

การเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยพหุเมื่อตัวแปรตามมีการสูญหายแบบสุ่ม

Comparison of Missing Data Estimation Methods for the Multiple Regression Analysis with Missing at Random Dependent Variable

เรืองลักษณ์ หล้าใจชื่อ*, อำไพ ทองธีรภาพ และจุฑาภรณ์ สิ้นสมบุญทอง

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

แขวงลาดยาว เขตจตุจักร กรุงเทพมหานคร 10900

Rueangluck Lamjaisue*, Ampai Thongteeraparp and Juthaphorn Sinsomboonthong

Department of Statistics, Faculty of Science, Kasetsart University,

Ladyao, Chatuchak, Bangkok 10900

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยพหุเมื่อตัวแปรตามมีการสูญหายอย่างสุ่ม วิธีการประมาณค่าสูญหายที่ศึกษาในงานวิจัยนี้ประกอบด้วยวิธีการประมาณค่าสูญหายแบบเดี่ยว 4 วิธี ได้แก่ วิธี regression imputation (RI) วิธี stochastic regression imputation (SRI) วิธี K-nearest neighbor (KNN) วิธี EM algorithm (EM) และวิธีการประมาณค่าสูญหายแบบรวม 2 วิธี ได้แก่ วิธี K-nearest regression imputation with equivalent weighted (KREW) และวิธี K-nearest stochastic regression imputation with equivalent weighted (KSEW) ซึ่งเป็นวิธีการประมาณค่าสูญหายที่ได้จากการรวมวิธีการประมาณค่าสูญหายแบบเดี่ยว 2 วิธี คือ วิธี KNN กับวิธี RI และวิธี KNN กับวิธี SRI ตามลำดับ โดยใช้กับฟังก์ชันการถ่วงน้ำหนักด้วยวิธีการให้น้ำหนักเท่ากัน (EW) จำลองข้อมูลด้วยวิธีมอนติคาร์โล โดยกำหนดขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5, 10 และ 15 การสูญหาย 4 ระดับ คือ 10, 20, 30 และ 40 % เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย ผลการวิจัยพบว่าวิธี KSEW มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่าง 20 และ 30 วิธี SRI มีประสิทธิภาพดีที่สุด เมื่อขนาดตัวอย่าง 50 และ 100 และทุกวิธีการจะมีประสิทธิภาพลดลงเมื่อเปอร์เซ็นต์การสูญหายและค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้น

คำสำคัญ : ข้อมูลสูญหาย; การวิเคราะห์การถดถอยพหุ; วิธีการประมาณค่าสูญหาย; การประมาณค่าสูญหายแบบเดี่ยว; การประมาณค่าสูญหายแบบรวม

Abstract

The objective of this research is to compare the efficiency of missing data estimation methods for the multiple regression analysis with missing at random dependent variable. The missing data estimation methods considered in research are four single imputation methods—p regression imputation (RI), stochastic regression imputation (SRI), K-nearest neighbor (KNN), EM algorithm—and two composite imputation methods (KREW and KSEW). The KREW method is derived from a combination of K-nearest neighbor and regression imputation. The KSEW method is derived from a combination of K-nearest neighbor and stochastic regression imputation. The composite imputation methods were weighted by equivalent weighted method. For this study, the Monte Carlo simulation was done under the condition of sample sizes 20, 30, 50 and 100; the standard deviations of errors 5, 10 and 15; the missing percentage 10, 20, 30 and 40 %. The criterion of comparison the efficiency is the mean square error (MSE). The results show that the KSEW method performs best when the sample sizes 20 and 30. The SRI method performs best when the sample sizes 50 and 100. All estimation methods are less effective when the missing percentage and the standard deviation of errors increase.

Keywords: missing data; multiple regression analysis; missing data estimation; single imputation; composite imputation

1. บทนำ

การวิเคราะห์การถดถอยพหุ (multiple regression analysis) เป็นการวิเคราะห์ข้อมูลโดยใช้ตัวแปรอิสระ (independent variable) ที่มีมากกว่าหนึ่งตัวแปรขึ้นไปมาใช้ในการอธิบายตัวแปรตาม (dependent variable) ซึ่งตัวแปรอิสระและตัวแปรตามจะมีความสัมพันธ์กันในลักษณะใดลักษณะหนึ่ง ปัญหาอย่างหนึ่งที่มักเกิดขึ้นเสมอในการวิเคราะห์ข้อมูลคือ ข้อมูลไม่สมบูรณ์ (incomplete data) ซึ่งในงานวิจัยที่มีลักษณะเชิงสำรวจอาจเกิดปัญหาเนื่องจากผู้ตอบแบบสอบถามตอบคำถามไม่ครบหรือบางคำถามอาจไม่สามารถให้ข้อมูลได้ จึงทำให้เกิดข้อมูลสูญหาย (missing data) ซึ่งถือเป็นปัญหาที่มีความสำคัญเนื่องจากการวิเคราะห์ข้อมูล หากนำข้อมูลที่ไม่สมบูรณ์ไปวิเคราะห์อาจทำให้ได้ผลลัพธ์ที่คลาดเคลื่อนจาก

ความเป็นจริง [1] ส่งผลกระทบต่อการนำไปใช้ในการวางแผนตัดสินใจในงานต่างๆ ถ้าหากนำข้อมูลที่ไม่มีความสมบูรณ์ไปใช้ในการวิเคราะห์การถดถอยย่อมทำให้ประสิทธิภาพในการวิเคราะห์ข้อมูลลดลง ผลลัพธ์ที่ได้ อาจเกิดความเอนเอียง (bias)

