วารสารวิทยาศาสตร์บูรพา ปีที่ 26 (ฉบับที่ 2) พฤษภาคม – สิงหาคม พ.ศ. 2564
BURAPHA SCIENCE JOURNAL Volume 26 (No.2) May – August 2021
บทความวิจัย

การประมาณค่าสูญหายด้วยวิธีการถดถอยแบบเบย์-บูตสแตรป

Estimating Missing Data with Bayes Bootstrap Regression Imputation

ธรรมรัตน์ กลีบเมฆ[1] และ นพคุณ ทองมวล[2*]

Tammarat Kleebmek[1] and Noppakun Thongmual[2*]

[1] *สาขาคณิตศาสตร์และสถิติประยุกต์ คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน*

[2] *สาขาวิชาวิทยาศาสตร์และคณิตศาสตร์ คณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์*

[1] *Department of Mathematics and Statistics, Faculty of Sciences and Liberal Arts, Rajamangala University of Technology Isan*

[2] *Department of Science and Mathematics, Faculty of Science and Health Technology, Kalasin University*

Received : 4 March 2020

Revised : 30 April 2020

Accepted : 8 September 2020

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อประมาณค่าข้อมูลสูญหายเมื่อตัวแปรตาม Y มีความสัมพันธ์กับตัวแปรอิสระ X โดยที่ตัวแปร X และ Y มีการแจกแจงปรกติ โดยนำเสนอวิธีประมาณค่าข้อมูลสูญหาย คือ วิธีการถดถอยแบบเบย์-บูตสแตรป เปรียบเทียบกับวิธีประมาณค่าสูญหายด้วยวิธีถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด โดยใช้เกณฑ์ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเพื่อวัดความแม่นยำ การเปรียบเทียบการประมาณค่าข้อมูลสูญหายใช้เทคนิคการจำลองแบบมอนติคาร์โล ผลการศึกษาพบว่า วิธีการถดถอยแบบเบย์-บูตสแตรปและวิธีการถดถอยมีความแม่นยำมากกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุดในทุกกรณี แต่มีบางกรณีที่ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอยแบบเบย์-บูตสแตรปมีค่าต่ำสุด ดังนั้นผู้วิจัยจึงแนะนำวิธีการถดถอยแบบเบย์-บูตสแตรปสำหรับการประมาณค่าข้อมูลสูญหายเมื่อทราบค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรตาม Y และตัวแปรอิสระ X และตัวแปรทั้งสองมีการแจกแจงปรกติ

คำสำคัญ : ข้อมูลสูญหาย ; วิธีการถดถอยแบบเบย์-บูตสแตรป ; วิธีการถดถอย ; วิธีการถดถอยด้วยระยะทางต่ำที่สุด

Abstract

This research is about estimating missing data when dependent variable Y is correlated with independent variable X, and X and Y are distributed as normal. The proposed method for estimating missing data is Bayes-Bootstrap regression imputation method (BRI) that is compared with regression imputation method (RI) and distance regression imputation method (DRI). The measurement criteria is mean absolute error (MAE). Comparing of estimating missing data used the Monte Carlo simulation technique. The results of study indicate that BRI and RI are more accuracy than DRI for all cases, but BRI presents the lowest mean absolute error in some case. Therefore, researchers introduce the BRI method for estimating missing data when the correlation coefficient between dependent variable Y and independent variable X is known and both variable distributions are normal distributions.

Keywords : : missing data ; Bayes-Bootstrap regression imputation ; Distance regression imputation ; regression imputation

