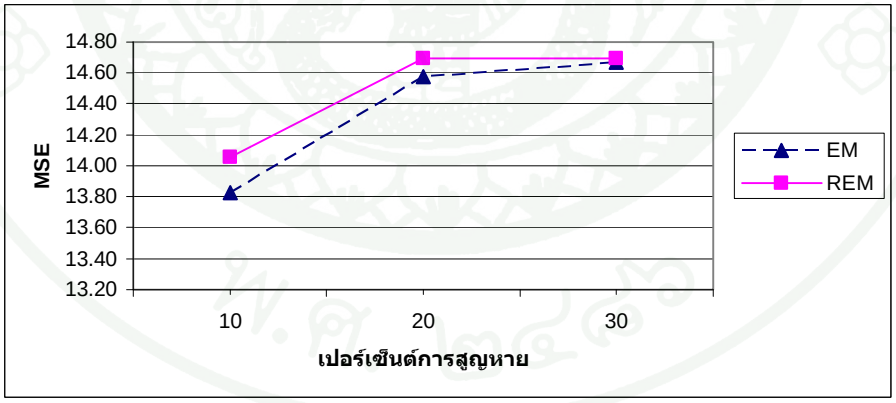
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

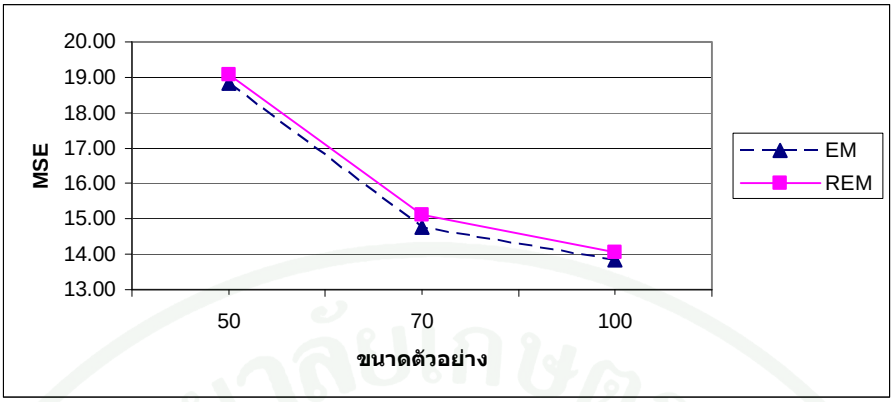


(ข) ขนาดตัวอย่างเท่ากับ 70

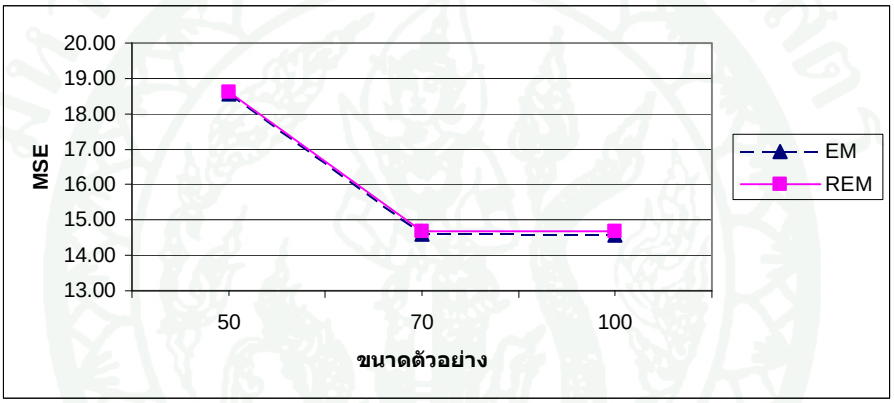


(ค) ขนาดตัวอย่างเท่ากับ 100

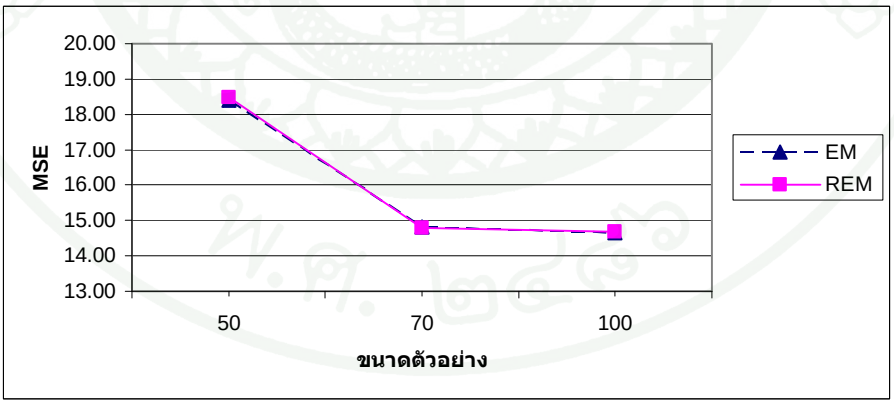
ภาพที่ 11 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 3, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 12 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 3, Y= 2 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

6. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 4 ตัว และตัวแปรตาม (Y) 2 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

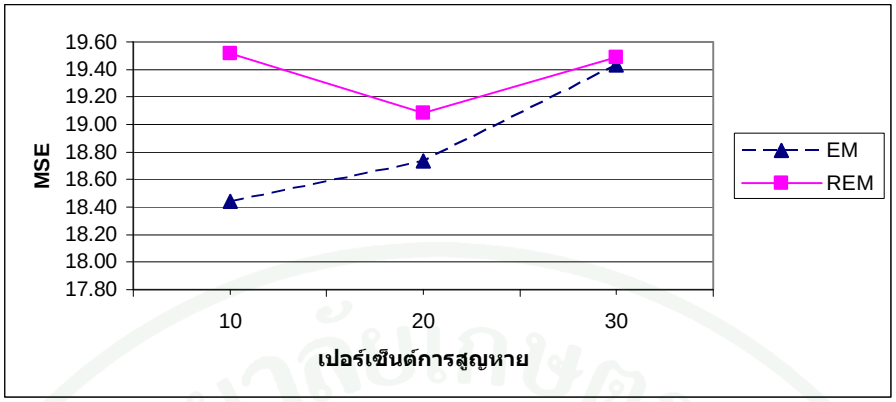
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 6 และภาพที่ 13 สำหรับขนาดตัวอย่างเท่ากับ 50 และ 70 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี EM ให้ค่า MSE ต่ำกว่าวิธี REM สำหรับขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM ที่เปอร์เซ็นต์การสูญหายเท่ากับ 30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 6 และภาพที่ 14 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 ทุกขนาดตัวอย่าง พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 30 ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อขนาดตัวอย่างเท่ากับ 100 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM สรุปโดยรวมได้ว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลงทั้ง 2 วิธี

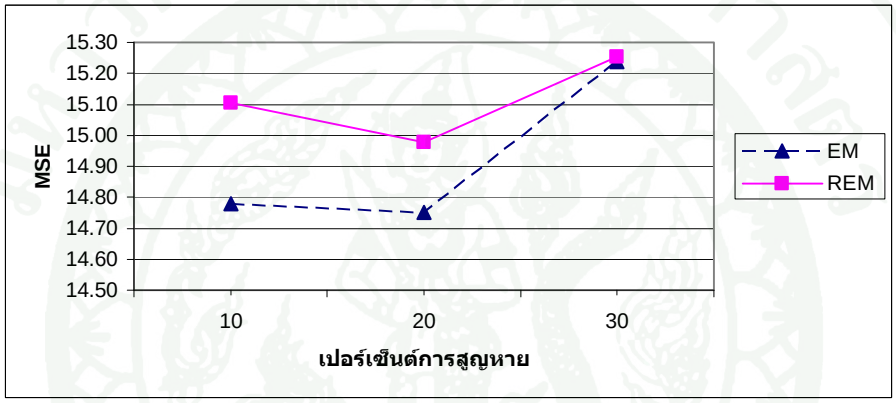
ตารางที่ 6 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 4, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	18.4424*	19.5157
	20	18.7327*	19.0857
	30	19.4335*	19.4873
70	10	14.7785*	15.1032
	20	14.7525*	14.9779
	30	15.2380*	15.2539
100	10	14.2687*	14.3160
	20	14.5933*	14.6239
	30	14.7682	14.7433*

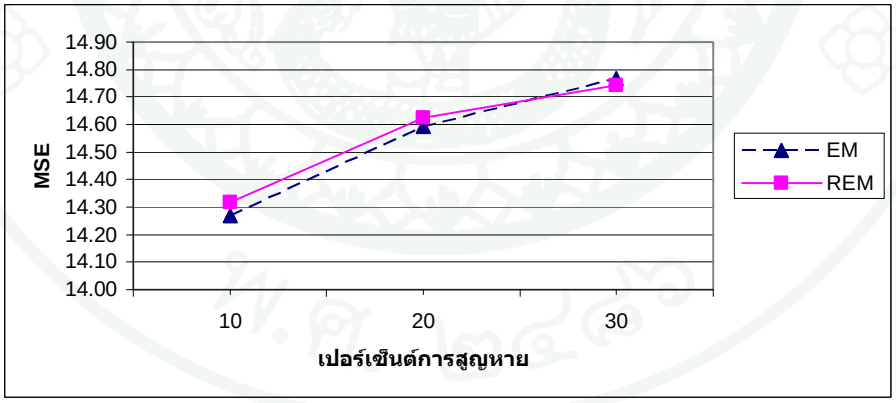
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

