

การเปรียบเทียบการประมาณข้อมูลสูญหายในการถดถอยเชิงเส้นอย่างง่าย
ระหว่างวิธีการของซิงห์และวิธีค่าคาดหวังสูงสุด
COMPARISON OF MISSING DATA ESTIMATION IN SIMPLE LINEAR
REGRESSION BETWEEN SINGH AND EXPECTATION
MAXIMIZATION ALGORITHM

พิชชา เครือแปง¹ บัณฑิตา พลับอินทร์^{2*} พุฒิพงษ์ พุกกะมาน² และมานะชัย รอดชื่น²
Phitcha Khrueapaeng¹, Bundhita Plubin^{2*}, Putipong Bookkamana²,
and Manachai Rodchuen²

¹Graduate School, Chiang Mai University

²Faculty of Science, Chiang Mai University

*corresponding author email: bandhita2459@gmail.com

บทคัดย่อ

การศึกษานี้มีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณข้อมูลสูญหายในการถดถอยเชิงเส้นอย่างง่ายประกอบไปด้วยตัวแปรตาม Y และตัวแปรอิสระ X โดยทำการประมาณค่าสูญหายด้วยวิธีการของซิงห์ และวิธีค่าคาดหวังสูงสุด (EM) โดยการจำลองข้อมูลซึ่งกำหนดขนาดตัวอย่าง 40, 100, 500 และ 1,000 ความแปรปรวนของตัวแปรอิสระเท่ากับ 1, 10 และ 50 ร้อยละการสูญหายของตัวแปรตามและตัวแปรอิสระเท่ากับ 5, 10 และ 15 สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเท่ากับ -0.3, -0.6, -0.9, 0.3, 0.6 และ 0.9 และข้อมูลมีการแจกแจงแบบปกติ เกณฑ์ในการเปรียบเทียบประสิทธิภาพในการประมาณค่าสูญหายจะใช้ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง (Root Mean Square Error: RMSE) จากผลการศึกษาสรุ้ได้ว่า วิธี EM เป็นวิธีการประมาณค่าสูญหายที่ดีกว่าวิธีการของซิงห์ เนื่องจากเป็นวิธีที่มีค่า RMSE ต่ำที่สุดในทุกระดับของสัมประสิทธิ์สหสัมพันธ์ ขนาดตัวอย่าง ความแปรปรวน และร้อยละการสูญหาย

คำสำคัญ: วิธีการของซิงห์ วิธีค่าคาดหวังสูงสุด ค่าความคลาดเคลื่อนเฉลี่ยกำลังสอง

Abstract

This study was focus on comparing the estimation methods for missing data in simple linear regression. The methods that used to estimate missing data are Singh method and Expectation Maximization Algorithm (EM). The comparison was done under condition of sample sizes 40, 100, 500 and 1,000; variances 1, 10 and 50; percentages of missing data 5%, 10% and 15%; the correlation coefficient levels between the dependent and independent variable are -0.3, -0.6, -0.9, 0.3, 0.6 and 0.9. The criterion of determination is Root Mean Square Error (RMSE). The results show that the EM method is a better estimation method than Singh method for simple linear

regression due to EM method give the lowest RMSE values for all levels of correlation coefficients, sample sizes, variances and percentages of missing data.

Keywords: Singh method, expectation maximization algorithm, root mean square error

บทนำ

การวิเคราะห์การถดถอย (regression analysis) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ ซึ่งเป็นเทคนิคในการวิเคราะห์ข้อมูลทางสถิติ โดยนำไปใช้ในการศึกษาและการวิจัยในด้านต่าง ๆ เช่น ด้านวิทยาศาสตร์ ด้านสังคมศาสตร์ เป็นต้น ซึ่งจากการสำรวจหรือการเก็บรวบรวมข้อมูลจากการวิจัยมักเกิดปัญหาข้อมูลสูญหาย โดยปัญหาข้อมูลสูญหายจะเกิดในลักษณะต่างๆ เช่น การไม่ตอบของหน่วยตัวอย่างบางหน่วยตัวอย่าง (unit nonresponse) หรือการสูญหายของข้อมูลที่เกิดจากการไม่ตอบเฉพาะบางคำถาม (item nonresponse) เป็นต้น ซึ่งหากนำชุดข้อมูลที่มีข้อมูลสูญหายนำไปวิเคราะห์ จะทำให้เกิดผลการวิเคราะห์ที่ผิดพลาดและเกิดความคลาดเคลื่อนมากขึ้น (Wararit, 2009)

