การเปรียบเทียบค่าประมาณสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้นบนข้อมูลระยะยาวที่มีค่าสูญหาย

Comparison of regression coefficient estimates of linear mixed model on missing longitudinal data

กาญจนารักษ์ ฤทธิรักษ์[1] ภาวินี แสนสุข[1] และ ใกล้รุ่ง สามารถ[1,2*]

[1]ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ อ.หาดใหญ่ จ.สงขลา 90110

[2]หน่วยวิจัยสถิติและการประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ อ.หาดใหญ่ จ.สงขลา 90110

Kanchanarak Ritthirak[1] Pawinee Saensuk[1] and Klairung Samart[1,2*]

[1]Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkla 90110 Thailand

[2]Statistics and Applications Research Unit, Faculty of Science, Prince of Songkla University, Hat Yai, Songkla 90110 Thailand

[*]Corresponding Author, E-mail: klairung.s@psu.ac.th

Received: 24 May 2019 | Revised: 2 August 2019 | Accepted: 26 December 2019

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบค่าประมาณของสัมประสิทธิ์การถดถอยตัวแบบผสมเชิงเส้นบนข้อมูลระยะยาวที่มีการสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing completely at random หรือ MCAR) การสูญหายแบบสุ่ม (Missing at random หรือ MAR) และการสูญหายแบบไม่สุ่ม (Missing not at random หรือ MNAR) ในกรณีศึกษานี้จะจำลองข้อมูลเริ่มต้นให้มีลักษณะเป็นข้อมูลระยะยาว โดยกำหนดให้ข้อมูลเริ่มต้นมีค่าคลาดเคลื่อนสุ่มที่มีการแจกแจงปรกติ กำหนดร้อยละการสูญหายของแต่ละประเภทการสูญหายเท่ากับ 10 และ 20 และขนาดตัวอย่าง (n) เท่ากับ 5 10 20 และ 50 โดยกำหนดจำนวนการวัดซ้ำของแต่ละตัวอย่างเท่ากับ 5 และทำการจำลองข้อมูลซ้ำจำนวน 1,000 ครั้ง โดยใช้โปรแกรม R ซึ่งใช้วิธีภาวะน่าจะเป็นสูงสุดในการประมาณค่าสัมประสิทธิ์การถดถอย ผลการวิจัยพบว่า โดยภาพรวม เมื่อพิจารณาทั้งค่าเอนเอียงเทียมและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียม วิธีการประมาณค่านี้เหมาะสมสำหรับข้อมูลที่มีกลไกการสูญหายแบบ MCAR เท่านั้น แต่ไม่เหมาะสมสำหรับข้อมูลมีการสูญหายแบบ MNAR โดยเฉพาะเมื่อตัวอย่างมีขนาดใหญ่

ABSTRACT

The aim of this research is to compare the regression coefficient estimates of linear mixed models on different mechanisms of missing longitudinal data namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In this study, longitudinal data were initially generated where the random error were normally distributed. Then, the comparison was carried out through the simulation study where the missing rates of each mechanism were set to be 10% and 20%, and the sample sizes were 5, 10, 20, and 50 with 5 repeated measures. The simulation process was replicated 1,000 times using R program in which

the maximum likelihood estimation was employed. In overall, by considering both pseudo-bias and pseudo-root mean square error (RMSE) we found that this estimation method is suitable only when the data were missing completely at random (MCAR), but not when the data were missing not at random (MNAR), especially in the large sample size.

คำสำคัญ: ข้อมูลสูญหาย ข้อมูลระยะยาว ตัวแบบผสมเชิงเส้น ความเอนเอียง
Keywords: Missing data, Longitudinal data, Linear mixed models, Bias