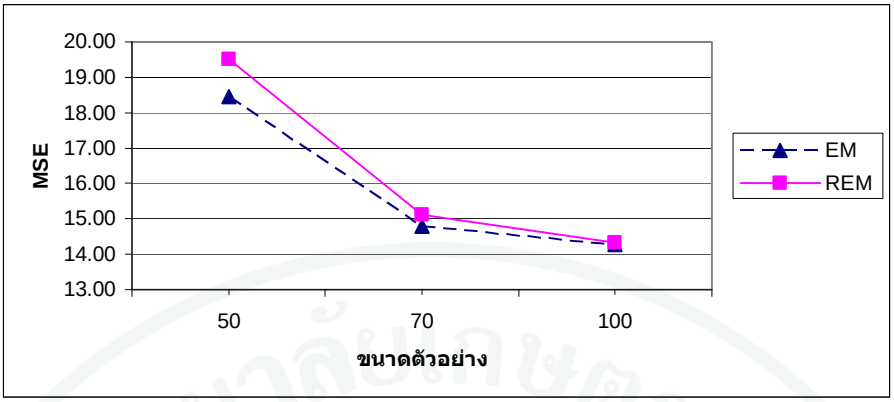


(ข) ขนาดตัวอย่างเท่ากับ 70

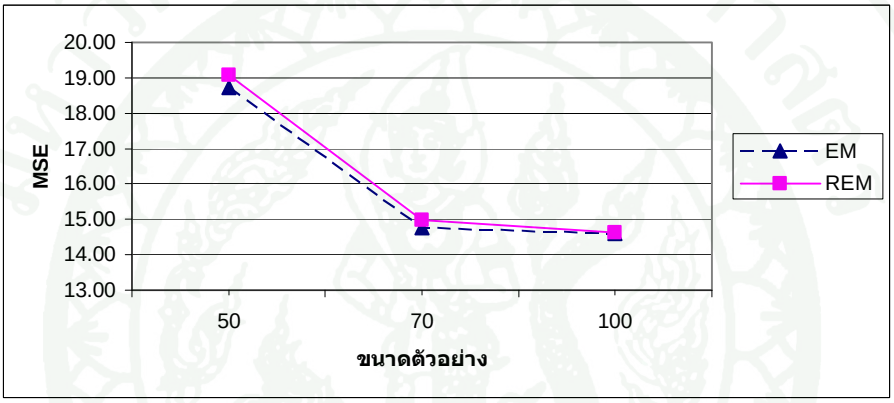


(ค) ขนาดตัวอย่างเท่ากับ 100

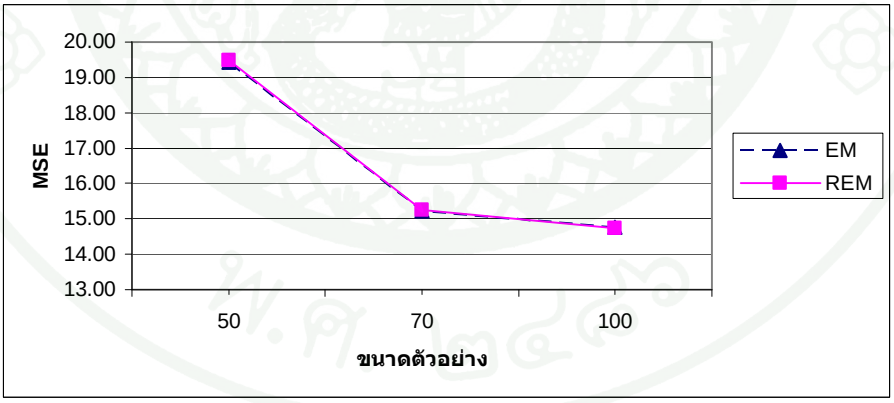
ภาพที่ 13 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 4, Y = 2$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 14 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 4, Y= 2 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

7. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 5 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

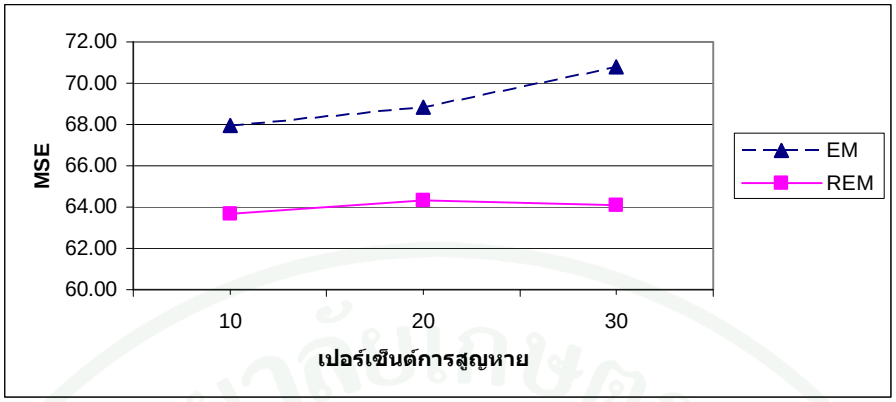
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 7 และภาพที่ 15 สำหรับขนาดตัวอย่างเท่ากับ 50 และ 70 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM ที่เปอร์เซ็นต์การสูญหายเท่ากับ 30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 7 และภาพที่ 16 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 พบว่า วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อขนาดตัวอย่างเท่ากับ 100 ค่า MSE ของวิธี EM ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 30 ทุกขนาดตัวอย่าง พบว่า วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สรุปโดยรวมได้ว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลงทั้ง 2 วิธี

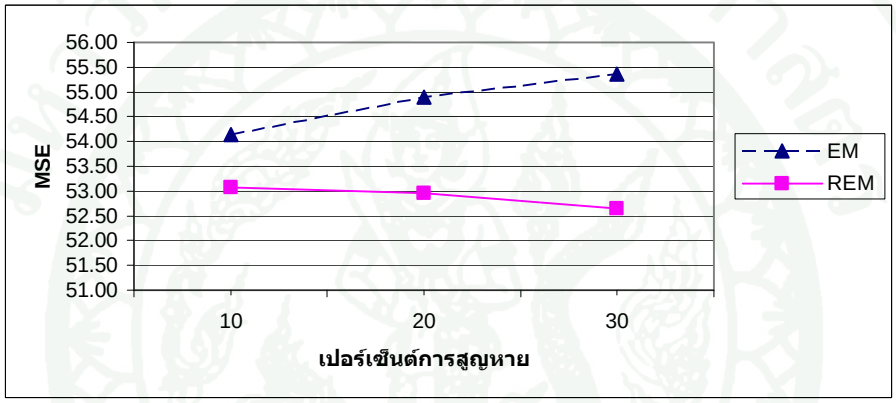
ตารางที่ 7 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 5, Y= 3, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	67.9531	63.6786*
	20	68.8602	64.3067*
	30	70.8009	64.1062*
70	10	54.1340	53.0799*
	20	54.9033	52.9534*
	30	55.3576	52.6496*
100	10	42.8445*	43.1575
	20	42.3615*	42.4346
	30	42.7057	42.1556*

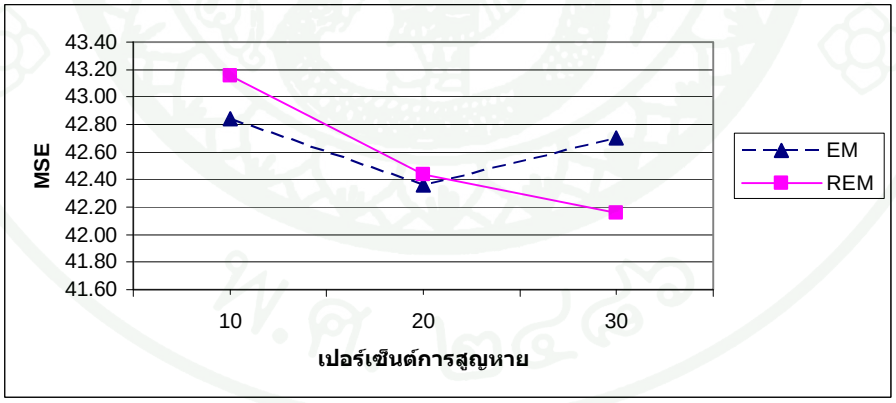
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

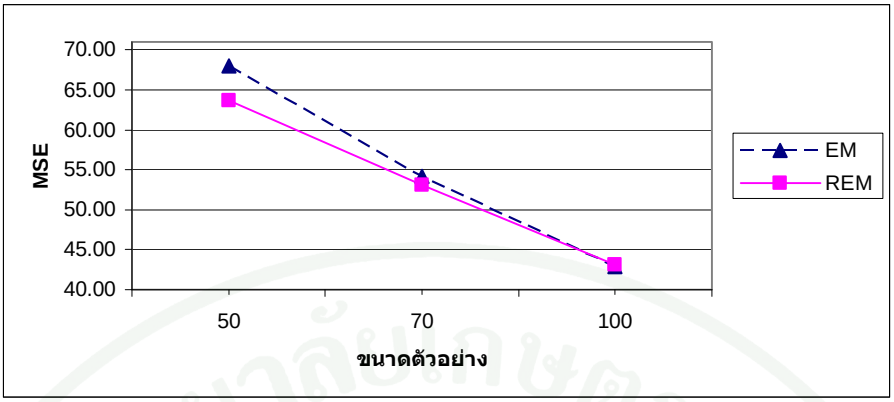


(ข) ขนาดตัวอย่างเท่ากับ 70

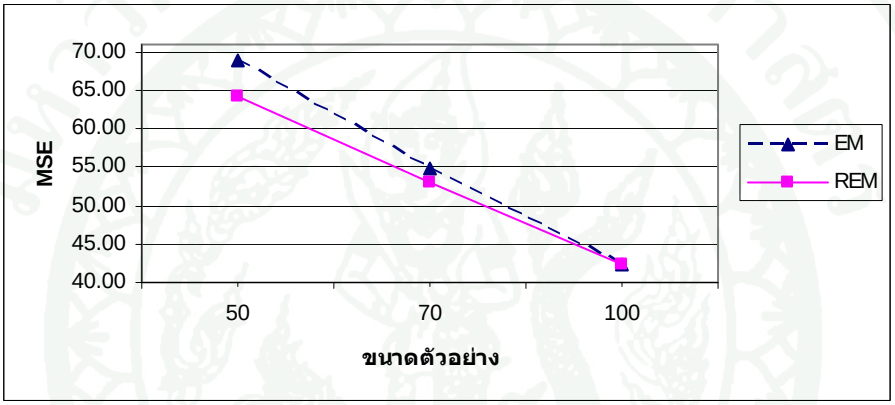


(ค) ขนาดตัวอย่างเท่ากับ 100

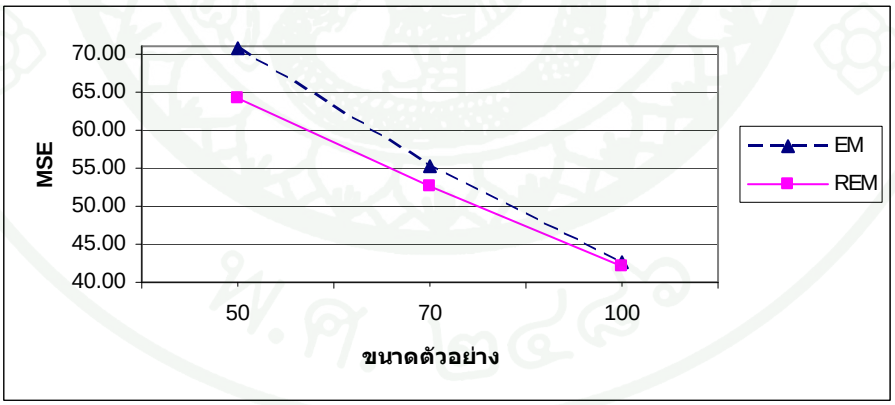
ภาพที่ 15 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 5, Y = 3,$ และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 16 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 5, Y= 3 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

8. กรณีความสัมพันธ์ระหว่างตัวแปรอิสระต่ำ เมื่อมีตัวแปรอิสระ (X) 7 ตัว และตัวแปรตาม (Y) 3 ตัว จำแนกตามขนาดตัวอย่างและเปอร์เซ็นต์การสูญหาย