การเลือกใช้วิธีการใด ๆ นั้นขึ้นอยู่กับลักษณะข้อมูลสูญหายที่เกิดขึ้น โดยลักษณะของข้อมูลสูญหายแบ่งออกเป็น 3 ประเภท (Little & Rubin, 1987) ได้แก่ 1) Missing completely at random (MCAR) เป็นลักษณะข้อมูลสูญหายในตัวแปรตามที่ไม่มีความสัมพันธ์กับตัวแปรตามเองและไม่มีความสัมพันธ์กับตัวแปรอื่น 2) Missing at random (MAR) เป็นลักษณะข้อมูลสูญหายในตัวแปรตามที่ไม่มีความสัมพันธ์กับตัวแปรตามเอง แต่มีความสัมพันธ์กับตัวแปรอื่น 3) Not missing at random (NMAR) เป็นลักษณะข้อมูลสูญหายในตัวแปรตามที่มีความสัมพันธ์กับตัวแปรตามเอง แต่ไม่มีความสัมพันธ์กับตัวแปรอื่น

การจัดการกับข้อมูลสูญหาย เมื่อเกิดข้อมูลสูญหายในบางตัวแปร แบ่งได้ออกเป็น 2 กลุ่ม (Laaksonen, 2000) คือ 1) Model-donor Imputation คือ การประมาณค่าที่ได้มาจากตัวแบบ (Model) โดยตรง ได้แก่ Mean Imputation, Regression Imputation, Ratio Imputation และ Multiple Imputation เป็นต้น 2) Real-donor Imputation คือ การประมาณค่าที่ได้จากเซตข้อมูลของค่าที่สังเกตได้ ได้แก่ Cold Deck Imputation, Hot Deck Imputation และ Nearest Neighbor Imputation เป็นต้น

การวิจัยครั้งนี้เป็นการประมาณข้อมูลสูญหายในการถดถอยเชิงเส้นอย่างง่าย ($Y = X\beta + e$) ด้วยวิธีการของSingh และวิธี Expectation Maximization Algorithm (EM) โดยกำหนดให้มีขนาดตัวอย่างทั้งหมด n หน่วย ข้อมูลตัวแปรอิสระ X มีขนาดสมบูรณ์จำนวน n หน่วย ข้อมูลตัวแปรตาม Y มีขนาดไม่สมบูรณ์จำนวน r หน่วย จึงเกิดค่าสูญหายจำนวน $n - r$ หน่วย

ตัวแปรตาม	y_1	y_2	$\dots$	y_r	y_{r+1}	$\dots$	y_n
ตัวแปรอิสระ	x_1	x_2	$\dots$	x_r	x_{r+1}	$\dots$	x_n

กำหนด (x_i, y_i) แทนค่าตัวแปรอิสระ x_i และตัวแปรตาม y_i ; $i = 1, \dots, r$

$(x_{i'}, y_{i'})$ แทนค่าตัวแปรอิสระ $x_{i'}$ และตัวแปรตาม $y_{i'}$ ซึ่งมีค่าสูญหาย; $i' = r + 1, \dots, n$

วิธีดำเนินการวิจัย

1. วิธีการประมาณค่าสูญหายด้วยวิธีการของ Singh

Singh et al. (2010) ได้เสนอตัวประมาณค่าสูญหายที่พัฒนามาจากวิธีการประมาณของ Singh & Horn (2000) และ Ahmed et al. (2006) ซึ่งเป็นการหาค่าเฉลี่ยของประชากรของตัวแปรตามที่มีการสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Complete at Random: MCAR) และพัฒนาจากวิธีการประมาณอัตราส่วน (ratio estimators) ซึ่งเป็นวิธีการประมาณค่าตัวแปรตาม Y ที่สูญหายโดยอาศัยความสัมพันธ์ระหว่างตัวแปรอิสระ X และตัวแปรตาม Y ที่เก็บค่าได้จำนวน r หน่วย ดังนี้ $\hat{Y} = \bar{X} \left(\frac{\bar{Y}_r}{\bar{X}_r} \right)$ ซึ่งมีเงื่อนไขว่าต้องทราบค่าเฉลี่ยของตัวแปรอิสระ X ซึ่งความแปรปรวนของ $\bar{X}_r$ และ $\bar{Y}_r$ คือ $Var(\bar{X}_r) = \frac{\sigma_X^2}{r}$ และ $Var(\bar{Y}_r) = \frac{\sigma_Y^2}{r}$ และความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error : MSE) ของ $\hat{Y}$ คือ

$$MSE_{\hat{Y}} \cong \frac{1-f}{n} \left(\sigma_Y^2 + \sigma_X^2 \left(\frac{\bar{Y}_r}{\bar{X}_r} \right)^2 (1-2d) \right)$$