***Corresponding author. E-mail** : nop_stat@hotmail.com

บทนำ

ในการศึกษาเชิงสำรวจมักมีข้อมูลสูญหายซึ่งเกิดจากสาเหตุสามประการ คือ 1. สำรวจไม่ครบตามจำนวนตัวอย่าง 2. ผู้ตอบแบบสอบถามไม่ตอบกลับทั้งหมด และ 3. ผู้ตอบแบบสอบถามไม่สามารถตอบได้ (Brick & Kalton, 1996) นอกจากนี้ Peng *et al* (2006) ตรวจสอบ 1,087 งานวิจัยที่ตีพิมพ์ในด้านการศึกษาและจิตวิทยาซึ่งมี 48% ที่มีข้อมูลสูญหาย ปัญหาการสูญหายของข้อมูลนั้นเกิดขึ้นในงานวิจัยหลายประเภท การสูญหายของข้อมูลทำให้คุณภาพของสารสนเทศลดลง และส่งผลกระทบเชิงลบต่อการวิเคราะห์ข้อมูลดังกล่าว ปัญหานี้นำไปสู่การวิจัยอย่างกว้างขวาง เพื่อหาวิธีการประมาณค่าข้อมูลสูญหาย การประมาณค่าข้อมูลสูญหายคือการพยายามเรียกคืนค่าที่หายไปในชุดข้อมูลนั้นโดยที่การประมาณค่าสูญหายใช้การวิเคราะห์เกี่ยวข้องกับคุณลักษณะ กฎและความสัมพันธ์ ที่ซ่อนอยู่ภายในชุดข้อมูล การใช้ชุดข้อมูลที่มีข้อมูลสูญหายอย่างไม่เหมาะสมอาจนำไปสู่การวิเคราะห์ข้อมูลที่ไม่ถูกต้องข้อสรุปที่ผิดและการคาดการณ์ที่ผิดพลาด (Troyanskaya *et al.*, 2001 & Lin *et al.*, 2017) ผู้วิจัยสนใจปัญหาการสูญหายของข้อมูลในงานวิจัยของ Pimchanok & Watchareeporn (2017) ซึ่งเกี่ยวข้องกับการเปรียบเทียบวิธีการประมาณค่าสูญหายในการสำรวจด้วยตัวอย่าง เมื่อตัวแปรตาม Y และตัวแปรอิสระ X มีความสัมพันธ์กันและตัวแปรทั้งสองมีการแจกแจงปรกติ พบว่า วิธีการประมาณค่าสูญหายด้วยวิธีการถดถอย (Regression Imputation : RI) วิธีการถดถอยด้วยระยะห่างต่ำสุด (Distance Regression Imputation : DRI) ให้ผลเป็นที่น่าพอใจในการประมาณค่าข้อมูลที่สูญหายโดยไม่มีผลต่อการเลือกวิธีการประมาณค่าสูญหายเมื่อข้อมูลมีความสัมพันธ์กัน

งานวิจัยนี้นำเสนอการประมาณค่าตัวใหม่โดยการประยุกต์ใช้ระหว่างวิธีการถดถอยและการประมาณค่าพารามิเตอร์ด้วยวิธีเบย์-บูตสแตรป (Rubin,1981) ผู้วิจัยได้นำวิธีเบย์-บูตสแตรปในการประมาณค่าพารามิเตอร์ในโมเดลการถดถอยเพื่อใช้ในการประมาณค่าข้อมูลที่สูญหายเมื่อตัวแปรตาม Y และตัวแปรอิสระ X มีความสัมพันธ์กันและตัวแปรทั้งสองมีการแจกแจงปรกติ ผู้วิจัยเรียกวิธีนี้ว่า วิธีการถดถอยแบบเบย์-บูตสแตรป (Bayes-Bootstrap regression Imputation : BRI) งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะห่างต่ำสุดและวิธีการถดถอยแบบเบย์-บูตสแตรปด้วยค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error) ซึ่งตัวชี้วัดนี้สะท้อนถึงความแม่นยำ (accuracy) ของวิธีการประมาณค่าสูญหาย

วิธีดำเนินการวิจัย

งานวิจัยนี้เปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของวิธีการถดถอยแบบเบย์-บูตสแตรป วิธีการถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด ซึ่งการประมาณค่าสูญหายด้วยวิธีการถดถอยแบบเบย์-บูตสแตรปเป็นวิธีที่ผู้วิจัยได้นำเสนอ วิธีการดำเนินงานวิจัยในครั้งนี้เป็นการจำลองข้อมูลแบบมอนติคาร์โล (Monte Carlo Simulation) ด้วยโปรแกรม R มีขั้นตอนดังต่อไปนี้

