วิทยานิพนธ์

การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย
ในการวิเคราะห์การถดถอยพหุคูณ

A COMPARATIVE STUDY OF MISSING DATA ESTIMATION METHODS IN MULTIPLE REGRESSION ANALYSIS

นางสาวจริยา แสงสุวรรณ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (สถิติ)

ปริญญา

สถิติ	สถิติ
สาขา	**ภาควิชา**

เรื่อง การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ

A Comparative Study of Missing Data Estimation Methods in Multiple Regression Analysis

นามผู้วิจัย นางสาวจริยา แสงสุวรรณ

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ....

(รองศาสตราจารย์ประสิทธิ์ พยัคฆพงษ์, M.S.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

หัวหน้าภาควิชา ....

(อาจารย์อำไพ ทองธีรภาพ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 15 **เดือน** มีนาคม **พ.ศ.** 2551

วิทยานิพนธ์

เรื่อง

การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ

A Comparative Study of Missing Data Estimation Methods in Multiple Regression Analysis

โดย

นางสาวจริยา แสงสุวรรณ

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ)
พ.ศ. 2551

จริยา แสงสุวรรณ 2551: การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย ในการวิเคราะห์การถดถอยพหุคูณ ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์ประสิทธิ์ พยัคฆพงษ์, M.S. 84 หน้า

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ โดยทำการประมาณค่าสูญหายของตัวแปรตาม 4 วิธี คือ วิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย และวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า (วิธีเอ็มไอ) เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) การเปรียบเทียบกระทำภายใต้สถานการณ์ของขนาดตัวอย่างเท่ากับ 50 70 100 และ 200 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 5 และ 15 เปอร์เซ็นต์การสูญหายของตัวแปรตามเท่ากับ 5% 10% 20% และ 30% ตามลำดับ และตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร ซึ่งระดับความสัมพันธ์ระหว่างตัวแปรอิสระมี 3 ระดับ คือ ระดับต่ำ (0.20) ระดับปานกลาง (0.50) และระดับสูง (0.70) ในการวิจัยครั้งนี้ได้ทำการจำลองด้วยวิธีมอนติคาร์โล ซึ่งกระทำซ้ำ 5,000 ครั้ง ในแต่ละสถานการณ์

ผลการวิจัยสรุปได้ดังนี้ เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีสมการถดถอยและวิธีเอ็มไอให้ค่าประมาณของ RMSE ลดลง และวิธีการประมาณค่าสูญหายทั้ง 4 วิธี ให้ค่าประมาณของ RMSE แตกต่างกัน วิธีสมการถดถอยและวิธีเอ็มไอ ให้ค่าประมาณของ RMSE ใกล้เคียงกัน แต่เนื่องจากวิธีสมการถดถอยเป็นวิธีที่ง่ายและไม่ซับซ้อน ดังนั้น วิธีสมการถดถอย จึงเป็นวิธีที่เหมาะสมสำหรับการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ

ลายมือชื่อนิสิต	ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	17 / มี.ค. / 2551

Chariya Saengsuwan 2008: A Comparative Study of Missing Data Estimation Methods in Multiple Regression Analysis. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Associate Professor Prasit Payakkapong, M.S. 84 pages.

The purpose of this research was to compare missing data estimation methods of the dependent variables in multiple regression analysis. Four estimation methods of missing data of the dependent variables were loss method, mean method, regression method and multiple imputation method (mi) . The criterion of comparison was the estimated squares root of mean squares error (RMSE). The comparisons were done under the conditions of sample sizes 50, 70, 100 and 200; standard deviations of error 1, 5 and 15; missing percentage 5%, 10%, 20% and 30% and distribution of independent variables were multivariate normal distribution with the levels of correlations among independent variables classified into 3 levels : low levels (0.20), middle levels (0.50) and high levels (0.70). The study used the monte carlo simulation method with repeated 5,000 times under each situation.

The results of the study are concluded as follows. When missing percentage are increased the estimated values of RMSE of regression method and mi method are decreased and four estimation methods of missing data give different increasing values of RMSE . The estimated values of RMSE of regression method and mi method give similar results but regression method is easier and less complicate, so that regression method is suitable method for estimation the missing data of dependent variables in multiple regression analysis.

Chariya Saengsuwan — Student's signature

Prasit Payakkapong — Thesis Advisor's signature

17 / March / 2008

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ รศ. ประสิทธิ์ พยัคฆพงษ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ที่ได้ช่วยเหลือในการวางแผนงานวิจัยในวิทยานิพนธ์ฉบับนี้ ตลอดจนการให้คำปรึกษา แนะนำ และตรวจแก้ไขข้อบกพร่องต่าง ๆ ขอกราบขอบพระคุณ อ. ดร. อำไพ ทองธีรภาพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้คำแนะนำและช่วยเหลือในการทำวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี ขอขอบพระคุณ รศ. เปรมใจ ตรีสรานุวัฒนา ประธานกรรมการสอบ และ ผศ. ดร. กุศยา ปลั่งพงษ์พันธ์ กรรมการสอบ และท่านอาจารย์ทุกท่านที่ประสิทธิ์ประสาทความรู้ให้แก่ข้าพเจ้าตลอดมา

ขอขอบพระคุณวารสารวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ที่ให้ความอนุเคราะห์ในการตีพิมพ์วิทยานิพนธ์

ขอขอบพระคุณ คุณวิลาวรรณ ศิริงามเพ็ญ ผู้อำนวยการกองวิชาการและวางแผน และ คุณบุญลักษณ์ อนันต์วัฒนชัย หัวหน้าฝ่ายสถิติ กรมการขนส่งทางน้ำและพาณิชยนาวี ที่คอยให้ความช่วยเหลือและสนับสนุนการทำวิทยานิพนธ์

ขอขอบพระคุณ แม่ น้าสพ น้านึก พี่หลวง อวยและครอบครัว ไหม นันต์ ตา อ้วน และเต๋า ที่สนับสนุนด้านกำลังทรัพย์ในขณะศึกษา

ขอขอบคุณ พี่ ๆ เพื่อน ๆ และน้อง ๆ ทุกคนที่ให้ความช่วยเหลือและเป็นกำลังใจขณะทำวิทยานิพนธ์

สุดท้ายนี้ คุณค่าและประโยชน์อันจะพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่แม่ บูรพาจารย์ และผู้มีพระคุณทุกท่าน

จริยา แสงสุวรรณ

มีนาคม 2551

สารบัญ

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

คำอธิบายสัญลักษณ์และคำย่อ

LOSS	=	วิธีสูญหาย
MEAN	=	วิธีค่าเฉลี่ย
REG	=	วิธีสมการถดถอย
MI	=	วิธีเอ็มไอ
n	=	ขนาดตัวอย่าง
PM	=	เปอร์เซ็นต์การสูญหาย
ρ	=	ความสัมพันธ์ระหว่างตัวแปรอิสระ
σ	=	ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน
RMSE	=	ค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์

การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย
ในการวิเคราะห์การถดถอยพหุคูณ

A Comparative Study of Missing Data Estimation Methods
in Multiple Regression Analysis

คำนำ

ในการศึกษาวิจัยทั่ว ๆ ไป บ่อยครั้งที่ผู้วิจัยพบว่า ข้อมูลที่รวบรวมมาได้มีคำตอบไม่ครบถ้วนสมบูรณ์ ข้อมูลได้รับคำตอบเพียงบางส่วน ซึ่งส่งผลให้ไม่สามารถใช้ประโยชน์จากข้อมูลนั้นได้เต็มที่ ซึ่งปัญหาข้อมูลสูญหายอาจไม่ถือว่าเป็นปัญหาที่รุนแรงหรืออาจถือว่าเป็นเรื่องเล็กน้อยถ้าการวิเคราะห์ข้อมูลกระทำด้วยสถิติที่วิเคราะห์ข้อมูลรายตัวแปร เช่น ร้อยละ หรือค่าเฉลี่ย หรือ สถิติพรรณนาอื่น แต่ถ้าการวิเคราะห์ข้อมูลนั้นจำเป็นต้องใช้วิธีวิเคราะห์พหุ เช่น การวิเคราะห์การถดถอยพหุคูณ การวิเคราะห์เส้นทาง การวิเคราะห์ปัจจัย การวิเคราะห์กลุ่มและการวิเคราะห์การจำแนกพหุ ในกรณีการสูญหายของข้อมูลจะมีผลกระทบที่รุนแรง เพราะถ้าพบว่าหน่วยวิเคราะห์ใดมีตัวแปรใดขาดข้อมูลไปแม้เพียงตัวเดียวก็จะตัดหน่วยวิเคราะห์นั้นทิ้งทั้งหน่วยโดยไม่สนใจว่าจะยังมีตัวแปรตัวอื่นอีกมากที่มีข้อมูลครบถ้วนหรือไม่ (Heeringa, 2000)

การวิเคราะห์การถดถอยพหุคูณ (Multiple Regression Analysis) เป็นเทคนิคหนึ่งที่นำมาใช้ในการพยากรณ์กันมาก โดยการใช้ค่าของตัวแปรอิสระ (Independent Variables) ตั้งแต่ 2 ตัวขึ้นไปมาพยากรณ์ค่าของตัวแปรตาม (Dependent Variable) ซึ่งตัวแปรทั้งสองประเภทนี้มีความสัมพันธ์กันในลักษณะใดลักษณะหนึ่ง ตัวอย่างเช่น ราคาสินค้าขึ้นอยู่กับต้นทุนสินค้า รายได้ประชาชาติและส่วนแบ่งการตลาด เป็นต้น

โดยปกติในการวิเคราะห์การถดถอย จะใช้ข้อมูลหรือค่าสังเกตที่ครบสมบูรณ์ของแต่ละตัวแปรที่พิจารณา ถ้าหากข้อมูลที่เก็บรวบรวมมาเพื่อทำการวิเคราะห์เป็นไปตามข้อตกลงเบื้องต้นและมีข้อมูลครบสมบูรณ์ทุกตัวในขอบเขตของการพิจารณาก็ไม่เกิดปัญหาในการวิเคราะห์ แต่ถ้าข้อมูลที่รวบรวมได้นั้นบางตัวค่าสูญหายไป ไม่ว่าจะด้วยสาเหตุ การสูญหายนี้เกิดขึ้นโดยไม่ได้ตั้งใจ หรือเนื่องจากไม่ได้เก็บค่าโดยจงใจ หรืออาจเนื่องมาจากค่าใช้จ่ายในการเก็บตัวอย่างมีจำนวน

จำกัด หรืออาจเกิดจากเวลา หรือสภาวะแวดล้อมที่ต้องทำให้ค่าสังเกตบางค่านั้นหายไป และไม่สามารถตามไปเก็บเพิ่มเติมได้ ทำให้ข้อมูลของตัวอย่างไม่สมบูรณ์อาจจะทำให้เกิดปัญหาในการวิเคราะห์ หากผู้วิจัยเลือกแก้ปัญหาโดยการตัดค่าสังเกตชุดนั้นทิ้งไป ในกรณีนี้อาจมีผลทำให้ขนาดตัวอย่างมีจำนวนน้อยลง และส่งผลให้ค่าพยากรณ์มีความคลาดเคลื่อนสูง และที่สำคัญยิ่งก็คือทำให้สูญเสียรายละเอียดบางอย่างไป ซึ่งอาจมีผลกระทบต่อผลสรุปของการวิเคราะห์นั้น ๆ ได้

โดยทั่วไปการวิเคราะห์นั้น วิธีที่นิยมใช้ในการประมาณพารามิเตอร์หรือสัมประสิทธิ์การถดถอย (Regression Coefficient) คือวิธีกำลังสองน้อยที่สุด (Least Squares Method) ซึ่งเป็นตัวประมาณที่ไม่เอนเอียง (Unbiased) และมีความแปรปรวนต่ำสุด (Minimum Variance) แต่เมื่อค่าสังเกตสูญหายไปจะไม่สามารถประมาณได้ด้วยวิธีดังกล่าว วิธีการแก้ปัญหาทางหนึ่งก็คือการตัดค่าสังเกตชุดนั้นทิ้งไป แต่การแก้ปัญหาด้วยวิธีนี้จะมีผลทำให้จำนวนค่าสังเกตน้อยลงและสูญเสียรายละเอียดบางอย่างไป วิธีการแก้ปัญหาอีกทางหนึ่งคือทำการประมาณค่าสังเกตที่สูญหายด้วยวิธีการต่าง ๆ ก่อนที่จะใช้วิธีกำลังสองน้อยที่สุด

ในการวิจัยครั้งนี้ ผู้วิจัยได้ทำการศึกษาวิธีการหลายวิธีในการประมาณค่าที่สูญหายของตัวแปรตามในการวิเคราะห์การถดถอยพหุคูณ และใช้วิธีกำลังสองน้อยที่สุดหาสัมประสิทธิ์การถดถอยเพื่อหาสมการถดถอยเชิงเส้นในการพยากรณ์ วิธีการประมาณค่าสูญหาย 4 วิธีที่ผู้วิจัยทำการศึกษาเปรียบเทียบคือ

1. วิธีสูญหาย (loss method)
2. วิธีค่าเฉลี่ย (mean method)
3. วิธีสมการถดถอย (regression method)
4. วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่า (วิธีเอ็มไอ) (multiple imputation method: mi)

วัตถุประสงค์

1. เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตาม 4 วิธี คือ วิธีสูญหาย วิธีค่าเฉลี่ย วิธีสมการถดถอย และวิธีเอ็มไอ ด้วยวิธีการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง

2. เพื่อเลือกวิธีการประมาณค่าสูญหายที่ดีที่สุดในแต่ละสถานการณ์

ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลที่ได้จากการจำลองสถานการณ์โดยใช้วิธีมอนติคาร์โล (Monte Carlo Method) เขียนด้วยโปรแกรม SAS (Statistical Analysis System) โดยทำการจำลองข้อมูลซ้ำ 5,000 ครั้งในแต่ละสถานการณ์

2. ข้อมูลตัวแปรอิสระมี 3 ตัวแปรคือ X_{1j} X_{2j} และ X_{3j} มีการแจกแจงแบบปกติหลายตัวแปร $N_3(\boldsymbol{\mu},\boldsymbol{\Sigma})$ ด้วยค่าเฉลี่ย $\boldsymbol{\mu} = \mathbf{0}$ และเมทริกซ์ความแปรปรวนร่วม (covariance matrix) $\boldsymbol{\Sigma}$ ซึ่ง

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

เมื่อ σ_{ij} เป็นความแปรปรวนร่วมระหว่างตัวแปร X_i และ X_j

σ_{ii} เป็นความแปรปรวนของตัวแปร X_i

โดยที่ $\sigma_{ij} = \rho_{ij}\left(\sigma_i \sigma_j\right)$; i, j = 1,2,3 (วรรัตน์, 2547)

และ ρ_{ij} เป็นสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร X_i และ X_j

3. กำหนดระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (ρ) 3 ระดับ คือระดับต่ำ (.20) ระดับปานกลาง (.50) และ ระดับสูง (.70)

4. ค่าความคลาดเคลื่อนมีการแจกแจงปกติด้วยค่าเฉลี่ย 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 5 และ 15

5. เปอร์เซ็นต์การสูญหายของตัวแปรตาม 5 10 20 และ 30

6. ขนาดตัวอย่าง 50 70 100 และ 200

7. กำหนดการสูญหายของข้อมูลเฉพาะข้อมูลของตัวแปรตามและมีการสูญหายอย่างสุ่ม

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สามารถแก้ปัญหาข้อมูลที่สูญหายในการวิเคราะห์การถดถอยพหุคูณได้อย่างถูกต้องเหมาะสม

2. เป็นแนวทางสำหรับผู้วิจัยในการเลือกวิธีการจัดการข้อมูลที่สูญหาย ความสัมพันธ์ระหว่างตัวแปร และจำนวนข้อมูลสูญหายให้เหมาะสมกับสภาพของการวิจัย

การตรวจเอกสาร

การตรวจเอกสารจะแยกออกเป็น 2 ส่วน คือ ส่วนแรกกล่าวถึงผลงานวิจัยที่เกี่ยวข้อง และส่วนที่ 2 กล่าวถึงวิธีการทางสถิติที่ใช้ในการวิจัย

ผลงานวิจัยที่เกี่ยวข้อง

ชะไมพร (2522) ศึกษาเกี่ยวกับวิธีการประมาณค่าข้อมูลสูญหายในการวิเคราะห์การถดถอยวิธีต่าง ๆ 6 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีอันดับศูนย์หรือวิธีใช้ค่าเฉลี่ยจากกลุ่มตัวอย่าง วิธีอันดับศูนย์ดัดแปลง วิธีถดถอยอันดับหนึ่งหรือวิธีสมการถดถอย วิธีถดถอยสองขั้น และวิธีผสมหรือวิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยจากกลุ่มตัวอย่างและสมการถดถอย ศึกษาจากข้อมูลทุติยภูมิ โดยนำข้อมูลที่สมบูรณ์มาจัดกระทำให้สูญหายแบบสุ่ม เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าสัมประสิทธิ์การตัดสินใจ (R^2) ใช้เป็นดัชนีในการตัดสินใจวิธีประมาณค่าใดสามารถประมาณค่าได้ใกล้เคียงกับค่าที่สูญหายมากกว่ากัน ผลการศึกษาพบว่า วิธีที่ให้ค่า R^2 สูงกว่าวิธีอื่น ๆ มีอยู่ 3 วิธี เรียงตามลำดับจากมากไปน้อยคือ วิธีถดถอยสองขั้น วิธีถดถอยอันดับหนึ่งหรือวิธีใช้สมการถดถอย และวิธีกำลังสองน้อยที่สุด

ชุติมา (2533) ศึกษาเปรียบเทียบการประมาณข้อมูลสูญหายในการวิเคราะห์การถดถอยเชิงซ้อน 4 วิธี คือ วิธีสมการถดถอย วิธีเม็กซิมั่มไลลิฮูด วิธีค่าเฉลี่ย และวิธีค่ามัธยฐาน จากกลุ่มตัวอย่าง 30 70 และ 100 การกระจายข้อมูล 3 ระดับ โดยใช้ C.V. เป็นตัวกำหนด คือ .05 .20 และ 1.00 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 เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของสมการถดถอยของวิธีที่ไม่มีข้อมูลสูญหาย ผลการศึกษาพบว่า วิธีการประมาณข้อมูลสูญหายในการวิเคราะห์การถดถอยเชิงซ้อนทั้ง 4 วิธี ให้ผลต่างกันตามสถานการณ์ต่าง ๆ ซึ่งโดยส่วนใหญ่วิธีค่าเฉลี่ยให้ผลดีที่สุด ยกเว้นเมื่อมีขนาดตัวอย่างน้อยและจำนวนตัวแปรอิสระมาก วิธีสมการถดถอยจะให้ผลดีที่สุด แต่ถ้าตัวอย่างมีขนาดใหญ่และจำนวนตัวแปรอิสระมีน้อย การตัดชุดของข้อมูลสูญหายทิ้งจะไม่มีผลกระทบต่อการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด