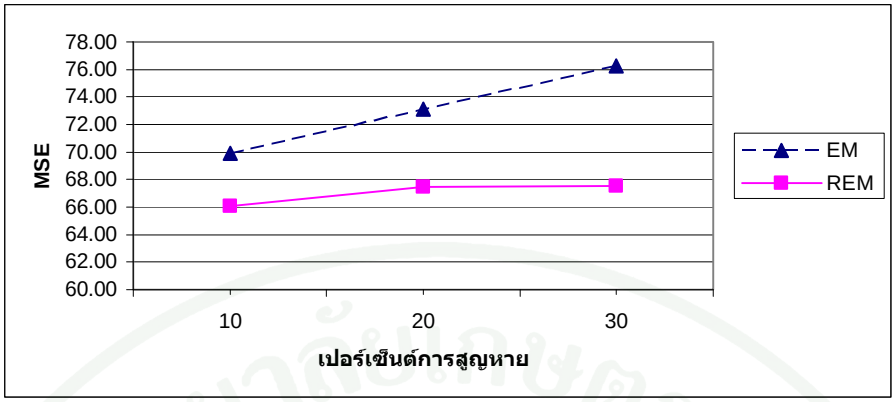
เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามขนาดตัวอย่าง จากตารางที่ 8 และภาพที่ 17 สำหรับขนาดตัวอย่างเท่ากับ 50 และ 70 ทุกระดับการสูญหายของข้อมูล พบว่า วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM สำหรับขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 พบว่า วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 20 และ30 ค่า MSE ของวิธี REM ต่ำกว่าวิธี EM

เมื่อพิจารณาค่า MSE ของแต่ละวิธี จำแนกตามเปอร์เซ็นต์การสูญหาย จากตารางที่ 8 และภาพที่ 18 สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 10 ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 พบว่า วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM และเมื่อขนาดตัวอย่างเท่ากับ 100 ค่า MSE ของวิธี EM ต่ำกว่าวิธี REM สำหรับเปอร์เซ็นต์การสูญหายเท่ากับ 20 และ 30 ทุกขนาดตัวอย่าง พบว่า วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM สรุปโดยรวมได้ว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า MSE มีแนวโน้มลดลง ทั้ง 2 วิธี

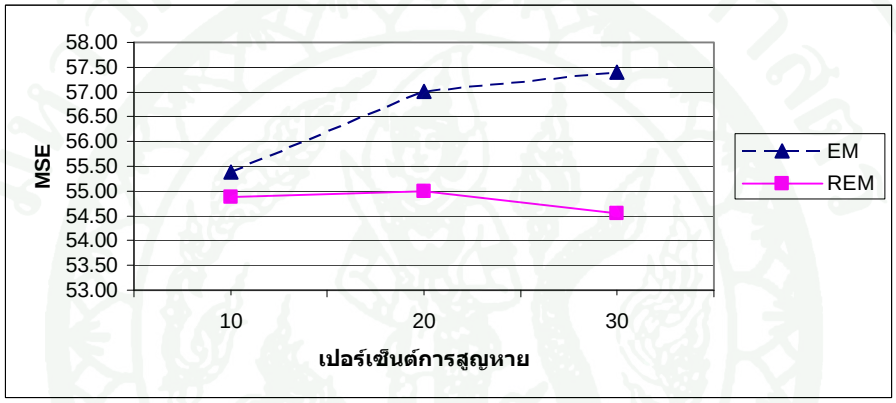
ตารางที่ 8 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 7, Y= 3, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
50	10	69.9140	66.0411*
	20	73.0969	67.4771*
	30	76.2210	67.5202*
70	10	55.3842	54.8800*
	20	57.0171	55.0054*
	30	57.4077	54.5536*
100	10	42.9180*	43.2574
	20	43.8865	43.8060*
	30	44.4505	44.2994*

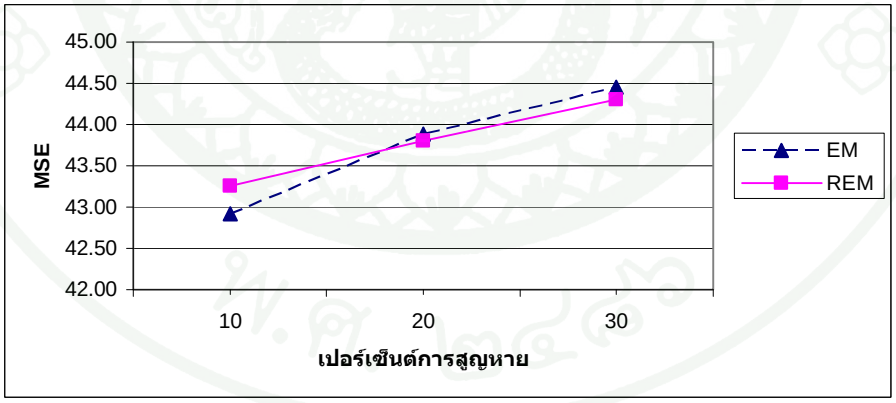
หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ขนาดตัวอย่างเท่ากับ 50

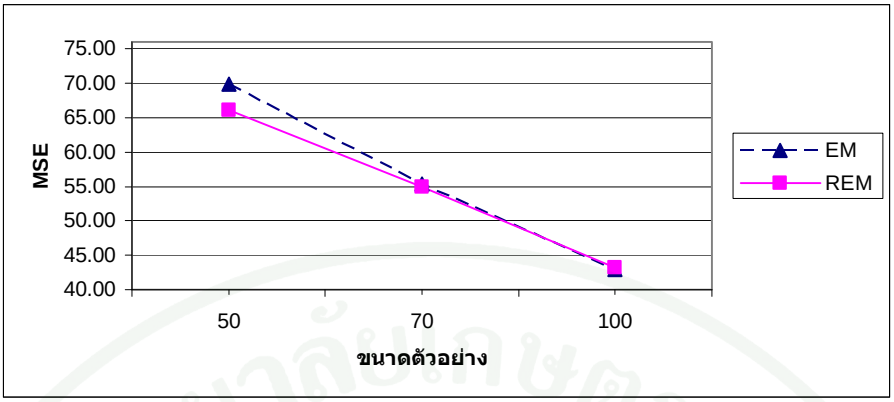


(ข) ขนาดตัวอย่างเท่ากับ 70

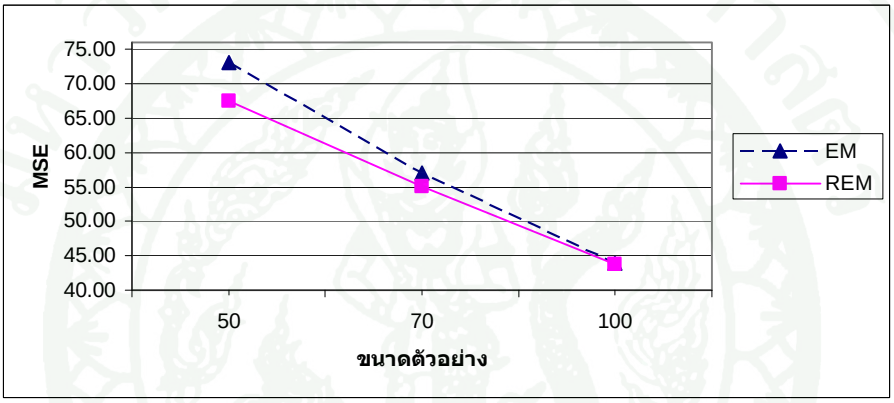


(ค) ขนาดตัวอย่างเท่ากับ 100

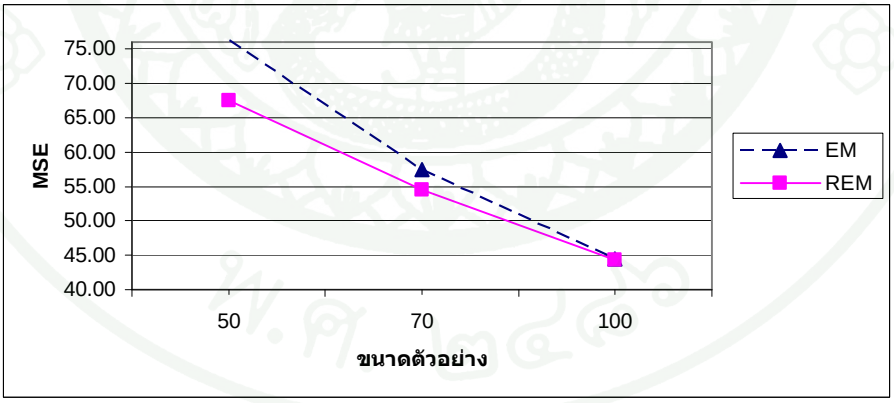
ภาพที่ 17 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, $X = 7$, $Y = 3$, และเปอร์เซ็นต์การสูญหายเท่ากับ 10, 20, 30 โดยจำแนกตามขนาดตัวอย่าง



(ก) เปอร์เซ็นต์การสูญหายเท่ากับ 10



(ข) เปอร์เซ็นต์การสูญหายเท่ากับ 20



(ค) เปอร์เซ็นต์การสูญหายเท่ากับ 30

ภาพที่ 18 ค่า MSE ของ วิธี EM และ วิธี REM กรณี $\rho = 0.1 - 0.3$, X = 7, Y= 3 , และขนาดตัวอย่างเท่ากับ 50, 70, 100 โดยจำแนกตามเปอร์เซ็นต์การสูญหาย

ส่วนที่ 2 ผลการศึกษาจากข้อมูลจริง

ข้อมูลจริงที่นำมาใช้ในการศึกษาครั้งนี้มาจากข้อมูลผลการตรวจอากาศชั้นบน (upper air observation) ระหว่างวันที่ 1 มีนาคม 2549 - 31 ตุลาคม 2551 จากสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา สำนักฝนหลวงและการบินเกษตร กระทรวงเกษตรและสหกรณ์ เป็นตัวแปรเชิงปริมาณ 49 ตัวแปร แสดงรายละเอียดดังตารางที่ 9