เมื่อ σ_X^2 แทนความแปรปรวนของตัวแปรอิสระ
 σ_Y^2 แทนความแปรปรวนของตัวแปรตาม
 d แทนค่าคงที่

สมการของ Singh เป็นการใช้อัตราส่วนของตัวแปรอิสระ X และตัวแปรตาม Y ที่เก็บค่าได้จำนวน r หน่วย โดยอาศัยสมมติฐานความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ และการชักตัวอย่าง โดยสมการของ Singh เป็นดังนี้

$$\hat{Y} = \begin{cases} y_i & \text{if } i \in R \\ \frac{1}{(n-r)} \bar{Y}_r \left[\frac{n((A+C)\bar{X} + fB\bar{X}_r)}{(A+fB)\bar{X} + C\bar{X}_r} \right] - r & \text{if } i \in R^c \end{cases}$$

เมื่อ R แทนชุดข้อมูลตัวแปรตามที่ไม่สูญหาย
 R^c แทนชุดข้อมูลตัวแปรตามที่มีค่าสูญหาย
 $\bar{X}$ แทนค่าเฉลี่ยของตัวแปรอิสระ n หน่วย
 $\bar{X}_r$ แทนค่าเฉลี่ยของตัวแปรอิสระ r หน่วย
 $\bar{Y}_r$ แทนค่าเฉลี่ยของตัวแปรตาม r หน่วย
 $f = \frac{n}{N}$ แทนสัดส่วนของขนาดตัวอย่างที่ถูกสุ่มจากประชากร
 $A = (d-1)(d-2)$, $B = (d-1)(d-4)$ และ
 $C = (d-2)(d-3)(d-4)$

แทนค่าคงที่ และค่า d มีค่ามากกว่า 0 ($d > 0$)

การหาค่า d

สามารถหาได้จากสมการดังนี้

$$-\hat{\rho}_{XY} \frac{C_Y}{C_X} = \frac{fB - C}{A + fB + C}$$

$$-\hat{\rho}_{XY} \frac{C_Y}{C_X} = \frac{\frac{n}{N}[(d-1)(d-4)] - [(d-2)(d-3)(d-4)]}{[(d-1)(d-2)] + \frac{n}{N}[(d-1)(d-4)] + [(d-2)(d-3)(d-4)]}$$

เมื่อ C_Y แทนสัมประสิทธิ์การแปรผันของข้อมูลของตัวแปรตาม
 C_X แทนสัมประสิทธิ์การแปรผันของข้อมูลของตัวแปรอิสระ
 $\hat{\rho}_{XY}$ แทนสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ

จากนั้นหาค่าคงที่ A, B, C โดยนำค่า d ที่ได้ แทนลงในสมการ ดังนี้

$$A = (d-1)(d-2)$$

$$B = (d-1)(d-4)$$

$$C = (d-2)(d-3)(d-4)$$

จากนั้นทำการแทนค่าตัวแปรตามที่สุดยหาย ดังนี้

$$\hat{Y}_{r,s} = T_{d1} = \bar{y}_r \left[\frac{(A+C)\bar{X} + fB\bar{X}_r}{(A+fB)\bar{X} + C\bar{X}_r} \right]$$

เมื่อ $\hat{Y}_{r,s}$ แทนค่าประมาณที่ได้จากวิธีการของ Singh
 $\bar{X}$ แทนค่าเฉลี่ยของตัวแปรอิสระ n หน่วย
 $\bar{X}_r$ แทนค่าเฉลี่ยของตัวแปรอิสระ r หน่วย
 $\bar{Y}_r$ แทนค่าเฉลี่ยของตัวแปรตาม r หน่วย

2. วิธีการประมาณค่าสูญหายด้วยวิธี Expectation Maximization Algorithm (EM)

Little & Rubin (1987) ได้ประยุกต์วิธี EM มาใช้ในการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้น โดยเป็นวิธีการที่พัฒนาเพิ่มเติมจากวิธีกำลังสองน้อยที่สุด (ordinary least square estimation) โดยจะทำการประมาณสัมประสิทธิ์การถดถอย β ซ้ำ ๆ จนพารามิเตอร์ β มีค่าที่คงที่

หลักการของวิธีการนี้

ประมาณสัมประสิทธิ์ β เริ่มต้นจากจำนวนข้อมูลที่เก็บค่าได้จำนวน r หน่วย โดยใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Square Estimation)

ขั้นตอน E-Step เป็นขั้นตอนที่ทำการหาสมการถดถอยด้วยการแทนค่าตัวแปรตามที่สุดยหาย ด้วยค่าสัมประสิทธิ์การถดถอย β ตัวปัจจุบัน