เชาว์ (2547) ศึกษาการพัฒนาวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอี และตรวจสอบความแม่นยำ และอำนาจการทดสอบที่ได้จากวิธีการจัดการข้อมูลสูญหายแบบอีพีเอสเอสอีกับแบบอีเอ็มและแบบลิสท์ไวส์ ตามวิธีการสุ่มตัวอย่างแบบแบ่งชั้น แบบกลุ่ม และแบบหลายขั้นตอน ที่ความสัมพันธ์ระหว่างตัวแปรระดับต่ำ (r = .30) ปานกลาง (r = .50) และสูง (r = .70) และจำนวนข้อมูลสูญหาย 5% 10% 20% และ 30% และศึกษาปฏิสัมพันธ์ระหว่างวิธีการสุ่มตัวอย่าง วิธีการประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย และความสัมพันธ์ระหว่างตัวแปร ที่มีต่อความแม่นยำของค่าเฉลี่ยเลขคณิต ความแปรปรวน และสัมประสิทธิ์สหสัมพันธ์ ข้อมูลที่ใช้ศึกษามีลักษณะการแจกแจงแบบปกติสองตัวแปร และใช้วิธีมอนติคาร์โล จำลองการทดลองด้วยเครื่องคอมพิวเตอร์ ผลการศึกษาพบว่า วิธีการประมาณค่าสูญหายโดยการแทนค่าแบบอีพีเอสเอสอีได้ค่าความแม่นยำของค่าเฉลี่ยเลขคณิต ไม่แตกต่างจากวิธีอีเอ็ม อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 วิธีประมาณค่าสูญหายโดยการตัดออกแบบลิสท์ไวส์ ได้ค่าความแม่นยำของความแปรปรวนแตกต่างจากวิธีอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05 และวิธีการประมาณค่าสูญหายแบบลิสท์ไวส์ ได้ค่าความแม่นยำของสัมประสิทธิ์สหสัมพันธ์แตกต่างจากวิธีการประมาณค่าสูญหายแบบอื่น ๆ อย่างมีนัยสำคัญทางสถิติที่ระดับ .05

ถวัลย์ (2531) ศึกษาเปรียบเทียบความแม่นยำของวิธีประมาณค่าข้อมูลสูญหาย 3 วิธี คือ วิธีค่าเฉลี่ย วิธีสมการถดถอย และ วิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยและสมการถดถอย จากกลุ่มตัวอย่างขนาด 5 10 และ 15 กำหนดจำนวนข้อมูลที่สูญหายครั้งละ 1 ค่า และ 2 ค่า โดยใช้ข้อมูลที่มีลักษณะการแจกแจงแบบปกติและในกรณีที่ใช้สมการถดถอยช่วยในการประมาณค่านั้นได้กำหนดสัมประสิทธิ์สหสัมพันธ์ของตัวแปรเกณฑ์กับตัวแปรพยากรณ์เท่ากับ .20 .40 และ .60 ทำการศึกษาด้วยวิธีมอนติคาร์โล โดยจำลองการทดลองด้วยคอมพิวเตอร์ ผลการศึกษาพบว่า วิธีใช้ค่าเฉลี่ยระหว่างค่าเฉลี่ยและสมการถดถอยให้ผลการประมาณค่าดีที่สุดเมื่อกลุ่มตัวอย่างเป็น 5 ในทุกกรณี และเมื่อกลุ่มตัวอย่างเพิ่มขึ้นเป็น 10 วิธีนี้จะประมาณค่าได้ดีเฉพาะข้อมูลสูญหายครั้งละ 2 ค่าที่มีค่าสัมประสิทธิ์สหสัมพันธ์เป็น 0.2 หรือ 0.4 เท่านั้น วิธีสมการถดถอยให้ผลการประมาณค่าดีที่สุดแทบทุกกรณีเมื่อขนาดตัวอย่างเป็น 10 หรือ 15 โดยที่จำนวนข้อมูลที่สูญหายครั้งละ 1 ค่า แต่ในจำนวนข้อมูลที่สูญหายเป็นครั้งละ 2 ค่า วิธีนี้จะประมาณได้ดีเฉพาะที่ตัวแปรเกณฑ์และตัวแปรพยากรณ์มีความสัมพันธ์กันสูงระดับ 0.6 ส่วนวิธีค่าเฉลี่ยให้ผลการประมาณค่าไม่ดีเท่ากับสองวิธีดังกล่าวในทุกกรณี

บวรวรรณ (2543) ศึกษาการประยุกต์ใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในข้อมูลทุติยภูมิผู้ขับขี่จักรยานยนต์ที่บาดเจ็บเข้ารับการรักษาที่แผนกฉุกเฉิน โรงพยาบาลราชวิถี จำนวน 2,668 ราย ตั้งแต่วันที่ 1 มกราคม – 31 ธันวาคม 2538 โดยใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าแบบสุ่มอย่างง่าย (Multiple Hot – Deck Imputation) ซึ่งเป็นวิธีการใส่ค่าแทนข้อมูลที่สูญหายอย่างง่าย โดยการสุ่มเลือกค่าที่จะนำไปใส่แทนค่าที่สูญหายซึ่งได้จากค่าที่มีคำตอบจากหน่วยตัวอย่างซึ่งจับคู่กันกับตัวแปรที่มีค่าสังเกต โดยจะทำการสุ่มค่าที่จะนำไปแทนข้อมูลที่สูญหายของหน่วยตัวอย่างนั้นหลายครั้ง จำนวนค่าที่ใส่แทนค่าสูญหายแต่ละค่า M = 3 และ 5 กับการใส่ค่าเพียงค่าเดียว M = 1 โดยใช้ข้อตกลงเบื้องต้นว่าข้อมูลมีการสูญหายแบบเชิงสุ่ม (MAR) ผลการศึกษาพบว่า ค่าความคลาดเคลื่อนของสัมประสิทธิ์การถดถอยแบบลอจิสติคเมื่อใช้ M =3 และ 5 มีค่ามากกว่าการใส่ค่าแทนข้อมูลที่สูญหายเพียงค่าเดียว M = 1

พรศิริ (2529) ศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลที่สูญหายในการวิเคราะห์ตัวแปรพหุคูณ 4 วิธี คือ วิธีค่าเฉลี่ย วิธีวิเคราะห์การถดถอยพหุคูณเชิงเส้น วิธีวิเคราะห์ความถดถอยเชิงเส้นดัดแปลง และวิธีวิเคราะห์ส่วนประกอบหลัก จากกลุ่มตัวอย่าง 30 50 70 100 และ 200 จำนวนตัวแปรเท่ากับ 3 5 7 และ 10 ตัวแปร ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรเท่ากับ .10 .20 … .90 และกำหนดสัดส่วนข้อมูลที่สูญหายของแต่ละตัวแปรมีค่าเท่ากับ 10% ทำการศึกษาโดยใช้วิธีมอนติคาร์โล เกณฑ์ที่ใช้ในการเปรียบเทียบคือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (MSE) ผลการศึกษาพบว่า วิธีประมาณค่าข้อมูลสูญหายในการวิเคราะห์ตัวแปรพหุคูณทั้ง 4 วิธี ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองไม่แตกต่างกันอย่างมีนัยสำคัญที่ระดับ .05 ไม่ว่าจะเป็นสถานการณ์ใดก็ตามที่มีข้อมูลสูญหายเกิดขึ้น

วารุณี (2538) ศึกษาการพยากรณ์ด้วยวิธีการถดถอยเชิงเส้นพหุคูณ เมื่อตัวแปรตามมีค่าสูญหาย วิธีที่ใช้ประมาณค่าตัวแปรตามเมื่อมีข้อมูลสูญหาย คือ วิธีค่าเฉลี่ย วิธีสมการถดถอย วิธีอีเอ็ม (EM algorithm) และวิธีของฮันท์ (Hunt’s Method) ใช้กลุ่มตัวอย่างขนาด 10 20 30 50 70 ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 5 10 15 20 และ 25 สัดส่วนการสูญหายของตัวแปรตาม 10% 20% 30% 40% 50% 60% และ 70% ทำการศึกษาด้วยวิธีมอนติคาร์โล และหารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ ผลการศึกษาพบว่า วิธีการของฮันท์เป็นวิธีการที่ดี เมื่อกลุ่มตัวอย่างมีขนาดเล็ก ความคลาดเคลื่อนน้อย และสัดส่วนการสูญหายมาก แต่ถ้าความคลาดเคลื่อนสูง วิธีค่าเฉลี่ยจะเป็นวิธีที่ดีในทุกสัดส่วนการสูญหายของตัวแปรตาม ส่วนในสถานการณ์ที่ขนาดตัวอย่างมีขนาดใหญ่ วิธีสูญหายจะเหมาะสมเกือบทุกกรณี

Crawford *et al.* (1995) ศึกษาเปรียบเทียบวิธีการจัดการข้อมูลที่สูญหาย 4 วิธี คือ วิธีการใส่ค่าเพียงค่าเดียว วิธีเอ็มไอ วิธีค่าเฉลี่ย และวิธีสูญหาย ข้อมูลที่ใช้ในการศึกษามาจากโครงการสุขภาพคนชราของมลรัฐแมชซะ- ซูเซ็ท วัตถุประสงค์เพื่อเปรียบเทียบวิธีการจัดการข้อมูลสูญหายทั้ง 4 วิธี กระทำภายใต้แบบแผนการสูญหายต่างกัน ผลการศึกษาพบว่า วิธีสูญหาย และวิธีค่าเฉลี่ย จะให้ค่าประมาณที่มีความเอนเอียง ยกเว้นเฉพาะกรณีการสูญหายของผลลัพธ์ไม่มีความสัมพันธ์กับตัวแปรร่วม วิธีค่าเฉลี่ย และ วิธีการใส่ค่าเพียงค่าเดียวจะให้ค่าความแปรปรวนต่ำกว่าความเป็นจริง ส่วน วิธีเอ็มไอ จะให้ค่าเฉลี่ยของประชากรไม่มีความเอนเอียง

Enders (2001) ศึกษาความสามารถของการประมาณค่าแบบเอฟไอเอ็มแอล (Full information maximum likelihood: FIML) ในการวิเคราะห์การถดถอยพหุคูณเมื่อมีข้อมูลสูญหาย ตัวแปรที่ศึกษามี 4 ตัวแปรคือ วิธีการประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย ขนาดของกลุ่มตัวอย่าง และขนาดของความสัมพันธ์ระหว่างตัวแปร ใช้วิธีมอนติคาร์โล จำลองข้อมูลที่มีรูปแบบของการสูญหายแตกต่างกัน 3 แบบคือ การสูญหายแบบสุ่มสมบูรณ์ การสูญหายแบบสุ่ม และการสูญหายแบบไม่สุ่ม ตัวแปรตามที่ศึกษาคือ สัมประสิทธิ์การถดถอย สัมประสิทธิ์การตัดสินใจ และประสิทธิภาพ นอกจากนั้นยังได้ศึกษาปฏิสัมพันธ์ของ วิธีประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย และขนาดของความสัมพันธ์ระหว่างตัวแปร ผลการวิจัยพบว่าการประมาณค่าข้อมูลสูญหายแบบเอฟไอเอ็มแอลดีกว่าการตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ แพร์ไวส์ การแทนข้อมูลด้วยวิธีค่าเฉลี่ย และการประมาณค่าข้อมูลสูญหายแบบเอฟไอเอ็มแอลมีความเอนเอียงน้อย และมีปฏิสัมพันธ์ระหว่างวิธีประมาณค่าสูญหาย จำนวนข้อมูลสูญหาย และขนาดความสัมพันธ์ระหว่างตัวแปร

Heitjan and Little (1991) ศึกษาการประยุกต์ใช้วิธีเอ็มไอในฐานข้อมูลการบาดเจ็บและตายด้วยอุบัติเหตุของโครงการระบบรายงานการตายด้วยอุบัติเหตุ (The Fatal Accident Reporting System: FARS) ปี 1985 โดยใช้วิธีประมาณค่าสูญหายด้วยวิธีการจับคู่ค่าเฉลี่ยทำนาย ซึ่งเป็นวิธีการใช้ค่าในตัวแปรที่ไม่มีคำตอบ จับคู่กับค่าของตัวแปรที่มีคำตอบของเมทริกซ์บางตัวแปร ได้ใช้จำนวนค่าที่ใส่แทนข้อมูลที่สูญหายแต่ละค่า M = 1 M = 3 และ M = 5 ในการศึกษานี้ได้ทำการจำลองวิธีการศึกษาโดยใช้วิธีการทดลองด้วยวิธีมอนติคาร์โลในการตรวจสอบคุณสมบัติด้านการแจกแจงของวิธีการ โดยจำลองข้อมูลให้มีลักษณะคล้ายกับข้อมูลของโครงการ FARS แล้วทำการศึกษาทดลองโดยวิธีการสุ่มซ้ำ และไม่ใช้วิธีการสุ่มซ้ำ ผลการศึกษาพบวิธีเอ็มไอจะให้ผลดีกว่าการใส่ค่าสูญหายเพียงค่าเดียว

Huang and Carriere (2006) ศึกษาเปรียบเทียบวิธีการจัดการข้อมูลสูญหายในข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก โดยมีวัตถุประสงค์ เพื่อ พัฒนาวิธีเอ็มไอ ของข้อมูลแบบวัดซ้ำในกลุ่มตัวอย่างขนาดเล็ก กระทำภายใต้ข้อตกลงเบื้องต้นข้อมูลมีการแจกแจงปกติหลายตัวแปร (multivariate normal distribution) ทำการศึกษาโดยใช้สถานการณ์จำลอง (simulation) และเปรียบเทียบวิธีการจัดการข้อมูลที่สูญหายระหว่างวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยใช้วิธีเม็กซิมั่มไลลิฮูด กับวิธีใส่ค่าแทนข้อมูลที่สูญหายด้วย วิธีเอ็มไอ (ใช้จำนวนค่าที่ใส่แทนข้อมูลที่สูญหาย M = 5) เพื่อทดสอบสมมติฐานของอิทธิพลของทรีทเมนต์ และอิทธิพลของความคลาดเคลื่อนของทรีทเมนต์ ผลการศึกษาปรากฏว่า วิธีการใส่ค่าแทนข้อมูลที่สูญหายด้วยวิธีเอ็มไอจะให้ผลการทดสอบไม่แตกต่างจากวิธีไม่ใส่ค่าแทนข้อมูลที่สูญหายโดยวิธีเม็กซิมั่มไลลิฮูด

Viragoontavan (2000) ศึกษาเปรียบเทียบประสิทธิภาพสัมพัทธ์ (relative effectiveness) ของวิธีการจัดการข้อมูลสูญหาย 6 วิธี คือ การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์ การแทนค่าด้วยค่าเฉลี่ยของกลุ่ม การแทนค่าด้วยวิธีการถดถอย การแทนค่าด้วยวิธีฮอทเดค (hot – deck imputation) การแทนค่าวิธีเอ็มไอด้วยโปรแกรม SOLAS และการแทนค่าวิธีเอ็มไอด้วยโปรแกรม NORM ข้อมูลได้จากสถานการณ์จำลองตามเงื่อนไขต่อไปนี้ ความสัมพันธ์ระหว่างตัวแปร ต่ำ ปานกลาง และ สูง ขนาดกลุ่มตัวอย่าง 100 200 และ 500 และสัดส่วนข้อมูลสูญหาย .05 .10 และ .20 จัดกระทำสมบูรณ์ให้เป็นข้อมูลสูญหายแบบสุ่มตามสัดส่วนที่กำหนดแล้วจัดกระทำข้อมูลสูญหายด้วยวิธีการทั้ง 6 วิธี นำข้อมูลไปวิเคราะห์เชิงจำแนกแล้วประเมินประสิทธิภาพสัมพัทธ์โดยพิจารณาฮิทเรท (hit rate) และอำนาจการจำแนก (discriminating power) ของสมการจำแนก ผลการวิจัยพบว่า วิธีการจัดการข้อมูลสูญหายแบบเอ็มไอทั้งสองวิธีมีประสิทธิภาพมากที่สุด การตัดข้อมูลสูญหายออกแบบลิสท์ไวส์มีประสิทธิภาพต่ำสุด วิธีการจัดการข้อมูลสูญหายทั้งหมดเมื่อข้อมูลมีความสัมพันธ์กันน้อยประมาณค่าได้ถูกต้องมากกว่าเมื่อข้อมูลมีความสัมพันธ์กันสูง การประมาณค่าฮิทเรทและอำนาจการจำแนกได้ถูกต้องมากขึ้นเมื่อจำนวนกลุ่มตัวอย่างมากขึ้นและจำนวนข้อมูลสูญหายน้อย

วิธีการทางสถิติ

การแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution)

ให้เวกเตอร์ของตัวแปร $\mathbf{X}$ หรือ $\mathbf{X}' = (X_1, X_2, ..., X_p)$ มีการแจกแจงแบบปกติหลายตัวแปรที่มีเวกเตอร์ค่าเฉลี่ยเป็น $\boldsymbol{\mu}$ และ เมตริกซ์ความแปรปรวนร่วม (Covariance matrix) คือ $\boldsymbol{\Sigma}$ โดยที่ฟังก์ชันความน่าจะเป็นของเวกเตอร์ตัวแปร $\mathbf{X}$ คือ

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

โดยที่ P เป็นจำนวนตัวแปร หรือเขียนย่อ ๆ ว่า $\mathbf{X} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

คุณสมบัติของเวกเตอร์ตัวแปรที่มีการแจกแจงแบบปกติหลายตัวแปร

ถ้าเวกเตอร์ $\mathbf{X} = (X_1, X_2, ..., X_p)'$ มีการแจกแจงแบบปกติ นั่นคือ

$$\mathbf{X} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

คุณสมบัติของเวกเตอร์ $\mathbf{X}$ เป็นดังนี้ (กัลยา, 2548)

1. ถ้า $\mathbf{a} = (a_1, ...a_p)'$ เป็นเวกเตอร์ของค่าคงที่ $\mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + ... + a_pX_p$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\mathbf{a}'\boldsymbol{\mu}$ และความแปรปรวน $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ หรือ กล่าวได้ว่า

ถ้า $\mathbf{X} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ แล้วจะได้ว่า $\mathbf{a}'\mathbf{X} \sim N_P(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$

โดยที่ $E(\mathbf{a}'\mathbf{X}) = \mathbf{a}'E(\mathbf{X}) = \mathbf{a}'\boldsymbol{\mu}$

$$Var(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \boldsymbol{\sigma}^2_{\mathbf{a}'\mathbf{x}}$$

2. ถ้า $\mathbf{A}$ เป็นเมทริกซ์ของค่าคงที่ ซึ่งมีขนาด $q \times p$ ซึ่งมี rank เท่ากับ q โดยที่ $q \le p$ จะได้ว่า $\mathbf{AX}$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\mathbf{A\mu}$ และความแปรปรวน $\mathbf{A\Sigma A'}$ เขียนแทนด้วย

$$\mathbf{AX} \sim N_q(\mathbf{A\mu}, \mathbf{A\Sigma A'})$$

โดยที่ $E(\mathbf{AX}) = \mathbf{A\mu}$

$$\mathrm{Cov}(\mathbf{AX}) = \mathbf{A\Sigma A'}$$

3. ตัวแปรปกติมาตรฐาน

กำหนดให้ $\mathbf{Z} = \mathbf{\Sigma}^{\frac{-1}{2}}(\mathbf{X} - \boldsymbol{\mu})$

โดยที่ $\mathbf{\Sigma}^{\frac{1}{2}}$ เป็นค่ารากที่สองของเมทริกซ์สมมาตร (Symmetric square root matrix) ที่ทำให้ $\mathbf{\Sigma} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}$ จะได้ว่า เวกเตอร์ตัวแปรปกติมาตรฐาน $\mathbf{Z}$ จะมีการแจกแจงปกติที่มีเวกเตอร์ค่าเฉลี่ยเป็นศูนย์ 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 ของเวกเตอร์ตัวแปรปกติมาตรฐาน $\mathbf{Z}$ จะได้ว่าตัวแปร Z_i เป็นตัวแปรที่มีการแจกแจงแบบปกติมาตรฐานที่มีค่าเฉลี่ยเป็นศูนย์ ($\boldsymbol{\mu}_{Z_i} = 0$) และความแปรปรวนเป็นหนึ่ง ($\boldsymbol{\sigma}^2_{Z_i} = 1$) นั่นคือ $Z_i \sim N(0,1)$