การจัดการกับปัญหาการสูญหายนั้นสามารถทำได้หลายวิธีโดยวิธีที่นิยมใช้ คือ วิธี mean หรือวิธี mode และวิธีที่ง่ายที่สุด คือ การตัดข้อมูลที่สูญหายทิ้งและนำข้อมูลที่สมบูรณ์เท่านั้นมาวิเคราะห์ โดยวิธีนี้จะทำให้ขนาดของข้อมูลลดน้อยลงและสูญเสียรายละเอียดบางอย่างไป ทำให้มีผู้คิดค้นศึกษาและพัฒนาวิธีการจัดการข้อมูลสูญหายขึ้นมาใหม่ เพื่อเพิ่มประสิทธิภาพในการประมาณค่าสูญหายที่ดีกว่าวิธีการเดิม และจากการศึกษางานวิจัยที่เกี่ยวข้องกับการประมาณค่าสูญหายของตัวแปรตาม พบว่า จีรกานต์

(2552) [2] ได้ประยุกต์โดยการนำวิธีการประมาณค่าสูญหายแบบดั้งเดิมที่มีอยู่มาประมาณค่าสูญหายร่วมกันแล้วนำมาใช้กับฟังก์ชันการถ่วงน้ำหนัก โดยได้นำเสนอการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบเดียว ซึ่งได้แก่ วิธี mean วิธี regression imputation (RI) วิธี multiple imputation (MI) และประยุกต์วิธีการประมาณค่าสูญหายขั้นใหม่ เป็นการประมาณค่าสูญหายแบบรวมที่ใช้ค่าประมาณสูญหายจากทั้งสามวิธีข้างต้นร่วมกัน แล้วนำมาใช้กับฟังก์ชันการถ่วงน้ำหนักด้วยกัน 3 แบบ คือ การให้น้ำหนักที่เท่ากัน (equivalent weighted method, EW) การถ่วงน้ำหนักด้วยค่าสัมบูรณ์ต่ำสุด (least absolute value method, LAV) และการถ่วงน้ำหนักด้วยความแปรปรวนต่ำสุด (minimum variance method, MV) ซึ่งผลการศึกษพบว่าวิธีการประมาณค่าสูญหายแบบรวมที่ถ่วงน้ำหนักด้วยวิธี LAV มีประสิทธิภาพในการประมาณค่าสูญหายดีที่สุด รองลงมาคือการใช้ฟังก์ชันการถ่วงน้ำหนักด้วยวิธี EW และ ศศิธร (2555) [3] ได้นำเสนอการเปรียบเทียบวิธีการประมาณค่าสูญหายแบบรวมที่ได้ประยุกต์ขั้นใหม่ 2 วิธี คือ วิธี K-nearest regression multiple imputation 1 (KRMI1) และวิธี K-nearest regression multiple imputation 2 (KRMI2) ซึ่งเป็นวิธีที่พัฒนาจากวิธีการประมาณค่าสูญหายแบบรวมจากงานวิจัยของ จีรกันต์ (2552) [2] โดยเลือกใช้วิธี K-nearest neighbor imputation (KNN) แทนวิธี mean เพื่อช่วยลดข้อจำกัดในกรณีที่ข้อมูลมีค่าผิดปกติ (outlier) โดยที่วิธี KRMI1 เป็นวิธีประมาณค่าสูญหายที่ได้จากการรวมวิธีการประมาณค่าสูญหาย 3 วิธี ได้แก่ วิธี KNN วิธี RI และวิธี MI ที่นำมาถ่วงน้ำหนักด้วยวิธี LAV และวิธี KRMI2 เป็นวิธีประมาณค่าสูญหายที่ได้จากการรวมวิธีการประมาณค่าสูญหาย 3 วิธี ได้แก่ วิธี KNN วิธี RI และวิธี MI ที่นำมาถ่วงน้ำหนักด้วยวิธี EW ซึ่งพบว่า

วิธีการประมาณค่าสูญหายวิธี KRMI2 ที่ถ่วงน้ำหนักด้วยวิธี EW มีประสิทธิภาพในการประมาณค่าสูญหายดีกว่าวิธี KRMI1 ที่ถ่วงน้ำหนักด้วยวิธี LAV

จากงานวิจัยที่กล่าวมาข้างต้นพบว่าการใช้ฟังก์ชันการถ่วงน้ำหนักด้วยวิธี EW นั้นมีประสิทธิภาพเพียงพอที่จะใช้ในการประมาณค่าสูญหายและมีขั้นตอนการคำนวณที่ไม่ซับซ้อน สะดวกในการใช้งานจริง ดังนั้นผู้วิจัยจึงสนใจศึกษาวิธีการประมาณค่าสูญหายแบบรวม 2 วิธี ที่ใช้ฟังก์ชันการถ่วงน้ำหนักด้วยวิธี EW คือ วิธี K-nearest regression imputation with equivalent weighted (KREW) เป็นวิธีประมาณค่าสูญหายที่ได้จากการรวมวิธีการประมาณค่าสูญหาย 2 วิธี ได้แก่ วิธี KNN และวิธี RI แล้วนำมาถ่วงน้ำหนักด้วยวิธี EW ส่วนวิธีที่ 2 คือ วิธี K-nearest stochastic regression imputation with equivalent weighted (KSEW) เป็นวิธีประมาณค่าสูญหายที่ได้จากการรวมวิธีการประมาณค่าสูญหาย 2 วิธี ได้แก่ วิธี KNN และวิธี stochastic regression imputation (SRI) แล้วนำมาถ่วงน้ำหนักด้วยวิธี EW ดังนั้นในงานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหายของตัวแปรตาม จำนวน 6 วิธี ได้แก่ วิธี RI วิธี SRI วิธี KNN วิธี EM algorithm (EM) วิธี KREW และวิธี KSEW ในกรณีที่ตัวแปรตามมีการสูญหายแบบสุ่มซึ่งเป็นการสูญหายที่มีความน่าจะเป็นขึ้นอยู่กับตัวแปรตัวอื่นที่ทราบค่าหรือสามารถทำนายตัวแปรอื่นได้ โดยเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error, MSE)