(1) กำหนดระดับสหสัมพันธ์ (ρ) ระหว่าง X และ Y คือ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 และ 0.9

(2) สร้างประชากรขนาด 5,000 หน่วย ประกอบด้วยตัวแปรอิสระ (X) และตัวแปรตาม (Y) ที่มีการแจกแจงปรกติมีค่าเฉลี่ย 10 และความแปรปรวน 1 ที่มีระดับสหสัมพันธ์ (ρ) ระหว่าง X และ Y ตามข้อที่ 1

(3) เลือกตัวอย่างสุ่มแบบง่าย (Simple Random Sampling) จากประชากรมาขนาด 30, 50, 70 และ 90

(4) กำหนดค่าสูญหายให้กับตัวแปรตาม Y จำนวน 5% 10% และ 15%

(5) ประมาณค่าสูญหายด้วยวิธีการถดถอยแบบเบย์-บูตสแตรป วิธีการถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด

(6) หาค่าคลาดเคลื่อนสัมบูรณ์

$$AE = \sum_{i=r+1}^{n} \left| \hat{y}_i - y_i \right| \quad (1)$$

(7) ทำซ้ำ 10,000 รอบ จากขั้นตอนที่ (3) – (6) จากนั้นหาค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยจาก

$$MAE = \frac{\sum_{i=1}^{10000} AE_i}{10000} \quad (2)$$

วิธีการประมาณค่าสูญหายทั้ง 3 วิธี มีวิธีการคำนวณดังนี้

(1) วิธีการถดถอย (regression imputation : RI)

การประมาณค่าสูญหายด้วยวิธีการถดถอย เป็นการประมาณค่า ตัวแปรที่ต้องการศึกษาโดยใช้ความสัมพันธ์ระหว่างตัวแปรอิสระ (X) กับตัวแปรตาม (Y) มาช่วยในการประมาณค่าโดยกำหนดให้ $y_1, y_2, y_3, \ldots, y_r$ เป็นค่าข้อมูลที่สมบูรณ์และ $y_{r+1}, y_2, y_3, \ldots, y_n$ เป็นค่าข้อมูลที่สูญหายและค่า x_i ทั้งหมดมีค่าที่สมบูรณ์ เมื่อ $i = 1, 2, 3, \ldots, n$ นำข้อมูลที่สมบูรณ์ $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_r, y_r)$ เพื่อประมาณค่าพารามิเตอร์ในตัวแบบการถดถอย คือ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ดังนั้น จึงได้สมการถดถอยเพื่อประมาณค่าสูญหายเป็นดังนี้ (Jitthavech, 2015)

$$\hat{y}_i = b_0 + b_1 x_i \quad (3)$$

โดยที่ b_0 , b_1 เป็นค่าประมาณของ β_0 และ β_1 ในตัวแบบการถดถอย

b_0 , b_1 สามารถคำนวณ ได้ดังนี้

$$b_1 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{r} (x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1 \bar{x}, \quad \bar{x} = \frac{\sum_{i=1}^{r} x_i}{r} \text{ และ } \bar{y} = \frac{\sum_{i=1}^{r} y_i}{r}$$

จากสมการที่ (3) นำค่า x_i เป็นค่าสังเกตใช้ประมาณค่าสูญหายของ y_i โดยที่ $\hat{y}_i$ คือ ค่าประมาณของ y_i ที่สูญหาย เมื่อ $i = r+1, r+2, \ldots, n$

(2) วิธีการถดถอยด้วยระยะทางต่ำที่สุด (distance regression imputation : DRI)