4. การแจกแจงแบบไคกำลังสอง

ถ้า $\mathbf{Z}$ เป็นเวกเตอร์ตัวแปรปกติมาตรฐานที่เป็นอิสระจากกัน จะได้ว่า $\sum_{i=1}^{p} \mathbf{Z}_i^2 = \mathbf{Z'Z} = (\mathbf{X} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ จะมีการแจกแจงแบบไคกำลังสองที่องศาอิสระ p ($\boldsymbol{\chi}^2_p$)

5. การแจกแจงแบบปกติของเซตย่อยสำหรับตัวแปรพหุคูณ

5.1 ถ้า $\mathbf{X} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 $\mathbf{X}$

$$\text{ดังนั้น} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{และ} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

โดยที่ $\mathbf{X}_1$ และ $\boldsymbol{\mu}_1$ เป็นเวกเตอร์ขนาด r x 1 ส่วน $\boldsymbol{\Sigma}_{11}$ เป็นเมทริกซ์ขนาด r x r จะได้ว่า $\mathbf{X}_1$ มีการแจกแจงแบบ $N_r(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ และ $\mathbf{X}_2$ และ $\boldsymbol{\mu}_2$ เป็นเวกเตอร์ขนาด p-r และ $\boldsymbol{\Sigma}_{22}$ เป็นเมทริกซ์ขนาด (p-r) x (p-r) ซึ่งจะได้ว่า $\mathbf{X}_2$ มีการแจกแจงแบบ $N_{P-r}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$

ตัวอย่างเช่น กำหนดเวกเตอร์ $\mathbf{X}_3' = (X_1, X_4, X_9, X_{10})$ เป็นเซตย่อยของเวกเตอร์ $\mathbf{X}$ โดย $\mathbf{X} \sim N_P(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ จะได้ว่า $\mathbf{X_3}$ จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\boldsymbol{\mu}_3$ และความแปรปรวนร่วม $\boldsymbol{\Sigma}_{33}$ หรือ $\mathbf{X_3} \sim N_4(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_{33})$

5.2 ถ้า X_i เป็นตัวแปรที่ i ในเวกเตอร์ นั่นคือ X_i เป็นตัวแปรเดี่ยวที่มีการแจกแจงแบบปกติมีค่าเฉลี่ย μ_i ความแปรปรวน σ_i^2 หรือเขียนแทนด้วย $X_i \sim N(\mu_i, \sigma_i^2)$ แล้วไม่จำเป็นที่เวกเตอร์ $\mathbf{X}$ จะต้องมีการแจกแจงแบบปกติ

6. ความเป็นอิสระระหว่างเวกเตอร์ของตัวแปรสุ่ม

กำหนดให้ $\mathbf{X}$ และ $\mathbf{Y}$ เป็นเวกเตอร์ย่อย 2 เวกเตอร์ จะได้ว่า

$$E\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \qquad Cov\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} \sum_{XX} & \sum_{XY} \\ \sum_{YX} & \sum_{YY} \end{bmatrix}$$

6.1 $\mathbf{X}$ และ $\mathbf{Y}$ จะเป็นอิสระกัน ถ้า $\sum_{XY} = 0$

6.2 ตัวแปร $\mathbf{X_i}$ และ $\mathbf{X_j}$ จะเป็นอิสระกัน ถ้า $\sigma_{ij} = 0$

7. การแจกแจงแบบมีเงื่อนไข

ถ้า $\mathbf{X}$ และ $\mathbf{Y}$ ไม่เป็นอิสระกัน จะทำให้ $\sum_{XY} \neq 0$ การแจกแจงแบบมีเงื่อนไขของ $\mathbf{X}$ โดยกำหนด y คือ $f(x|y)$ จะมีการแจกแจงแบบปกติของหลายตัวแปรที่มีค่าคาดหวังและค่าแปรปรวนร่วมเป็น

$$E(\mathbf{X}|\mathbf{y}) = \boldsymbol{\mu}_x + \sum_{XY}\sum_{YY}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_Y)$$

$$\mathrm{Cov}(\mathbf{X}|\mathbf{y}) = \sum_{XX} - \sum_{XY}\sum_{YY}^{-1}\sum_{YX}$$

ค่า $E(\mathbf{X}|\mathbf{y})$ เป็นฟังก์ชันเชิงเส้นของเวกเตอร์ $\mathbf{Y}$ ในขณะที่ $\mathrm{Cov}(\mathbf{X}|\mathbf{y})$ ไม่ขึ้นกับ $\mathbf{Y}$

8. การแจกแจงของผลบวกของเวกเตอร์ 2 เวกเตอร์

ถ้า $\mathbf{X}$ และ $\mathbf{Y}$ เป็นเวกเตอร์ของตัวแปรที่มีขนาดเท่ากันคือ $p \times 1$ และเป็นอิสระต่อกันจะได้ว่า

$$\mathbf{X} + \mathbf{Y} \sim N_p\left(\boldsymbol{\mu}_\mathbf{X} + \boldsymbol{\mu}_Y, \sum_{\mathbf{XX}} + \sum_{\mathbf{YY}}\right)$$

$$\mathbf{X} - \mathbf{Y} \sim N_p\left(\boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}_Y, \sum_{\mathbf{XX}} + \sum_{\mathbf{YY}}\right)$$

ถ้าเวกเตอร์ $\mathbf{X}$ และ $\mathbf{Y}$ มีการแจกแจงแบบปกติหลายตัวแปร จะทำให้เวกเตอร์ $\mathbf{X} \pm \mathbf{Y}$ จะมีการแจกแจงแบบปกติหลายตัวแปรด้วย

วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares: OLS)

วิธีการหาสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด คือหาค่าประมาณของพารามิเตอร์ที่ทำให้ผลบวกกำลังสองของผลต่างระหว่างค่าสังเกตกับค่าคาดหวังของตัวแปรมีค่าต่ำที่สุด

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

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

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

เมื่อ $\mathbf{Y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n\times 1$

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

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด $(p+1)\times 1$

n เป็นขนาดตัวอย่าง

p เป็นจำนวนตัวแปรอิสระ

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อน ขนาด $n\times 1$

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

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

การหาค่ากำลังสองน้อยที่สุดของผลบวกกำลังสองของความคลาดเคลื่อน ทำได้โดยหาอนุพันธ์เทียบกับ $\hat{\boldsymbol{\beta}}$ แล้วกำหนดให้เท่ากับศูนย์

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

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

$$\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{Y}$$

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

ดังนั้นสมการถดถอยที่ใช้พยากรณ์คือ

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

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

ทฤษฎีที่เกี่ยวข้องกับวิธีเอ็มไอ

ทฤษฎีของเบย์

ให้เวกเตอร์ของตัวแปรสุ่ม $\mathbf{Y}$ หรือ $\mathbf{Y}' = \left(Y_1,...,Y_n\right)$ มีการแจกแจงความน่าจะเป็น $p\left(y|\theta\right)$ ขึ้นอยู่กับพารามิเตอร์ k ค่า $\boldsymbol{\theta}' = \left(\theta_1,...,\theta_k\right)$ ถ้า $\boldsymbol{\theta}$ มีการแจกแจงความน่าจะเป็น $p\left(\theta\right)$ แล้ว

$$p\left(y|\theta\right)p\left(\theta\right) = p\left(y,\theta\right) = p\left(\theta|y\right)p\left(y\right)$$

การแจกแจงแบบมีเงื่อนไขของ $\boldsymbol{\theta}$ เมื่อกำหนด y มีค่าเท่ากับ

$$p\left(\theta|y\right) = \frac{p\left(y|\theta\right)p\left(\theta\right)}{p\left(y\right)}$$

ซึ่ง

$$p\left(y\right) = Ep\left(y|\theta\right) = c^{-1} = \begin{cases} \int p\left(y|\theta\right)p\left(\theta\right)d\theta & ;\ \theta\ \text{ต่อเนื่อง} \\ \sum p\left(y|\theta\right)p\left(\theta\right) & ;\ \theta\ \text{ไม่ต่อเนื่อง} \end{cases}$$

ดังนั้นการแจกแจงแบบมีเงื่อนไขของ $\boldsymbol{\theta}$ เมื่อกำหนด y เขียนได้ใหม่เป็น

$$p\left(\theta|y\right) = cp\left(y|\theta\right)p\left(\theta\right)$$

โดยที่ $p\left(\theta\right)$ เรียกว่า การแจกแจงเริ่มต้นของ $\boldsymbol{\theta}$ และ $p\left(\theta|y\right)$ เรียกว่า การแจกแจงภายหลังของ $\boldsymbol{\theta}$ เมื่อกำหนดค่า y และ c เป็นค่าคงที่ (George, 1973)

ทฤษฎีของเบย์และฟังก์ชันไลลิฮูดส์

ฟังก์ชันไลลิฮูดส์ของ $\boldsymbol{\theta}$ เมื่อกำหนด y เขียนแทนด้วย $l(\theta|y)$ ดังนั้นเขียนได้คือ (Fisher, 1922)

$$p(\theta|y) = l(\theta|y)p(\theta)$$

ดังนั้นการแจกแจงภายหลังของ $\boldsymbol{\theta}$ เมื่อกำหนดค่า y จะแปรผันตามผลคูณของการแจกแจงเริ่มต้นกับฟังก์ชันไลลิฮูดส์ของ $\boldsymbol{\theta}$ เมื่อกำหนด y นั่นคือ

$$\text{posterior distribution} \propto \text{likelihood} \times \text{pior distribution}$$

ทฤษฎีการแจกแจงภายหลังของ σ^2

กำหนดให้ s^2 มีการแจกแจงแบบ $(\sigma^2/\nu)\chi^2_\nu$ และ การแจกแจงเริ่มต้นของ $\log\sigma$ มีการแจกแจงแบบยูนิฟอร์ม $[0,1]$ ดังนั้นการแจกแจงภายหลังของ σ^2 มีการแจกแจงแบบ $(\nu s^2)/\chi^2_\nu$

สำหรับรายละเอียดเกี่ยวกับวิธีประมาณค่าสูญหายของตัวแปรตาม ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณแต่ละวิธีเป็นดังนี้

วิธีสูญหาย (Loss Method)

วิธีการทำได้โดยการตัดชุดข้อมูลที่สูญหายออกไป หลังจากนั้นก็จะประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS ดังนี้

$$\hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$$

เมื่อ $\mathbf{X}^*$ และ $\mathbf{Y}^*$ เป็นชุดข้อมูลที่เหลืออยู่ทั้งของ X และ Y

ข้อดีของวิธีสูญหาย

Little and Rubin (1987) กล่าวไว้คือ “ความง่าย” เพราะว่าข้อมูลที่เหลืออยู่จะเป็นข้อมูลที่สมบูรณ์ สามารถใช้โปรแกรมทางสถิติทุกโปรแกรมวิเคราะห์ข้อมูลทางสถิติได้ นอกจากนี้ยังสามารถ “เปรียบเทียบ” ระหว่างตัวแปรได้เพราะว่าหน่วยตัวอย่างจะเท่ากันสำหรับทุกตัวแปรที่มีการคำนวณค่าสถิติ

ข้อเสียของวิธีสูญหาย

ทำให้เสียเวลาในการเก็บข้อมูลเบื้องต้น เพราะมีการตัดทิ้งไปภายหลังจากที่เก็บข้อมูลได้บางส่วน นอกจากนี้การตัดหน่วยตัวอย่างที่ไม่สมบูรณ์ออกไปทำให้ขนาดตัวอย่างลดลงก่อให้เกิดปัญหาเกี่ยวกับอำนาจการทดสอบ รวมทั้งปัญหาการอ้างอิงไปถึงประชากร

วิธีค่าเฉลี่ย (Mean Method)

วิธีการทำได้โดยการประมาณค่าสูญหายตัวแปรตามที่สูญหายโดยใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม นั่นคือ

$$\overline{Y}^* = \frac{\sum_{j=1}^{n_1} Y_j}{n_1}$$

เมื่อ $\overline{Y}^*$ เป็นค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม

n_1 เป็นจำนวนข้อมูลที่ไม่สูญหายของตัวแปรตาม

เมื่อแทนข้อมูลที่สูญหายด้วยค่าเฉลี่ยแล้ว จะทำการประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS

ข้อดีของวิธีค่าเฉลี่ย

เป็นวิธีที่ง่ายต่อการเข้าใจและคำนวณ ทำให้เป็นวิธีที่ดึงดูดความสนใจของนักวิจัย

ข้อเสียของวิธีค่าเฉลี่ย

ถ้าหากว่ามีจำนวนข้อมูลสูญหายมากเท่าไร ค่าความแปรปรวนของข้อมูลทั้งหมด (ค่าที่มีอยู่เดิมรวมทั้งค่าที่นำไปแทน) ก็ยิ่งจะลดลงมากเท่านั้น เพราะค่าที่เอาไปแทนค่าที่สูญหายเป็นค่าคงที่

วิธีสมการถดถอย (Regression Method)

วิธีนี้จะเป็นวิธีการประมาณค่าสูญหายของตัวแปรตาม มีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 ใช้ข้อมูล (X,Y) ที่เหลืออยู่ประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธี OLS ดังนี้

$$\hat{\boldsymbol{\beta}}^{*} = \left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^{*}$$

เมื่อ $\mathbf{X}^{*}$ และ $\mathbf{Y}^{*}$ เป็นชุดข้อมูลที่เหลืออยู่ทั้งของ X และ Y

ขั้นตอนที่ 2 นำค่าที่ได้จากขั้นตอนที่ 1 มาประมาณค่าสูญหายของตัวแปรตามโดยพิจารณาจากสมการถดถอยเชิงเส้นพหุคูณ

$$\hat{Y}_j = \mathbf{x}_j\hat{\boldsymbol{\beta}}^{*}$$

$$= \begin{bmatrix}1 & X_{1j} & X_{2j} & X_{3j}\end{bmatrix}\hat{\boldsymbol{\beta}}^{*}$$

เมื่อ $\hat{Y}_j$ เป็นค่าประมาณของค่าสูญหายตัวที่ j

X_{1j}, X_{2j}, X_{3j} เป็น ค่าสังเกตชุดที่ j ของตัวแปรอิสระตัวที่ 1 2 และ 3

ขั้นตอนที่ 3 นำค่า $\hat{Y}_j$ แทนในค่าสูญหายตัวที่ j ของตัวแปรตาม แล้วทำการประมาณสัมประสิทธิ์การถดถอย โดยวิธี OLS

ข้อดีของวิธีสมการถดถอย

ทำให้ความแปรปรวนของข้อมูลมีค่าความแปรปรวนมากกว่าค่าความแปรปรวนที่ได้จากวิธีค่าเฉลี่ย

ข้อเสียของวิธีสมการถดถอย

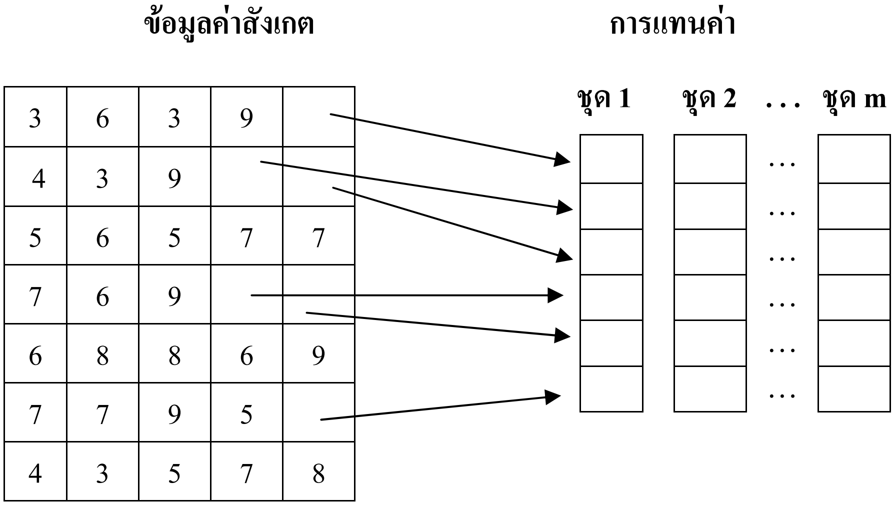
ทำให้ค่าที่เกี่ยวกับความสัมพันธ์ระหว่างตัวแปรบิดเบือนไป เพราะว่าสมการการประมาณค่าที่สูญหายไปนั้น สร้างจากความสัมพันธ์ระหว่างตัวแปร

วิธีเอ็มไอ (Multiple Imputation Method)

Rubin ได้เสนอเป็นคนแรกโดยกล่าวว่าวิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าเป็นเทคนิคในการแทนค่าสูญหายแต่ละค่าโดยใช้ค่าที่นำไปแทนตั้งแต่สองค่าขึ้นไป ให้ชื่อวิธีการนี้ว่าวิธีเอ็มไอ ซึ่งมีชื่อเต็ม ๆ ว่า multiple imputation ซึ่งการใส่ค่าสูญหายซ้ำหลายครั้งเป็นการสุ่มค่าสูญหายจากการแจกแจงภายหลังของค่าพารามิเตอร์ ประกอบด้วย 3 ขั้นตอนหลักคือ (Little and Rubin, 1987)

ขั้นตอนที่ 1 สร้างค่าสำหรับแทนค่าสูญหาย

เป็นการสร้างชุดข้อมูลสมบูรณ์ $Y^{(i)} = (Y_{obs}, Y^{(i)}_{miss})$ คือการรวมเซตข้อมูลค่าสังเกต Y_{obs} และเซตข้อมูลค่าสูญหาย $Y^{(i)}_{miss}$ จำนวน m ชุด เมื่อ i = 1,2, . . . ,m นิยมใช้ m = 3,5 (Huang and Carriere, 2006) โดยที่ m คือจำนวนครั้งของการแทนค่า และ 1< m < n จากภาพที่ 1 เป็นตัวอย่างการสร้างเซตข้อมูลสมบูรณ์ m ชุด ซึ่งค่าสูญหายแต่ละค่าจะถูกแทนด้วยจุดที่ชี้ไปยังเวกเตอร์ของการแทนค่า ซึ่งค่า m ค่า จะจัดเรียงเป็นลำดับ โดยที่ค่าแรกของเวกเตอร์เมื่อนำมาแทนค่าสูญหายก็จะให้ข้อมูลหนึ่งชุด เมื่อนำค่าที่สองของเวกเตอร์ไปใส่แทนค่าที่สูญหายก็จะให้ข้อมูลชุดที่สอง เป็นเช่นนี้ไปเรื่อย ๆ จนถึงข้อมูลชุดที่ m



ภาพที่ 1 แสดงตัวอย่างแบบแผนของวิธีเอ็มไอแทนข้อมูลสูญหาย เมื่อ m คือจำนวนของการแทนค่า

หมายเหตุ เมื่อช่องว่างของข้อมูลค่าสังเกต หมายถึงตำแหน่งที่มีข้อมูลสูญหาย

ขั้นตอนการสร้างค่าสำหรับแทนค่าสูญหายที่ใกล้เคียงกับการแจกแจงภายหลังของค่า Y_{miss} ภายใต้แบบจำลองของเบย์ มีขั้นตอนดังนี้

1) กำหนดตัวแบบทำนาย เป็นขั้นตอนการสร้างแบบจำลองการทำนายค่าของตัวแปรตาม **Y** ซึ่งขึ้นอยู่กับตัวแปรอิสระ **X** สำหรับตัวแบบการถดถอยเชิงเส้นทั่วไปในการพยากรณ์ค่าของตัวแปรตาม **Y** จากตัวแปรอิสระ **X** จากข้อกำหนดเบื้องต้น กำหนดให้ตัวแปรตาม **Y** มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ $x_i\boldsymbol{\beta}$ และความแปรปรวนเท่ากับ $\boldsymbol{\sigma}^2$ และกำหนดเวกเตอร์ของพารามิเตอร์ $\boldsymbol{\theta} = (\boldsymbol{\beta}, \log\boldsymbol{\sigma})$ โดยมีการแจกแจงเริ่มต้นของ $\boldsymbol{\theta}$ แปรผันตามค่าคงที่ (Rubin, 2004)