ตารางที่ 9 รายละเอียดของตัวแปรในข้อมูลผลการตรวจอากาศชั้นบน

ตัวแปร	ชื่อ	คำอธิบาย	หน่วย
x_1	PRECIPITABLE_WATER_SFC850	ปริมาณน้ำในบรรยากาศตั้งแต่ผิวพื้นถึง 850 มิลลิบาร์	ลบ.ซม.
x_2	ISOTHERM_HEIGHTS_0c	ความสูงของระดับอุณหภูมิเท่า 0 องศาเซลเซียส	ฟุต
x_3	PRECIPITABLE_WATER_SFC700	ปริมาณน้ำในบรรยากาศตั้งแต่ผิวพื้นถึง 700 มิลลิบาร์	ลบ.ซม.
x_4	ISOTHERM_HEIGHTS_minus5c	ความสูงของระดับอุณหภูมิเท่า -5 องศาเซลเซียส	ฟุต
x_5	PRECIPITABLE_WATER_SFC500	ปริมาณน้ำในบรรยากาศตั้งแต่ผิวพื้นถึง 500 มิลลิบาร์	ลบ.ซม.
x_6	ISOTHERM_HEIGHTS_minus10c	ความสูงของระดับอุณหภูมิเท่า -10 องศาเซลเซียส	ฟุต
x_7	PRECIPITABLE_WATER_TOTAL	ปริมาณน้ำในบรรยากาศรวม	ลบ.ซม.
x_8	ISOTHERM_HEIGHTS_minus15c	ความสูงของระดับอุณหภูมิเท่า -15 องศาเซลเซียส	ฟุต
x_9	MEAN_MIXING_RATIO_LOWEST_100_MB	อัตราส่วนไอน้ำในอากาศเฉลี่ย จากผิวพื้นขึ้นไป 100 มิลลิบาร์	g/kg
x_{10}	MEAN_MIXING_RATIO_LOWEST_50_MB	อัตราส่วนไอน้ำในอากาศเฉลี่ย จากผิวพื้นขึ้นไป 50 มิลลิบาร์	g/kg
x_{11}	MEAN_DEW_POINT_TEMP_LOWEST_50_MB	อุณหภูมิจุดน้ำค้างเฉลี่ยจากผิวพื้นขึ้นไป 50 มิลลิบาร์	เซลเซียส

ตารางที่ 9 (ต่อ)

ตัวแปร	ชื่อ	คำอธิบาย	หน่วย
x_{12}	RAOB_SURFACE_TEMPERATURE	RAOB_SURFACE_TEMPERATURE	เซลเซียส
x_{13}	SFC_TEMP_RISE_REQUIRED_FOR_CCL	SFC_TEMP_RISE_REQUIRED_FOR_CCL	เซลเซียส
x_{14}	SURFACE_CONVECTIVE_TEMPERATURE	อุณหภูมิของอากาศผิวพื้น	เซลเซียส
x_{15}	CLOUD_BASE_PRESSURE_CCL	ความกดอากาศที่ระดับฐานเมฆ CCL	มิลลิบาร์
x_{16}	CLOUD_BASE_HEIGHT_CCL	ความสูงของฐานเมฆ CCL	ฟุต
x_{17}	CLOUD_BASE_TEMPERATURE_CCL	อุณหภูมิที่ระดับฐานเมฆ CCL	เซลเซียส
x_{18}	SUB_CLOUD_MIXING_RATIO_CCL	SUB_CLOUD_MIXING_RATIO_CCL	g/kg
x_{19}	CLOUD_BASE_HEIGHT_LCL	ความสูงของฐานเมฆ LCL	ฟุต
x_{20}	CLOUD_BASE_TEMPERATURE_LCL	อุณหภูมิที่ระดับฐานเมฆ LCL	เซลเซียส
x_{21}	LIFTED_INDEX_100_MBAR_LAYER_ADIABATIC	ดัชนีการยกตัวของมวลอากาศ ที่ลากจากค่าเฉลี่ยของอากาศผิวพื้นหนา 100 มิลลิบาร์	ไม่มี
x_{22}	LIFTED_INDEX_50_MBAR_LAYER_MEAN_VALUES	ดัชนีการยกตัวของมวลอากาศ ที่ลากจากค่าเฉลี่ยของอากาศผิวพื้นหนา 50 มิลลิบาร์	ไม่มี
x_{23}	SHOWALTER_INDEX	ดัชนีแสดงค่าเสถียรของอากาศ (+ แสดงถึงอากาศอยู่ในสภาวะเสถียร, - แสดงถึงอากาศอยู่ในสภาวะไม่เสถียร)	ไม่มี
x_{24}	TOTAL_TOTALS_INDEX	ดัชนีแสดงค่าผลรวมเสถียรของอากาศทั้งแนวตั้งและแนวนอน	ไม่มี
x_{25}	K_INDEX	ดัชนีวัดโอกาสความเป็นไปได้ของการเกิดพายุฝนฟ้าคะนอง (probability of thunderstorms) จากการยกตัวของกลุ่มอากาศทางแนวตั้ง (คล้ายดัชนี LI)	ไม่มี
x_{26}	SWEAT_INDEX	ดัชนีที่บอกความแรงของพายุ	ไม่มี
x_{27}	POTENTIAL_BUOYANCY_INDEX	POTENTIAL_BUOYANCY_INDEX	ไม่มี
x_{28}	LEVEL_OF_FREE_CONVECTION_LFC_wrtTma	ความสูงของระดับยกตัวอิสระ	มิลลิบาร์

ตารางที่ 9 (ต่อ)

ตัวแปร	ชื่อ	คำอธิบาย	หน่วย
x_{29}	TOP_OF_LATENT_INSTABILITY_ LAYER_wrtTma	TOP_OF_LATENT_INSTABILITY_ LAYER_wrtTma	มิลลิบาร์
x_{30}	LEVEL_OF_NEUTRAL_BUOYANCY_ LNB	ความสูงของระดับสมดุลของแรงลอยตัว	มิลลิบาร์
x_{31}	CCL_HEIGHT	ค่าพยากรณ์ความสูงของฐานเมฆคิวมูลัส	ฟุต
x_{32}	RH_AT_THE_CCL	ความชื้นสัมพัทธ์ที่ระดับฐานเมฆ CCL	เปอร์เซ็นต์
x_{33}	AVG_RH_0_10000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 0-10,000 ฟุต	เปอร์เซ็นต์
x_{34}	AVG_RH_10_18000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 10,000 -18,000 ฟุต	เปอร์เซ็นต์
x_{35}	AVG_5_10K_FT_WD_SP	ค่าเฉลี่ยของความเร็วลมที่ระดับความสูง 5,000-10,000 ฟุต	knot
x_{36}	CONVECTIVE_TEMPERATURE	อุณหภูมิที่เมฆจะยกตัว	เซลเซียส
x_{37}	TIME_CONV_TEMP_REACHED	ค่าพยากรณ์เวลาที่เกิดเมฆ	
x_{38}	MEAN_WIND_1000_5000_FT_DEG	ทิศทางลมเฉลี่ยจาก 1,000-5,000 ฟุต	degree
x_{39}	MEAN_WIND_5000_10000_FT_DEG	ทิศทางลมเฉลี่ยจาก 5,000-10,000 ฟุต	degree
x_{40}	MEAN_WIND_10000_15000_FT_ DEG	ทิศทางลมเฉลี่ยจาก 10,000-15,000 ฟุต	degree
x_{41}	MEAN_WIND_20000_25000_FT_ DEG	ทิศทางลมเฉลี่ยจาก 20,000-25,000 ฟุต	degree
x_{42}	MEAN_WIND_1000_5000_FT_KTS	ความเร็วลมเฉลี่ยจาก 1,000-5,000 ฟุต	knot
x_{43}	MEAN_WIND_5000_10000_FT_KTS	ความเร็วลมเฉลี่ยจาก 5,000-10,000 ฟุต	knot
x_{44}	MEAN_WIND_10000_15000_FT_KTS	ความเร็วลมเฉลี่ยจาก 10,000-15,000 ฟุต	knot
x_{45}	MEAN_WIND_20000_25000_FT_KTS	ความเร็วลมเฉลี่ยจาก 20,000-25,000 ฟุต	knot
x_{46}	MEAN_RH_1000_5000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 1,000-5,000 ฟุต	เปอร์เซ็นต์
x_{47}	MEAN_RH_5000_10000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 5,000-10,000 ฟุต	เปอร์เซ็นต์
x_{48}	MEAN_RH_10000_15000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 10,000-15,000 ฟุต	เปอร์เซ็นต์
x_{49}	MEAN_RH_20000_25000_FT	ความชื้นสัมพัทธ์เฉลี่ยจาก 20,000-25,000 ฟุต	เปอร์เซ็นต์

ที่มา : วราวุธ (2539)

ซึ่งมีข้อมูลทั้งสิ้นจำนวน 401 ชุด และผู้วิจัยคัดเลือกเฉพาะชุดข้อมูลที่สมบูรณ์ (มีตัวแปรครบทั้ง 49 ตัวแปร) มาใช้ในการศึกษาครั้งนี้ จำนวน 87 ชุด จากนั้นคำนวณค่าสถิติเชิงพรรณนา แสดงดังตารางที่ 10

ตารางที่ 10 ค่าสถิติเชิงพรรณาของตัวแปรในข้อมูลผลการตรวจอากาศชั้นบน

ตัวแปร	N	Range	Minimum	Maximum	Mean	SE.	SD.	Variance
x_1	87	.76	1.39	2.15	1.8259	.01732	.16151	.026
x_2	87	3866.00	13647.00	17513.00	15922.4368	85.75332	799.85373	639765.993
x_3	87	1.42	2.43	3.85	3.2543	.03335	.31108	.097
x_4	87	3329.00	16878.00	20207.00	19146.6207	76.44728	713.05279	508444.285
x_5	87	2.49	2.67	5.16	4.2540	.05623	.52452	.275
x_6	87	3117.00	20099.00	23216.00	22024.0000	65.72943	613.08334	375871.186
x_7	87	2.95	2.72	5.67	4.5594	.06908	.64431	.415
x_8	87	2461.00	23444.00	25905.00	24726.2874	63.35727	590.95728	349230.509
x_9	87	5.86	10.33	16.19	13.8202	.12979	1.21059	1.466
x_{10}	87	5.96	11.11	17.07	14.5059	.12477	1.16381	1.354
x_{11}	87	6.75	14.67	21.42	18.7929	.13805	1.28763	1.658
x_{12}	87	5.30	22.50	27.80	24.8816	.11232	1.04765	1.098
x_{13}	87	9.30	2.20	11.50	7.5713	.21856	2.03863	4.156
x_{14}	87	10.50	26.30	36.80	32.4552	.23611	2.20231	4.850
x_{15}	87	180.00	740.00	920.00	812.9885	3.92769	36.63506	1342.128
x_{16}	87	6040.00	2836.00	8876.00	6202.4023	138.64721	1293.21504	1672405.150
x_{17}	87	8.60	11.50	20.10	16.4540	.19647	1.83251	3.358
x_{18}	87	5.66	11.56	17.22	14.7321	.12145	1.13284	1.283
x_{19}	87	3018.00	1324.00	4342.00	2703.4943	67.08678	625.74383	391555.346
x_{20}	87	5.60	16.00	21.60	19.3115	.14579	1.35984	1.849
x_{21}	87	7.00	-3.80	3.20	-.2897	.13837	1.29067	1.666
x_{22}	87	6.70	-2.70	4.00	.5310	.12719	1.18639	1.408
x_{23}	87	7.00	-1.10	5.90	1.8690	.17001	1.58579	2.515
x_{24}	87	11.70	36.40	48.10	42.4345	.25682	2.39548	5.738
x_{25}	87	30.20	6.90	37.10	31.0494	.50693	4.72829	22.357