ขั้นตอน M-Step เป็นขั้นตอนที่นำสมการถดถอยที่ได้จากขั้นตอน E-Step นำไปประมาณสัมประสิทธิ์การถดถอย β ตัวถัดไปแล้วเปรียบเทียบกับ β ตัวก่อน

โดยทำซ้ำจนกระทั่งได้สมการถดถอยและค่าสัมประสิทธิ์การถดถอย β มีค่าที่คงที่

วิธีการดำเนินการ

1. เริ่มต้นจะนำข้อมูลตัวแปรอิสระและตัวแปรตามที่เก็บค่าได้ r หน่วย จากนั้นทำการประมาณพารามิเตอร์ β โดยวิธี OLS ดังนี้ $\tilde{\beta}^{(0)} = (X_r' X_r)^{-1} X_r' Y_r$

2. เข้าสู่ขั้นตอน E-Step ทำการประมาณค่าตัวแปรตามที่สูญหาย ในรอบที่ 1

$$Y^{(1)} = X_{n-r} \tilde{\beta}^{(0)} \text{ โดย } Y^{(1)} = \begin{bmatrix} y_r \\ M \\ \hat{y}_{n-r}^{(1)} \end{bmatrix}$$

3. เข้าสู่ขั้นตอน M-step ทำการประมาณสัมประสิทธิ์การถดถอย ในรอบที่ 1

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

4. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยเริ่มต้น ($\tilde{\beta}^{(0)}$) กับค่าสัมประสิทธิ์การถดถอยรอบที่ 1 ($\tilde{\beta}^{(1)}$) ดังนี้ $|\tilde{\beta}^{(1)} - \tilde{\beta}^{(0)}|$

5. เปรียบเทียบค่าในขั้นตอนที่ 4 ถ้าไม่มากกว่า 0.0001 ตัวแปรตามที่มีประมาณได้จะเป็นค่าประมาณค่าสูญหาย แต่ถ้ามากกว่า 0.0001 ให้ทำขั้นตอนต่อไป

6. เข้าสู่ขั้นตอน E-Step ในการทำซ้ำรอบที่ t ; $t = 2, 3, \dots$

$$Y^{(t)} = X_{n-r} \tilde{\beta}^{(t-1)} \text{ โดย } Y^{(t)} = \begin{bmatrix} y_r \\ M \\ \hat{y}_{n-r}^{(t)} \end{bmatrix}$$

7. เข้าสู่ขั้นตอน M-Step ในการทำซ้ำรอบที่ t ; $t = 2, 3, \dots$

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

8. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ $t-1$ ($\tilde{\beta}^{(t-1)}$) กับค่าสัมประสิทธิ์การถดถอยรอบที่ t ($\tilde{\beta}^{(t)}$) ดังนี้ $|\tilde{\beta}^{(t)} - \tilde{\beta}^{(t-1)}|$

9. เปรียบเทียบค่าในขั้นตอนที่ 8 ถ้าไม่มากกว่า 0.0001 ตัวแปรตามที่มีประมาณได้จะเป็นค่าประมาณค่าสูญหาย แต่ถ้ามากกว่า 0.0001 ให้ทำขั้นตอนที่ 6 ถึงขั้นตอนที่ 8 ทำซ้ำจนกระทั่งสัมประสิทธิ์การถดถอยตัวปัจจุบันคงที่ (จนกว่าจะผ่านเงื่อนไข)

ดังนั้น จะได้ค่าประมาณตัวแปรตามที่สูญหาย ดังนี้

$$\hat{Y}_{i',EM} = X_{n-r} \tilde{\beta}^{(t-1)}$$

เมื่อ $\hat{Y}_{i',EM}$ แทนเวกเตอร์ของตัวแปรตามที่สูญหายขนาด $r \times 1$

X_{n-r} แทนเมทริกซ์ของตัวแปรอิสระขนาด $r \times 2$

$\tilde{\beta}^{(t-1)}$ แทนเวกเตอร์ของค่าประมาณสัมประสิทธิ์การถดถอยรอบที่ $t-1$ ขนาด 2×1

3. เกณฑ์ในการเปรียบเทียบ

เกณฑ์ในการเปรียบเทียบว่าการประมาณค่าสูญหายด้วยวิธีใดใช้ได้ดีกว่า จะเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริงในรูปแบบรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error: RMSE) โดยสามารถคำนวณจากสมการดังนี้

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n-r} (y_{i'} - \hat{y}_{i'})^2}{n-r}}$$