2) การประมาณค่า เป็นวิธีการประมาณค่าการแจกแจงภายหลังของค่าพารามิเตอร์ ซึ่งได้จากตัวแบบทำนายที่มีการแจกแจงเริ่มต้นของ $\boldsymbol{\theta}$ แปรผันตามค่าคงที่ ประกอบกับทฤษฎีการแจกแจงภายหลังของ σ^2 จะได้การแจกแจงภายหลังของ σ^2 มีการแจกแจงแบบ $\hat{\sigma}_1^2\left(n_1 - p\right)/\chi^2_{n_1-p}$

ซึ่ง
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n_1}\left(Y_i^* - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^2}{\left(n_1 - p\right)} \qquad \text{(Walczak and Massart, 2001)}$$

เมื่อ $\hat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$

$\mathbf{X}^*$ และ $\mathbf{Y}^*$ เป็นชุดข้อมูลที่เหลืออยู่ทั้งของ X และ Y

n_1 เป็นจำนวนข้อมูลที่ไม่สูญหาย

p เป็นจำนวนสัมประสิทธ์การถดถอย

3) การใส่ค่าสูญหาย เป็นการสุ่มค่าของข้อมูลจากการแจกแจงภายหลังของค่าสูญหาย โดยขั้นแรกสุ่มค่า 1 ค่า จากการแจกแจงภายหลังของ σ^2 แล้วนำมาคำนวณหาค่าสูญหาย Y จะทำให้ได้ค่าเพื่อนำไปแทนที่ค่าสูญหาย Y 1 ค่า หลังจากนั้นทำการสุ่มค่าใหม่สำหรับการแจกแจงภายหลังของ σ^2 แล้วนำมาคำนวณหาค่าสูญหาย Y ทำเช่นนี้ไปเรื่อย ๆ จนครบ m ครั้ง มีรายละเอียดดังนี้

3.1) สุ่มค่า k จากตัวแปรสุ่มที่มีการแจกแจงแบบไคสแควร์ด้วยองศาแห่งความอิสระเท่ากับ n_1-p แล้วหาการแจกแจงภายหลังของ σ^2 จากสูตร

$$\sigma_*^2 = \hat{\sigma}_1^2\left(n_1 - p\right)/k \sim \chi^2$$

3.2) สุ่มค่า Z_j จำนวนเท่ากับจำนวนสัมประสิทธิ์การถดถอย จากตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน แล้วประมาณค่าสัมประสิทธิ์การถดถอยตัวที่ j เมื่อ j= 1,…, p จากสูตร

$$\beta_*^{(j)} = \hat{\beta}^{*(j)} + \sigma_*\left(SE^{(j)}/\hat{\sigma}_1\right)Z_j$$

เมื่อ $SE^{(j)}$ เป็นความคลาดเคลื่อนมาตรฐานของสัมประสิทธิ์การถดถอยตัวที่ j จากข้อมูลที่เหลืออยู่ ซึ่งหาได้จากสูตร

$$SE^{(j)} = \sqrt{\hat{\sigma}_1^2 c_{jj}}$$

c_{jj} เป็นสมาชิกที่ j บนเส้นทแยงมุมหลักของเมทริกซ์ $\left(\mathbf{x}^{*\prime}\mathbf{x}^*\right)^{-1}$ โดยที่ (วิรัชช, 2545)

$$\left(\mathbf{x}^{*\prime}\mathbf{x}^*\right)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} & \cdots & C_{0k} \\ C_{10} & C_{11} & C_{12} & \cdots & C_{1k} \\ C_{20} & C_{21} & C_{22} & \cdots & C_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ C_{k0} & C_{k1} & C_{k2} & & C_{kk} \end{bmatrix}$$

เนื่องจากค่า $\hat{\beta}^{*(j)}$ เป็นค่าสถิติได้มาจากกลุ่มตัวอย่างที่มีข้อมูลสมบูรณ์ซึ่งตัดหน่วยตัวอย่างที่มีข้อมูลสูญหายออกไป การประมาณค่าสัมประสิทธิ์การถดถอยด้วยค่าเหล่านี้จึงมีความเอนเอียง เพื่อลดปัญหาดังกล่าวจึงมีการบวกค่าความคลาดเคลื่อนแบบสุ่ม $\sigma_*\left(SE^{(j)}/\hat{\sigma}_1\right)Z_j$ รวมเข้ากับ $\hat{\beta}^{*(j)}$ เพื่อให้ได้ค่าสัมประสิทธิ์การถดถอย $\beta_*^{(j)}$ มีความถูกต้องมากขึ้น (Crawford *et al.*, 1995)

3.3) สุ่มค่า Z_i จำนวนเท่ากับจำนวนสัมประสิทธิ์การถดถอย จากตัวแปรสุ่มที่มีการแจกแจงแบบปกติมาตรฐาน แล้วแทนที่ข้อมูลที่สูญหายด้วย Y_{i*}

$$Y_{i*} = \mathbf{X_i}\boldsymbol{\beta}_* + z_i\sigma_*$$

การประมาณค่าสูญหาย Y_{i*} จะใช้สมการถดถอยในการทำนาย โดยมี $z_i\sigma_*$ เป็นค่าความคลาดเคลื่อนแบบสุ่มซึ่งเทียบได้กับค่าความคลาดเคลื่อน (residual) ในตัวแบบการถดถอย (Rubin, 2004)

ทำซ้ำขั้นตอนที่ 3.1-3.3 จนครบ m ครั้ง ทำให้ได้ชุดข้อมูลสมบูรณ์ m ชุด สำหรับงานวิจัยครั้งนี้ใช้ m=5 ในทางปฏิบัติวิธีเอ็มไอจะใช้ได้ผลดีเมื่อ m มีค่าไม่มากคืออยู่ระหว่าง 2 ถึง 10 (บวรวรรณ, 2543)

ขั้นตอนที่ 2 การวิเคราะห์ชุดข้อมูลที่สมบูรณ์

นำข้อมูลสมบูรณ์แต่ละชุดมาประมาณค่าสัมประสิทธิ์การถดถอย $\boldsymbol{\beta}_l$ เมื่อ l=1,…,m ด้วยวิธี OLS

ขั้นตอนที่ 3 การอนุมานค่า

นำค่าสัมประสิทธิ์การถดถอยมาหาค่าเฉลี่ยเพื่อทำการอนุมานค่าสัมประสิทธิ์การถดถอยสุดท้ายเพื่อใช้สำหรับสมการพยากรณ์ จากสูตร

$$\boldsymbol{\beta}=\left(\boldsymbol{\beta}_1+\boldsymbol{\beta}_2+\ldots+\boldsymbol{\beta}_m\right)/m$$

ข้อดีของวิธีเอ็มไอ
การประมาณค่าความคลาดเคลื่อนมาตรฐานพิสูจน์ได้โดยวิธีการทางคณิตศาสตร์
ข้อเสียของวิธีเอ็มไอ
ขั้นตอนการทำงานต้องใช้เวลามาก

เกณฑ์การเปรียบเทียบ

เกณฑ์การเปรียบเทียบว่าการประมาณค่าสูญหายด้วยวิธีใดใช้ได้ดีกว่า จะพิจารณาโดยการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปของค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (The Estimated Squares Root of Mean Squares Error : RMSE) วิธีการใดให้ค่าประมาณของ RMSE ต่ำกว่าจะเป็นวิธีการประมาณที่ดีกว่า โดยคำนวณจากสูตร

$$\mathrm{RMSE} = \frac{1}{5,000}\sum_{i=1}^{5,000}\sqrt{\frac{\sum_{j=1}^{n}\left(y_{ji}-\hat{y}_{ji}\right)^2}{n-p}}$$

โดยที่ y_{ji} เป็นค่าสังเกตของข้อมูลตัวที่ j ในการทำซ้ำรอบที่ i

$\hat{y}_{ji}$ เป็นค่าพยากรณ์ของข้อมูลตัวที่ j ในการทำซ้ำรอบที่ i

i เป็นจำนวนรอบของการทำซ้ำ i=1,2,…,5,000

n เป็นขนาดตัวอย่าง

p เป็นจำนวนสัมประสิทธิ์การถดถอย

อุปกรณ์และวิธีการ

อุปกรณ์

เครื่องไมโครคอมพิวเตอร์ที่มีหน่วยความจำขนาด 40 GB มีความเร็วในการประมวลผล 256 MHz ที่ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ และโปรแกรม SAS (Statistical Analysis System) เวอร์ชัน 9.1

วิธีการ

ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลอง (Simulation) โดยวิธีมอนติคาร์โล กระทำซ้ำ 5,000 ครั้ง ในแต่ละสถานการณ์ มีขั้นตอนดังนี้

1. จำลองข้อมูลตัวแปรอิสระที่มีการแจกแจงแบบปกติหลายตัวแปร ตามขนาดตัวอย่างและระดับความสัมพันธ์ตามที่กำหนดไว้ในขอบเขตการวิจัย

2. จำลองข้อมูลค่าความคลาดเคลื่อนที่มีการแจกแจงแบบปกติ ตามที่กำหนดไว้ในขอบเขตการวิจัย

3. กำหนด $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$ การกำหนดค่า β เพื่อให้เกิดค่าความคลาดเคลื่อนของตัวประมาณน้อยที่สุดค่า β ควรมีค่าระหว่าง -1 ถึง 1 (Bolch and Huang, 1974)

4. สร้างตัวแปรตามที่มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ ค่าสัมประสิทธิ์การถดถอยและค่าความคลาดเคลื่อน โดยใช้รูปแบบความสัมพันธ์เชิงเส้น

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \varepsilon_j \qquad ; j = 1, \ldots, n$$

5. คำนวณหาจำนวนข้อมูลที่สูญหายและสุ่มตำแหน่งที่สูญหายของข้อมูล

5.1 คำนวณหาจำนวนข้อมูลที่สูญหาย จาก

$$\text{จำนวนข้อมูลที่สูญหาย} = \frac{\text{ขนาดตัวอย่าง X \% การสูญหาย}}{100}$$

ถ้าค่าที่คำนวณได้เป็นเลขทศนิยม จะใช้เลขจำนวนเต็มที่น้อยที่สุดที่มีค่ามากกว่าค่านั้น ยกตัวอย่างกรณี n = 50 สัดส่วนการสูญหาย = 5%

$$\text{จำนวนข้อมูลที่สูญหาย} = \frac{50 \times 5}{100} = 2.5 \simeq 3$$

5.2 ทำการสุ่มตำแหน่งที่สูญหาย โดยใช้โปรแกรม SAS ด้วยคำสั่ง PROC SURVEYSELECT

6. ประมาณค่าสูญหาย ทั้ง 4 วิธี

7. ประมาณค่าพารามิเตอร์ที่มีข้อมูลสมบูรณ์ โดยนำค่า Y_j X_{1j} X_{2j} และ X_{3j} ที่ครบสมบูรณ์ ; $j = 1,\ldots,n$ คำนวณค่าสัมประสิทธิ์การถดถอยจากสมการถดถอยเชิงเส้นพหุคูณ ด้วยวิธีกำลังสองน้อยที่สุด

8. หาค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์

8.1 ประมาณค่า $\hat{Y}$ ด้วยสมการถดถอยเชิงเส้นพหุคูณที่ใช้ในการพยากรณ์

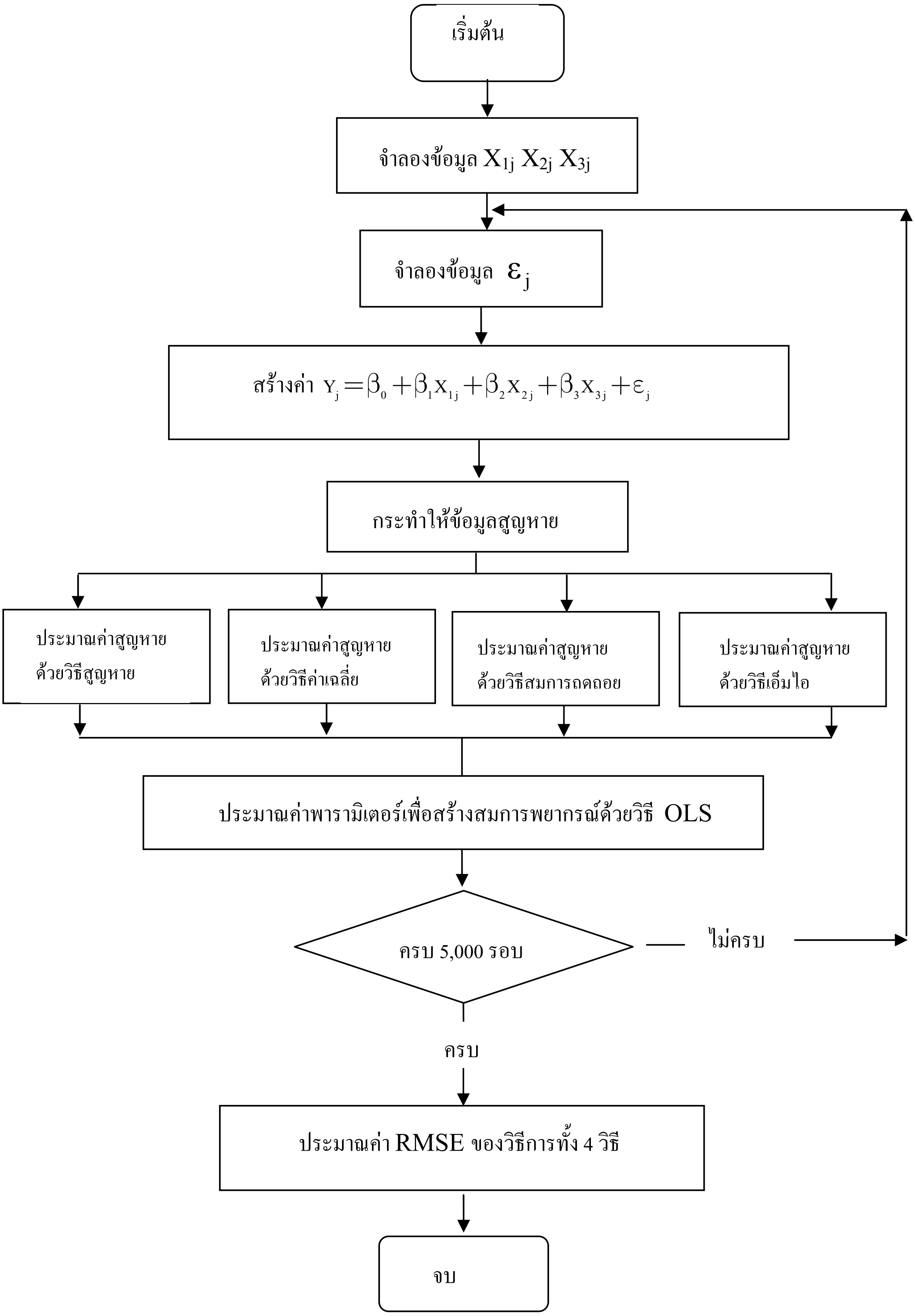
8.2 หาค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ ในการทำซ้ำรอบที่ i โดยสูตร

$$\mathrm{RMSE}_i = \sqrt{\frac{\sum_{j=1}^{n}\left(Y_{ji} - \hat{Y}_{ji}\right)^2}{n-p}}$$

8.3 หาค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์โดยทำการเฉลี่ยใน 5,000 ครั้ง ดังสูตร

$$\mathrm{RMSE} = \frac{\sum_{i=1}^{5,000} \mathrm{RMSE}_i}{5,000}$$

8.4 เปรียบเทียบค่าประมาณของ RMSE ของวิธีการประมาณค่าสูญหายทั้ง 4 วิธี



ภาพที่ 2 แสดงผังงานของขั้นตอนการดำเนินงาน

ผลและวิจารณ์

ผล

เนื่องจากความสัมพันธ์ระหว่างตัวแปรอิสระมี 3 ระดับ คือระดับต่ำ (0.20) ระดับปานกลาง (0.50) และระดับสูง (0.70) ดังนั้นผลการเปรียบเทียบค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนของค่าพยากรณ์ที่ได้จากการประมาณค่าสูญหายของตัวแปรตามทั้ง 4 วิธี แยกเป็น 3 กรณี ดังนี้

1. กรณีที่ตัวแปรอิสระมีความสัมพันธ์ระดับต่ำ

ในทุกขนาดตัวอย่าง (50, 70, 100 และ 200) วิธีสมการถดถอย ให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีเอ็มไอ ยกเว้นกรณีขนาดตัวอย่างเท่ากับ 50 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 5 และขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 30 พบว่า วิธีเอ็มไอให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีสมการถดถอย ส่วนกรณีขนาดตัวอย่างเท่ากับ 70 เมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15 และเปอร์เซ็นต์การสูญหายเท่ากับ 5 วิธีสมการถดถอย ให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีค่าเฉลี่ย

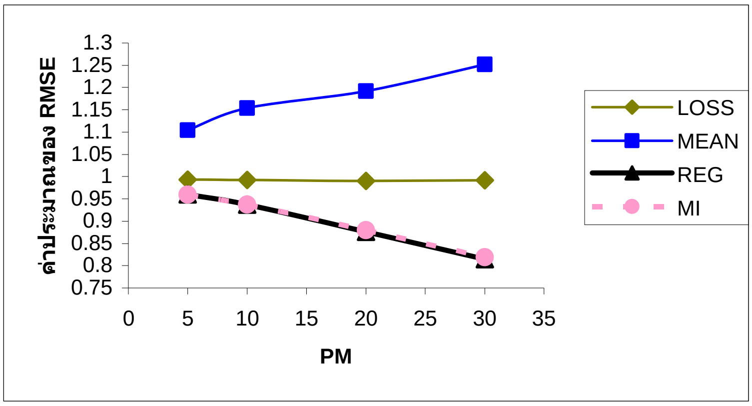
ในทุกกรณีของค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 วิธีค่าเฉลี่ยให้ค่าประมาณของ RMSE สูงที่สุด ในขณะเมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 และ 15 วิธีสูญหาย ให้ค่าประมาณของ RMSE สูงที่สุด และเมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้น ค่าประมาณของ RMSE เพิ่มขึ้น ในทุกขนาดตัวอย่าง

รายละเอียดแสดงดังตารางที่ 1-4 และ ภาพที่ 3- 6

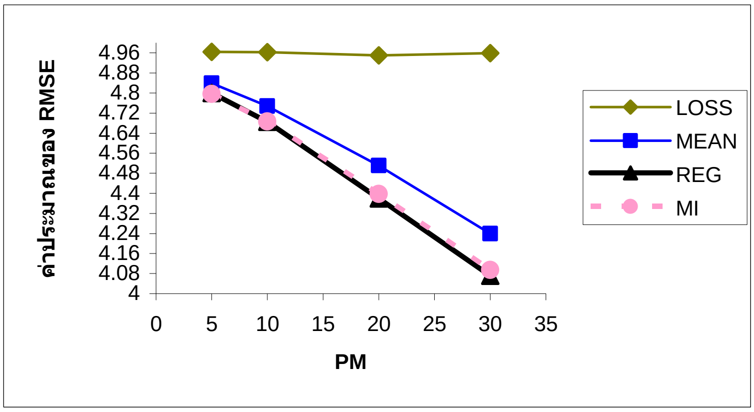
ตารางที่ 1 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9928771	1.1041548	0.9599548	0.9591345*
	10	0.9926826	1.1540439	0.9371809*	0.9373779
	20	0.9901757	1.1914518	0.8759607*	0.8796082
	30	0.9918234	1.2521368	0.8142097*	0.8188514
5	5	4.9643855	4.8395241	4.7997742	4.7956725*
	10	4.9634129	4.7479976	4.6859043*	4.6868895
	20	4.9508786	4.5113892	4.3798034*	4.3980412
	30	4.9591171	4.2399999	4.0710486*	4.0942568
15	5	14.893157	14.436313	14.399323	14.387018*
	10	14.890239	14.115485	14.057713*	14.060669
	20	14.852636	13.343506	13.13941*	13.194124
	30	14.877351	12.437421	12.213146*	12.28277