2. วิธีการวิจัย

การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายสำหรับการวิเคราะห์การถดถอยพหุ เมื่อตัวแปรตามมีการสูญหายอย่างสุ่ม โดยการจำลองข้อมูล

ด้วยวิธีมอนติคาร์โล ภายใต้สถานการณ์ที่แตกต่างกัน ดังนี้

2.1 กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติ (normal distribution) ในการศึกษานี้จะใช้ตัวแปรอิสระ 5 ตัวแปรโดยที่ $x_p \sim N(0,1)$ เมื่อ $p = 1, \dots, 5$

2.2 กำหนดค่าความคลาดเคลื่อนให้มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 0 ส่วนเบี่ยงเบนมาตรฐานเท่ากับ 5, 10 และ 15 ตามลำดับ

2.3 สร้างตัวแปรตามที่มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ ซึ่งใช้รูปแบบความสัมพันธ์เชิงเส้นดังนี้

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$
เมื่อ $i = 1, \dots, m, m+1, \dots, n$; กำหนดให้ ค่าสัมประสิทธิ์การถดถอย $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 1$ เพื่อควบคุมความแปรปรวนของตัวแปรตามที่สร้างขึ้นมาจากตัวแปรอิสระแต่ละตัวที่ใช้ในงานวิจัยให้มีขนาดเท่ากัน โดยควรกำหนดค่าสัมประสิทธิ์การถดถอยให้มีค่าระหว่างอยู่ระหว่าง -1 ถึง 1 ทั้งนี้เพื่อให้เกิดค่าความคลาดเคลื่อนของตัวประมาณน้อยที่สุด [4] โดยที่ ตัวแปรอิสระไม่มีการสูญหาย $y_1, \dots, y_m$ เป็นข้อมูลของตัวแปรตามที่ไม่มีการสูญหาย และ $y_{m+1}, \dots, y_n$ เป็นข้อมูลของตัวแปรตามที่มีการสูญหาย

2.4 กำหนดให้ตัวแปรตามมีการสูญหายแบบสุ่ม โดยมีเปอร์เซ็นต์การสูญหายของตัวแปรตาม 4 ระดับ คือ 10, 20, 30 และ 40 %

2.5 กำหนดขนาดตัวอย่างเท่ากับ 20, 30, 50 และ 100

2.6 วิธีการประมาณค่าสูญหายที่ศึกษาในงานวิจัยประกอบด้วย 6 วิธี ดังนี้

2.6.1 วิธี RI เป็นวิธีการที่ใช้สมการถดถอยประมาณค่าสูญหาย โดยใช้ชุดข้อมูลที่ทราบค่า (x_i, y_i) ; $i = 1, 2, \dots, m$ นำมาคำนวณค่าสัมประสิทธิ์

การถดถอยด้วยวิธีกำลังสองน้อยที่สุด เพื่อมาประมาณค่าสูญหายของตัวแปรตาม $(\hat{y}_j)$; $j = m + 1, \dots, n$

2.6.2 วิธี SRI [5] เป็นวิธีการที่ใช้สมการถดถอย จากชุดข้อมูลที่ทราบค่า (x_i, y_i) ; $i = 1, 2, \dots, m$ โดยจะแตกต่างจากวิธี RI ที่ในการประมาณค่าสูญหายจะเพิ่มเทอมความคลาดเคลื่อนสุ่มเข้ามาในสมการถดถอย เพื่อมาประมาณค่าสูญหายของตัวแปรตาม (y_j^*) ; $j = m + 1, \dots, n$

2.6.3 วิธี KNN เป็นการประมาณค่าสูญหายด้วยค่าเฉลี่ยของข้อมูลที่ทราบค่า จำนวน K ตัว ที่ลักษณะของตัวแปรที่ไม่เกิดการสูญหายมีความคล้ายคลึงกับหน่วยตัวอย่างที่เกิดค่าสูญหายมากที่สุด ซึ่งโดยทั่วไปจะกำหนดให้ $K \approx \sqrt{m}$ โดย K เป็นจำนวนเต็มที่มีค่าใกล้เคียงกับ $\sqrt{m}$ มากที่สุด เมื่อ m เป็นจำนวนข้อมูลที่สมบูรณ์ [6] กำหนดให้ $\tilde{y}_j$ โดยที่ $j = m+1, \dots, n$ แทนค่าประมาณของข้อมูลตัวแปรตามที่สูญหายด้วยวิธี KNN พิจารณาจากระยะห่างยูคลิด (Euclidean Distance) ของตัวแปรที่ไม่เกิดการสูญหาย [7] ซึ่งหาได้จากสมการต่อไปนี้