ตารางที่ 10 (ต่อ)

ตัวแปร	N	Range	Minimum	Maximum	Mean	SE.	SD.	Variance
x_{26}	87	203.40	78.40	281.80	190.2943	2.92529	27.28531	744.488
x_{27}	87	6.90	-1.80	5.10	1.9632	.13718	1.27955	1.637
x_{28}	87	430.00	480.00	910.00	697.4713	9.45482	88.18873	7777.252
x_{29}	87	640.00	230.00	870.00	573.3333	13.95056	130.12218	16931.783
x_{30}	87	480.00	370.00	850.00	560.5747	10.04581	93.70111	8779.898
x_{31}	87	6040.30	2835.90	8876.20	6202.3782	138.64818	1293.22413	1672428.653
x_{32}	87	61.10	31.40	92.50	71.4368	1.22091	11.38785	129.683
x_{33}	87	45.50	46.70	92.20	70.2368	.94112	8.77820	77.057
x_{34}	87	73.50	16.10	89.60	63.6253	1.73757	16.20701	262.667
x_{35}	87	22.70	1.20	23.90	8.8701	.54557	5.08875	25.895
x_{36}	87	10.50	26.30	36.80	32.4552	.23611	2.20231	4.850
x_{37}	87	7.16	8.23	15.39	12.2721	.16884	1.57483	2.480
x_{38}	87	325.80	26.30	352.10	222.8874	7.35059	68.56171	4700.708
x_{39}	87	344.00	14.70	358.70	207.5080	8.73952	81.51677	6644.984
x_{40}	87	323.90	26.60	350.50	202.2908	9.36254	87.32798	7626.176
x_{41}	87	346.60	5.30	351.90	189.4632	8.73149	81.44194	6632.790
x_{42}	87	22.60	.20	22.80	9.1563	.50234	4.68552	21.954
x_{43}	87	22.70	1.20	23.90	8.8701	.54557	5.08875	25.895
x_{44}	87	29.70	.30	30.00	9.9218	.67937	6.33670	40.154
x_{45}	87	471.50	.50	472.00	17.0103	5.35010	49.90241	2490.250
x_{46}	87	43.60	48.30	91.90	71.6713	.91036	8.49127	72.102
x_{47}	87	53.90	39.90	93.80	71.2644	1.28035	11.94231	142.619
x_{48}	87	70.00	19.80	89.80	65.2563	1.59999	14.92374	222.718
x_{49}	87	76.80	5.60	82.40	47.1690	2.54844	23.77030	565.027

จากข้อมูลผลการตรวจอากาศชั้นบน ซึ่งประกอบด้วยตัวแปรเชิงปริมาณ 49 ตัวแปร วราวุธ (2539) และดลพรรษ (2552) ได้แบ่งข้อมูลเหล่านั้นออกเป็น 5 กลุ่ม คือ

1. ตัวแปรกลุ่มอุณหภูมิ ได้แก่ X_{12}, X_{13}, X_{14}, X_{17}
2. ตัวแปรกลุ่มความชื้น ได้แก่ X_1, X_5, X_7, X_9, X_{10}, X_{11}, X_{18}, X_{33}, X_{46}
3. ตัวแปรกลุ่มความสูงและความกดอากาศ ได้แก่ X_{15}, X_{16}, X_{28}, X_{29}
4. ตัวแปรกลุ่มกระแสลม ได้แก่ X_{35}, X_{38}, X_{39}, X_{40}, X_{41}, X_{42}, X_{43}, X_{44}, X_{45}
5. ตัวแปรกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ ได้แก่ X_{21}, X_{22}, X_{23}, X_{24}, X_{27}, X_{31}, X_{37}

ผู้วิจัยทดสอบการแจกแจงปกติของตัวแปรแต่ละตัว และคัดเลือกชุดข้อมูลจริงที่ตัวแปรมีการแจกแจงแบบปกติ นำมาทดลองกับวิธีการประมาณค่าสูญหายทั้ง 2 วิธี จำนวน 2 ชุด คือ 1) ข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ สำหรับศึกษากรณีตัวแปรอิสระ 3 ตัว ตัวแปรตาม 2 ตัว และ 2) ข้อมูลกลุ่มความชื้น สำหรับกรณีตัวแปรอิสระ 5 ตัว ตัวแปรตาม 3 ตัว จากนั้นคำนวณค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรแต่ละกลุ่มดังนี้

ชุดที่ 1 กลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์อากาศ

	x_{22}	x_{23}	x_{24}
x_{22}	1	0.619381	-0.6407
x_{23}	0.619381	1	-0.90534
x_{24}	-0.6407	-0.90534	1

ชุดที่ 2 กลุ่มความชื้น

	x_1	x_5	x_7	x_{33}	x_{46}
x_1	1	0.805334	0.780154	0.787179	0.837099
x_5	0.805334	1	0.988088	0.902584	0.798298
x_7	0.780154	0.988088	1	0.882782	0.797284
x_{33}	0.787179	0.902584	0.882782	1	0.871613
x_{46}	0.837099	0.798298	0.797284	0.871613	1

ข้อสมมติเบื้องต้นสำหรับการประมาณข้อมูลสูญหายด้วยวิธี EM มีดังนี้

1. ข้อมูลต้องมีการแจกแจงแบบปกติพหุ (Multivariate Normal distribution)

2. การสูญหายของข้อมูลเป็นแบบสุ่ม (Schafer, 1997)

ดังนั้น ในการศึกษาครั้งนี้ผู้วิจัยจึงต้องตรวจสอบการแจกแจงแบบปกติพหุของตัวแปร ซึ่งถ้าสามารถตรวจสอบได้ว่าเวกเตอร์ของตัวแปรมีการแจกแจงแบบปกติพหุจะสามารถสรุปได้ว่าตัวแปรแต่ละตัวมีการแจกแจงแบบปกติด้วย (กัลยา, 2552) วิธีการตรวจสอบการแจกแจงแบบปกติพหุมีหลายวิธี ในที่นี้ผู้วิจัยเลือกใช้วิธีของมาร์ดินา (Mardina) ที่ใช้การตรวจสอบด้วยค่าความเบ้ และความโด่ง ที่ระดับนัยสำคัญ 0.01 แสดงผลการตรวจสอบดังตาราง

ตารางที่ 11 ค่า P- value ในการตรวจสอบการแจกแจงแบบปกติพหุด้วยวิธีของมาร์ดินา

ชุดข้อมูล	ตัวแปรอิสระ	ตัวแปรตาม	ค่า P- value Based on skewness	ค่า P- value Based on kurtosis
1	X_{22}, X_{23}, X_{24}	X_{31}, X_{37}	0.0144756	0.0908735
2	$X_1, X_5, X_7, X_{33}, X_{46}$	X_9, X_{10}, X_{11}	0.0321868	0.8308719

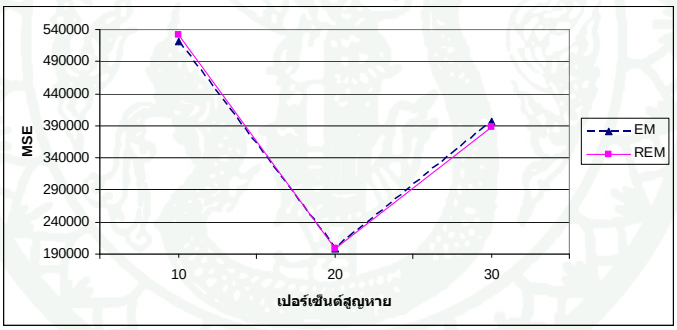
จากตารางที่ 11 ผลการตรวจสอบการแจกแจงแบบปกติพหุของตัวแปรอิสระทั้ง 2 ชุดข้อมูล พบว่า ข้อมูลทั้ง 2 ชุดมีการแจกแจงแบบปกติพหุ อย่างมีนัยสำคัญที่ระดับ 0.01 โดยมีค่า P- value เกิน 0.01

จากนั้นทำให้ตัวแปรตามของชุดข้อมูลแต่ละชุดมีการสูญหายแบบสุ่ม ที่ระดับการสูญหาย 10%, 20% และ 30% ของขนาดตัวอย่าง และใช้การประมาณค่าข้อมูลสูญหายทั้ง 2 วิธี พบว่า เมื่อข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์ของอากาศ มีการสูญหายแบบสุ่ม ที่ระดับการสูญหาย 10% วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM และเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธี REM ให้ค่า MSE ต่ำกว่าวิธี EM และสำหรับข้อมูลกลุ่มความชื้น มีการสูญหายแบบสุ่ม ที่ระดับการสูญหาย 10% และ20% วิธี REMให้ค่า MSE ต่ำกว่าวิธี EM ส่วนที่ระดับการสูญหาย 30% วิธี EMให้ค่า MSE ต่ำกว่าวิธี REM แสดงดังตารางที่ 12 และภาพที่ 19

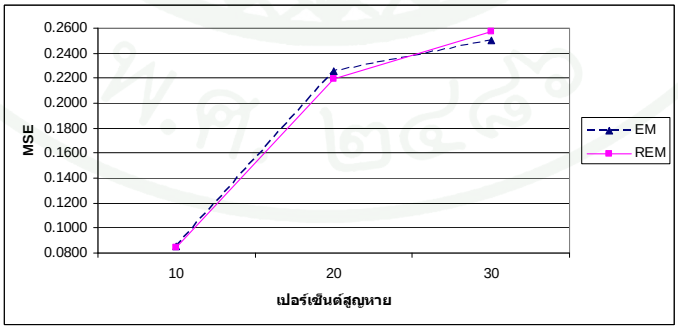
ตารางที่ 12 ค่า MSE ของ วิธี EM และ วิธี REM ที่ทดลองกับชุดข้อมูลจริง เปอร์เซ็นต์การสูญหาย เท่ากับ 10, 20, 30 จำแนกตามชุดข้อมูล

ชุดข้อมูล	เปอร์เซ็นต์การสูญหาย	ค่า MSE	
		EM	REM
1	10	520679.2155*	530788.0139
	20	199080.0440	198973.7993*
	30	396659.3135	388453.1800*
2	10	0.0849	0.0848*
	20	0.2258	0.2197*
	30	0.2502*	0.2575

หมายเหตุ * ค่าMSE ที่มีค่าต่ำสุด



(ก) ชุดข้อมูลที่ 1



(ข) ชุดข้อมูลที่ 2

ภาพที่ 19 ค่า MSE ของ วิธี EM และ วิธี REM ที่ทดลองกับชุดข้อมูลจริง เปอร์เซ็นต์การสูญหาย เท่ากับ 10, 20, 30 จำแนกตามชุดข้อมูล