เมื่อ $y_{i'}$ แทนค่าจริงของข้อมูลตัวแปรตามตัวที่ i'
 $\hat{y}_{i'}$ แทนค่าประมาณของข้อมูลตัวแปรตามตัวที่ i'
 $n-r$ แทนขนาดของตัวอย่างที่สูญหาย

ในการเปรียบเทียบวิธีการประมาณ ถ้าวิธีการประมาณใดให้ค่าเฉลี่ย RMSE ต่ำกว่า แสดงว่าเป็นวิธีการประมาณที่ดีกว่า

4. ข้อมูลที่ใช้ในการวิจัย

4.1 ประชากร

จำลองข้อมูลตัวแปรคู่ของประชากรขนาด N เท่ากับ 20,000 ซึ่งประกอบด้วยตัวแปรอิสระ X ซึ่งเป็นตัวแปรกำหนดและตัวแปรตาม Y ซึ่งเป็นตัวแปรสุ่มที่มีการแจกแจงปกติที่มีค่าความแปรปรวนเท่ากับ 1, 10 และ 50 ตามลำดับ โดยตัวแปรสุ่ม Y จะถูกสร้างให้มีความสัมพันธ์กับตัวแปรอิสระ X ที่ระดับสัมประสิทธิ์สหสัมพันธ์เท่ากับ -0.9, -0.6, -0.3, 0.3, 0.6 และ 0.9 ตามลำดับ

4.2 ตัวอย่าง

ชักตัวอย่างขนาด n เท่ากับ 0.2%, 0.5%, 2.5% และ 5% ของขนาดประชากรจากแต่ละประชากรด้วยวิธีชักตัวอย่างแบบสุ่มชนิดไม่ใส่คืน (Simple Random Sampling Without Replacement หรือ SRSWOR)

4.3 ข้อมูลสูญหาย

จำลองข้อมูลสูญหายโดยการสุ่มตำแหน่งการสูญหายของตัวแปรตามและกำหนดขนาดข้อมูลสูญหายเท่ากับ 5%, 10% และ 15% ของขนาดตัวอย่าง n โดยทำการสุ่มค่าสูญหายแต่ละขนาดจำนวน 1,000 รอบ

ผลการวิจัยและอภิปรายผล

จากการพิจารณาค่า RMSE ณ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X ที่ระดับ -0.3 จากวิธีการประมาณค่าสูญหายทั้ง 2 วิธี พบว่าค่า RMSE จะมีแนวโน้มเพิ่มขึ้น เมื่อข้อมูลมีความแปรปรวนเพิ่มขึ้น ขนาดตัวอย่างเพิ่มขึ้น และร้อยละการสูญหายของข้อมูลเพิ่มขึ้น

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และวิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า -0.3 โดยอาศัยสถิติทดสอบ t-test พบว่า ที่ระดับนัยสำคัญ 0.05 ในกรณีขนาดตัวอย่าง 40 ร้อยละการสูญหาย 5% ความแปรปรวน 1 และ 10 และกรณี ขนาดตัวอย่าง 40 ร้อยละการสูญหาย 10% ความแปรปรวน 10 วิธีการประมาณ 2 วิธีให้ค่า RMSE ไม่แตกต่างกันส่วนกรณีอื่นวิธีการประมาณของ 2 วิธีมีความแตกต่างกัน โดยกรณีตัวอย่างขนาดตั้งแต่ 100 ขึ้นไป วิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน และค่าร้อยละการสูญหาย ดังตารางที่ 1 (Table 1)

Table 1 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when -0.3 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	0.9308	0.8897	2.9191	2.6729	6.7926*	6.0855*
	10%	0.9762*	0.9396*	3.0103	2.8567	7.3930*	6.3517*
	15%	0.9801*	0.9344*	3.3113*	2.9025*	7.0144*	6.6400*
100	5%	0.9707*	0.9263*	3.0939*	2.8345*	7.2358*	6.4990*
	10%	1.0000*	0.9489*	3.1962*	2.9293*	7.4894*	6.6552*
	15%	1.0160*	0.9481*	3.2244*	2.9778*	7.2328*	6.6629*
500	5%	1.0333*	0.9483*	3.2097*	2.9776*	7.1508*	6.7531*
	10%	1.0564*	0.9566*	3.3943*	3.0032*	7.1120*	6.7152*
	15%	1.0409*	0.9546*	3.2578*	3.0120*	7.4965*	6.7542*
1,000	5%	1.0738*	0.9600*	3.2060*	3.0024*	7.1476*	6.7711*
	10%	1.0582*	0.9563*	3.5066*	2.9955*	7.5006*	6.7387*
	15%	1.0411*	0.9551*	3.7587*	3.0003*	7.1572*	6.7687*