บทนำ

เนื่องจากในปัจจุบันการวิเคราะห์ข้อมูลทางสถิติมีความสำคัญอย่างมากทั้งในด้านการศึกษา การเงิน การแพทย์ สาธารณสุข การระบาดวิทยา การวิจัยทางคลินิก และเศรษฐศาสตร์ เป็นต้น โดยข้อมูลที่ถูกนำมาใช้มากในด้านต่าง ๆ เหล่านี้จะเป็นข้อมูลระยะยาว ซึ่งเป็นข้อมูลที่เก็บหรือวัดมาจากหน่วยตัวอย่างเดิม ซ้ำ ๆ กันหลายครั้งตามระยะเวลา ข้อมูลที่ได้มาจะไม่เป็นอิสระต่อกัน ยกตัวอย่าง เช่น การเก็บข้อมูลน้ำหนักรายสัปดาห์ของผู้ที่เข้าร่วมโปรแกรมลดน้ำหนักที่คลินิกแห่งหนึ่ง น้ำหนักของแต่ละคนจะถูกวัดซ้ำ ๆ กัน ตลอดจนจบโปรแกรม ซึ่งข้อมูลที่ได้มาจากคน ๆ เดียวกันจะไม่เป็นอิสระต่อกัน หรือ การสอบวัดระดับวิชาคณิตศาสตร์ของนักเรียนในโรงเรียนแห่งหนึ่งเป็นประจำทุกปีตั้งแต่ชั้นปีที่ 1 ถึง 6 เพื่อดูแนวโน้มผลการเรียนวิชาคณิตศาสตร์ของนักเรียนแต่ละคน เป็นต้น

การเก็บข้อมูลแบบระยะยาวนั้นมีประโยชน์สำหรับงานวิจัยในหลาย ๆ สาขาที่ต้องการศึกษาการเปลี่ยนแปลงของตัวแปรตาม หรือตัวแปรที่สนใจตามระยะเวลาที่เปลี่ยนไป ดังนั้นเวลาเป็นตัวแปรหนึ่งที่สำคัญในการเก็บข้อมูลลักษณะนี้ และลำดับของเวลาก็เป็นสิ่งที่สำคัญอีกอันหนึ่งเพราะค่าสังเกตที่ได้ในช่วงระยะเวลาที่ใกล้เคียงกันจะมีค่าใกล้เคียงกันมากกว่าค่าสังเกตที่ได้จากช่วงเวลาที่ห่างกัน (ใกล้รุ่ง, 2558) โดยในระหว่างการเก็บข้อมูลระยะยาวอาจจะเกิดการสูญหายของข้อมูลได้ ซึ่งเป็นเรื่องปกติที่พบได้ในการเก็บข้อมูลลักษณะนี้ หากเรานำข้อมูลนั้นมาวิเคราะห์ทางสถิติจะทำให้ผลการวิเคราะห์ การสรุปผล และการวิจารณ์ผลไม่น่าเชื่อถือเท่าที่ควร ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาเกี่ยวกับความแตกต่างของกลไกการสูญหายของข้อมูลระยะยาวที่มีต่อค่าประมาณของสัมประสิทธิ์การถดถอยตัวแบบผสมเชิงเส้น ซึ่งกลไกการสูญหายของข้อมูลจะแบ่งเป็น 3 ประเภท ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing completely at random: MCAR) การสูญหายแบบสุ่ม (Missing at random: MAR) และการสูญหายแบบไม่สุ่ม (Missing not at random: MNAR)

ดังนั้น วัตถุประสงค์ของงานวิจัยนี้จึงเพื่อศึกษาผลกระทบของการสูญหายของข้อมูลแต่ละประเภทในขนาดตัวอย่างและอัตราการสูญหายที่แตกต่างกันโดยใช้เกณฑ์ความเอนเอียงเทียม (Pseudo-bias) และรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียม (Pseudo-root mean square error: RMSE) ในการเปรียบเทียบ

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. กลไกการสูญหายของข้อมูล

กลไกการสูญหายของข้อมูลสามารถจำแนกได้ทั้งหมด 3 ประเภท (ทัตดา, 2555; Ibrahim and Molenberghs, 2009; Rubin, 1976; Weiss, 2005) ได้แก่