วิจารณ์

ผลการศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุจากการจำลองข้อมูลโดยเทคนิคมอนติคาร์โล ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ 2 ระดับ คือระดับต่ำ (0.10 - 0.30) และระดับสูง (0.7 - 0.9) พบว่า ปัจจัยที่มีผลต่อการประมาณค่าข้อมูลสูญหายด้วยวิธี EM และวิธี REM คือ

1. จำนวนตัวแปรอิสระ กรณีที่จำนวนตัวแปรอิสระน้อย (X=3, 4) การประมาณค่าข้อมูลสูญหายด้วยวิธี EM และวิธี REM จะให้ค่า MSE ใกล้เคียงกัน แต่วิธี EM เป็นวิธีที่มีการคำนวณซับซ้อนน้อยกว่า วิธี REM ส่วนกรณีที่จำนวนตัวแปรอิสระมาก (X=5, 7) ความซับซ้อนของข้อมูลมีมากกว่า การประมาณค่าข้อมูลสูญหายด้วยวิธี REM จะเหมาะสมกว่าวิธี EM

2. ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระสูง การประมาณค่าข้อมูลสูญหายด้วยวิธี REM จะเหมาะสมกว่าวิธี EM ยกเว้นกรณีขนาดตัวอย่าง50 และจำนวนตัวแปรอิสระน้อย (X=3, 4) ส่วนกรณีระดับความสัมพันธ์ระหว่างตัวแปรอิสระต่ำการประมาณค่าข้อมูลสูญหายด้วยวิธี REM จะเหมาะสมกว่าวิธี EM ที่ขนาดตัวอย่าง 50, 70และจำนวนตัวแปรอิสระมาก (X=5, 7)

3. เปอร์เซ็นต์การสูญหาย เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีทั้งสองให้ค่า MSE เพิ่มขึ้น เนื่องจากเมื่อจำนวนชุดข้อมูลสูญหายมากขึ้น การประมาณค่าสูญหายจึงมีความคลาดเคลื่อนสูงขึ้น ยกเว้นกรณีตัวแปรอิสระ 3 ตัว ที่ขนาดตัวอย่างเท่ากับ 50 และ 70 ค่า MSE ของทั้งสองวิธีจะลดลงเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น และกรณีตัวแปรอิสระ 5 ตัว ที่ขนาดตัวอย่างเท่ากับ 70 และ 100 กับกรณีตัวแปรอิสระ 7 ตัว ที่ขนาดตัวอย่างเท่ากับ 70 ค่า MSE ของวิธี REM จะลดลงเมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น

4. ขนาดตัวอย่าง เมื่อขนาดตัวอย่างเพิ่มขึ้น พบว่า วิธีทั้งสองให้ค่า MSE ลดลง เนื่องจากเมื่อขนาดตัวอย่างเพิ่มในขณะที่เปอร์เซ็นต์การสูญหายเท่าเดิม จะพบว่าสัดส่วนของข้อมูลที่เหลืออยู่มีมากกว่าข้อมูลที่สูญหายไป

สรุปและข้อเสนอแนะ

สรุป

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีประมาณค่าสูญหายแบบ EM Algorithm และแบบ REM Algorithm ในชุดข้อมูลตัวแปรพหุ ที่มีระดับความสัมพันธ์กันสูงและต่ำ โดยพิจารณาจากค่า MSE พร้อมทั้งได้นำข้อมูลจริง จำนวน 2 ชุดมาทดลองใช้วิธีการประมาณค่าสูญหายทั้ง 2 วิธี ผลการศึกษาสรุปได้ดังนี้

ส่วนที่ 1 ผลการศึกษาข้อมูลจากการจำลอง

ผลการเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลตัวแปรพหุทั้ง 2 วิธี คือ EM และ REM นั้นมีความเหมาะสมในสถานการณ์ต่าง ๆ ที่แตกต่างกัน แบ่งออกเป็น 2 กรณี ดังนี้

1.1. กรณีความสัมพันธ์ของตัวแปรระดับสูง

กรณีที่จำนวนตัวแปรอิสระน้อย (X=3, 4) ขนาดตัวอย่าง 50 ที่ทุกระดับการสูญหาย วิธี EM เป็นวิธีที่การคำนวณซับซ้อนน้อยกว่าและเหมาะสม ส่วนขนาดตัวอย่าง 70 และ 100 ที่ทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม กรณีที่จำนวนตัวแปรอิสระมาก (X=5, 7) ที่ทุกขนาดตัวอย่าง และเกือบทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม รายละเอียดแสดงดังตารางที่ 13

1.2. กรณีความสัมพันธ์ของตัวแปรระดับต่ำ

กรณีที่จำนวนตัวแปรอิสระน้อย (X=3, 4) ที่ทุกขนาดตัวอย่าง เกือบทุกระดับการสูญหาย วิธี EM เป็นวิธีที่การคำนวณซับซ้อนน้อยกว่า และเหมาะสม กรณีที่จำนวนตัวแปรอิสระมาก (X=5, 7) ที่ทุกขนาดตัวอย่าง เกือบทุกระดับการสูญหาย วิธี REM เป็นวิธีที่เหมาะสม รายละเอียดแสดงดังตารางที่ 13

ตารางที่ 13 วิธีที่เหมาะสมที่สุดในแต่ละสถานการณ์ของการประมาณค่าสูญหาย
สำหรับข้อมูลตัวแปรพหุ กรณีตัวแปรอิสระมีความสัมพันธ์กันสูงและต่ำ

ρ	รูปแบบชุดข้อมูล		เปอร์เซ็นต์การสูญหาย	ขนาดตัวอย่าง		
	จำนวนตัวแปรอิสระ(X)	จำนวนตัวแปรตาม (Y)		50	70	100
0.7-0.9	3	2	10	EM	EM	REM
			20	EM	REM	REM
			30	EM	REM	REM
	4	2	10	EM	REM	REM
			20	EM	REM	REM
			30	EM	REM	REM
	5	3	10	REM	EM	EM
			20	REM	REM	EM
			30	REM	REM	REM
	7	3	10	EM	EM	REM
			20	REM	EM	REM
			30	REM	REM	REM
0.1-0.3	3	2	10	EM	EM	EM
			20	EM	EM	EM
			30	EM	REM	EM
	4	2	10	EM	EM	EM
			20	EM	EM	EM
			30	EM	EM	REM
	5	3	10	REM	REM	EM
			20	REM	REM	EM
			30	REM	REM	REM
	7	3	10	REM	REM	EM
			20	REM	REM	REM
			30	REM	REM	REM

หมายเหตุ EM แทน Expectation Maximization Algorithm,
REM แทน Regularized Expectation Maximization Algorithm
ในที่นี้ผู้วิจัยทำการสรุปผลการศึกษาโดยพิจารณาจากค่า MSE (ทศนิยม 3 ตำแหน่ง)

ส่วนที่ 2 ผลการศึกษาจากข้อมูลจริง

การนำวิธีประมาณค่าสูญหายแบบ EM Algorithm และแบบ REM Algorithm มาทดลองใช้กับข้อมูลผลการตรวจอากาศชั้นบน (upper air observation) ระหว่างวันที่ 1 มีนาคม 2549 - 31 ตุลาคม 2551 จากสถานีเรดาร์ อำเภอพิมาย จังหวัดนครราชสีมา สำนักฝนหลวงและการบินเกษตร กระทรวงเกษตรและสหกรณ์ พบว่า ในชุดข้อมูลที่ 1 และชุดข้อมูลที่ 2 วิธี REM มีประสิทธิภาพดีกว่า วิธี EM เนื่องจากชุดข้อมูลจริงดังกล่าว มีตัวแปรที่มีความสัมพันธ์กันสูง ซึ่งส่วนใหญ่สอดคล้องกับผลการศึกษาจากข้อมูลจำลอง ยกเว้นกรณีชุดข้อมูลที่ 2 ระดับการสูญหาย 30 เปอร์เซ็นต์ ซึ่งค่า MSE ของวิธี EM ต่ำกว่าวิธี REM เล็กน้อย

ตารางที่ 14 วิธีที่เหมาะสมที่สุดในแต่ละสถานการณ์ของการประมาณค่าสูญหายสำหรับข้อมูลจริง จำแนกตามชุดข้อมูล

ชุดข้อมูล	เปอร์เซ็นต์การสูญหาย	วิธีที่เหมาะสมที่สุดของการประมาณค่าสูญหาย
1	10	EM
	20	REM
	30	REM
2	10	REM
	20	REM
	30	EM

หมายเหตุ EM แทน Expectation Maximization Algorithm,
REM แทน Regularized Expectation Maximization Algorithm
ในที่นี้ผู้วิจัยทำการสรุปผลการศึกษาโดยพิจารณาจากค่า MSE (ทศนิยม 3 ตำแหน่ง)

ข้อเสนอแนะ

จากผลการศึกษาวิธีการประมาณค่าข้อมูลสูญหายในชุดข้อมูลตัวแปรพหุ ด้วยวิธี EM และวิธี REM ผู้วิจัยมีข้อเสนอแนะดังต่อไปนี้

ข้อเสนอแนะเพื่อการนำผลการวิจัยไปใช้

1. กรณีจำนวนตัวแปรอิสระน้อย (X=3, 4) วิธี EM เป็นวิธีการคำนวณซับซ้อนน้อยกว่าและให้ค่า MSE ใกล้เคียงกับวิธี REM จึงเหมาะสมในการใช้ประมาณค่าข้อมูลสูญหาย

2. กรณีจำนวนตัวแปรอิสระมาก (X=5, 7) ซึ่งความซับซ้อนของข้อมูลมีมากกว่า วิธี REM จึงเหมาะสมในการใช้ประมาณค่าข้อมูลสูญหายมากกว่า วิธี EM

ข้อเสนอแนะในการวิจัยครั้งต่อไป

1.ในการศึกษาครั้งนี้ผู้วิจัยใช้ค่าเฉลี่ยฮาร์โมนิกในการคำนวณค่า c เริ่มต้นในวิธี REM ดังนั้น ในการศึกษาครั้งต่อไปอาจใช้ค่าเฉลี่ยเลขคณิต หรือ ค่าเฉลี่ยเรขาคณิต ในการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายเพิ่มเติม

2. ในการศึกษาครั้งนี้ผู้วิจัยศึกษาในกรณีเปอร์เซ็นต์การสูญหายของตัวแปรตามอยู่ในระดับ 10 - 30 เปอร์เซ็นต์ ในการศึกษาครั้งต่อไปอาจเพิ่มระดับการสูญหายให้สูงขึ้น เช่น 50 - 70 เปอร์เซ็นต์ หรือ มีการสูญหายของข้อมูลในตัวแปรอิสระ

3. ในการศึกษาครั้งนี้ผู้วิจัยศึกษากรณีที่ข้อมูลเชิงปริมาณ ในการศึกษาครั้งต่อไปอาจศึกษาเมื่อเป็นข้อมูลเชิงคุณภาพ เช่น การวิเคราะห์การถดถอยแบบลอจิสติค การใช้ตัวแปรหุ่นในสมการถดถอย หรือ การวิเคราะห์ข้อมูลลักษณะอื่น เช่น การวิเคราะห์อนุกรมเวลา เป็นต้น

เอกสารและสิ่งอ้างอิง

กัลยา วานิชย์บัญชา. 2552. **การวิเคราะห์ข้อมูลหลายตัวแปร.** บริษัท ธรรมสาร จำกัด, กรุงเทพฯ.