Chaimongkol (2005) ได้เสนอวิธีการประมาณค่าสูญหายด้วยวิธีการถดถอยระยะห่างต่ำสุดโดยมีรายละเอียดดังนี้ กำหนดให้ $y_1, y_2, y_3, \ldots, y_r$ เป็นค่าข้อมูลที่สมบูรณ์และ $y_{r+1}, \ldots, y_n$ เป็นค่าข้อมูลที่สูญหายและค่า x_i, $i = 1, 2, 3, \ldots, n$ ทั้งหมดมีค่าที่สมบูรณ์ เมื่อ $i = 1, 2, 3, \ldots, r$ ให้ $d_{xi} = x_{i+1} - x_i$ และ $d_{yi} = y'_{i+1} - y'_i$ เมื่อ y'_i คือค่าของ y ที่สอดคล้องกับสถิติลำดับของ $x_{(i)}$ ดังนั้น สมการถดถอยเชิงเส้นอย่างง่ายระหว่าง d_{xi} กับ d_{yi} คือ

$$\hat{d}_y = b'_0 + b'_1 d_x \tag{4}$$

เมื่อ $b'_1 = \dfrac{\sum_{i=1}^{r-1}(d_{xi} - \bar{d}_x)(d_{yi} - \bar{d}_y)}{\sum_{i=1}^{r-1}(d_{xi} - \bar{d}_x)^2}$, $b'_0 = \bar{d}_y - b'_1\bar{d}_x$, $\bar{d}_y = \dfrac{\sum_{i=1}^{r-1} d_{yi}}{r-1}$ และ $\bar{d}_x = \dfrac{\sum_{i=1}^{r-1} d_{xi}}{r-1}$

สำหรับหน่วยที่มีค่าสูญหาย $j = r+1, \ldots, n$

ให้ $m_j = \left|x_j - x_{(i)}\right| = \min_{1 \le k \le r}\left|x_i - x_{(k)}\right|$ สำหรับบางค่า k เมื่อ $1 \le k \le r$ ดังนี้

$$\Delta y_j = b'_0 + b'_1 m_j \tag{5}$$

เมื่อ Δy_j เป็นค่าประมาณระยะห่างของ y_j ดังนั้น ค่าสูญหายจะถูกแทนที่ด้วยค่าประมาณ $y'_j = y'_k + \Delta y_i$ เมื่อ y'_k เป็นค่าของ y ที่สอดคล้องกับค่า $x_{(k)}$ โดยโครงสร้างของวิธี DRI ดังตาราง

$x_{(i)}$	d_{xi}	y'_i	d_{yi}	Δy_j	y'_j
$x_{(1)}$	$x_{i+1} - x_i$	y'_1	$y'_2 - y'_1$		y'_1
$x_{(2)}$	$x_{i+1} - x_i$	y'_2	$y'_3 - y'_2$		y'_2
...	...	...	...		...
$x_{(r-1)}$	$x_{i+1} - x_i$	y'_{r-1}	$y'_r - y'_{r-1}$		y'_{r-1}
$x_{(r)}$	-	y'_r	-		y'_r
$x_{(r+1)}$				Δy_{r+1}	$\hat{y}'_k + \Delta y_{r+1}$
...				...	...
$x_{(n)}$				Δy_n	$\hat{y}'_k + \Delta y_{r+1}$

ตัวประมาณค่าสูญหายของ y_j เมื่อ $j = r+1, \ldots, n$ แสดงผลได้ดังนี้

กรณี $\hat{b}_1' > 0$

$$\hat{y}_j' = \begin{cases} \hat{y}_k' + \Delta y_j, x_j \geq x_{(k)} \\ \hat{y}_k' - \Delta y_j, x_j < x_{(k)} \end{cases}$$

และกรณี $\hat{b}_1' < 0$

$$\hat{y}_j' = \begin{cases} \hat{y}_k' + \Delta y_j, x_j \leq x_{(k)} \\ \hat{y}_k' - \Delta y_j, x_j > x_{(k)} \end{cases}$$

(3) วิธีการถดถอยแบบเบย์-บูตสแตรป (Bayes-Bootstrap regression Imputation : BRI)

การประมาณค่าสูญหายวิธีนี้จะนำหลักการของวิธีการถดถอยมาประยุกต์ใช้โดยที่จะแทนตัวประมาณในสมการที่ (3) ด้วยตัวประมาณแบบเบย์-บูตสแตรป ซึ่งมีขั้นตอนการประมาณค่าดังนี้