1. การสูญหายแบบสุ่มสมบูรณ์ (Missing Completely at Random หรือ MCAR) เป็นการสูญหายที่ข้อมูลสูญหายที่จุดใด ๆ ไม่มีความสัมพันธ์กับค่าของข้อมูลอื่น ๆ ยกตัวอย่างเช่น ในการเก็บข้อมูลน้ำหนักรายสัปดาห์ของผู้ที่เข้าร่วมโปรแกรมลดน้ำหนักข้างต้น ข้อมูลสูญหายประเภทนี้อาจเกิดจากการการที่ผู้เข้าร่วมโปรแกรมบางคนเดินทางไปต่างจังหวัด หรือย้ายถิ่นฐาน จึงไม่ได้มาชั่งน้ำหนักตามกำหนดเวลา ซึ่งการสูญหายของข้อมูลไม่มีความสัมพันธ์กับข้อมูลน้ำหนักที่เคยบันทึกไว้

2. การสูญหายแบบสุ่ม (Missing at Random หรือ MAR) เป็นการสูญหายที่ข้อมูลสูญหายขึ้นอยู่กับค่าของข้อมูลอื่น ๆ ที่ทราบค่าข้อมูล แต่ไม่ขึ้นกับค่าของข้อมูลที่สูญหายของตัวเอง จากตัวอย่างการเก็บข้อมูลน้ำหนักรายสัปดาห์ของผู้ที่เข้าร่วมโปรแกรม

ลดน้ำหนักข้างต้น ถ้าผู้เข้าร่วมโปรแกรมมาชั่งน้ำหนักที่คลินิกแล้วพบว่าน้ำหนักตัวเพิ่มขึ้น จึงตัดสินใจไม่มาร่วมโปรแกรมลดน้ำหนักอีกต่อไป ดังนั้น ข้อมูลที่สูญหายจึงขึ้นอยู่กับข้อมูลน้ำหนักที่บันทึกล่าสุด แต่ไม่ขึ้นอยู่กับข้อมูลที่สูญหายไป

3. การสูญหายแบบไม่สุ่ม (Missing Not at Random หรือ MNAR) เป็นการสูญหายที่ข้อมูลสูญหายที่ตำแหน่งใด ๆ ขึ้นอยู่กับค่าตำแหน่งนั้น ๆ จากตัวอย่างการเก็บข้อมูลน้ำหนักรายสัปดาห์ของผู้ที่เข้าร่วมโปรแกรมลดน้ำหนักข้างต้น ถ้าผู้เข้าร่วมโปรแกรมรู้ตัวว่าน้ำหนักจะต้องเพิ่มขึ้นอย่างแน่นอนก่อนมาชั่งน้ำหนักที่คลินิก จึงตัดสินใจไม่มาร่วมโปรแกรมลดน้ำหนักอีกเลย โดยที่คลินิกไม่ได้บันทึกข้อมูลน้ำหนักที่เพิ่มขึ้น ดังนั้น ข้อมูลที่สูญหายจึงขึ้นอยู่กับข้อมูลน้ำหนักที่สูญหายไป

โดยในบทความวิจัยของ Ibrahim and Molenberghs (2009) และ Nakai and Ke (2011) ยังได้กล่าวอีกว่า เราสามารถเพิกเฉยต่อข้อมูลที่มีการสูญหายแบบ MCAR และ MAR ได้ นั่นคือ เราสามารถใช้สถิติอนุมานในการวิเคราะห์ข้อมูลที่เหลืออยู่ต่อไปได้ แต่ถ้าข้อมูลมีการสูญหายแบบ MNAR เรามิอาจจะเพิกเฉยได้ ซึ่งในการประมาณค่าพารามิเตอร์ของข้อมูลสูญหายประเภทนี้จะมีความซับซ้อนมากกว่า

2. ตัวแบบผสมเชิงเส้น

การถดถอยผสมเชิงเส้น (Linear mixed effect regression) เป็นวิธีการทางสถิติที่ใช้ในการวิเคราะห์ข้อมูลระยะยาว เพราะข้อมูลลักษณะนี้จะไม่มีความเป็นอิสระต่อกัน โดยการถดถอยผสมเชิงเส้นเป็นส่วนขยายของตัวแบบการถดถอยเชิงเส้น ตัวแบบทั้งสองมีการตีความที่คล้ายกันในรูปแบบที่มีอิทธิพลคงที่ (fixed effect) และค่าคลาดเคลื่อนสุ่ม ความแตกต่างก็คือ การวิเคราะห์การถดถอยผสมเชิงเส้นจะเพิ่มอิทธิพลสุ่ม (random effect) เข้าไปในตัวแบบ (Long, 2012) ซึ่งตัวแบบการถดถอยเชิงเส้นเชิงเดียวของประชากรขนาด N คือ