$$D_{ij} = \sqrt{\sum_{p=1}^5 (x_{ip} - x_{jp})^2}$$

เมื่อ $i = 1, 2, \dots, m$ และ $j = m+1, \dots, n$ โดยวิธี KNN มีขั้นตอนดังนี้ (1) คำนวณหาค่า K ตัว และหาค่าระยะห่างยูคลิด D_{ij} ที่มีค่าต่ำที่สุดจำนวน K ตัว สำหรับแต่ละชุดตัวอย่างที่ j (2) คำนวณหาค่าเฉลี่ยของตัวแปรตาม y_i สอดคล้องกับค่าต่ำสุด K ตัว ของชุดตัวอย่างที่ j โดยกำหนดให้เป็น $\bar{y}_j^*$ และ (3) จะได้ค่าประมาณของข้อมูลตัวแปรตามที่สูญหายวิธี KNN คือ $\tilde{y}_j = \bar{y}_j^*$

2.6.4 วิธี EM เป็นกระบวนการวนซ้ำเพื่อใช้สำหรับหาค่าประมาณภาวะน่าจะเป็นสูงสุด (maximum likelihood) ของพารามิเตอร์ เมื่อมีข้อมูล

บางส่วนเกิดการสูญหาย การประมาณค่าสูญหายสามารถแบ่งออกเป็น 2 ขั้นตอน คือ ขั้นตอน E-step เป็นขั้นตอนในการหาค่าคาดหวังของค่าข้อมูลสูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ไม่มีการสูญหายและพารามิเตอร์ตัวปัจจุบัน เพื่อนำค่าคาดหวังที่ได้ไปประมาณค่าข้อมูลที่สูญหาย และขั้นตอน M-step เป็นขั้นตอนการประมาณค่าพารามิเตอร์ที่จะเป็นสูงสุดของพารามิเตอร์ จากข้อมูลที่ไม่มีการสูญหาย ซึ่งจะได้การแทนค่าข้อมูลสูญหายที่ได้จากการประมาณในขั้น E-step และในปีค.ศ. 2002 Little และ Rubin [8] ได้ประยุกต์วิธี EM โดยการประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุ

จัดชุดข้อมูลให้อยู่ในรูป $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

โดยแบ่งข้อมูลของตัวแปรตามออกเป็น 2 ส่วน คือ ส่วนที่มีข้อมูลสมบูรณ์และส่วนที่มีข้อมูลไม่สมบูรณ์ ดังนี้

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

เมื่อ $\mathbf{y}_1$ เป็นเวกเตอร์ของตัวแปรตามที่ไม่มีการสูญหาย ขนาด $m \times 1$; $\mathbf{y}_2$ เป็นเวกเตอร์ของตัวแปรตามที่มีการสูญหายขนาด $(n-m) \times 1$; $\mathbf{X}_1$ เป็นเมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามทราบค่าขนาด $m \times (p+1)$; $\mathbf{X}_2$ เป็นเมทริกซ์ของตัวแปรอิสระที่ชุดข้อมูลของตัวแปรตามที่มีการสูญหายขนาด $(n-m) \times (p+1)$; $\boldsymbol{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ขนาด $(p+1) \times 1$; $\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อน ขนาด $n \times 1$

มีขั้นตอนดังนี้

(1) ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดจากชุดข้อมูลที่ไม่มีการสูญหาย จะได้ค่าสัมประสิทธิ์การถดถอยตัวเริ่มต้น $\hat{\boldsymbol{\beta}}^{(0)}$

(2) เข้าสู่ขั้นตอน E-step โดยการนำค่า

$\hat{\boldsymbol{\beta}}^{(0)}$ ที่ได้จากขั้นตอนที่ 1 มาหาค่าคาดหวัง เพื่อประมาณค่าข้อมูลสูญหายของตัวแปรตามทีสูญหาย ในการทำซ้ำรอบที่ 1 ดังนี้

$$E(\mathbf{y}_i | \mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)}) = \begin{cases} \mathbf{y}_i & ; i=1, 2, \dots, m \\ \hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \hat{\beta}_2^{(0)} x_{i2} + \dots + \hat{\beta}_p^{(0)} x_{ip} & ; i=m+1, \dots, n \end{cases}$$

(3) เข้าสู่ขั้นตอน M-step โดยการแทนค่าข้อมูลสูญหายจากค่าที่ประมาณได้ในขั้นตอนที่ 2 แล้วนำมาคำนวณหาค่าประมาณสัมประสิทธิ์การถดถอยในการทำซ้ำรอบที่ 1 ($\hat{\boldsymbol{\beta}}^{(1)}$)

(4) ประมาณค่าสูญหายใหม่อีกรอบซึ่งกลับเข้าสู่ขั้นตอน E-step ในการทำซ้ำรอบที่ t ; $t = 2, 3, \dots$

โดยใช้ค่าสัมประสิทธิ์การถดถอยใหม่หาคำนวณได้จากการแทนค่าข้อมูลสูญหายแล้ว แล้วเข้าสู่ขั้นตอน M-step รอบที่ t คำนวณหาค่าสัมประสิทธิ์การถดถอยใหม่ ทำวนซ้ำเช่นนี้ไปเรื่อย ๆ จนกระทั่งค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยมีค่าน้อยกว่าหรือเท่ากับ 0.001 จึงหยุด