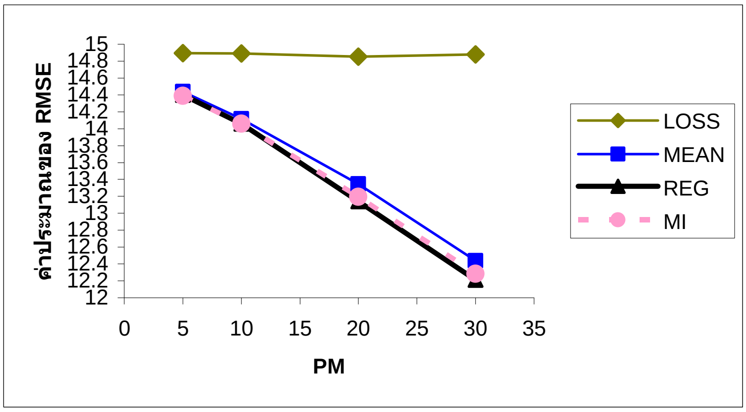
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



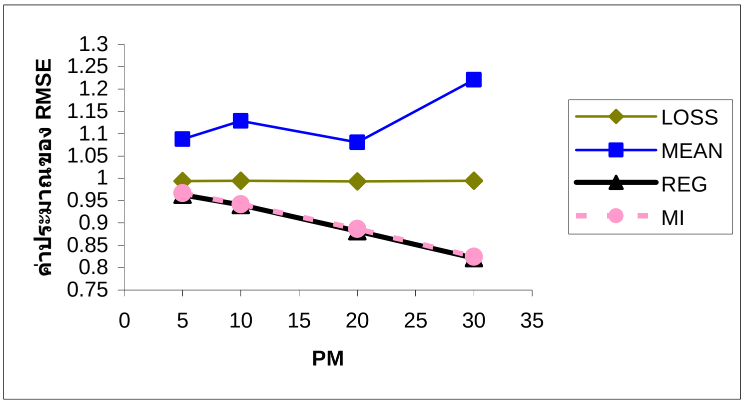
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 3 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20

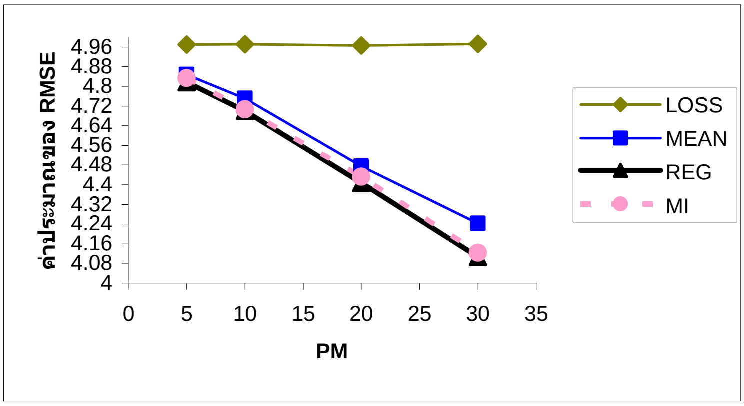
ตารางที่ 2 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 70 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9939331	1.0874036	0.9633432*	0.9667356
	10	0.9943183	1.1284247	0.9401118*	0.9414116
	20	0.9931452	1.0807006	0.8815409*	0.8863872
	30	0.994522	1.2202864	0.8211995*	0.8247019
5	5	4.9696654	4.8482981	4.8167158*	4.8336781
	10	4.9715916	4.7511545	4.7005588*	4.7070582
	20	4.965726	4.4754479	4.4077044*	4.4319361
	30	4.97261	4.2424829	4.1059976*	4.1235093
15	5	14.908996	14.475017	14.450148*	14.501034
	10	14.914775	14.144156	14.101676*	14.121175
	20	14.897178	13.311912	13.223113*	13.295808
	30	14.91783	12.472467	12.317993*	12.370528

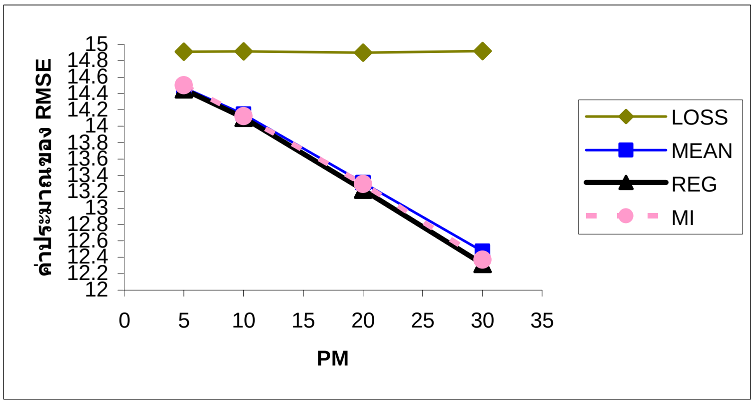
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



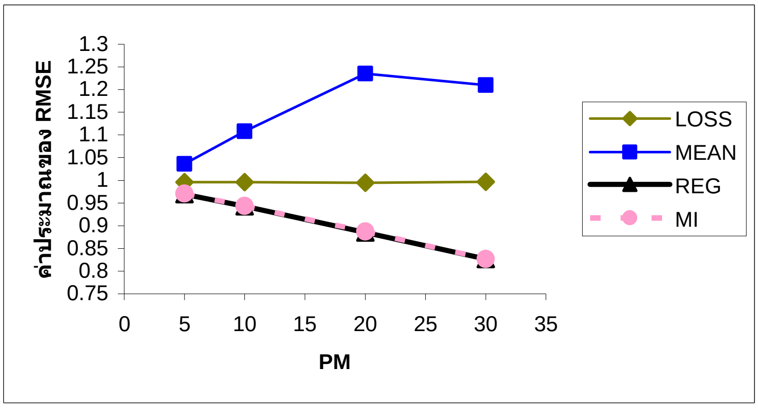
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 4 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 70 และความสัมพันธ์ระหว่างตัวแปรอิสระ เท่ากับ 0.20

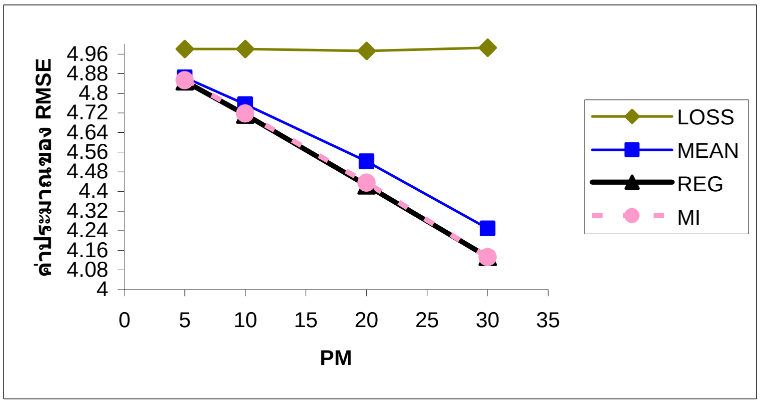
ตารางที่ 3 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 100 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.995995	1.036147	0.9697108*	0.9705151
	10	0.9960632	1.1077576	0.9427586*	0.9436432
	20	0.9945607	1.2353594	0.8849168*	0.8870763
	30	0.9970381	1.2103027	0.8267004	0.8264721*
5	5	4.979975	4.8651965	4.8485541*	4.8525755
	10	4.9803159	4.7554785	4.713793*	4.718216
	20	4.9728033	4.5227343	4.4245842*	4.4353815
	30	4.9851907	4.2494427	4.1335018	4.1323603*
15	5	14.939925	14.559328	14.545662*	14.557727
	10	14.940948	14.171491	14.141379*	14.154648
	20	14.91841	13.348159	13.273753*	13.306144
	30	14.955572	12.50342	12.400505	12.397081*

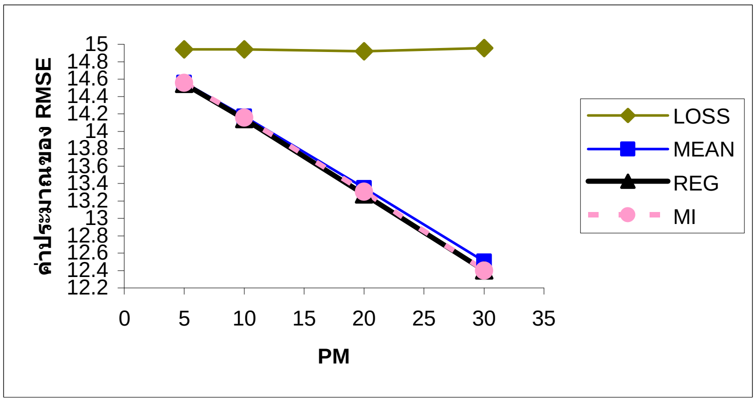
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



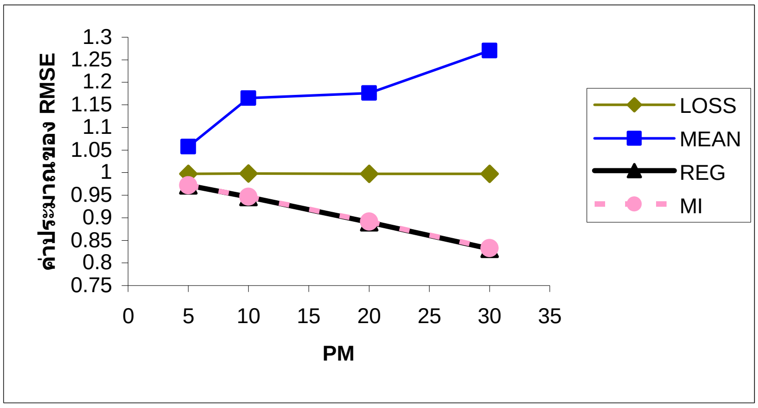
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 5 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 100 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20

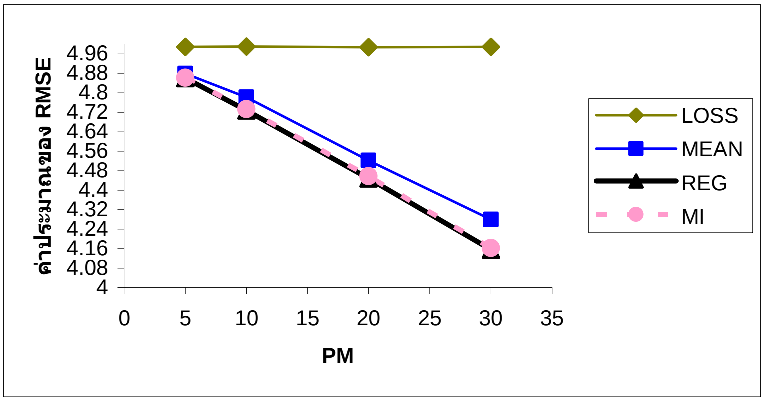
ตารางที่ 4 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่าง เท่ากับ 200 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9977364	1.0572612	0.9719508*	0.9722698
	10	0.9978585	1.1648775	0.9455778*	0.9464493
	20	0.9973185	1.1759373	0.8897503*	0.8916089
	30	0.9976854	1.270276	0.8310651*	0.8325844
5	5	4.9886821	4.8787605	4.8597538*	4.8613492
	10	4.9892924	4.7805203	4.727888*	4.7322466
	20	4.9865925	4.521571	4.4487514*	4.4580443
	30	4.9884272	4.2793663	4.1553256*	4.1629218
15	5	14.966046	14.588903	14.579261*	14.584048
	10	14.967877	14.212132	14.183667*	14.19674
	20	14.959777	13.390256	13.346254*	13.374133
	30	14.965282	12.547431	12.465977*	12.488766

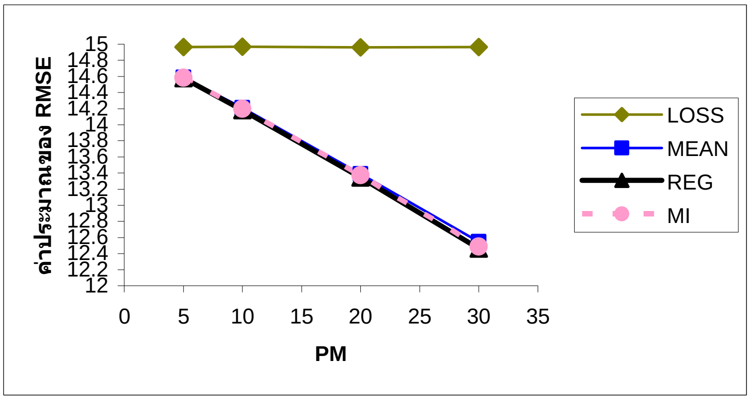
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 6 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 200 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.20

2. กรณีที่ตัวแปรอิสระมีความสัมพันธ์ระดับปานกลาง

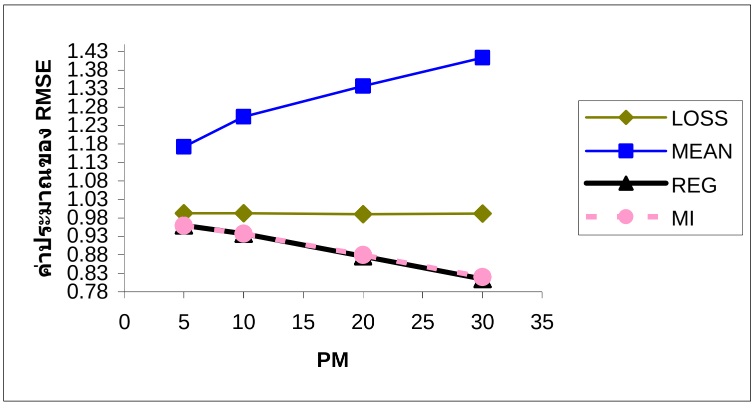
ผลการวิจัยสรุปผลได้ทำนองเดียวกันกับกรณีที่ตัวแปรอิสระมีความสัมพันธ์ระดับต่ำ คำอธิบายดังหน้า 28

รายละเอียดแสดงดังตารางที่ 5-8 และ ภาพที่ 7-10

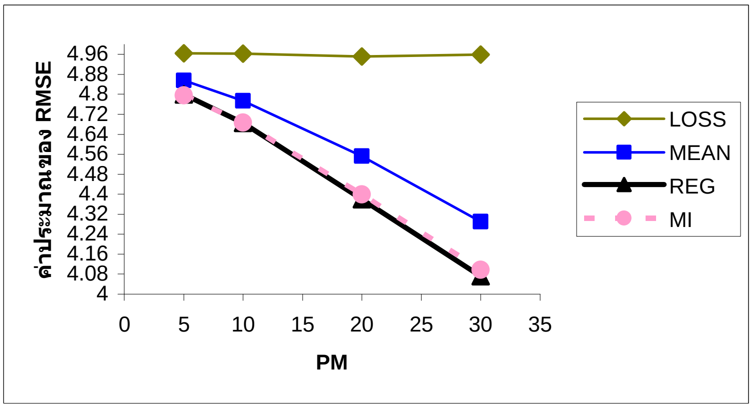
ตารางที่ 5 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9928771	1.1727372	0.9599548	0.9591632*
	10	0.9926826	1.2537077	0.9371809*	0.9374654
	20	0.9901757	1.3365067	0.8759607*	0.8798227
	30	0.9918234	1.413904	0.8142097*	0.8193439
5	5	4.9643855	4.8557686	4.7997742	4.7958162*
	10	4.9634129	4.7734903	4.6859043*	4.6873268
	20	4.9508786	4.5521415	4.3798034*	4.3991135
	30	4.9591171	4.2904101	4.0710486*	4.0967195
15	5	14.893157	14.441824	14.399323	14.387449*
	10	14.890239	14.124225	14.057713*	14.06198
	20	14.852636	13.357491	13.13941*	13.19734
	30	14.877351	12.454503	12.213146*	12.290158

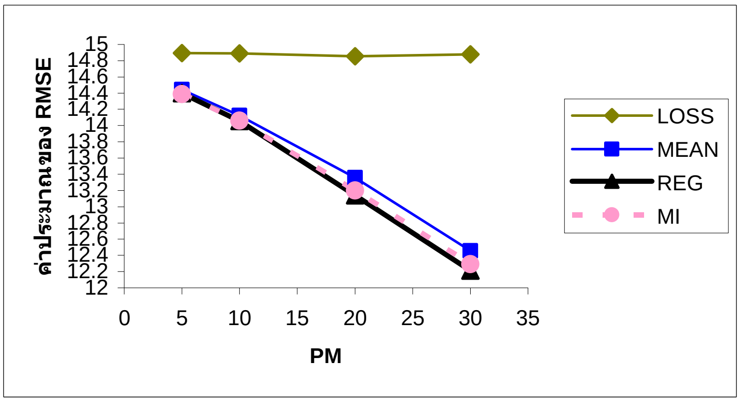
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



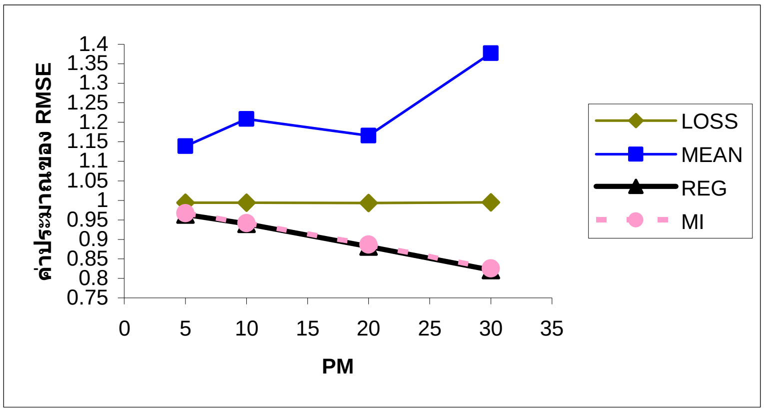
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 7 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50

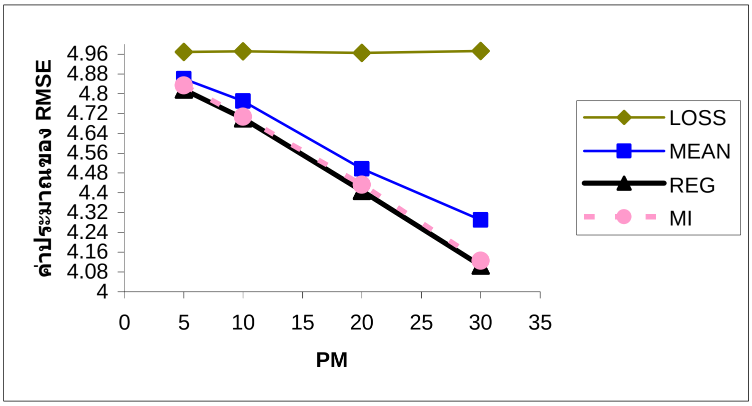
ตารางที่ 6 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่าง เท่ากับ 70 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50 จำแนกตามเปอร์เซ็นการสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9939331	1.138909	0.9633432*	0.9667495
	10	0.9943183	1.2084694	0.9401118*	0.9414486
	20	0.9931452	1.1656381	0.8815409*	0.8864303
	30	0.994522	1.3769104	0.8211995*	0.825028
5	5	4.9696654	4.8601771	4.8167158*	4.8337474
	10	4.9715916	4.7709257	4.7005588*	4.7072431
	20	4.965726	4.49697	4.4077044*	4.4321516
	30	4.97261	4.2902188	4.1059976*	4.12514
15	5	14.908996	14.47903	14.450148*	14.501242
	10	14.914775	14.150862	14.101676*	14.121729
	20	14.897178	13.319313	13.223113*	13.296455
	30	14.91783	12.488837	12.317993*	12.37542