Remark $p = -0.3$, *Sig. Level 0.05

จากการพิจารณาค่า RMSE ณ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X ที่ระดับ -0.6 จากวิธีการประมาณค่าสูญหายทั้ง 2 วิธี พบว่าค่า RMSE จะมีแนวโน้มเพิ่มขึ้น เมื่อข้อมูลมีความแปรปรวนเพิ่มขึ้น ขนาดตัวอย่างเพิ่มขึ้น และร้อยละการสูญหายของข้อมูลเพิ่มขึ้น

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และวิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า -0.6 โดยอาศัยสถิติทดสอบพบว่า ที่ระดับนัยสำคัญ 0.05 วิธีการประมาณ 2 วิธีให้ค่า RMSE แตกต่างกันทุกกรณีของระดับความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย โดยวิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย ดังตารางที่ 2 (Table 2)

Table 2 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when -0.6 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	0.9845*	0.7125*	2.9184*	2.1922*	8.4108*	4.9831*
	10%	1.0080*	0.7617*	3.1167*	2.4357*	8.5360*	5.3601*
	15%	0.9691*	0.7881*	3.1998*	2.4527*	8.6123*	5.5891*
100	5%	0.9793*	0.7676*	3.2234*	2.4115*	7.8720*	5.4985*
	10%	1.0011*	0.7865*	3.1763*	2.4754*	8.1899*	5.5077*
	15%	1.0144*	0.7895*	3.1692*	2.5119*	8.0892*	5.6366*

Table 2 (cant)

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
500	5%	1.0962*	0.7929*	3.2139*	2.4790*	7.0200*	5.5731*
	10%	1.0257*	0.8000*	3.4169*	2.4960*	7.0846*	5.6333*
	15%	1.0530*	0.7986*	3.2276*	2.5047*	7.0942*	5.6328*
1,000	5%	1.1039*	0.7901*	3.4556*	2.4985*	7.0532*	5.6203*
	10%	1.2718*	0.8009*	3.4110*	2.4962*	7.0592*	5.6238*
	15%	1.2215*	0.8004*	3.5052*	2.5070*	7.0653*	5.6218*

Remark $p = -0.6$, *Sig. Level 0.05

จากการพิจารณาค่า RMSE ณ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X ที่ระดับ -0.9 จากวิธีการประมาณค่าสูญหายทั้ง 2 วิธี พบว่าค่า RMSE จะมีแนวโน้มเพิ่มขึ้น เมื่อข้อมูลมีความแปรปรวนเพิ่มขึ้น ขนาดตัวอย่างเพิ่มขึ้น และร้อยละการสูญหายของข้อมูลเพิ่มขึ้น

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และวิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า -0.9 โดยอาศัยสถิติทดสอบ พบว่าที่ระดับนัยสำคัญ 0.05 วิธีการประมาณ 2 วิธีให้ค่า RMSE แตกต่างกันทุกกรณีของระดับความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย โดยวิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย ดังตารางที่ 3 (Table 3)

Table 3 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when -0.9 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	0.8811*	0.3887*	2.9115*	1.2378*	6.3578*	2.8564*
	10%	0.9443*	0.4228*	3.1175*	1.3044*	6.6519*	2.9611*
	15%	0.9583*	0.4238*	3.0791*	1.3322*	7.0761*	3.0143*
100	5%	0.9524*	0.4121*	3.2203*	1.3294*	6.7943*	2.9287*
	10%	0.9822*	0.4275*	3.1384*	1.3669*	6.9502*	3.0648*
	15%	1.0713*	0.4315*	3.1799*	1.3843*	7.0171*	3.0519*
500	5%	0.9973*	0.4365*	3.1233*	1.3665*	7.0470*	3.0717*
	10%	0.9951*	0.4361*	3.1979*	1.3634*	7.0821*	3.0812*
	15%	1.0132*	0.4360*	3.1933*	1.3690*	7.0782*	3.0814*
1,000	5%	1.0112*	0.4366*	3.2089*	1.3751*	7.3181*	3.0871*
	10%	1.0582*	0.9563*	3.5066*	2.9955*	7.5006*	6.7387*
	15%	1.0411*	0.9551*	3.7587*	3.0003*	7.1572*	6.7687*