2.6.5 วิธี KREW เป็นวิธีที่ประยุกต์การใช้ตัวถ่วงน้ำหนักด้วยวิธี EW ให้กับค่าประมาณที่ได้จากวิธีการประมาณค่าสูญหาย 2 วิธี ได้แก่ วิธี KNN และวิธี RI โดยมีขั้นตอนดังนี้

(1) ประมาณค่าสูญหายด้วยวิธี KNN และวิธี RI จากข้อมูลที่ทราบค่า

(2) ถ่วงน้ำหนักด้วยวิธีการให้น้ำหนักที่เท่ากัน ให้กับค่าประมาณที่ได้จากวิธีการประมาณค่าสูญหายวิธี KNN และวิธี RI ได้ดังนี้

$$\hat{\mathbf{y}}_j = \mathbf{W}_M (\hat{\mathbf{y}}_j + \tilde{\mathbf{y}}_j)$$

เมื่อวิธีการให้น้ำหนักที่เท่ากัน [9] คำนวณได้จาก

$$\mathbf{W}_M = \frac{1}{M} ; \mathbf{W}_M \text{ คือ ตัวถ่วงน้ำหนักของวิธีการ}$$

ประมาณค่าสูญหาย M วิธี โดยกำหนดให้ $\hat{\mathbf{y}}_j$; $j = m+1, \dots, n$ คือ ตัวประมาณค่าสูญหายของวิธี KREW

2.6.6 วิธี KSEW เป็นวิธีที่ประยุกต์การใช้ตัวถ่วงน้ำหนักด้วยวิธี EW ให้กับค่าประมาณที่ได้จากวิธีการประมาณค่าสูญหาย 2 วิธี ได้แก่ วิธี KNN และวิธี SRI โดยมีขั้นตอนดังนี้

(1) ประมาณค่าสูญหายด้วยวิธี KNN และวิธี SRI จากข้อมูลที่ทราบค่า

(2) ถ่วงน้ำหนักด้วยวิธีการให้น้ำหนักที่เท่ากัน ให้กับค่าประมาณที่ได้จากวิธีการประมาณค่าสูญหายวิธี KNN และวิธี SRI ได้ดังนี้

$$\tilde{y}_j^* = W_M(y_j^* + \tilde{y}_j)$$

เมื่อวิธีการให้น้ำหนักที่เท่ากัน [9] คำนวณได้จาก

$$W_M = \frac{1}{M}; W_M \text{ คือ ตัวถ่วงน้ำหนักของวิธีการ}$$

ประมาณค่าสูญหาย M วิธี โดยกำหนดให้ $\tilde{y}_j^*$; $j = m+1, \dots, n$ คือ ตัวประมาณค่าสูญหายของวิธี KSEW

2.7 ประมาณค่าสัมประสิทธิ์การถดถอยที่มีข้อมูลครบสมบูรณ์จากการแทนค่าข้อมูลสูญหายจากค่าประมาณที่ได้ในแต่ละวิธีการประมาณค่าสูญหายด้วยวิธีกำลังสองน้อยที่สุดเพื่อสร้างสมการถดถอยเชิงเส้นพหุ

2.8 คำนวณค่า MSE จากวิธีการประมาณค่าสูญหายทั้ง 6 วิธี โดยการทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์ เพื่อเปรียบเทียบประสิทธิภาพการประมาณค่าสูญหายจากค่า MSE ของแต่ละวิธีการประมาณค่าสูญหาย โดยวิธีใดที่ให้ค่า MSE ต่ำที่สุดจะเป็นวิธีที่มีประสิทธิภาพในการประมาณค่าสูญหายดีที่สุด

3. ผลการวิจัยและวิจารณ์

3.1 ผลการวิจัย

เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 ขนาดตัวอย่างเท่ากับ 20 และ 30 ที่ทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูลวิธีที่ให้ค่า

MSE ต่ำที่สุดเกือบทุกสถานการณ์ คือ วิธี KSEW ยกเว้นที่ขนาดตัวอย่าง 30 เมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเป็น 10 % และเมื่อขนาดตัวอย่าง 50 และ 100 พบว่าวิธี SRI มีแนวโน้มให้ค่า MSE ต่ำที่สุดเกือบทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ขนาดตัวอย่าง 50 เมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเป็น 30 และ 40 % แสดงดังตารางที่ 1 และจากรูปที่ 1 และ 2 เมื่อขนาดตัวอย่างเท่ากับ 20 และ 30 ตามลำดับ พบว่าจากการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 พิจารณาที่ขนาดตัวอย่างเดียวกัน เปอร์เซ็นต์การสูญหายเพิ่มขึ้นพบว่าค่า MSE ของวิธีการประมาณค่าสูญหายทุกวิธีมีแนวโน้มที่เพิ่มขึ้น เช่น เมื่อพิจารณาที่ขนาดตัวอย่างเท่ากับ 20 และ 30