ขั้นตอนที่ 1 กำหนดให้ r คือจำนวนข้อมูลที่สมบูรณ์ สุ่มความน่าจะเป็นภายหลัง (Posterior probability) จากการแจกแจงเอกรูปที่อยู่ในช่วง 0 และ 1 มา $r-1$ ตัว คือ $u_1, u_2, \ldots, u_{r-1}$ จากนั้นให้เรียงลำดับและคำนวณ $g_i = u_{(i)} - u_{(i-1)}$, $i = 1, 2, \ldots, r-1$ โดยที่ $u_{(0)} = 0$ และ $u_{(r)} = 1$ แล้ว $g = (g_1, g_2, \ldots, g_r)$ นี้คือเวกเตอร์ของความน่าจะเป็น

ขั้นตอนที่ 2 ทำการสุ่มแบบใส่คืนจากข้อมูล $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_r, y_r)$ จะได้ตัวอย่างบูตสแตรป (Bootstrap sample) คือ $(x_1^B, y_1^B), (x_2^B, y_2^B), (x_3^B, y_3^B), \ldots, (x_r^B, y_r^B)$ และให้น้ำหนักด้วยความน่าจะเป็นจากขั้นตอนที่ 1 จะได้ว่า $(g_1 x_1^B, g_1 y_1^B), (g_2 x_2^B, g_2 y_2^B), (g_3 x_3^B, g_3 y_3^B), \ldots, (g_r x_r^B, g_r y_r^B)$

ขั้นตอนที่ 3 จากนั้นคำนวณค่าสัมประสิทธิ์การถดถอยในแต่ละชุด จะได้ว่า

$$b_{1_j}^B = \frac{\sum_{i=1}^{r} (g_i x_i^B - \bar{x}^B)(g_i y_i^B - \bar{y}^B)}{\sum_{i=1}^{r} (g_i x_i^B - \bar{x}^B)^2}, \quad b_{0_j}^B = \bar{y}^B - b_1^B \bar{x}^B, \quad \bar{x}^B = \frac{\sum_{i=1}^{r} g_i x_i^B}{r} \text{ และ } \bar{y}^B = \frac{\sum_{i=1}^{r} g_i y_i^B}{r}$$

ขั้นตอนที่ 4 ทำซ้ำขั้นตอนที่ 2 - 3 จำนวน 1,000 รอบ จะได้ $b_{1_j}^B$ และ $b_{0_j}^B$ จำนวน $1{,}000$ ค่า

ขั้นตอนที่ 5 คำนวณ b_{1B_i} และ b_{0B_i}

$$b_{1B_i} = \frac{\sum_{j=1}^{1000} b_{1_j}^B}{1000} \text{ และ } b_{0B_i} = \frac{\sum_{j=1}^{1000} b_{0_j}^B}{1000}$$

ขั้นตอนที่ 6 ทำซ้ำจากขั้นตอนที่ 2 – 5 จำนวน 5,000 รอบ แล้วหาค่าเฉลี่ยดังนี้

$$b_{0BB} = \frac{\sum_{i=1}^{5000} b_{1B_i}}{5000} \text{ และ } b_{1BB} = \frac{\sum_{i=1}^{5000} b_{1B_i}}{5000}$$

ประมาณค่าข้อมูลสูญหายด้วยสมการ

$$\hat{y}_i^* = b_{0BB} + b_{1BB} x_i \quad (6)$$

ผลการวิจัย

ผลการวิเคราะห์การเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายโดยใช้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยทั้ง 3 วิธี ดังนี้