จริยา แสงสุวรรณ. 2551. **การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ**.วิทยาพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

ชะไมพร ธรรมวัฒน์ไพศาล. 2522. **วิธีการประมาณค่าที่ขาดหายไปในการวิเคราะห์การถดถอย**. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

ชุติมา ชัยมุสิก. 2533. **การวิเคราะห์การถดถอยเชิงซ้อนเมื่อข้อมูลของตัวแปรอิสระสูญหาย.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

เชาว์ อินใย. 2547. **การพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีและการตรวจสอบความแม่นยำและอำนาจการทดสอบเปรียบเทียบกับวิธีอีเอ็มและลิสท์ไวท์: เทคนิคมอนติคาร์โล.** วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยนเรศวร.

ดลพรรษ พันธุ์พาณิชย์. 2552**. การวิเคราะห์ปัจจัยสำหรับตัวแบบคาดการณ์โอกาสการเกิดฝน ภาคตะวันออกเฉียงเหนือของประเทศไทย.** วิทยาพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

ดวงภรณ์ โปทาวี. 2552. **การประมาณค่าข้อมูลสูญหายในการวัดซ้ำด้วยวิธีมาร์คอฟเชนมอนติคาร์โลและวิธีคอปูลาส์.** วิทยาพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

ถวัลย์ จันทร์เพ็ง. 2531. **การเปรียบเทียบความแม่นยำของการประมาณค่าข้อมูลสูญหายสามวิธีในกลุ่มตัวอย่างขนาดเล็ก**. วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

ประลองพล ประสงค์พร. 2551. **การประมาณค่าพารามิเตอร์ในตัวแบบความถดถอยโลจิสติกเมื่อมีค่าสูญหาย.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

ปิยะภรณ์ ประสิทธิ์วัฒนเสรี และ สุคนธ์ ประสิทธิ์วัฒนเสรี. 2551. ข้อมูลสูญหายและแนวทางการจัดการ. **Data Management & Biostatistics Journal** Vol.4 No.3 : 52-61.

พรศิริ หมื่นไชยศรี. 2529. **การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ตัวแปรพหุ.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

เพียงออ ยี่สา. 2551. **การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

มนูญ บุบผาพ่วง. 2550. **การศึกษาเปรียบเทียบวิธีริดจ์รีเกรสชั่นและวิธีเจเนอรัลไรซ์แมกซิมั่มเอนโทรฟี่ เมื่อเกิดภาวะร่วมเส้นตรงหลายตัวแปร**. วิทยาพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

วารุณี ตรีบำรุงศักดิ์. 2538. **การพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุเมื่อตัวแปรตามมีค่าสูญหาย.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

วรรัตน์ ราชกิจจา. 2547. **การเปรียบเทียบการประมาณค่าพารามิเตอร์ในการวิเคราะห์ความถดถอยพหุคูณโดยวิธีเอ็มพิริคัลเบส์ริดจ์รีเกรสชันเมื่อเกิดพหุสัมพันธ์.** วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย.

วราวุธ ขันติยานันท์. 2539. **วิเคราะห์ผลพยากรณ์อากาศโดย GPCM.** สำนักฝนหลวงและการบินเกษตร สำนักปลัดกระทรวงเกษตรและสหกรณ์, กรุงเทพฯ.

Bolch, B.W. and C.J. Huang 1974. **Multivariate Statistical for Business and Economics**. Prentice-Hall, Inc. Englewood Clilfs.

Chantala , K and C. Suchindran, n.d. **Multiple Imputation for Missing Data**. Center for Population Studies, University of North Carolina, Chapel Hill. <hpp://www.cpc.unc.edu/services/computer/presentation/mi presentation 2.pdf>

Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. Maximum Likelihood From Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society.** B39:1-38.

Enders, C.K. 2001. The Performance of the Full Information Maximum Likelihoods Estimator in Multiple Regression Models with Missing Data. **Educational and Psychological Measurement.** 61: 713-740.

Heeringa, S. 2000. **Multivariate Imputation of Coarsened Survey on Household Wealth**. Ph.D. thesis, Michigan University.

Hoerl, A.E. and R.W. Kennard. 1970. Ridge Regression: Biased Estimator for Nonorthogonal Problems. **Technometrics.** 12 (1): 55-67.

Huang, R. and K.C. Carriere. 2006. Comparison of Methods for Incomplete Repeated Measures Data Analysis in Small Samples. **Statistical Planning and Inference.** 136: 235-247.

Little, R.J.A. and D.B. Rubin. 2002. **Statistical Analysis with Missing Data**. John Wiley and Sons, Inc. New York.

Raaijimakers, Q.A.W. 1999. Effectiveness of Different Missing Data Treatments in Surveys with Likert – Type Data: Introducing the Relative Mean Substitution Approach. **Educational and Psychological Measurement.** 59: 725-748.

Roth, P.L. 1994. Missing Data: A Conceptual Review for Applied Psychologists. **Personnel Psychology.** 47: 537-560.

Schafer, J.L. 1997. **Analysis of Incomplete Multivariate Data.** Chapman and Hall, Inc.,London.

Schneider, T. 2001. Analysis of incomplete climate data: Estimation of mean values and Covariance matrices and imputation of missing values. **American Meteorological Society.** 14:853-871.

Truxillo, C. 2005. **Maximum Likelihood Parameter Estimation with Incomplete Data.** SAS Institute, Inc., Cary, NC.

Yuan, Y.C. 2000. **Multiple Imputation for Missing Data: Concepts and New Development.** SAS Institute, Inc., Rockville, MD.

ภาคผนวก

ภาคผนวก ก

ข้อมูลจริงที่ใช้ในการศึกษา

1. ชุดข้อมูลกลุ่มการทรงตัวของบรรยากาศและค่าพยากรณ์ของอากาศ

ลำดับที่	ตัวแปรอิสระ			ตัวแปรตาม	
	x22	x23	x24	x31	x37
1	0.3	0.4	45.8	7139.2	15.04
2	-1.2	-1	47.1	6345.3	12.20
3	1	2.9	41.8	6955.5	12.52
4	0.8	5.9	38	7366.2	14.38
5	0.9	2.4	43.6	6677.6	11.31
6	3.1	5.8	37.6	7751.5	13.00
7	0.4	0.9	44.8	7398.2	13.12
8	2.5	3.1	41.8	8876.2	14.21
9	0.6	0.5	44.2	7774.8	13.59
10	0.2	2.5	42.4	6010.3	12.03
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
81	1.1	0.5	43.4	5907	11.19
82	0.9	2.1	42.6	7400	14.39
83	0.1	1.7	43	5945	11.50
84	-1.5	-0.4	46	5604.1	11.25
85	0.3	0.8	44	7025.1	14.01
86	0.1	3.9	39.4	6012.9	15.20
87	2.4	3.9	39.6	6191.8	14.34

2. ชุดข้อมูลกลุ่มความชื้น

ลำดับที่	ตัวแปรอิสระ					ตัวแปรตาม		
	x1	x5	x7	x33	x46	x9	x10	x11
1	1.74	3.47	3.7	61.9	63.9	12.67	13.3	17.48
2	1.87	3.98	4.06	69.3	72.1	13.81	14.32	18.64
3	1.7	3.98	4.11	65.4	66.6	13.17	13.8	18.07
4	1.73	3.38	3.49	50.4	63.8	13.65	14.75	19.1
5	1.44	3.64	3.96	63.1	63	10.77	12.28	16.21
6	1.49	3.16	3.22	55.1	62.6	11.39	12.4	16.37
7	1.56	4.04	4.17	63.5	59	11.72	13.09	17.21
8	1.39	3.62	3.82	56	48.3	10.33	11.11	14.67
9	1.77	4.08	4.27	66.9	60.8	13.09	13.48	17.7
10	1.89	4.42	4.78	75	73.1	14.42	14.75	19.11
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
81	1.85	4.71	5.17	76	76.6	13.85	14.11	18.41
82	1.76	4.58	5.04	71.3	74	13.22	13.72	17.96
83	1.79	4.37	4.79	69.7	69.8	13.7	14.63	18.97
84	1.87	4.66	5.13	74.9	73.7	14.16	15.08	19.44
85	1.82	4.35	4.6	67.1	69.4	13.7	14.31	18.64
86	1.9	4.28	4.66	67.6	75.8	14.41	15.07	19.44
87	1.66	4.37	4.9	71.6	67.9	13.04	13.96	18.24