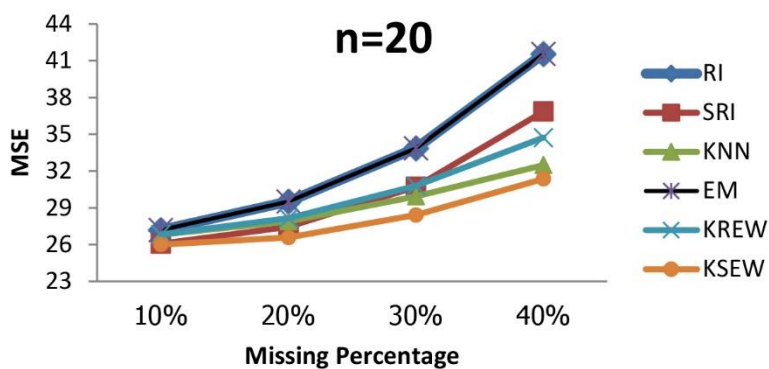
นอกจากนี้พบว่าทุกสถานการณ์ของการเปรียบเทียบค่า MSE ของวิธีการประมาณค่าสูญหายที่ทุกค่าของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนทุกวิธีการประมาณค่าสูญหายจะมีแนวโน้มให้ค่า MSE เพิ่มขึ้นเช่นกัน จึงแสดงกราฟเพียงกรณีค่าความคลาดเคลื่อนเท่ากับ 5 เท่านั้น

เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 ขนาดตัวอย่างเท่ากับ 20 และ 30 ที่ทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูลวิธีที่ให้ค่า MSE ต่ำที่สุดเกือบทุกสถานการณ์ คือ วิธี KSEW ยกเว้นที่ขนาดตัวอย่าง 30 เมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเป็น 10 % และเมื่อขนาดตัวอย่าง 50 และ 100 พบว่าวิธี SRI มีแนวโน้มให้ค่า MSE ต่ำที่สุดเกือบทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูล ยกเว้นที่ขนาดตัวอย่าง 50 เมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเป็น 30 และ 40 % และที่ขนาดตัวอย่าง 100 เมื่อเปอร์เซ็นต์การสูญหายของข้อมูลเป็น 40 % แสดงดังตารางที่ 2

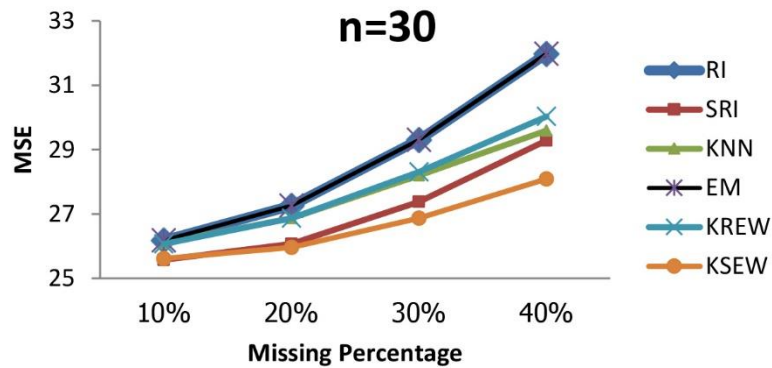
ตารางที่ 1 ค่า MSE ของวิธีการประมาณค่าสูญหาย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 จำแนกตามขนาดตัวอย่าง (n)

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	วิธีการประมาณค่าสูญหาย					
		RI	SRI	KNN	EM	KREW	KSEW
20	10	27.184	26.077	26.809	27.184	26.821	25.995
	20	29.502	27.454	27.888	29.502	28.147	26.572
	30	33.860	30.719	29.921	33.860	30.804	28.415
	40	41.565	36.857	32.487	41.565	34.732	31.358
30	10	26.177	25.577	26.096	26.177	26.069	25.615
	20	27.262	26.071	26.872	27.262	26.867	25.957
	30	29.299	27.381	28.191	29.299	28.308	26.870
	40	31.979	29.284	29.593	31.979	30.028	28.088
50	10	25.381	25.041	25.388	25.381	25.355	25.096
	20	26.118	25.398	26.058	26.118	26.008	25.484
	30	26.623	25.690	26.473	26.623	26.377	25.634
	40	27.953	26.559	27.416	27.953	27.367	26.319
100	10	25.311	25.157	25.327	25.311	25.308	25.191
	20	25.508	25.190	25.531	25.508	25.486	25.251
	30	25.708	25.213	25.755	25.708	25.657	25.290
	40	26.318	25.667	26.321	26.318	26.189	25.701

หมายเหตุ : ตัวหนาหมายถึง MSE ต่ำที่สุด



รูปที่ 1 ค่า MSE ของวิธีการประมาณค่าสูญหาย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 ขนาดตัวอย่าง (n) เท่ากับ 20



รูปที่ 2 ค่า MSE ของวิธีการประมาณค่าสูญหาย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 5 ขนาดตัวอย่าง (n) เท่ากับ 30

ตารางที่ 2 ค่า MSE ของวิธีการประมาณค่าสูญหาย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 10 จำแนกตามขนาดตัวอย่าง (n)

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	วิธีการประมาณค่าสูญหาย					
		RI	SRI	KNN	EM	KREW	KSEW
20	10	105.719	101.592	103.485	105.719	104.063	101.116
	20	119.549	111.285	111.681	119.549	113.710	107.398
	30	133.323	119.875	116.559	133.323	121.205	111.391
	40	170.495	152.523	127.784	170.495	139.419	126.121
30	10	102.939	100.525	102.324	102.939	102.411	100.603
	20	108.882	103.834	106.852	108.882	107.226	103.363
	30	117.198	109.700	111.740	117.198	113.134	107.478
	40	126.079	115.742	114.444	126.079	117.684	110.201
50	10	102.235	100.865	102.134	102.235	102.104	101.076
	20	104.004	101.170	103.605	104.004	103.578	101.430
	30	107.606	103.574	106.317	107.606	106.484	103.424
	40	113.356	107.748	109.804	113.356	110.578	106.402
100	10	100.647	99.980	100.638	100.647	100.614	100.119
	20	101.665	100.341	101.583	101.665	101.543	100.560
	30	102.568	100.614	102.294	102.568	102.249	100.788
	40	104.367	102.003	103.784	104.367	103.753	101.939