Remark $p = -0.9$, *Sig. Level 0.05

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และ วิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า 0.3 โดยอาศัยสถิติทดสอบ พบว่าที่ระดับนัยสำคัญ 0.05 ในกรณีขนาดตัวอย่าง 40 ร้อยละการสูญหาย 5% ความแปรปรวน 1 และ 10 วิธีการประมาณของ 2 วิธีให้ค่า RMSE ไม่แตกต่างกัน ส่วนกรณีอื่นวิธีการประมาณของ 2 วิธีให้ค่า RMSE แตกต่างกัน โดยกรณีตัวอย่างขนาดตั้งแต่ 100 ขึ้นไป วิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน และค่าร้อยละการสูญหายดังตารางที่ 4 (Table 4)

Table 4 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when 0.3 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	0.9617	0.8889	3.3457	2.6462	6.3787*	6.1797*
	10%	0.9689*	0.9094*	3.1880*	2.9327*	6.7030*	6.4357*
	15%	1.0321*	0.9280*	3.5391*	2.9206*	6.9183*	6.6957*
100	5%	0.9679*	0.9237*	3.2303*	2.9065*	6.8233*	6.5420*
	10%	0.9832*	0.9407*	3.4057*	2.9948*	7.0544*	6.7156*
	15%	1.0051*	0.9556*	3.4466*	2.9628*	6.9570*	6.6713*
500	5%	0.9845*	0.9374*	3.3280*	3.0042*	6.9439*	6.6283*
	10%	0.9914*	0.9461*	3.3692*	3.0097*	7.0425*	6.6878*
	15%	0.9991*	0.9523*	3.2854*	3.0239*	7.1154*	6.6947*
1,000	5%	0.9971*	0.9499*	3.1968*	3.0098*	7.1179*	6.6818*
	10%	1.0030*	0.9553*	3.3229*	3.0118*	7.0633*	6.7370*
	15%	1.0049*	0.9577*	3.2191*	3.0275*	7.0772*	6.7215*

Remark $p = 0.3$, *Sig. Level 0.05

จากการพิจารณาค่า RMSE ณ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X ที่ระดับ 0.6 จากวิธีการประมาณค่าสูญหายทั้ง 2 วิธี พบว่าค่า RMSE จะมีแนวโน้มเพิ่มขึ้น เมื่อข้อมูลมีความแปรปรวนเพิ่มขึ้น ขนาดตัวอย่างเพิ่มขึ้น และร้อยละการสูญหายของข้อมูลเพิ่มขึ้น

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และวิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า 0.6 โดยอาศัยสถิติทดสอบ พบว่าที่ระดับนัยสำคัญ 0.05 วิธีการประมาณ 2 วิธีให้ค่า RMSE แตกต่างกัน ทุกกรณีของระดับความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย โดยวิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหายดังตารางที่ 5 (Table 5)

Table 5 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when 0.6 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	1.0048*	0.7306*	2.8398*	2.2429*	7.8780*	5.1135*
	10%	0.9667*	0.7683*	3.0464*	2.4656*	7.9133*	5.3676*
	15%	1.0003*	0.7860*	3.1434*	2.4474*	7.6643*	5.4533*
100	5%	1.0789*	0.7775*	3.0406*	2.4026*	7.6068*	5.4383*
	10%	1.0690*	0.7795*	3.1068*	2.5027*	7.8233*	5.5315*
	15%	1.0054*	0.7875*	3.2576*	2.5346*	7.8940*	5.6026*
500	5%	1.0702*	0.7927*	3.1765*	2.5232*	7.4612*	5.6150*
	10%	1.0296*	0.7961*	3.1684*	2.5369*	7.4189*	5.6135*
	15%	1.0415*	0.7939*	3.2178*	2.5456*	7.4830*	5.6097*
1,000	5%	1.3521*	0.7948*	3.1577*	2.5388*	7.1328*	5.6207*
	10%	1.2106*	0.7988*	3.2175*	2.5452*	7.2855*	5.5998*
	15%	1.2124*	0.7992*	3.1704*	2.5375*	7.0985*	5.6201*

Remark $p = 0.6$, *Sig. Level 0.05

จากการพิจารณาค่า RMSE ณ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X ที่ระดับ -0.9 จากวิธีการประมาณค่าสูญหายทั้ง 2 วิธี พบว่าค่า RMSE จะมีแนวโน้มเพิ่มขึ้น เมื่อข้อมูลมีความแปรปรวนเพิ่มขึ้น ขนาดตัวอย่างเพิ่มขึ้น และร้อยละการสูญหายของข้อมูลเพิ่มขึ้น