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; i=1,2,...,N \tag{1}$$

โดยที่ Y_i คือ ค่าสังเกตที่ i ของตัวแปรตาม

β_0 และ β_1 คือ พารามิเตอร์ของตัวแบบหรือสัมประสิทธิ์การถดถอย

x_i คือ ค่าสังเกตที่ i ของตัวแปรอิสระ

ε_i คือ พจน์ค่าคลาดเคลื่อนสุ่ม ; $\varepsilon_{ij} \sim N(0,1)$

ส่วนรูปแบบมาตรฐานอย่างง่ายของตัวแบบการถดถอยผสมเชิงเส้นเชิงเดียวคือ

$$Y_{ij} = (\beta_0 + b_{0i}) + \beta_1 x_{ij} + \varepsilon_{ij} \; ; i=1,2,...,N \; ; j=1,2,...,n_i \tag{2}$$

โดยที่ Y_{ij} คือ ค่าสังเกตของ subject (ประชากรที่เราศึกษา) ที่ i ณ เวลาที่ j ของตัวแปรตาม

β_0 และ β_1 คือ พารามิเตอร์ของตัวแบบหรือสัมประสิทธิ์การถดถอย

x_{ij} คือ ค่าสังเกตของ subject ที่ i ณ เวลาที่ j ของตัวแปรอิสระ

b_{0i} คือ ค่าสัมประสิทธิ์อิทธิพลสุ่มของ subject ที่ i

ε_{ij} คือ พจน์ค่าคลาดเคลื่อนสุ่มของ subject ที่ i ณ เวลาที่ j

n_i คือ จำนวนครั้งที่วัดซ้ำของ subject ที่ i

ยกตัวอย่างเช่น ในการเก็บข้อมูลน้ำหนักรายสัปดาห์ของผู้ที่เข้าร่วมโปรแกรมลดน้ำหนักที่คลินิกแห่งหนึ่งดังที่ได้กล่าวไว้แล้วข้างต้น ผู้วิจัยได้ทำการทดลองกับอาสาสมัคร (subject) 10 คน โดยให้แต่ละคนรับประทานยาลดน้ำหนักและชั่งน้ำหนักทุกสัปดาห์เป็นเวลา 8 สัปดาห์ (n_i) เพื่อดูการเปลี่ยนแปลงของน้ำหนัก เราจะได้ข้อมูลมาทั้งหมด 80 ค่าสังเกตที่ไม่เป็นอิสระต่อกันเพราะข้อมูล 8 ค่าถูกสังเกตมาจากคนคนเดียวกัน ดังนั้นจึงไม่สามารถใช้การวิเคราะห์การถดถอยเชิงเส้นเชิงเดียวได้

จากสมการ (2) เราสามารถเขียนตัวแบบการถดถอยผสมเชิงเส้นในรูปของเมทริกซ์ ดังนี้

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon} \tag{3}$$

โดยที่ $\mathbf{y}$ เป็นเวกเตอร์ของค่าสังเกต Y_{ij} ขนาด q x 1 ; $q = \sum_{i=1}^{N} n_i$

$\mathbf{X}$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด q x 2

$\boldsymbol{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด 2 x 1

$\mathbf{Z}$ เป็นเมทริกซ์แผนแบบของอิทธิพลสุ่มขนาด q x N

$\mathbf{b}$ เป็นเวกเตอร์ของอิทธิพลสุ่มขนาด N x 1 โดยที่ $E(\mathbf{b}) = \mathbf{0}$, $Var(\mathbf{b}) = \sigma_b^2 \mathbf{I}_N$