หมายเหตุ : ตัวหนาหมายถึง MSE ต่ำที่สุด

เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15 ขนาดตัวอย่างเท่ากับ 20 และ 30 ที่ทุกระดับเปอร์เซ็นต์การสูญเสียของข้อมูลวิธีที่ให้ค่า MSE ต่ำที่สุดเกือบทุกสถานการณ์ คือ วิธี KSEW ยกเว้นที่ขนาดตัวอย่าง 30 เมื่อเปอร์เซ็นต์การสูญเสียของข้อมูลเป็น 10 % และเมื่อขนาดตัวอย่าง 50 และ

100 พบว่าวิธี SRI มีแนวโน้มให้ค่า MSE ต่ำที่สุดเกือบทุกระดับเปอร์เซ็นต์การสูญเสียของข้อมูล ยกเว้นที่ขนาดตัวอย่าง 50 เมื่อเปอร์เซ็นต์การสูญเสียของข้อมูลเป็น 30 และ 40 % และที่ขนาดตัวอย่าง 100 เมื่อเปอร์เซ็นต์การสูญเสียของข้อมูลเป็น 40 % แสดงดังตารางที่ 3

ตารางที่ 3 ค่า MSE ของวิธีการประมาณค่าสูญเสีย 6 วิธี เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 15 จำแนกตามขนาดตัวอย่าง (n)

ขนาด ตัวอย่าง	เปอร์เซ็นต์การสูญเสีย	วิธีการประมาณค่าสูญเสีย					
		RI	SRI	KNN	EM	KREW	KSEW
20	10	244.429	233.671	239.159	244.429	240.534	232.794
	20	258.718	240.069	245.685	258.718	249.059	234.989
	30	312.619	283.987	268.435	312.619	280.902	259.197
	40	380.349	339.608	287.400	380.349	314.274	284.541
30	10	231.945	226.452	230.941	231.945	231.023	226.814
	20	244.619	233.269	240.052	244.619	241.107	232.584
	30	255.764	239.889	243.676	255.764	246.886	234.756
	40	279.541	257.699	255.989	279.541	262.329	245.601
50	10	228.549	225.307	228.212	228.549	228.217	225.808
	20	233.735	227.532	232.537	233.735	232.657	228.031
	30	241.515	232.255	238.119	241.515	238.829	231.828
	40	253.242	240.467	245.432	253.242	247.412	238.002
100	10	225.074	223.723	225.051	225.074	225.005	223.991
	20	228.331	225.425	228.047	228.331	228.028	225.860
	30	232.164	227.700	231.276	232.164	231.386	228.066
	40	235.084	229.391	233.133	235.084	233.454	229.207

หมายเหตุ : ตัวหนาหมายถึง MSE ต่ำที่สุด

3.2 วิจัยรณ

การเปรียบเทียบค่า MSE ของวิธีการประมาณค่าสูญเสียทั้ง 6 วิธี พบว่าเกือบทุกระดับเปอร์เซ็นต์การสูญเสียของข้อมูล ในกรณีที่ขนาด

ตัวอย่างเท่ากับ 20 และ 30 วิธี KSEW มีแนวโน้มให้ค่า MSE ต่ำที่สุด เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 พบว่าวิธี SRI จะมีแนวโน้มให้ค่า MSE ต่ำที่สุด และในทุกสถานการณ์วิธีสมการถดถอย และวิธี EM

algorithm มีประสิทธิภาพในการประมาณค่าสูญหาย ไม่แตกต่างกัน ซึ่งสอดคล้องกับงานวิจัยของวารุณีและเพียงอ [10,11] นอกจากนี้วิธีการประมาณค่าสูญหายทั้ง 6 วิธี จะมีประสิทธิภาพในการประมาณค่าสูญหายลดลง เมื่อเปอร์เซ็นต์การสูญหายเพิ่มขึ้นและค่าส่วน

เบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้น แต่พบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้นวิธีการประมาณค่าสูญหายทุกวิธีจะมีประสิทธิภาพในการประมาณค่าสูญหายเพิ่มขึ้น เมื่อพิจารณาที่เปอร์เซ็นต์การสูญหายระดับเดียวกัน

ตารางที่ 4 วิธีการประมาณค่าสูญหายที่ให้ค่า MSE ต่ำที่สุดในแต่ละสถานการณ์ของการประมาณค่าสูญหายของตัวแปรตาม

ขนาดตัวอย่าง	เปอร์เซ็นต์การสูญหาย	ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนกลุ่ม		
		5	10	15
20	10	KSEW	KSEW	KSEW
	20	KSEW	KSEW	KSEW
	30	KSEW	KSEW	KSEW
	40	KSEW	KSEW	KSEW
30	10	SRI	SRI	SRI
	20	KSEW	KSEW	KSEW
	30	KSEW	KSEW	KSEW
	40	KSEW	KSEW	KSEW
50	10	SRI	SRI	SRI
	20	SRI	SRI	SRI
	30	KSEW	KSEW	KSEW
	40	KSEW	KSEW	KSEW
100	10	SRI	SRI	SRI
	20	SRI	SRI	SRI
	30	SRI	SRI	SRI
	40	SRI	KSEW	KSEW