จากการเปรียบเทียบความแตกต่างของค่า RMSE ระหว่างวิธี EM และวิธี Singh ที่ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าสังเกตของตัวแปรตาม Y และค่าสังเกตของตัวแปรอิสระ X มีค่า 0.9 โดยอาศัยสถิติทดสอบ พบว่าที่ระดับนัยสำคัญ 0.05 วิธีการประมาณ 2 วิธีให้ค่า RMSE แตกต่างกันทุกกรณีของระดับความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหาย โดยวิธี EM ให้ค่า RMSE ต่ำกว่าวิธี Singh ในทุกระดับของความแปรปรวน ขนาดตัวอย่าง และค่าร้อยละการสูญหายดังตารางที่ 6 (Table 6)

Table 6 Root mean square error (RMSE) with sample sizes, percentages of missing data, variances when 0.9 correlation coefficient

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
40	5%	0.9364*	0.3922*	2.8306*	1.2540*	6.2374*	2.6910*
	10%	1.0213*	0.4236*	3.0108*	1.3225*	6.6263*	2.9379*
	15%	0.9949*	0.4373*	3.0541*	1.3285*	6.9264*	2.9453*
100	5%	1.0879*	0.4227*	3.0314*	1.3433*	6.8013*	2.9421*
	10%	1.0120*	0.4351*	3.0706*	1.3596*	7.0572*	3.0366*
	15%	1.0808*	0.4362*	3.0890*	1.3398*	7.1236*	3.0379*

Table 6 (cant)

Sample sizes	Percentages of missing data	Variance = 1		Variance = 10		Variance = 50	
		Singh	EM	Singh	EM	Singh	EM
50	5%	1.0574*	0.4326*	3.1326*	1.3653*	7.0157*	3.0159*
	10%	1.0239*	0.4353*	3.1519*	1.3701*	7.1002*	3.0453*
	15%	1.0423*	0.4369*	3.1593*	1.3651*	7.0676*	3.0539*
1,000	5%	1.0132*	0.4349*	3.2598*	1.3703*	7.0267*	3.0303*
	10%	1.0188*	0.4382*	3.2788*	1.3669*	7.0739*	3.0529*
	15%	1.0560*	0.4383*	3.1661*	1.3702*	7.0993*	3.0560*

Remark $p = 0.9$, *Sig. Level 0.05

สรุปผลการวิจัย

วิธี EM เป็นวิธีการประมาณค่าสูญหายที่มีประสิทธิภาพดีกว่าวิธีการของ Singh สำหรับการถดถอยเชิงเส้นอย่างง่าย เนื่องจากวิธี EM ให้ค่า RMSE ต่ำกว่าวิธีการของ Singh ในทุกระดับของสัมประสิทธิ์สหสัมพันธ์ ความแปรปรวน ขนาดตัวอย่าง และร้อยละการสูญหาย และจากการทดสอบโดยใช้สถิติทดสอบ ที่ระดับนัยสำคัญ 0.05 พบว่าวิธี EM และวิธี Singh ไม่มีความแตกต่างกันในการประมาณค่าสูญหายในกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์อยู่ในระดับต่ำ (-0.3, 0.3) และขนาดตัวอย่างเล็ก ($n = 40$) ส่วนในกรณีอื่น ๆ วิธี EM และวิธี Singh มีความแตกต่างกันการประมาณค่าสูญหาย

กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ และบัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่ที่ให้การสนับสนุน ขอขอบคุณเจ้าของแนวคิดและทฤษฎีต่าง ๆ ของงานวิจัย วารสารและบทความที่ถูกนำมาอ้างอิงในการทำวิจัยฉบับนี้ และขอขอบคุณผู้ทรงคุณวุฒิที่ได้อ่านผลงานวิจัยและได้กรุณาให้คำแนะนำในการแก้ไขข้อบกพร่องต่าง ๆ จนทำให้ผลงานวิจัยนี้มีความถูกต้องและสมบูรณ์มากยิ่งขึ้น

เอกสารอ้างอิง

- Ahmed MS, AL-Titi O, AL-Rawi Z. et al. Estimation of a population mean using different imputation methods, *Statistics in Transition*. 2006: 7(6); 1247-1264.
- Laaksonen S. Regression-Based nearest neighbor hot decking, *Computation Statistics*. 2000: 15(1); 65-71.
- Little Roderick JA, Rubin Donald B. *Statistical Analysis with Missing Data*. New York: John Wiley, 1987.
- Singh S, Horn S. Compromised imputation in survey sampling, *Metrika*. 2000: 51; 266-276.
- Singh GN, Kumari P, Jong MK. Estimation of population mean using imputation techniques in sample surveys, *Journal of the Korean Statistic Society*. 2010: 39: 67-74.
- Wararit P. The monte carlo simulation for estimating the coefficients of skewness when observations are inverse gaussian distributed, *Kasalongkham Research Journal*. 2009; 3(1): 14-23.