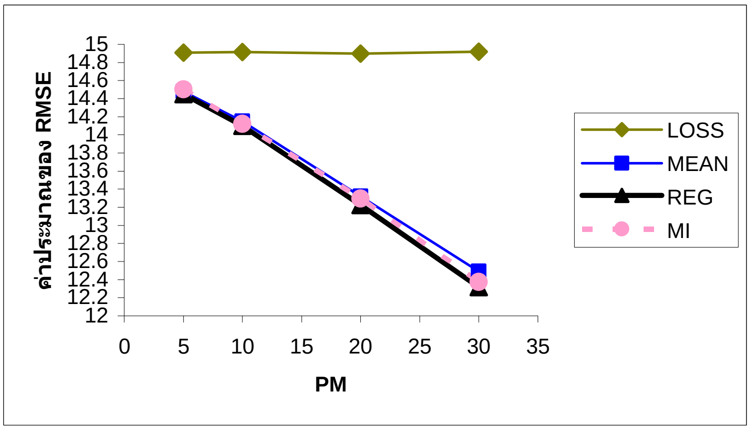
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



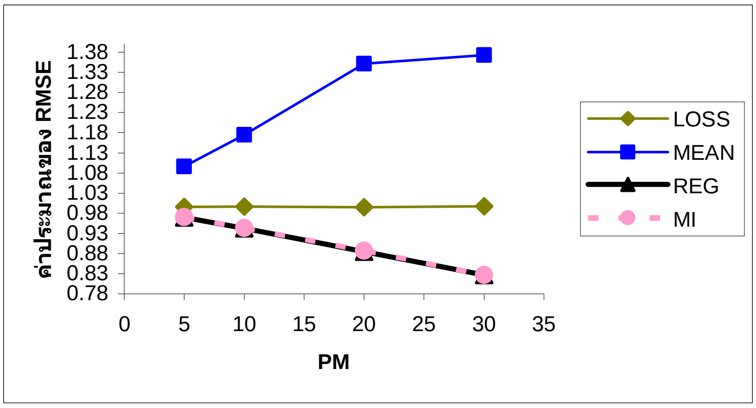
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 8 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 70 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50

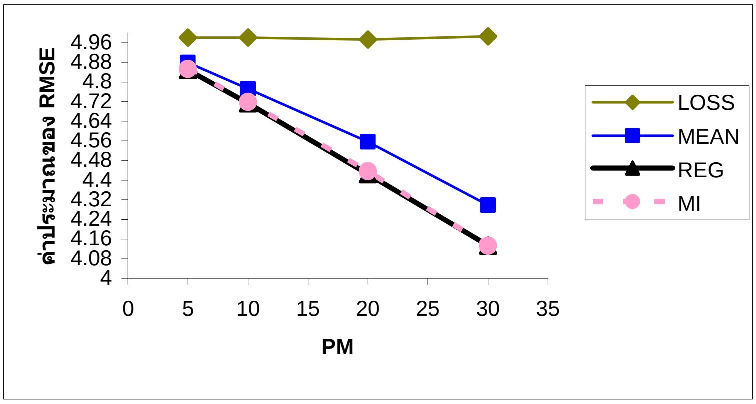
ตารางที่ 7 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 100 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.995995	1.0966582	0.9697108*	0.9705243
	10	0.9960632	1.1749744	0.9427586*	0.9436582
	20	0.9945607	1.3515765	0.8849168*	0.8871783
	30	0.9970381	1.3726646	0.8267004	0.8266654*
5	5	4.979975	4.8784555	4.8485541*	4.8526215
	10	4.9803159	4.7716297	4.713793*	4.7182909
	20	4.9728033	4.5559692	4.4245842*	4.4358914
	30	4.9851907	4.2984936	4.1335018	4.1333268*
15	5	14.939925	14.56376	14.545662*	14.557864
	10	14.940948	14.176926	14.141379*	14.154873
	20	14.91841	13.359498	13.273753*	13.307674
	30	14.955572	12.520133	12.400505	12.399998*

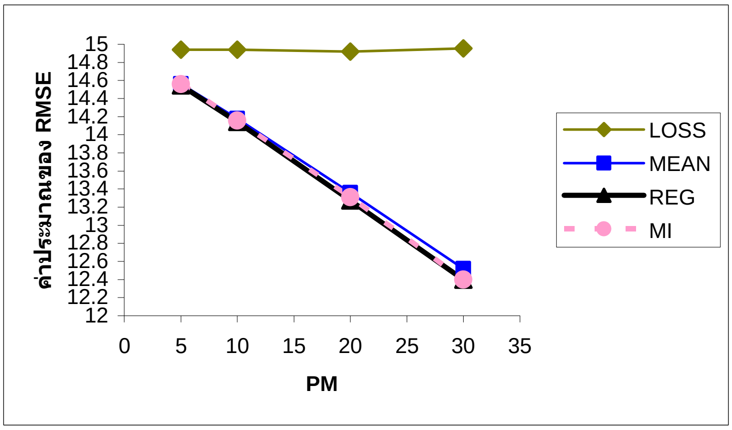
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



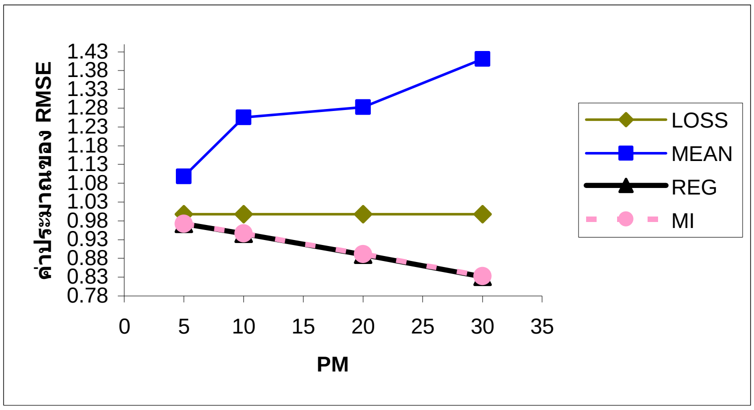
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 9 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อ ขนาดตัวอย่างเท่ากับ 100 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50

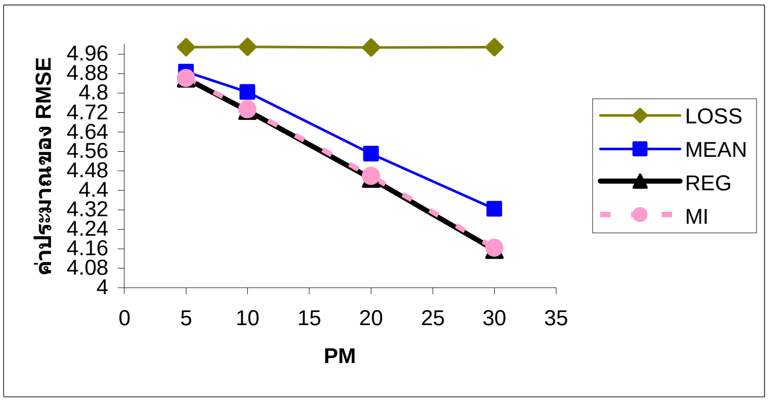
ตารางที่ 8 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 200 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9977364	1.0982198	0.9719508*	0.9722725
	10	0.9978585	1.2552306	0.9455778*	0.9464649
	20	0.9973185	1.282774	0.8897503*	0.8916327
	30	0.9976854	1.4111979	0.8310651*	0.8326729
5	5	4.9886821	4.887805	4.8597538*	4.8613624
	10	4.9892924	4.8033434	4.727888*	4.7323247
	20	4.9865925	4.5504683	4.4487514*	4.4581636
	30	4.9884272	4.323327	4.1553256*	4.1633644
15	5	14.966046	14.591928	14.579261*	14.584087
	10	14.967877	14.21981	14.183667*	14.196974
	20	14.959777	13.399962	13.346254*	13.374491
	30	14.965282	12.56249	12.465977*	12.490093

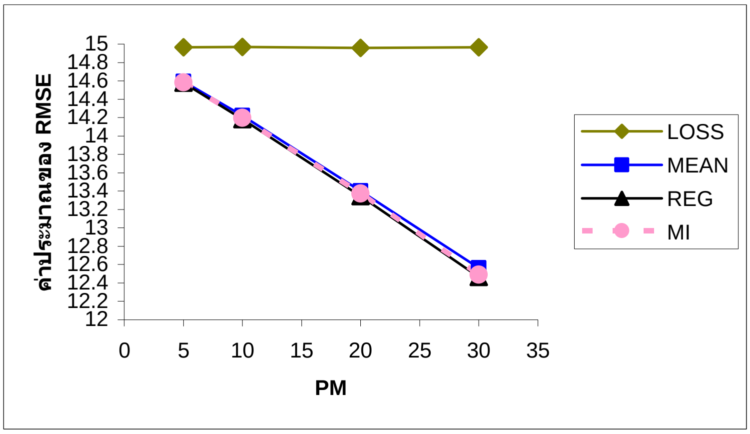
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 10 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 200 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.50

3. กรณีที่ตัวแปรอิสระมีความสัมพันธ์ระดับสูง

ในทุกขนาดตัวอย่าง (50, 70, 100 และ 200) วิธีสมการถดถอย ให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีเอ็มไอ ยกเว้นกรณีขนาดตัวอย่างเท่ากับ 50 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 5 พบว่าวิธีเอ็มไอ ให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีสมการถดถอย ส่วนกรณีขนาดตัวอย่างเท่ากับ 70 เมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15 และเปอร์เซ็นต์การสูญหายเท่ากับ 5 พบว่าวิธีสมการถดถอย ให้ค่าประมาณของ RMSE ต่ำที่สุด รองลงมาคือ วิธีค่าเฉลี่ย

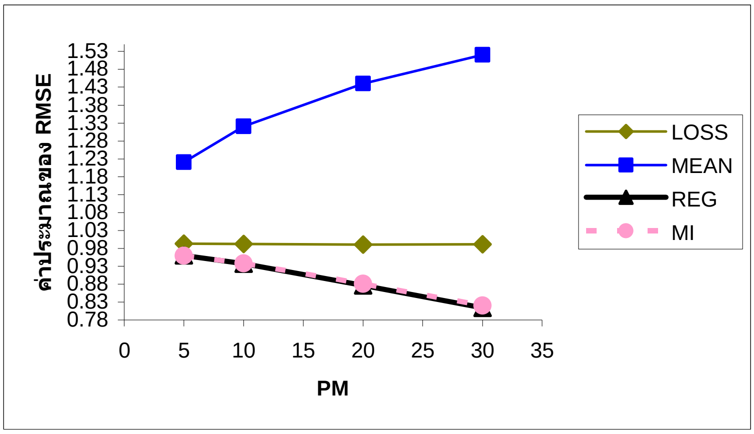
ในทุกกรณีของค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 วิธีค่าเฉลี่ย ให้ค่าประมาณของ RMSE สูงที่สุด ในขณะเมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 และ 15 พบว่าวิธีสูญหาย ให้ค่าประมาณของ RMSE สูงที่สุด และเมื่อค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้น ค่าประมาณของ RMSE เพิ่มขึ้น ในทุกขนาดตัวอย่าง

รายละเอียดแสดงดังตารางที่ 9-12 และ ภาพที่ 11-14

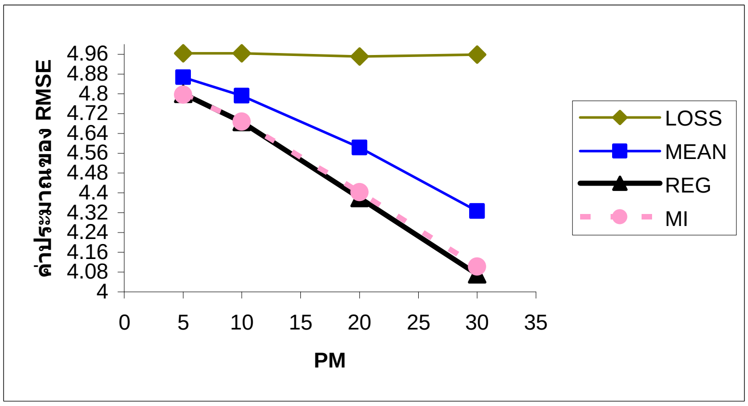
ตารางที่ 9 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9928771	1.2209286	0.9599548	0.9592228*
	10	0.9926826	1.3206338	0.9371809*	0.937642
	20	0.9901757	1.4404723	0.8759607*	0.8804994
	30	0.9918234	1.5202516	0.8142097*	0.8204333
5	5	4.9643855	4.8677192	4.7997742	4.7961139*
	10	4.9634129	4.7916777	4.6859043*	4.6882099
	20	4.9508786	4.583856	4.3798034*	4.402497
	30	4.9591171	4.3264436	4.0710486*	4.1021665
15	5	14.893157	14.445869	14.399323	14.388342*
	10	14.890239	14.13045	14.057713*	14.06463
	20	14.852636	13.368385	13.13941*	13.207491
	30	14.877351	12.466777	12.213146*	12.306499

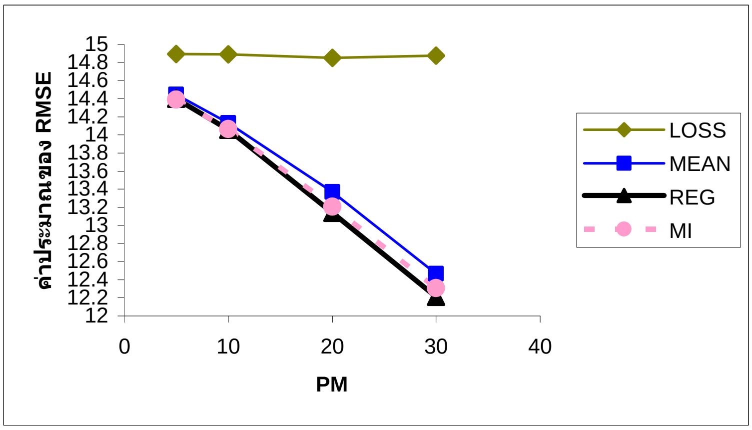
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



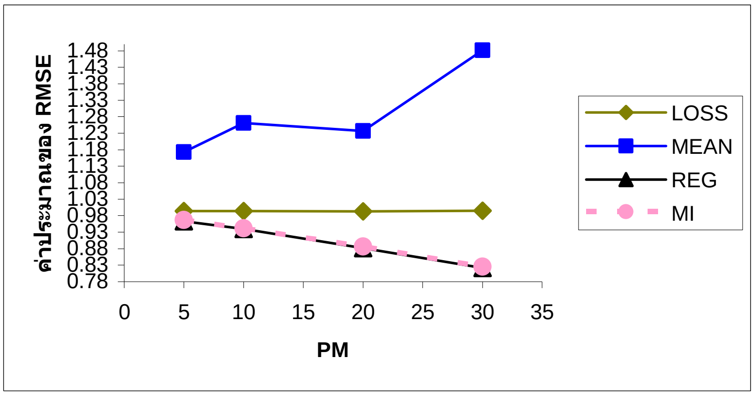
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 11 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 50 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70

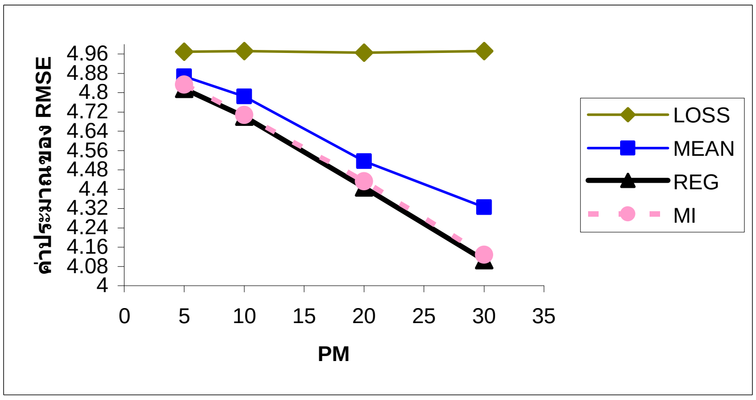
ตารางที่ 10 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 70 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9939331	1.1726852	0.9633432*	0.9667768
	10	0.9943183	1.2605017	0.9401118*	0.9415271
	20	0.9931452	1.2363431	0.8815409*	0.88656
	30	0.994522	1.4820049	0.8211995*	0.8257178
5	5	4.9696654	4.8682517	4.8167158*	4.8338838
	10	4.9715916	4.7844589	4.7005588*	4.7076355
	20	4.965726	4.5160394	4.4077044*	4.4328002
	30	4.97261	4.3251315	4.1059976*	4.1285888
15	5	14.908996	14.481761	14.450148*	14.501651
	10	14.914775	14.155463	14.101676*	14.122907
	20	14.897178	13.325896	13.223113*	13.298401
	30	14.91783	12.500912	12.317993*	12.385766

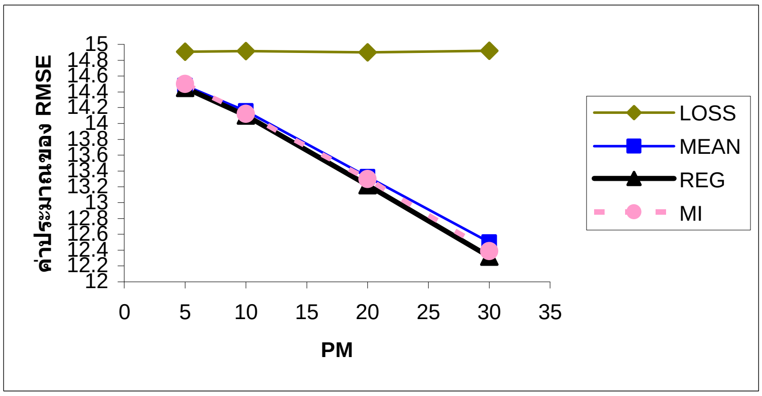
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



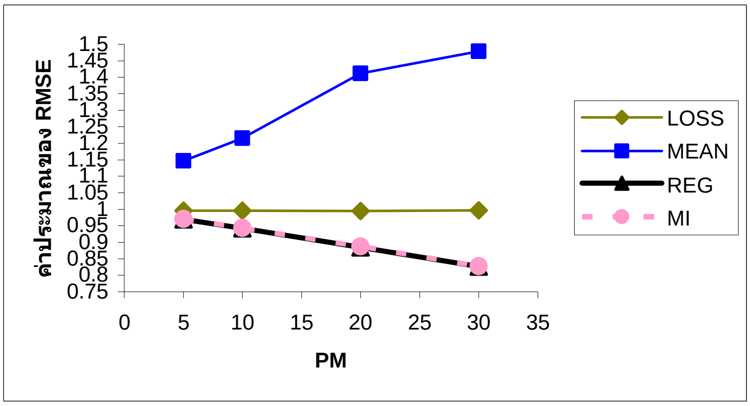
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 12 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 70 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70