4. สรุป

ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย 6 วิธี สำหรับกรณีวิเคราะห์การถดถอยเมื่อตัวแปรตามมีการสูญหายอย่างสุ่ม โดยพิจารณาจากค่า MSE พบว่าทุกค่าส่วนเบี่ยงเบน

มาตรฐานของความคลาดเคลื่อนที่ขนาดตัวอย่าง 20 และ 30 เกือบทุกระดับเปอร์เซ็นต์การสูญหายของตัวแปรตาม วิธี KSEW มีประสิทธิภาพในการประมาณค่าสูญหายดีที่สุด ยกเว้นที่ขนาดตัวอย่าง 30 เมื่อเปอร์เซ็นต์การสูญหาย 10 % วิธี SRI มีประสิทธิภาพ

ในการประมาณค่าสูญหายดีกว่าวิธี KSEW และเมื่อขนาดตัวอย่าง 50 วิธี SRI มีประสิทธิภาพดีที่สุดเมื่อเปอร์เซ็นต์การสูญหายเท่ากับ 10 และ 20 % และเมื่อขนาดตัวอย่างมีขนาดใหญ่ 100 พบว่าวิธี SRI มีประสิทธิภาพดีที่สุดทุกระดับเปอร์เซ็นต์การสูญหาย แต่เมื่อค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้นเป็น 10 และ 15 ที่เปอร์เซ็นต์การสูญหายเท่ากับ 40 % วิธี KSEW จะมีประสิทธิภาพดีกว่าวิธี SRI นอกจากนี้ยังพบว่าวิธีการประมาณค่าสูญหายทั้ง 6 วิธี จะมีประสิทธิภาพในการประมาณค่าสูญหายดีขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้นที่เปอร์เซ็นต์การสูญหายของตัวแปรตามเดียวกันและประสิทธิภาพในการประมาณค่าสูญหายลดลงเมื่อเปอร์เซ็นต์การสูญหายของตัวแปรตามและค่าส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเพิ่มขึ้นที่ขนาดตัวอย่างเดียวกัน รายละเอียดแสดงดังตารางที่ 4

5. ข้อเสนอแนะ

จากงานวิจัยนี้ ผู้วิจัยได้กำหนดให้ตัวแปรตามมีการสูญหายอย่างสุ่ม พบว่าในกรณีที่ขนาดตัวอย่างปานกลาง (20, 30) ทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูล ผู้ศึกษาอาจเลือกใช้วิธี KSEW และโดยส่วนใหญ่ที่เกือบทุกระดับเปอร์เซ็นต์การสูญหายของข้อมูลอาจเลือกใช้วิธี SRI เมื่อมีขนาดตัวอย่างใหญ่ (50, 100) ซึ่งจากงานวิจัยนี้ผู้ที่สนใจอาจจะศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหาย ภายใต้สถานการณ์อื่น ๆ ต่อไป เช่น ศึกษาในกรณีที่สมการถดถอยพหุมีข้อมูลสูญหายเกิดขึ้นกับตัวแปรตามและตัวแปรอิสระบางตัว

6. รายการอ้างอิง

- [1] ไกรรุ่ง เสงพระพรหม, 2553, การเลือกคุณลักษณะสำหรับการพยากรณ์ค่าที่ได้ขาดหายสำหรับข้อมูลหลายมิติและการประยุกต์สำหรับข้อมูลไมโครอาร์เรย์, วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 107 น.
- [2] จิรกานต์ นวลละออง, 2552, การเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับตัวแบบพยากรณ์, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 137 น.
- [3] ศศิธร สมพวงค์นวกิจ, 2555, การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบร่วม, วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ, 88 น.
- [4] Bolch, B.W. and Huang, C.J., 1974, Multivariate Statistical Method for Business and Economics, Prentice-Hall, Inc., Englewood Cliffs, NJ, 329 p.
- [5] Chaimongkol, W., 2005, Three Composite Imputation Methods for Item Nonresponse Estimation in Sample Surveys, Ph.D. Thesis, National Institute of Development Administration, Bangkok, 141 p.
- [6] Jönsson, P. and Wohlin, C., 2004, An Evaluation of K-Nearest Neighbor Imputation Using Likert Data, pp. 1530-1435, International Symposium on Software Metrics (METRICS'04), Chicago, Illinois.
- [7] อุษณีย์ วงศ์อำมาตย์, 2555, การเปรียบเทียบวิธีการประมาณค่าสูญหายแบบนอนอินกอร์เรเบิลในการวิเคราะห์การถดถอยเส้นพหุ, วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย

- ลัย, กรุงเทพฯ, 93 น.
- [8] Little, R.J.A. and Rubin, D.B., 2002, Statistical Analysis with Missing Data, 4th Ed., John Wiley & Sons, Inc., New York.
- [9] เสาวณิต สุขภารังษี, 2546, การใช้เทคนิคการพยากรณ์ร่วมด้วยตัวถ่วงน้ำหนัก, ว.พัฒนาเทคนิคศึกษา 15(16): 60-65.
- [10] วารุณี ตรีบำรุงศักดิ์, 2538, การพยากรณ์ด้วย
- วิธีการถดถอยเชิงเส้นพหุเมื่อตัวแปรตามมีค่าสูญหาย, วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 158 น.
- [11] เพียงอ อธิสา, 2551, การเปรียบเทียบวิธีประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น, วิทยานิพนธ์ปริญญาโท, จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 106 น.