ตารางที่ 1 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเมื่อการสูญหาย 5% ด้วย n = 30, 50, 70, 90 และ ρ = 0.1, 0.2, 0.3,…, 0.9

n		30			50			70			90		
วิธี		RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI
ρ	0.1	0.0552	0.0411	0.0412	0.0212	0.0211	0.0333	0.0246	0.0245	0.0365	0.0162	0.0161	0.0218
	0.2	0.0379	0.0378	0.0533	0.0220	0.0219	0.0322	0.0227	0.0227	0.0315	0.0181	0.0181	0.0247
	0.3	0.0324	0.0323	0.0464	0.0213	0.0213	0.0281	0.0220	0.0220	0.0341	0.0163	0.0162	0.0221
	0.4	0.0348	0.0348	0.0521	0.0213	0.0213	0.0279	0.0220	0.0220	0.0285	0.0156	0.0156	0.0212
	0.5	0.0329	0.0329	0.0493	0.0201	0.0201	0.0264	0.0208	0.0208	0.0269	0.0141	0.0141	0.0219
	0.6	0.0320	0.0320	0.0414	0.0183	0.0183	0.0257	0.0193	0.0193	0.0264	0.0131	0.0131	0.0191
	0.7	0.0289	0.0289	0.0373	0.0164	0.0164	0.0231	0.0164	0.0164	0.0262	0.0122	0.0122	0.0151
	0.8	0.0252	0.0252	0.0347	0.0143	0.0143	0.0237	0.0147	0.0147	0.0194	0.0101	0.0101	0.0134
	0.9	0.0174	0.0174	0.0234	0.0092	0.0092	0.0139	0.0113	0.0113	0.0135	0.0073	0.0073	0.0109

จากตารางที่ 1 พิจารณาวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรป ที่เปอร์เซ็นต์การสูญหายของข้อมูล 5% เมื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายด้วยค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่า ในกรณีที่ขนาดตัวอย่าง n = 30, 50, 70, 90 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.1$ และกรณีที่ขนาดตัวอย่าง n = 30, 50 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.2$ และกรณีที่ขนาดตัวอย่าง n = 30, 90 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.3$ วิธีการถดถอยแบบเบย์-บูตสแตรปมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยที่สุดเมื่อเปรียบเทียบกับวิธีการถดถอย และวิธีการถดถอยด้วยระยะทางต่ำที่สุด เมื่อพิจารณาในกรณีอื่น ๆ

ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอยและวิธีการถดถอยแบบเบย์-บูตสแตรปใกล้เคียงกันและทั้งสองวิธีมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุด นอกจากนี้ยังพบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรปมีแนวโน้มลดลงในทุกกรณี

ตารางที่ 2 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเมื่อการสูญหาย 10% ด้วย n = 30, 50, 70, 90 และ ρ = 0.1, 0.2, 0.3,…, 0.9

n		30			50			70			90		
วิธี		RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI
ρ	0.1	0.0488	0.0488	0.0740	0.0399	0.0399	0.0557	0.0359	0.0358	0.0449	0.0279	0.0278	0.0353
	0.2	0.0480	0.0479	0.0600	0.0342	0.0341	0.0583	0.0363	0.0363	0.0402	0.0274	0.0274	0.0402
	0.3	0.0486	0.0486	0.0600	0.0330	0.0330	0.0495	0.0270	0.0270	0.0435	0.0288	0.0288	0.0392
	0.4	0.0424	0.0424	0.0602	0.0305	0.0305	0.0426	0.0284	0.0284	0.0372	0.0262	0.0262	0.0397
	0.5	0.0406	0.0406	0.0506	0.0331	0.0331	0.0486	0.0304	0.0304	0.0363	0.0222	0.0222	0.0326
	0.6	0.0382	0.0382	0.0553	0.0308	0.0308	0.0406	0.0252	0.0252	0.0360	0.0215	0.0215	0.0296
	0.7	0.0366	0.0366	0.0499	0.0263	0.0263	0.0360	0.0207	0.0207	0.0288	0.0195	0.0195	0.0310
	0.8	0.0302	0.0302	0.0376	0.0243	0.0243	0.0325	0.0184	0.0184	0.0247	0.0173	0.0173	0.0225
	0.9	0.0213	0.0213	0.0309	0.0152	0.0152	0.0200	0.0128	0.0128	0.0209	0.0134	0.0134	0.0166