$\boldsymbol{\varepsilon}$ เป็นเวกเตอร์ของค่าคลาดเคลื่อนสุ่ม ε_{ij} ขนาด q x 1 โดยที่ $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $Var(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_q$

ในการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้นนั้น โดยทั่วไปจะใช้วิธีภาวะน่าจะเป็นสูงสุด (maximum likelihood method) (Bates et al., 2015; Searle et al., 2006) โดยมีฟังก์ชันล็อกภาวะน่าจะเป็น ดังนี้

$$\log L = -\frac{1}{2} q \log 2\pi - \frac{1}{2} \log|\mathbf{V}| - \frac{1}{2} (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \quad (4)$$

โดยที่ $V=Var(\mathbf{y}) = \sigma_b^2 ZZ' + \sigma_\varepsilon^2 \mathbf{I}_q$ ซึ่งจะทำให้ได้ตัวประมาณภาวะน่าจะเป็นสูงสุด ดังนี้

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}) \quad (5)$$

โดยวิธีการประมาณค่าดังกล่าวเป็นวิธีการที่ใช้ในคำสั่ง lmer ในการวิเคราะห์การถดถอยผสมเชิงเส้นในโปรแกรม R ซึ่งผู้วิจัยใช้ในการศึกษาครั้งนี้

วิธีการดำเนินการวิจัย

การวิจัยใช้การจำลองข้อมูลบนโปรแกรม R โดยกำหนดขนาดตัวอย่าง (n) ซึ่งมี 4 ระดับ ดังนี้ คือ n เท่ากับ 5 10 20 และ 50 โดยในแต่ละระดับของขนาดตัวอย่างจะมีจำนวนการวัดซ้ำ 5 ครั้ง และมีข้อมูลทั้งหมดเท่ากับ n x 5 โดยมีขั้นตอนการดำเนินการดังนี้

1. จำลองตัวแปรตาม Y จากสมการ $Y_{ij} = b_{0i} + \varepsilon_{ij}$ โดยที่ $b_{0i} \sim N(0,100)$ และ $\varepsilon_{ij} \sim N(0,1)$ ซึ่งในการศึกษานี้ Y คือข้อมูลระยะยาวที่มีความสมบูรณ์หรือข้อมูลที่ยังไม่เกิดการสูญหาย

2. นำข้อมูลที่มีความสมบูรณ์หรือข้อมูลที่ยังไม่เกิดการสูญหายนั้นมาทำให้เกิดการสูญหายตามลักษณะของกลไกการสูญหายของข้อมูล 3 ประเภท โดยกำหนดร้อยละการสูญหายของแต่ละประเภทการสูญหายเท่ากับ 10 และ 20 โดยเป็นค่าร้อยละเทียบกับข้อมูลที่ยังไม่เกิดการสูญหาย

3. หาค่าประมาณสัมประสิทธิ์ถดถอยผสมเชิงเส้น $\hat{\beta}_0^*$ และ $\hat{\beta}_1^*$ ของข้อมูลเริ่มต้นที่ยังไม่เกิดการสูญหาย และหาค่าประมาณสัมประสิทธิ์ถดถอยผสมเชิงเส้น $\hat{\beta}_0$ และ $\hat{\beta}_1$ ของกลไกการสูญหายแต่ละแบบ ในโปรแกรม R โดยใช้คำสั่ง lmer ซึ่งอยู่ใน package lme4 จากตัวแบบถดถอยผสมเชิงเส้น (2) โดยให้ $x_{ij} = j \; ; \; j = 1, 2, ..., 5$

4. คำนวณผลต่างระหว่างค่าประมาณสัมประสิทธิ์ถดถอยผสมเชิงเส้นของกลไกการสูญหายแต่ละแบบและค่าประมาณสัมประสิทธิ์ถดถอยผสมเชิงเส้นของข้อมูลเริ่มต้นที่ยังไม่เกิดการสูญหาย

5. ทำซ้ำ 1,000 ครั้ง

6. เปรียบเทียบค่าประมาณสัมประสิทธิ์การถดถอยผสมเชิงเส้นของแต่ละกลไกการสูญหายของข้อมูลโดยใช้เกณฑ์ในการเปรียบเทียบ ดังนี้