ภาคผนวก ข

โปรแกรม R ที่ใช้ในการศึกษา

1. โปรแกรมสำหรับจำลองชุดข้อมูลตัวแปรพหุ

```
library(MASS)
options(digits=20)

ini.mu   <- 0              #mean of independent variables
num.ind  <- c(3, 4, 5, 7)   #number of independent variables
sam.size <- c(50, 70, 100)  #sample size
error.mu <- c(0, 0, 0)      #mean of error term
error.sigma <- c(1, 5, 10)  #standard deviation of error term
max.dep  <- 3              #maximum number of dependent variables
beta3x <- c(0, 1, 1, 1)
beta4x <- c(0, 1, 1, 1, 1)
beta5x <- c(0, 1, 1, 1, 1, 1)
beta7x <- c(0, 1, 1, 1, 1, 1, 1, 1)

pr.miss <- c(10, 20, 30) #percentage number of missing value base on row missing

mum.loop <- 1000
num.seed <- 17031980
max.loop <- 30
opt.error <- 0.0001

sigma3x <- matrix(c(1.0, 0.7, 0.8, 0.7, 1.0, 0.9, 0.8, 0.9, 1.0), nrow=3)
fm3x2y  <- formula(cbind(y1, y2) ~ x1 + x2 + x3)
sigma4x <- matrix(c(1.0, 0.7, 0.8, 0.7, 0.7, 1.0, 0.7, 0.8, 0.8, 0.7, 1.0, 0.7,
          0.7, 0.8, 0.7, 1.0), nrow=4)
fm4x2y  <- formula(cbind(y1, y2) ~ x1 + x2 + x3 + x4)
sigma5x <- matrix(c(1.0, 0.7, 0.8, 0.7, 0.7, 0.7, 1.0, 0.7, 0.8, 0.8, 0.8, 0.7,
          1.0, 0.7, 0.7, 0.7, 0.8, 0.7, 1.0, 0.8, 0.7, 0.8, 0.7, 0.8, 1.0), nrow=5)
```

```
fm5x3y  <- formula(cbind(y1, y2, y3) ~ x1 + x2 + x3 + x4 + x5)
sigma7x <- matrix(c(1.0, 0.7, 0.8, 0.7, 0.7, 0.8, 0.7, 0.7, 1.0, 0.7, 0.8, 0.8,
          0.7, 0.7, 0.8, 0.7, 1.0, 0.7, 0.7, 0.7, 0.8, 0.7, 0.8, 0.7, 1.0,
          0.8, 0.7, 0.7, 0.7, 0.8, 0.7, 0.8, 1.0, 0.7, 0.7, 0.8, 0.7, 0.7,
          0.7, 0.7, 1.0, 0.8, 0.7, 0.7, 0.8, 0.7, 0.7, 0.8, 1.0), nrow=7)
fm7x3y  <- formula(cbind(y1, y2, y3) ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)

#-------------------------------------------------------------------------#
#                          PROGRAM START                                  #
#-------------------------------------------------------------------------#
#Random Number Generator of X, Error and Y
#Generate error terms: c(list1, list2, list3) =c (n50, n70, n100)
set.seed(num.seed)
error <- error.gen(sam.size, error.mu, error.sigma)

#Generate x: n=50, p=3 with correlation matrix: use in case 1, 2, 3
set.seed(num.seed)
x3n050 <- x.gen(sam.size[1], ini.mu, num.ind[1], sigma3x, nloop=mum.loop,
        seed=num.seed, c("x1","x2","x3"))
#Generate x: n=70, p=3 with correlation matrix: use in case 4, 5, 6
set.seed(num.seed)
x3n070 <- x.gen(sam.size[2], ini.mu, num.ind[1], sigma3x, nloop=mum.loop,
        seed=num.seed, c("x1","x2","x3"))
#Generate x: n=100, p=3 with correlation matrix: use in case 7, 8, 9
set.seed(num.seed)
x3n100 <- x.gen(sam.size[3], ini.mu, num.ind[1], sigma3x, nloop=mum.loop,
        seed=num.seed, c("x1","x2","x3"))
```

2. โปรแกรมสำหรับจำลองชุดข้อมูลตัวแปรตาม

```
#Generate y=2 from x=3, n=50, beta and error term
y2x3n050 <- lapply(x3n050, FUN=y.gen, beta3x, error[[1]][,-ncol(error[[1]])],
        c("y1","y2"))
#Generate y=2 from x=3, n=70, beta and error term
y2x3n070 <- lapply(x3n070, FUN=y.gen, beta3x, error[[2]][,-ncol(error[[2]])],
        c("y1","y2"))
#Generate y=2 from x=3, n=100, beta and error term
y2x3n100 <- lapply(x3n100, FUN=y.gen, beta3x, error[[3]][,-ncol(error[[3]])],
        c("y1","y2"))
```

3. โปรแกรมสำหรับสุ่มชุดข้อมูลสูญหาย

```
doMissing <- function(org.response, percent, cnames=NULL){
   NRow <- nrow(org.response)
   nMissing <- floor(NRow*percent/100)
   rowMissing <- sample(c(1:NRow), nMissing)
   dataMissing <- org.response
   dataMissing[rowMissing, ] <- NA
   output <- dataMissing
   colnames(output) <- cnames
   return(output)
}
```

4. โปรแกรมสำหรับการประมาณค่าสูญหายด้วยวิธี EM

```
OEM <- function(list.data, Formula, maxLoop=30, error=0.01){
  RowMissing <- which(apply(list.data[[1]], 1, FUN=function(x){sum(is.na(x))})!=0)
  dtFitModel <- list.data[[1]][-RowMissing, ]
  lmModel <- lm(Formula, data=as.data.frame(dtFitModel))
  betaEst <- lmModel$coefficients
  #Call Function "yPredict"
  yEst <- yPredict(RowMissing, list.data[[2]], betaEst)

  betaTemp <- new("list")
  i    <- 2
  nLoop <- 2
  while(i<=maxLoop){
    yNew <- list.data[[3]]
    yNew[RowMissing, ] <- yEst
    dataAdjust <- cbind(list.data[[2]], yNew)
    ModelTemp  <- lm(Formula, data=as.data.frame(dataAdjust))
    betaTemp[[i]] <- ModelTemp$coefficients
    #Call Function "yPredict"
    yEstTemp  <- yPredict(RowMissing, list.data[[2]], betaTemp[[i]])
    if(all(abs(betaTemp[[i]]-betaEst)<=error)){
      i    <- i+maxLoop
      nLoop <- nLoop
    }else{
      betaEst <- betaTemp[[i]]
      nLoop   <- nLoop+1
      i      <- i+1
    }
  }
```

```
    betaFinal <- betaTemp[[length(betaTemp)]]
    output <- list(yEstTemp, betaFinal, nLoop, RowMissing)
    names(output) <- c("y.estimated","beta.estimated","loops","row.missing")
    return(output)
}
```

5. โปรแกรมสำหรับการประมาณค่าสูญหายด้วยวิธี REM

```
REM <- function(list.data, Formula, maxLoop=30, error=0.01){
  RowMissing <- which(apply(list.data[[1]], 1, FUN=function(x){sum(is.na(x))})!=0)
  dtFitModel <- list.data[[1]][-RowMissing, ]
  lmModel <- lm(Formula, data=as.data.frame(dtFitModel))
  betaEst <- lmModel$coefficients
  #Calculate Y estimate (Yhat) from initial beta estimate and x exclude row y missing
  xComp <- cbind(1, list.data[[2]][-RowMissing, ])
  colnames(xComp) <- rownames(betaEst)
  yHat  <- xComp%*%betaEst  #Calc YHat
  #Calculate Error: Diff Yorg (excl. missing) and Yhat
  sqE <- (list.data[[3]][-RowMissing, ]-yHat)^2
  #Calculate Variance
  varT <- apply(sqE, 2, sum)/(nrow(xComp)-(ncol(xComp)-1)-1)
  #Matrix multiplication beta transpose and beta
  BtBofY <- rep(NA, ncol(list.data[[3]]))
  for(i in 1:ncol(list.data[[3]])){
    BtBofY[i] <- t(betaEst[,i])%*%betaEst[,i]
  }
  #Calculate "c" values
  cT <- (as.vector(varT)*ncol(list.data[[2]]))/BtBofY
  #Calculate beta of ridge regression (BRR)
  XtX <- t(xComp)%*%xComp
  XtY <- t(xComp)%*%list.data[[3]][-RowMissing, ]
  #C*Identity Matrix
  CIM <- new("list")
  BRR <- new("list")
  for(i in 1:length(cT)){
    CIM[[i]] <- matrix(0, ncol(xComp), ncol(xComp))
```

```
    diag(CIM[[i]]) <- cT[i]
    CIM[[i]] <- solve(CIM[[i]]+XtX)
    BRR[[i]] <- CIM[[i]]%*%XtY[,i]
}
BRR <- matrix(unlist(BRR),nrow(XtY))

xAll <- cbind(1, list.data[[2]])
xMis <- cbind(1, list.data[[2]][RowMissing,])
colnames(xAll) <- rownames(betaEst)
colnames(xMis) <- rownames(betaEst)

BRRadj <- new("list")
i    <- 2
nLoop <- 2
while(i<=maxLoop){
    yHatAll <- xAll%*%BRR
    yHatMis <- xMis%*%BRR   #for update Ynew
    #Update y
    yUpdated <- list.data[[3]]
    yUpdated[RowMissing, ] <- yHatMis
    #Calc. Sqaure Error
    sqEnew <- (yHatAll-yUpdated)^2
    #Calc. Var.
    varTnew <- apply(sqEnew, 2, sum)/(nrow(xAll)-(ncol(xAll)-1)-1)
    #Matrix multiplication BRR transpose and BRR (in first loop)
    BRRtBRRofY <- rep(NA, ncol(list.data[[3]]))
    for(j in 1:ncol(list.data[[3]])){
        BRRtBRRofY[j] <- t(BRR[,j])%*%BRR[,j]
    }
```

```
#Calc. c-value second time
    cTnew <- (as.vector(varTnew)*ncol(list.data[[2]]))/BRRtBRRofY

    #If some of c-value equare Inf stop, because cannot calc. Inv matrix
    if(all(cTnew!=Inf)){

      XAtXA <- t(xAll)%*%xAll
      XAtYU <- t(xAll)%*%yUpdated

      CIMnew <- new("list")
      BRRnew <- new("list")
      for(k in 1:length(cTnew)){
        CIMnew[[k]] <- matrix(0, ncol(xAll), ncol(xAll))
        diag(CIMnew[[k]]) <- cTnew[k]
        CIMnew[[k]] <- solve(CIMnew[[k]]+XAtXA)
        BRRnew[[k]] <- CIMnew[[k]]%*%XAtYU[,k]
      }
      BRRadj[[i]] <- matrix(unlist(BRRnew),nrow(XAtYU))

      errorBRR <- as.vector(abs(BRR-BRRadj[[i]]))

      if(all(errorBRR<=error)){
        i     <- i+maxLoop
        nLoop <- nLoop
      }else{
        BRR <- BRRadj[[i]]
        nLoop   <- nLoop+1
        i       <- i+1
      }
    }else{
```

```
        i    <- i+maxLoop
        nLoop <- nLoop
      }
      #print(i)
    }
    betaFinal <- BRRadj[[length(BRRadj)]]
    output <- list(yUpdated[RowMissing,], betaFinal, nLoop, RowMissing)
    names(output) <- c("y.estimated","beta.estimated","loops","row.missing")
    return(output)
}
```

6. โปรแกรมสำหรับการคำนวณค่า MSE

```
mse.calc <- function(dataList){
   Yhat   <- dataList[[1]]$y.estimated
   RowMis <- dataList[[1]]$row.missing
   YorgMis <- dataList[[2]][RowMis, ]
   mse.loop <- sum((Yhat-YorgMis)^2)/length(Yhat)
   output <- mse.loop
   return(output)
}
```

ประวัติการศึกษา และการทำงาน

ชื่อ	นางสาวน้ำทิพย์ พนมไทย
เกิดวันที่	17 มีนาคม 2523
สถานที่เกิด	อำเภอบางขุนเทียน จังหวัดกรุงเทพมหานคร
ประวัติการศึกษา	วท.บ. (ฟิสิกส์) มหาวิทยาลัยเชียงใหม่
ตำแหน่งหน้าที่การงานปัจจุบัน	นักวิทยาศาสตร์
สถานที่ทำงานปัจจุบัน	สำนักฝนหลวงและการบินเกษตร กระทรวงเกษตรและสหกรณ์
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-