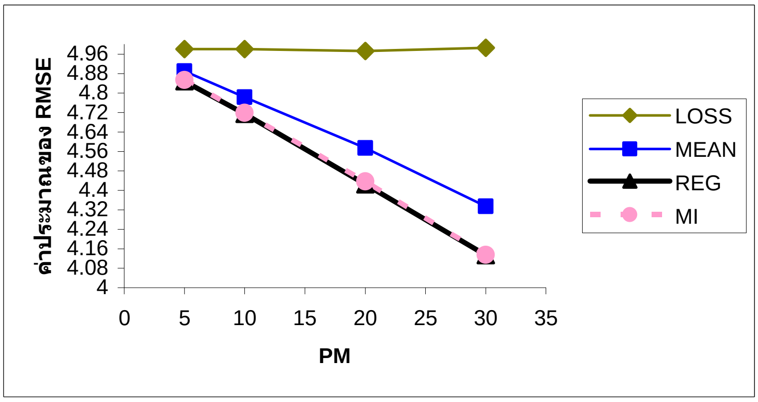
ตารางที่ 11 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 100 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70 จำแนกตามเปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.995995	1.1465769	0.9697108*	0.9705449
	10	0.9960632	1.2157308	0.9427586*	0.9436884
	20	0.9945607	1.4120036	0.8849168*	0.88735
	30	0.9970381	1.4791734	0.8267004*	0.8270655
5	5	4.979975	4.8899375	4.8485541*	4.8527246
	10	4.9803159	4.7818494	4.713793*	4.7184419
	20	4.9728033	4.5743161	4.4245842*	4.4367498
	30	4.9851907	4.333677	4.1335018*	4.1353276
15	5	14.939925	14.567607	14.545662*	14.558174
	10	14.940948	14.180365	14.141379*	14.155326
	20	14.91841	13.365775	13.273753*	13.310249
	30	14.955572	12.532234	12.400505*	12.405983

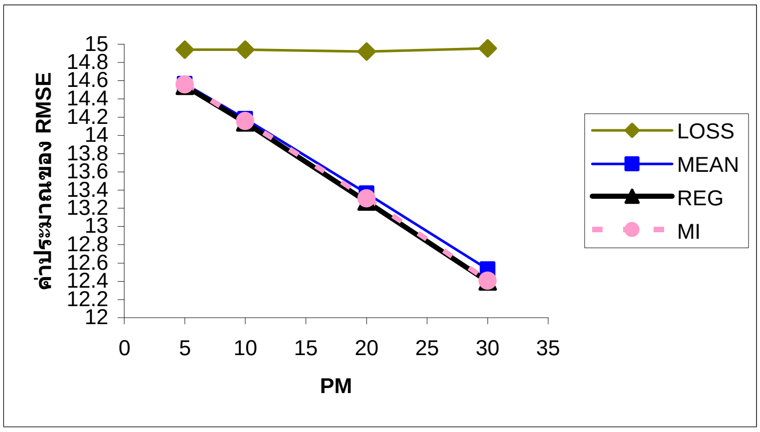
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



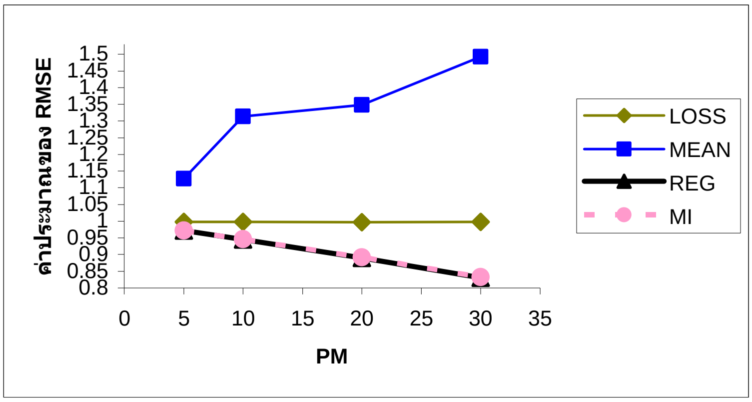
(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 13 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 100 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70

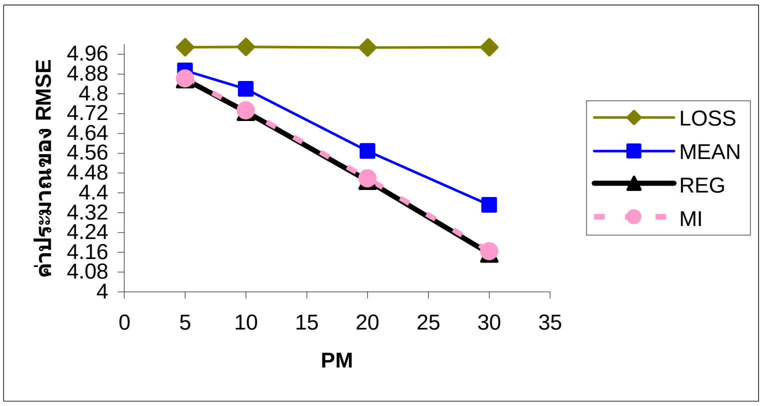
ตารางที่ 12 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่าง เท่ากับ 200 ความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70 จำแนกตาม เปอร์เซ็นต์การสูญหายของตัวแปรตาม และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน

σ	PM	ค่าประมาณของ RMSE			
		LOSS	MEAN	REG	MI
1	5	0.9977364	1.1270692	0.9719508*	0.9722774
	10	0.9978585	1.313628	0.9455778*	0.9464939
	20	0.9973185	1.3480058	0.8897503*	0.8916797
	30	0.9976854	1.4925311	0.8310651*	0.8328412
5	5	4.9886821	4.8943696	4.8597538*	4.8613871
	10	4.9892924	4.8189412	4.727888*	4.7324697
	20	4.9865925	4.5692644	4.4487514*	4.4583987
	30	4.9884272	4.3505354	4.1553256*	4.1642058
15	5	14.966046	14.594125	14.579261*	14.584161
	10	14.967877	14.225077	14.183667*	14.197409
	20	14.959777	13.406317	13.346254*	13.375196
	30	14.965282	12.571845	12.465977*	12.492617

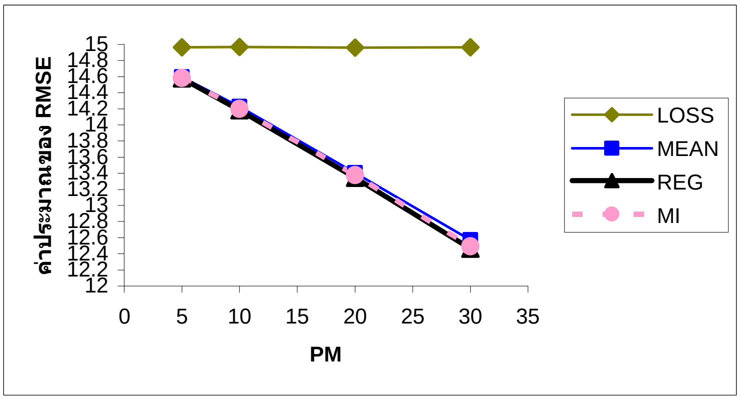
หมายเหตุ * ค่าประมาณของ RMSE ที่มีค่าต่ำสุด



(ก) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1



(ข) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5



(ค) ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15

ภาพที่ 14 แสดงค่าประมาณของรากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยพหุคูณ ทั้ง 4 วิธี เมื่อขนาดตัวอย่างเท่ากับ 200 และความสัมพันธ์ระหว่างตัวแปรอิสระเท่ากับ 0.70

วิจารณ์

ปัจจัยที่มีผลต่อการประมาณค่าสูญหายด้วยวิธีสมการถดถอยและวิธีเอ็มไอ

1. เปอร์เซ็นต์การสูญหาย เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีทั้งสองให้ค่าประมาณของ RMSE ลดลงเนื่องจากค่าประมาณของค่าสูญหายแต่ละค่ามีค่าใกล้เคียงกันมากกว่าค่าของข้อมูลจริง ทำให้ชุดข้อมูลสมบูรณ์ที่ได้จากกรณีเปอร์เซ็นต์การสูญหายน้อยกว่ามีความแปรปรวนมากกว่า ชุดข้อมูลสมบูรณ์เมื่อเปอร์เซ็นต์การสูญหายมากกว่า จากผลการวิจัยพบว่าในขนาดตัวอย่าง และค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนขนาดเดียวกัน เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้น วิธีการประมาณค่าสูญหายทั้ง 4 วิธี ให้ค่าประมาณของ RMSE แตกต่างกันมากขึ้น โดย วิธีสมการถดถอยและวิธีเอ็มไอ ให้ค่าประมาณของ RMSE ต่ำกว่าอีก 2 วิธีอย่างชัดเจน ซึ่งสอดคล้องกับงานวิจัยของ Roth (1994) ที่พบว่าการเลือกวิธีการประมาณค่าสูญหายมีความสำคัญ เมื่อเปอร์เซ็นต์การสูญหายอยู่ระหว่าง 15-20% และมีความสำคัญมากที่สุดเมื่อเปอร์เซ็นต์การสูญหายอยู่ระหว่าง 30-40% ที่ระดับนี้ การเลือกใช้วิธีการประมาณค่าสูญหายทำให้ผลลัพธ์แตกต่างกัน และสอดคล้องกับผลงานวิจัยของ Raaijimakers (1999) ที่พบว่าความแตกต่างของความคลาดเคลื่อนของวิธีการประมาณค่าสูญหายต่าง ๆ มีค่าน้อยเมื่อเปอร์เซ็นต์การสูญหายน้อย

2. ขนาดตัวอย่าง เมื่อขนาดตัวอย่างเพิ่มขึ้น พบว่า วิธีทั้งสอง ให้ค่าประมาณของ RMSE เพิ่มขึ้น เนื่องจากเมื่อขนาดตัวอย่างเพิ่มในขณะที่เปอร์เซ็นต์การสูญหายเท่าเดิม จะพบว่าสัดส่วนของข้อมูลที่เหลืออยู่กับข้อมูลที่สูญหายมากกว่าเมื่อตัวอย่างขนาดเล็ก

3. ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ เมื่อเพิ่มระดับความสัมพันธ์ระหว่างตัวแปรอิสระ ไม่มีผลต่อการเลือกวิธีการประมาณค่าสูญหาย

สรุปและข้อเสนอแนะ

สรุป

ในทุกขนาดตัวอย่าง (50, 70, 100 และ 200) เมื่อพิจารณาจากค่าประมาณของ RMSE พบว่า ในทุกสถานการณ์วิธีสมการถดถอย จะเป็นวิธีที่เหมาะสมที่สุด สำหรับการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอย ถึงแม้ว่าในบางสถานการณ์ต่อไปนี้คือ กรณีขนาดตัวอย่างเท่ากับ 50 เมื่อเปอร์เซ็นการสูญหายเท่ากับ 5 และกรณีขนาดตัวอย่างเท่ากับ 100 เมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 30 และระดับความสัมพันธ์ของตัวแปรอิสระระดับต่ำถึงปานกลาง พบว่า วิธีสมการถดถอยและวิธีเอ็มไอให้ค่าประมาณของ RMSE แตกต่างกันค่อนข้างน้อย โดยวิธีเอ็มไอให้ค่าประมาณของ RMSE ต่ำกว่า แต่เนื่องจากวิธีสมการถดถอยเป็นวิธีที่ง่ายและไม่ซับซ้อน ดังนั้นจะเลือกใช้วิธีสมการถดถอยแทนวิธีเอ็มไอ รายละเอียดแสดงดังตารางที่ 13

ตารางที่ 13 แสดงวิธีที่เหมาะสมที่สุดในแต่ละสถานการณ์ของการประมาณค่าสูญหาย ในการวิเคราะห์การถดถอยพหุคูณ

σ	PM	ขนาดตัวอย่าง			
		50	70	100	200
1	5	REG	REG	REG	REG
	10	REG	REG	REG	REG
	20	REG	REG	REG	REG
	30	REG	REG	REG	REG
5	5	REG	REG	REG	REG
	10	REG	REG	REG	REG
	20	REG	REG	REG	REG
	30	REG	REG	REG	REG
15	5	REG	REG	REG	REG
	10	REG	REG	REG	REG
	20	REG	REG	REG	REG
	30	REG	REG	REG	REG

หมายเหตุ REG แทน วิธีสมการถดถอย

ข้อเสนอแนะ

1. ศึกษาวิธีการประมาณค่าสูญหายกรณีการวิเคราะห์การถดถอยของข้อมูลลักษณะเชิงคุณภาพ เช่น การวิเคราะห์การถดถอยแบบลอจิสติค การใช้ตัวแปรหุ่นในสมการถดถอย หรือ การวิเคราะห์ข้อมูลลักษณะอื่น เช่น การวิเคราะห์อนุกรมเวลา

2. ศึกษาในกรณีที่ตัวแปรอิสระมีจำนวนมากกว่า 3 ตัวแปร เช่น 5 หรือ 7 ตัวแปร หรือการสูญหายของข้อมูลเกิดขึ้นในตัวแปรอิสระ

3. ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายด้วยวิธีการอื่น ๆ นอกเหนือจากที่ใช้ในการศึกษาในครั้งนี้ เช่น การประมาณค่าสูญหายโดยใช้ทฤษฎีโครงข่ายประสาทเทียม

เอกสารและสิ่งอ้างอิง

กัลยา วานิชย์บัญชา. 2548. **การวิเคราะห์ข้อมูลหลายตัวแปร.** บริษัท ธรรมสาร จำกัด, กรุงเทพฯ.

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

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

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

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

บวรวรรณ ดิเรกโรค. 2543. **การประยุกต์ใช้วิธีการใส่ค่าหลายค่าแทนข้อมูลที่สูญหายแต่ละค่าในการวิเคราะห์ข้อมูลอุบัติเหตุผู้ขับขี่จักรยานยนต์**. วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยมหิดล.

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

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

วิรัชช พาณิชวงศ์. 2545**. การวิเคราะห์การถดถอย.** สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ.

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

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

Crawford, L., L. Tennstedt and B. Mckinlay. 1995. Comparison of Analytic Method for Non-Random Missingness of Outcome Data. **J Clin Epidemiol**. 48: 209-219.

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.

Fisher, R.A. 1922. **On the Mathematical Foundations of Theoretical Statistics.** Roy, Inc. New York.

George, E.P. and C. George. 1973. **Bayesian Inference in Statistical Analysis**. Addison – Wesley Publishing Company, Inc. New York.

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

Heitjan, D.B. and R.J.A. Little. 1991. Multiple Imputation for the Fetal Accident Reporting System. **Appl Stat**. 40: 13-29.

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. 1987. **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.

Rubin, D.B. 2004. **Multiple Imputation for Nonresponse in Surveys**. John Wiley and Sons, Inc. New York.

Viragoontavan, S. 2000. **Comparing Six Missing Data Methods within the Discriminant Analysis Context: A Monte Carlo Study**. Ph.D. thesis, The Ohio State University.

Walczak, B. and D.L. Massart. 2001. Dealing with Missing Data: Part II. **Chemometrics and Intelligent Laboratory Systems.** 58: 29-42.

ภาคผนวก

โปรแกรมที่ใช้ในการวิจัยทั้งหมด เขียนด้วยโปรแกรม SAS (Statistical Analysis System) โดยใช้คำสั่งหลัก Proc iml (iml: Interactive Matrix Language) ซึ่งเป็นโปรแกรมภาษาสำหรับใช้กับข้อมูลที่จัดอยู่ในรูปเมทริกซ์ ส่วนโปรแกรมย่อยแสดงได้ดังรายละเอียด

<u>โปรแกรมสำหรับจำลองข้อมูลตัวแปรอิสระ X</u>

```
seed=54321;
n=50;
sigma={1  .2  .2,
       .2  1  .2,
       .2 .2   1};
mu={0, 0, 0};
p=nrow(sigma);
m=repeat(t(mu),n,1);
g=root(sigma);
gprime=t(g);
z=normal(repeat(seed,n,p));
x=z*g+m;
```

<u>โปรแกรมสำหรับจำลองค่าความคลาดเคลื่อน</u>

```
seed=54321;
n=50;
p1=1;
e=rannor(j(n,p1,seed));
```

<u>โปรแกรมสำหรับสร้างข้อมูลตัวแปรตาม Y</u>

```
b={1, 1, 1};
bo=j(n,p1,1);
xb=x*b;
y=bo+xb+e;
```

<u>โปรแกรมสำหรับสุ่มหาตำแหน่งสูญหายของข้อมูล</u>

```
data y50;
    input i;
datalines;
1 2 3 4 5 6 7 8 9 10
11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30
31 32 33 34 35 36 37 38 39 40
41 42 43 44 45 46 47 48 49 50;
run;
proc surveyselect data=y50 method=srs rep=1
   sampsize=2 seed=54321 out=y2;
run;
proc print data=y2;
run;
```

<u>โปรแกรมสำหรับกำหนดสัญลักษณ์ในตำแหน่งสูญหายของข้อมูล</u>

```
y50[21,]=.;
y50[31,]=.;
y50[41,]=.;
```

```
x50[21,]=.;
x50[31,]=.;
x50[41,]=.;
```

<u>โปรแกรมสำหรับตัดค่าสูญหายทิ้ง</u>

```
notmiss=loc(y[,1]^=.);
newy=y[notmiss,];
notmiss1=loc(xt[,1]^=.);
newx=xt[notmiss1,];
```

<u>โปรแกรมสำหรับหาค่าสัมประสิทธิ์การถดถอยของวิธีประมาณค่าสูญหายด้วยวิธีสูญหาย</u>

```
xpx=t(newx)*newx;
xpy=t(newx)*newy;
xpxi=inv(xpx);
beta=xpxi*xpy;
```

<u>โปรแกรมสำหรับหาค่าสัมประสิทธิ์การถดถอยของวิธีประมาณค่าสูญหายด้วยวิธีค่าเฉลี่ย</u>

```
sumy=newy[+,+];
meany=sumy/nrow(newy);
y[21,]=meany;y[31,]=meany;y[41,]=meany;
xpx=t(xt)*xt;
xpy=t(xt)*y;
xpxi=inv(xpx);
beta=xpxi*xpy;
```

<u>โปรแกรมสำหรับหาค่าสัมประสิทธิ์การถดถอยของวิธีประมาณค่าสูญหายด้วยวิธีสมการถดถอย</u>

```
xpx1=t(newx)*newx;
xpy1=t(newx)*newy;
xpxi1=inv(xpx1);
beta1=xpxi1*xpy1;
y[21,]=xt1[21,]*beta1;y[31,]=xt1[31,]*beta1;
y[41,]=xt1[41,]*beta1;
xpx=t(xt1)*xt1;
xpy=t(xt1)*y;
xpxi=inv(xpx);
beta=xpxi*xpy;
```