ความเอนเอียงเทียม (Pseudo-bias) หาได้จาก

$$\text{Pseudo-bias}_{\hat{\beta}_0^*}[\hat{\beta}_0] = \frac{\sum_{k=1}^{1000} (\hat{\beta}_{0k} - \hat{\beta}_{0k}^*)}{1000}$$

$$\text{Pseudo-bias}_{\hat{\beta}_1^*}[\hat{\beta}_1] = \frac{\sum_{k=1}^{1000}(\hat{\beta}_{1k} - \hat{\beta}_{1k}^*)}{1000}$$

ค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียม (Pseudo-RMSE)

$$\text{Pseudo-RMSE}\ [\hat{\beta}_0] = \sqrt{\frac{\sum_{k=1}^{1000}(\hat{\beta}_{0k} - \hat{\beta}_{0k}^*)^2}{1000}}$$

$$\text{Pseudo-RMSE}\ [\hat{\beta}_1] = \sqrt{\frac{\sum_{k=1}^{1000}(\hat{\beta}_{1k} - \hat{\beta}_{1k}^*)^2}{1000}}$$

ผลการวิจัย

จากการวิเคราะห์ค่าประมาณของสัมประสิทธิ์การถดถอยผสมเชิงเส้นเมื่อมีข้อมูลสูญหายแบบ MCAR, MAR และ MNAR ในการวิเคราะห์ข้อมูลระยะยาวผ่านวิธีการจำลองภายใต้ค่าคลาดเคลื่อนสุ่มและอิทธิพลสุ่มที่มีการแจกแจงปรกติ โดยกำหนดร้อยละการสูญหายของแต่ละประเภทการสูญหายเท่ากับ 10 และ 20 และขนาดของตัวอย่างที่ใช้ในการศึกษานี้คือ 5 10 20 และ 50 โดยทำการจำลองข้อมูลซ้ำจำนวน 1,000 ครั้ง พบว่าผลเป็นดังตารางที่ 1 และ 2

จากตารางที่ 1 ที่ร้อยละการสูญหายของข้อมูล เท่ากับ 10 พบว่าถ้ามีขนาดตัวอย่างเท่ากับ 5 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณจะมีค่าต่ำกว่าค่าจริง สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR จะทำให้มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ถ้ามีขนาดตัวอย่างเท่ากับ 10 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริง สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ถ้ามีขนาดตัวอย่างเท่ากับ 20 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริง สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่สูงว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ถ้ามีขนาดตัวอย่างเท่ากับ 50 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริงมาก สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น ข้อมูลที่มีการสูญหายแบบ MAR มีค่าน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีค่ามากที่สุด สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MCAR และ MNAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ตารางที่ 1 ค่า Pseudo-bias และ Pseudo-RMSE ของค่าประมาณสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้น ในแต่ละประเภทการสูญหาย เมื่อร้อยละการสูญหายเท่ากับ 10

ขนาดตัวอย่าง	ตัวประมาณ	ประเภทการสูญหาย	เกณฑ์การเปรียบเทียบ	
			Pseudo-bias	Pseudo-RMSE
n=5	$\hat{\beta}_0$	MCAR	0.0046	0.1942
		MAR	0.0038	0.1537
		MNAR	-0.0899	0.1956
	$\hat{\beta}_1$	MCAR	-0.0009	0.0588
		MAR	-0.0024	0.0906
		MNAR	0.0010	0.0540
n=10	$\hat{\beta}_0$	MCAR	0.0058	0.1365
		MAR	-0.0021	0.0903
		MNAR	-0.1211	0.1758
	$\hat{\beta}_1$	MCAR	-0.0010	0.0422
		MAR	0.0003	0.0530
		MNAR	0.0002	0.0369
n=20	$\hat{\beta}_0$	MCAR	-0.0007	0.0939
		MAR	0.0000	0.0602
		MNAR	-0.9476	1.0329
	$\hat{\beta}_1$	MCAR	0.0005	0.0287
		MAR	-0.0001	0.0355
		MNAR	-0.0003	0.0259
n=50	$\hat{\beta}_0$	MCAR	0.0002	0.0610
		MAR	0.0017	0.0384
		MNAR	-1.3679	1.4169
	$\hat{\beta}_1$	MCAR	-0.0001	0.0184
		MAR	-0.0010	0.0227
		MNAR	-0.0001	0.0153