จากตารางที่ 2 พิจารณาวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรป ที่เปอร์เซ็นต์การสูญหายของข้อมูล 10% เมื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายด้วยค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่า ในกรณีที่ขนาดตัวอย่าง n = 70, 90 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.1$ และกรณีที่ขนาดตัวอย่าง n = 30, 50 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.2$ วิธีการถดถอยแบบเบย์-บูตสแตรปมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยที่สุดเมื่อเปรียบเทียบกับวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด เมื่อพิจารณาในกรณีอื่น ๆ ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอยและวิธีการถดถอยแบบเบย์-บูตสแตรปใกล้เคียงกันและทั้งสองวิธีมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุด นอกจากนี้ยังพบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรปมีแนวโน้มลดลงในทุกกรณี

ตารางที่ 3 ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยเมื่อการสูญหาย 15% ด้วย n = 30, 50, 70, 90 และ ρ = 0.1, 0.2, 0.3,..., 0.9

n		30			50			70			90		
วิธี		RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI	RI	BRI	DRI
ρ	0.1	0.0635	0.0635	0.0805	0.0480	0.0480	0.0574	0.0321	0.0321	0.0507	0.0360	0.0360	0.0440
	0.2	0.0504	0.0503	0.0698	0.0485	0.0484	0.0715	0.0410	0.0410	0.0537	0.0357	0.0357	0.0486
	0.3	0.0610	0.0610	0.0828	0.0509	0.0509	0.0744	0.0362	0.0362	0.0536	0.0369	0.0369	0.0495
	0.4	0.0482	0.0482	0.0647	0.0493	0.0493	0.0773	0.0288	0.0288	0.0432	0.0329	0.0329	0.0468
	0.5	0.0488	0.0488	0.0711	0.0398	0.0398	0.0541	0.0325	0.0325	0.0493	0.0299	0.0299	0.0366
	0.6	0.0449	0.0449	0.0634	0.0435	0.0435	0.0543	0.0327	0.0327	0.0417	0.0302	0.0302	0.0394
	0.7	0.0419	0.0419	0.0598	0.0384	0.0384	0.0469	0.0269	0.0269	0.0396	0.0218	0.0218	0.0351
	0.8	0.0343	0.0343	0.0455	0.0268	0.0268	0.0429	0.0226	0.0226	0.0328	0.0181	0.0181	0.0300
	0.9	0.0237	0.0237	0.0323	0.0222	0.0222	0.0292	0.0186	0.0186	0.0234	0.0156	0.0156	0.0208

จากตารางที่ 3 พิจารณาวิธีการประมาณค่าสูญหายทั้ง 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรป ที่เปอร์เซ็นต์การสูญหายของข้อมูล 15% เมื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายด้วยค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่า ในกรณีที่ขนาดตัวอย่าง n = 30, 50 ด้วยค่าสัมประสิทธิ์สหสัมพันธ์ $\rho = 0.2$ วิธีการถดถอยแบบเบย์-บูตสแตรปมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยที่สุดเมื่อเปรียบเทียบกับวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด เมื่อพิจารณาในกรณีอื่น ๆ ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอยและวิธีการถดถอยแบบเบย์-บูตสแตรปใกล้เคียงกันและทั้งสองวิธีมีค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุด นอกจากนี้ยังพบว่าเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุด และวิธีการถดถอยแบบเบย์-บูตสแตรปมีแนวโน้มลดลงในทุกกรณี