<u>โปรแกรมสำหรับหาค่าสัมประสิทธิ์การถดถอยของวิธีประมาณค่าสูญหายด้วยวิธีเอ็มไอ</u>

```
xpx=t(newx)*newx;
xpy=t(newx)*newy;
xpxi=inv(xpx);
beta=xpxi*xpy;
yhatc=newx*beta;
residc2=newy-yhatc;
sres=ssq(residc2);
df=nrow(newy)-nrow(beta);
df2=df/2;
s2c=sres/df;
sc=sqrt(s2c);
        /* m=1*/
k1=2*rangam(54321,df2);
s21=s2c*df/k1;
```

```
s1=sqrt(s21);
c11=xpxi[1,1];
c22=xpxi[2,2];
c33=xpxi[3,3];
c44=xpxi[4,4];
se1=sqrt(s2c*c11);
se2=sqrt(s2c*c22);
se3=sqrt(s2c*c33);
se4=sqrt(s2c*c44);
se=se1//se2//se3//se4;
sec=se/sc;
ss1=s1*sec;
zj1= rannor (j(4,1,54321));
b11=beta[1,1]+(ss1[1,1]*zj1[1,1]);
b21=beta[2,1]+(ss1[2,1]*zj1[2,1]);
b31=beta[3,1]+(ss1[3,1]*zj1[3,1]);
b41=beta[4,1]+(ss1[4,1]*zj1[4,1]);
b1=b11//b21//b31//b41;
zi1= rannor (j(3,1,54321));
x21=xt1[21,1:4];
x31=xt1[31,1:4];
x41=xt1[41,1:4];
y21i1=(x21*b1)+(s1*zi1[1,1]);
y31i1=(x31*b1)+(s1*zi1[2,1]);
y41i1=(x41*b1)+(s1*zi1[3,1]);
ynew1=y[1:20,1]//y21i1//y[22:30,1]//y31i1//
        y[32:40,1]//y41i1//y[42:50,1];
xpx1=t(xt1)*xt1;
xpy1=t(xt1)*ynew1;
xpxi1=inv(xpx1);
```

```
beta1=xpxi1*xpy1;
        /*m=2*/
k2=2*rangam(543,df2);
s22=s2c*df/k2;
s2=sqrt(s22);
ss2=s2*sec;
zj2= rannor (j(4,1,543));
b12=beta[1,1]+(ss2[1,1]*zj2[1,1]);
b22=beta[2,1]+(ss2[2,1]*zj2[2,1]);
b32=beta[3,1]+(ss2[3,1]*zj2[3,1]);
b42=beta[4,1]+(ss2[4,1]*zj2[4,1]);
b2=b12//b22//b32//b42;
zi2= rannor (j(3,1,543));
y21i2=(x21*b2)+(s2*zi2[1,1]);
y31i2=(x31*b2)+(s2*zi2[2,1]);
y41i2=(x41*b2)+(s2*zi2[3,1]);
ynew2=y[1:20,1]//y21i2//y[22:30,1]//y31i2//
         y[32:40,1]//y41i2//y[42:50,1];
xpy2=t(xt1)*ynew2;
beta2=xpxi1*xpy2;
/*m=3*/
k3=2*rangam(123,df2);
s23=s2c*df/k3;
s3=sqrt(s23);
ss3=s3*sec;
zj3= rannor (j(4,1,123));
b13=beta[1,1]+(ss3[1,1]*zj3[1,1]);
b23=beta[2,1]+(ss3[2,1]*zj3[2,1]);
b33=beta[3,1]+(ss3[3,1]*zj3[3,1]);
b43=beta[4,1]+(ss3[4,1]*zj3[4,1]);
```

```
b3=b13//b23//b33//b43;
zi3= rannor (j(3,1,123));
y21i3=(x21*b3)+(s3*zi3[1,1]);
y31i3=(x31*b3)+(s3*zi3[2,1]);
y41i3=(x41*b3)+(s3*zi3[3,1]);
ynew3=y[1:20,1]//y21i3//y[22:30,1]//y31i3//
        y[32:40,1]//y41i3//y[42:50,1];
xpy3=t(xt1)*ynew3;
beta3=xpxi1*xpy3;
/*m=4*/
k4=2*rangam(321,df2);
s24=s2c*df/k4;
s4=sqrt(s24);
ss4=s4*sec;
zj4= rannor (j(4,1,321));
b14=beta[1,1]+(ss4[1,1]*zj4[1,1]);
b24=beta[2,1]+(ss4[2,1]*zj4[2,1]);
b34=beta[3,1]+(ss4[3,1]*zj4[3,1]);
b44=beta[4,1]+(ss4[4,1]*zj4[4,1]);
b4=b14//b24//b34//b44;
zi4= rannor (j(3,1,321));
y21i4=(x21*b4)+(s4*zi4[1,1]);
y31i4=(x31*b4)+(s4*zi4[2,1]);
y41i4=(x41*b4)+(s4*zi4[3,1]);
ynew4=y[1:20,1]//y21i4//y[22:30,1]//y31i4//
        y[32:40,1]//y41i4//y[42:50,1];
xpy4=t(xt1)*ynew4;
beta4=xpxi1*xpy4;
/*m=5*/
k5=2*rangam(12345,df2);
```

```
s25=s2c*df/k5;
s5=sqrt(s25);
ss5=s5*sec;
zj5= rannor (j(4,1,12345));
b15=beta[1,1]+(ss5[1,1]*zj5[1,1]);
b25=beta[2,1]+(ss5[2,1]*zj5[2,1]);
b35=beta[3,1]+(ss5[3,1]*zj5[3,1]);
b45=beta[4,1]+(ss5[4,1]*zj5[4,1]);
b5=b15//b25//b35//b45;
zi5= rannor (j(3,1,12345));
y21i5=(x21*b5)+(s5*zi5[1,1]);
y31i5=(x31*b5)+(s5*zi5[2,1]);
y41i5=(x41*b5)+(s5*zi5[3,1]);
ynew5=y[1:20,1]//y21i5//y[22:30,1]//y31i5//
        y[32:40,1]//y41i5//y[42:50,1];
xpy5=t(xt1)*ynew5;
beta5=xpxi1*xpy5;
betam=(beta1+beta2+beta3+beta4+beta5)/5;
```

โปรแกรมสำหรับคำนวณค่า RMSE

```
    trmse=0;
do I=1 to 5000;
    yhat=newx*beta;
    resid=newy-yhat;
    sse=ssq(resid);
    mse=sse/(nrow(newy)-4);
    rmsej=sqrt(mse);
    trmse=trmse+rmsej;
    rmse=trmse/5000;
```

```
 end;
            print rmse;
```

<u>โปรแกรมสำหรับตรวจสอบการแจกแจงแบบปกติหลายตัวแปร</u>

```
dm log 'clear' ;
dm output 'clear' ;
proc iml;
seed=54321;
n=42;
sigma={1  .2  .2,
     .2  1  .2,
            .2 .2   1};
mu={0, 0, 0};
p=nrow(sigma);
m=repeat(t(mu),n,1);
g=root(sigma);
gprime=t(g);
z=normal(repeat(seed,n,p));
x=z*g+m;
n=nrow(x);
p=ncol(x);
/*xbar=(1/n)*t(x)*j(n,1,1);*/
dfchi=p*(p+1)*(p+2)/6;
q=i(n)-(1/n)*j(n,n,1);
s=(1/(n))*t(x)*q*x;
s_inv=inv(s);
g_matrix=q*x*s_inv*t(x)*q;
beta1hat=(sum(g_matrix#g_matrix#g_matrix))/(n*n);
beta2hat=trace(g_matrix#g_matrix)/n;
```

```
kappa1=n*beta1hat/6;
kappa2=(beta2hat-p*(p+2))/sqrt(8*p*(p+2)/n);
pvalskew=1-probchi(kappa1,dfchi);
pvalkurt=2*(1-probnorm(abs(kappa2)));
p1=1;
e=rannor(j(n,p1,seed));
b={1, 1, 1};
bo=j(n,p1,1);
xb=x*b;
y=bo+xb+e;
print mu m g gprime z x e b bo xb y;
print s;
print s_inv;
print'Based on skewness:' beta1hat kappa1 pvalskew;
print'Based on kurtosis:' beta2hat kappa2 pvalkurt;
quit;
```

<u>โปรแกรมสำหรับตรวจสอบการแจกแจงแบบปกติ</u>

```
data test;
  input e @@;
 datalines;
 0.5395446    0.1515141    1.0530452
1.8522055     0.5306824    0.8530319
-1.671015     0.8321276    -0.546006
0.4369414     -0.021418    -1.290315
-1.407207     0.0552162    -0.056727
0.3682446     -0.15583     -0.331073
0.2903829     -0.162897    -0.129466
-0.497799     0.922835     -0.350373
1.2277315     -1.476466    -1.460598
0.5025386     -0.49003     -1.26465
1.4234359     0.1508395    0.3899138
-0.715347     -0.504273    -1.232963
0.8250716     0.4685437    -0.253821
-1.100393     -1.037903    -0.452661;
proc univariate data=test normal plot;
var e;
run ;
```

ตัวอย่างโปรแกรม กรณีขนาดตัวอย่างเท่ากับ 50 ค่าความเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 1 และเปอร์เซ็นต์การสูญหายเท่ากับ 5

1. การประมาณค่าสูญหายด้วยวิธีสูญหาย

```
dm log 'clear' ;
dm output 'clear' ;
proc iml;
        trmse=0;
        seed=54321;
        n=50;
        sigma={1  .2  .2,
                  .2  1  .2,
                  .2 .2   1};
         mu={0, 0, 0};
         p=nrow(sigma);
         m=repeat(t(mu),n,1);
         g=root(sigma);
         gprime=t(g);
         z=normal(repeat(seed,n,p));
         x=z*g+m;
         one=j(n,1,1);
         xt=one||x;
      do I=1 to 5000;
          p1=1;
          e=rannor(j(n,p1,seed));
          b={1, 1, 1};
          bo=j(n,p1,1);
          xb=x*b;
          y=bo+xb+e;
```

```
        y[21,]=.; y[31,]=.; y[41,]=.;
        xt[21,]=.; xt[31,]=.; xt[41,]=.;
        notmiss=loc(y[,1]^=.);
        newy=y[notmiss,];
        notmiss1=loc(xt[,1]^=.);
        newx=xt[notmiss1,];
        xpx=t(newx)*newx;
        xpy=t(newx)*newy;
        xpxi=inv(xpx);
        beta=xpxi*xpy;
        yhat=newx*beta;
        resid=newy-yhat;
        sse=ssq(resid);
        mse=sse/(nrow(newy)-4);
        rmsej=sqrt(mse);
        trmse=trmse+rmsej;
        rmse=trmse/5000;
    end;
        print rmse;
quit;
```

2. การประมาณค่าสูญหายด้วยวิธีค่าเฉลี่ย

```
dm log 'clear' ;
dm output 'clear' ;
proc iml;
    trmse=0;
    seed=54321;
    n=50;
```

```
sigma={1  .2  .2,
       .2  1  .2,
       .2 .2   1};
mu={0, 0, 0};
p=nrow(sigma);
m=repeat(t(mu),n,1);
g=root(sigma);
gprime=t(g);
z=normal(repeat(seed,n,p));
x=z*g+m;
one=j(n,1,1);
xt=one||x;
do I=1 to 5000;
   p1=1;
   e=rannor(j(n,p1,seed));
   b={1, 1, 1};
   bo=j(n,p1,1);
   xb=x*b;
   y=bo+xb+e;
   y[21,]=.;y[31,]=.;y[41,]=.;
   notmiss=loc(y[,1]^=.);
   newy=y[notmiss,];
   sumy=newy[+,+];
   meany=sumy/nrow(newy);
   y[21,]=meany;y[31,]=meany;y[41,]=meany;
   xpx=t(xt)*xt;
   xpy=t(xt)*y;
   xpxi=inv(xpx);
   beta=xpxi*xpy;
   yhat=xt*beta;
```

```
        resid=y-yhat;
        sse=ssq(resid);
        mse=sse/(n-4);
        rmsej=sqrt(mse);
        trmse=trmse+rmsej;
        rmse=trmse/5000;
    end;
        print  rmse;
quit;
```

3. การประมาณค่าสูญหายด้วยวิธีสมการถดถอย

```
dm log 'clear' ;
dm output 'clear' ;
proc iml;
    trmse=0;
    seed=54321;
     n=50;
    sigma={1  .2 .2,
           .2  1  .2,
           .2 .2   1};
    mu={0, 0, 0};
    p=nrow(sigma);
    m=repeat(t(mu),n,1);
    g=root(sigma);
    gprime=t(g);
    z=normal(repeat(seed,n,p));
    x=z*g+m;
    one=j(n,1,1);
    xt=one||x;
```

```
xt1=xt;
do I=1 to 5000;
    p1=1;
    e=rannor(j(n,p1,seed));
    b={1, 1, 1};
    bo=j(n,p1,1);
    xb=x*b;
    y=bo+xb+e;
    y[21,]=.; y[31,]=.; y[41,]=.;
    xt[21,]=.; xt[31,]=.; xt[41,]=.;
    notmiss=loc(y[,1]^=.);
    newy=y[notmiss,];
    notmiss1=loc(xt[,1]^=.);
    newx=xt[notmiss1,];
    xpx1=t(newx)*newx;
    xpy1=t(newx)*newy;
    xpxi1=inv(xpx1);
    beta1=xpxi1*xpy1;
    y[21,]=xt1[21,]*beta1;y[31,]=xt1[31,]*beta1;
    y[41,]=xt1[41,]*beta1;
    xpx=t(xt1)*xt1;
    xpy=t(xt1)*y;
    xpxi=inv(xpx);
    beta=xpxi*xpy;
    yhat=xt1*beta;
    resid=y-yhat;
    sse=ssq(resid);
    mse=sse/(nrow(y)-4);
    rmsej=sqrt(mse);
    trmse=trmse+rmsej;
```

```
        rmse=trmse/5000;
    end;
        print rmse;
quit;
```

4. การประมาณค่าสูญหายด้วยวิธีเอ็มไอ

```
dm log 'clear' ;
dm output 'clear' ;
proc iml;
    trmse=0;
    seed=54321;
     n=50;
    sigma={1  .2 .2,
           .2  1  .2,
           .2 .2   1};
    mu={0, 0, 0};
    p=nrow(sigma);
    m=repeat(t(mu),n,1);
    g=root(sigma);
    gprime=t(g);
    z=normal(repeat(seed,n,p));
    x=z*g+m;
    one=j(n,1,1);
    xt=one||x;
    xt1=xt;
    do I=1 to 5000;
        p1=1;
        e=rannor(j(n,p1,seed));
         b={1, 1, 1};
```

```
bo=j(n,p1,1);
xb=x*b;
y=bo+xb+e;
y[21,]=.; y[31,]=.; y[41,]=.;
xt[21,]=.; xt[31,]=.; xt[41,]=.;
notmiss=loc(y[,1]^=.);
newy=y[notmiss,];
notmiss1=loc(xt[,1]^=.);
newx=xt[notmiss1,];
xpx=t(newx)*newx;
xpy=t(newx)*newy;
xpxi=inv(xpx);
beta=xpxi*xpy;
yhatc=newx*beta;
residc2=newy-yhatc;
sres=ssq(residc2);
df=nrow(newy)-nrow(beta);
df2=df/2;
s2c=sres/df;
sc=sqrt(s2c);
        /* m=1*/
k1=2*rangam(54321,df2);
s21=s2c*df/k1;
s1=sqrt(s21);
c11=xpxi[1,1];
c22=xpxi[2,2];
c33=xpxi[3,3];
c44=xpxi[4,4];
se1=sqrt(s2c*c11);
se2=sqrt(s2c*c22);
```

```
se3=sqrt(s2c*c33);
se4=sqrt(s2c*c44);
se=se1//se2//se3//se4;
sec=se/sc;
ss1=s1*sec;
zj1= rannor (j(4,1,54321));
b11=beta[1,1]+(ss1[1,1]*zj1[1,1]);
b21=beta[2,1]+(ss1[2,1]*zj1[2,1]);
b31=beta[3,1]+(ss1[3,1]*zj1[3,1]);
b41=beta[4,1]+(ss1[4,1]*zj1[4,1]);
b1=b11//b21//b31//b41;
zi1= rannor (j(3,1,54321));
x21=xt1[21,1:4];
x31=xt1[31,1:4];
x41=xt1[41,1:4];
y21i1=(x21*b1)+(s1*zi1[1,1]);
y31i1=(x31*b1)+(s1*zi1[2,1]);
y41i1=(x41*b1)+(s1*zi1[3,1]);
ynew1=y[1:20,1]//y21i1//y[22:30,1]//y31i1//
        y[32:40,1]//y41i1//y[42:50,1];
xpx1=t(xt1)*xt1;
xpy1=t(xt1)*ynew1;
xpxi1=inv(xpx1);
beta1=xpxi1*xpy1;
        /*m=2*/
k2=2*rangam(543,df2);
s22=s2c*df/k2;
s2=sqrt(s22);
ss2=s2*sec;
zj2= rannor (j(4,1,543));
```

```
b12=beta[1,1]+(ss2[1,1]*zj2[1,1]);
b22=beta[2,1]+(ss2[2,1]*zj2[2,1]);
b32=beta[3,1]+(ss2[3,1]*zj2[3,1]);
b42=beta[4,1]+(ss2[4,1]*zj2[4,1]);
b2=b12//b22//b32//b42;
zi2= rannor (j(3,1,543));
y21i2=(x21*b2)+(s2*zi2[1,1]);
y31i2=(x31*b2)+(s2*zi2[2,1]);
y41i2=(x41*b2)+(s2*zi2[3,1]);
ynew2=y[1:20,1]//y21i2//y[22:30,1]//y31i2//
        y[32:40,1]//y41i2//y[42:50,1];
xpy2=t(xt1)*ynew2;
beta2=xpxi1*xpy2;
/*m=3*/
k3=2*rangam(123,df2);
s23=s2c*df/k3;
s3=sqrt(s23);
ss3=s3*sec;
zj3= rannor (j(4,1,123));
b13=beta[1,1]+(ss3[1,1]*zj3[1,1]);
b23=beta[2,1]+(ss3[2,1]*zj3[2,1]);
b33=beta[3,1]+(ss3[3,1]*zj3[3,1]);
b43=beta[4,1]+(ss3[4,1]*zj3[4,1]);
b3=b13//b23//b33//b43;
zi3= rannor (j(3,1,123));
y21i3=(x21*b3)+(s3*zi3[1,1]);
y31i3=(x31*b3)+(s3*zi3[2,1]);
y41i3=(x41*b3)+(s3*zi3[3,1]);
ynew3=y[1:20,1]//y21i3//y[22:30,1]//y31i3//
        y[32:40,1]//y41i3//y[42:50,1];
```

```
xpy3=t(xt1)*ynew3;
beta3=xpxi1*xpy3;
/*m=4*/
k4=2*rangam(321,df2);
s24=s2c*df/k4;
s4=sqrt(s24);
ss4=s4*sec;
zj4= rannor (j(4,1,321));
b14=beta[1,1]+(ss4[1,1]*zj4[1,1]);
b24=beta[2,1]+(ss4[2,1]*zj4[2,1]);
b34=beta[3,1]+(ss4[3,1]*zj4[3,1]);
b44=beta[4,1]+(ss4[4,1]*zj4[4,1]);
b4=b14//b24//b34//b44;
zi4= rannor (j(3,1,321));
y21i4=(x21*b4)+(s4*zi4[1,1]);
y31i4=(x31*b4)+(s4*zi4[2,1]);
y41i4=(x41*b4)+(s4*zi4[3,1]);
ynew4=y[1:20,1]//y21i4//y[22:30,1]//y31i4//
        y[32:40,1]//y41i4//y[42:50,1];
xpy4=t(xt1)*ynew4;
beta4=xpxi1*xpy4;
/*m=5*/
k5=2*rangam(12345,df2);
s25=s2c*df/k5;
s5=sqrt(s25);
ss5=s5*sec;
zj5= rannor (j(4,1,12345));
b15=beta[1,1]+(ss5[1,1]*zj5[1,1]);
b25=beta[2,1]+(ss5[2,1]*zj5[2,1]);
b35=beta[3,1]+(ss5[3,1]*zj5[3,1]);
```

```
            b45=beta[4,1]+(ss5[4,1]*zj5[4,1]);
            b5=b15//b25//b35//b45;
            zi5= rannor (j(3,1,12345));
            y21i5=(x21*b5)+(s5*zi5[1,1]);
            y31i5=(x31*b5)+(s5*zi5[2,1]);
            y41i5=(x41*b5)+(s5*zi5[3,1]);
            ynew5=y[1:20,1]//y21i5//y[22:30,1]//y31i5//
                      y[32:40,1]//y41i5//y[42:50,1];
            xpy5=t(xt1)*ynew5;
            beta5=xpxi1*xpy5;
            betam=(beta1+beta2+beta3+beta4+beta5)/5;
            yhat=xt1*betam;
           resid=y-yhat;
           sse=ssq(resid);
           mse=sse/(nrow(y)-4);
           rmsej=sqrt(mse);
           trmse=trmse+rmsej;
           rmse=trmse/5000;
    end;
           print rmse;
quit;
```

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวจริยา แสงสุวรรณ
วัน เดือน ปี ที่เกิด	28 เมษายน 2521
สถานที่เกิด	อำเภอระโนด จังหวัดสงขลา
ประวัติการศึกษา	วท.บ. (สถิติ) มหาวิทยาลัยทักษิณ
ตำแหน่งหน้าที่การงานปัจจุบัน	นักสถิติ
สถานที่ทำงานปัจจุบัน	กรมการขนส่งทางน้ำและพาณิชยนาวี กระทรวงคมนาคม