จากตารางที่ 2 ที่ร้อยละการสูญหายของข้อมูล เท่ากับ 20 พบว่าถ้ามีขนาดตัวอย่างเท่ากับ 5 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริง สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและค่าราก

ของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริง

ถ้ามีขนาดตัวอย่างเท่ากับ 10 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริงมาก สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่สูงกว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ถ้ามีขนาดตัวอย่างเท่ากับ 20 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริงมาก สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MCAR มีความเอนเอียงเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมมากที่สุด โดยให้ค่าประมาณที่สูงว่าค่าจริง สำหรับค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การสูญหายแบบ MNAR มีค่าน้อยที่สุด ในขณะที่การสูญหายแบบ MAR มีค่ามากที่สุด

ถ้ามีขนาดตัวอย่างเท่ากับ 50 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบนี้จะมีค่าต่ำกว่าค่าจริงมาก สำหรับค่าประมาณ $\hat{\beta}_1$ ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยให้ค่าประมาณที่ต่ำกว่าค่าจริง

ตารางที่ 2 ค่า Pseudo-bias และ Pseudo-RMSE ของค่าประมาณสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้น ในแต่ละประเภทการสูญหาย เมื่อร้อยละการสูญหายเท่ากับ 20

ขนาดตัวอย่าง	ตัวประมาณ	ประเภทการสูญหาย	เกณฑ์การเปรียบเทียบ	
			Pseudo-bias	Pseudo-RMSE
n=5	$\hat{\beta}_0$	MCAR	0.0070	0.2992
		MAR	0.0042	0.2125
		MNAR	-0.2319	0.3418
	$\hat{\beta}_1$	MCAR	-0.0025	0.0906
		MAR	-0.0034	0.1258
		MNAR	-0.0003	0.0733
n=10	$\hat{\beta}_0$	MCAR	-0.0034	0.2260
		MAR	-0.0020	0.1374
		MNAR	-1.7501	1.9125
	$\hat{\beta}_1$	MCAR	0.0003	0.0639
		MAR	0.0012	0.0813
		MNAR	0.0015	0.0538

ตารางที่ 2 ค่า Pseudo-bias และ Pseudo-RMSE ของค่าประมาณสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้น ในแต่ละประเภทการสูญหาย เมื่อร้อยละการสูญหายเท่ากับ 20 (ต่อ)

ขนาดตัวอย่าง	ตัวประมาณ	ประเภทการสูญหาย	เกณฑ์การเปรียบเทียบ	
			Pseudo-bias	Pseudo-RMSE
n=20	$\hat{\beta}_0$	MCAR	-0.0047	0.1537
		MAR	-0.0019	0.0968
		MNAR	-2.4055	2.5149
	$\hat{\beta}_1$	MCAR	0.0009	0.0434
		MAR	0.0013	0.0573
		MNAR	0.0018	0.0381
n=50	$\hat{\beta}_0$	MCAR	0.0041	0.0946
		MAR	0.0022	0.0609
		MNAR	-2.7327	2.7874
	$\hat{\beta}_1$	MCAR	-0.0010	0.0269
		MAR	-0.0011	0.0358
		MNAR	0.0003	0.0245