วิจารณ์ผลการวิจัย

จากการเปรียบเทียบวิธีการประมาณค่าสูญหายของวิธีการถดถอยแบบเบย์-บูตสแตรป วิธีการถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด ซึ่งความแม่นยำในการประมาณค่าสูญหายจะพิจารณาจากค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยน้อยที่สุด เมื่อขนาดตัวอย่าง n = 30, 50, 70 และ 90 ด้วยเปอร์เซ็นต์การสูญหายของข้อมูล 5%, 10% และ 15% และสัมประสิทธิ์สหสัมพันธ์ ρ = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 วิธีการถดถอยและวิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำใกล้เคียงกัน แต่เมื่อพิจารณากรณี ρ = 0.1, 0.2, 0.3 ขนาดตัวอย่าง n = 30, 50, 70 และ 90 ด้วยเปอร์เซ็นต์การสูญหายของข้อมูล 5%, 10% และ 15% การประมาณค่าสูญหายของวิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำค่อนข้างดีกว่าวิธีการถดถอย ผลจากตารางที่ 1-3 วิธีการถดถอยและวิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำกว่าวิธีการถดถอยด้วยระยะทางต่ำที่สุดทุก ๆกรณี วิธีการถดถอยแบบเบย์-บูตสแตรปค่อนข้างยากต่อการคำนวณแต่วิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำในการประมาณค่าสูญหายมากกว่าวิธีการถดถอยและวิธีการถดถอยด้วยระยะทางต่ำที่สุด ซึ่งสนับสนุนจากงานวิจัยของ Merlise และ Herbert (2000) กล่าวว่า การใช้วิธีแบบเบย์-บูตสแตรปมาช่วยในการประมาณค่า

จะมีความแม่นยำมาก ดังนั้น วิธีการถดถอยแบบเบย์-บูตสแตรป จึงเป็นวิธีทางเลือกในการประมาณค่าสูญหายเพื่อให้เกิดความแม่นยำ

สรุปผลการวิจัย

วิธีการประมาณค่าข้อมูลสูญหายในการสำรวจด้วยตัวอย่างเมื่อตัวแปรตาม Y และตัวแปรอิสระ X มีความสัมพันธ์กัน และตัวแปรทั้งสองมีการแจกแจงแบบปรกติด้วย วิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธีการถดถอย วิธีการถดถอยด้วยระยะทางต่ำที่สุดและวิธีการถดถอยแบบเบย์-บูตสแตรปภายใต้ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่า เมื่อ n = 30, 50, 70, 90 และสัมประสิทธิ์สหสัมพันธ์เพิ่มขึ้น วิธีการถดถอยแบบเบย์-บูตสแตรปและวิธีการถดถอยมีความแม่นยำในการประมาณค่าสูญหายมากกว่าวิธีการถดถอยด้วยระยะทางต่ำเนื่องจากค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยมีค่าน้อยสุด แต่เมื่อพิจารณาค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยในกรณี n = 30, 50 พบว่า วิธีการถดถอยแบบเบย์-บูตสแตรปมีความแม่นยำกว่าวิธีการถดถอยเล็กน้อย และเมื่อขนาดตัวอย่างเพิ่มขึ้น พบว่า ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ยของวิธีการประมาณค่าสูญหายทั้ง 3 วิธีลดลง

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยกาฬสินธุ์ที่อำนวยความสะดวกสถานที่และอุปกรณ์ในการทำวิจัยในครั้งนี้

เอกสารอ้างอิง

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*(3), 215-238.

Chaimongkol, W. (2005). Three composite imputation methods for item nonresponse estimation in sample surveys(Doctoral dissertation) Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok.

Jitthavech, J. (2015). *Regression Analysis* (1st ed.). Bangkok, Thailand: Academic Promotion and Development Program, National Institute of Development Administration. (in Thai)

Lin, J. Q., Wu, H. C., Chan, S. C. (2017). A new regularized recursive dynamic factor analysis with variable forgetting factor for wireless sensor networks with missing data. *IEEE International Symposium on Circuits and Systems,* 1–4.

Merlise, A. C., Herbert K. H. L. (2000). Bagging and the Bayesian Bootstrap. Retrieved Jan, 2019, form https://www.researchgate.net/publication/2469163.

Peng, C.Y.J., Harwell, M., Liou, S.M., Ehman. LH. (2006) Advances in missing data methods and implications for educational research. Real data analysis, 31–78.

Pimchanok, C., Watchareeporn, C. (2017). A comparison of the estimation methods for missing data in sample survey. *The Journal of Applied Science*, *16*(1), 60-73.

Rubin, D. (1981). The Bayesian bootstrap, *Annals of Statistics, 9*, 130-134.

Troyanskaya, O., Cantor, M., Sherlock, G. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525, 2001.