สรุปผลการวิจัย

การศึกษาเพื่อเปรียบเทียบค่าประมาณสัมประสิทธิ์การถดถอยของตัวแบบผสมเชิงเส้นบนข้อมูลระยะยาวที่มีค่าสูญหายแบบ MCAR, MAR และ MNAR ผ่านวิธีการจำลองด้วยโปรแกรม R ด้วยการใช้คำสั่ง lmer ซึ่งใช้วิธีภาวะน่าจะเป็นสูงสุด ในการประมาณค่าสัมประสิทธิ์การถดถอย โดยใช้ค่าเอนเอียงเทียมและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมในการเปรียบเทียบนั้น พบว่าไม่ว่าร้อยละการสูญหายของข้อมูลจะเท่ากับ 10 หรือ 20 ค่าประมาณ $\hat{\beta}_0$ ของข้อมูลที่มีการสูญหายแบบ MAR จะมีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MNAR มีความเอนเอียงเทียมและมีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุด โดยค่าประมาณของข้อมูลที่มีการสูญหายแบบ MNAR จะมีค่าต่ำกว่าค่าจริงมากขึ้นเรื่อย ๆ เมื่อตัวอย่างมีขนาดใหญ่ขึ้น และมีร้อยละการสูญหายของข้อมูลมากขึ้น ซึ่งสอดคล้องกับงานวิจัยของ Ibrahim and Molenberghs (2009) และ Nakai and Ke (2011) ดังนั้น การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดนี้จึงไม่เหมาะสมสำหรับข้อมูลที่มีการสูญหายแบบ MNAR

สำหรับค่าประมาณ $\hat{\beta}_1$ นั้น ไม่ว่าร้อยละการสูญหายของข้อมูลจะเท่ากับ 10 หรือ 20 ไม่มีรูปแบบที่แน่ชัดในเรื่องของค่าเอนเอียงเทียม แต่ในส่วนของค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น ข้อมูลที่มีการสูญหายแบบ MNAR จะได้ค่าประมาณที่มีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมน้อยที่สุด ในขณะที่ข้อมูลที่มีการสูญหายแบบ MAR จะได้ค่าประมาณที่มีค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมมากที่สุดในทุก ๆ ขนาดตัวอย่าง

โดยภาพรวม เมื่อพิจารณาทั้งค่าเอนเอียงเทียมและค่ารากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเทียมนั้น การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุดนี้เหมาะสมสำหรับข้อมูลที่มีกลไกการสูญหายแบบ MCAR เท่านั้น แต่ไม่เหมาะสมสำหรับข้อมูลมีการสูญหายแบบ MNAR โดยเฉพาะเมื่อตัวอย่างมีขนาดใหญ่

ในการศึกษาครั้งนี้ ผู้วิจัยใช้คำสั่ง lmer ในโปรแกรม R ซึ่งใช้วิธีภาวะน่าจะเป็นสูงสุดในการประมาณค่าสัมประสิทธิ์การถดถอย ในการศึกษาครั้งต่อไป เราอาจจะพิจารณาวิธีการประมาณค่าอื่น ๆ ที่อาจจะมีความเหมาะสมมากกว่าสำหรับข้อมูลที่มีการสูญหายแบบ MNAR และในกรณีที่ข้อมูลมีการสูญหายมากขึ้น

เอกสารอ้างอิง

ใกล้รุ่ง สามารถ. (2558). การวิเคราะห์ข้อมูลระยะยาว. วารสารคณิตศาสตร์ 60(680-682): 7-12.

ทัตดา หิรัญพต. (2555). การเปรียบเทียบประสิทธิภาพระหว่างวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยและวิธีการประมาณข้อมูลสูญหายด้วยค่าถดถอยแบบสโทแคสติก. ปัญหาพิเศษวิทยาศาสตรบัณฑิต, มหาวิทยาลัยบูรพา. ชลบุรี. 24 หน้า.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1): 1-48.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. TEST 18(1): 1-43.

Long, J. D. (2012). Longitudinal Data Analysis for the Behavioral Sciences Using R. Thousand Oaks: SAGE. pp. 160-163.

Nakai, M. and Ke, W. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. International Journal of Mathematical Analysis 5(1): 1-13.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3): 581-592.

Searle, S. R., Casella, G. and McCulloch, C. E. (2006). Variance Components. New Jersey: John Wiley and Sons. pp. 233-234.

Weiss, R. E. (2005). Modeling Longitudinal Data. Los Angeles: Springer. pp. 365-366.

❑❑❑❑❑