การศึกษาเปรียบเทียบการประมาณค่าจากตัวแบบการถดถอย สำหรับข้อมูลที่มีการแจกแจงแบบล็อกนอร์มอล ที่ถูกตัดปลายทางขวาแบบสุ่มที่มีการแจกแจงแบบเบตา

A Comparative Study on Estimation from Regression Mode for Data from Lognormal Distribution Under Random Right-Censoring from Beta Distribution

ธัญพิชชา ยอดแก้ว[1*] และอนุภาพ สมบูรณ์สวัสดี[2*]

Thunpischa Yodkaew[1*] and Anupap Somboonsavatdee[2*]

[1] นักศึกษาระดับปริญญาโท, วิทยาศาสตรมหาบัณฑิต, สถิติ, จุฬาลงกรณ์มหาวิทยาลัย

[1] Graduate student, Department of Statistics, Chulalongkorn Business School, Chulalongkorn University.

[2] ผู้ช่วยศาสตราจารย์ ดร., ภาควิชาสถิติ, คณะพาณิชยศาสตร์และการบัญชี, จุฬาลงกรณ์มหาวิทยาลัย

[2] Assistant Professor, Ph.D., Department of Statistics, Chulalongkorn Business School, Chulalongkorn University.

[*]**Corresponding author, E-mail:** 6181528126@student.chula.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบการประมาณค่าจากตัวแบบการถดถอย สำหรับข้อมูลที่มีการแจกแจงแบบล็อกนอร์มอล ที่ถูกตัดปลายทางขวาแบบสุ่มที่มีการแจกแจงแบบเบตา ด้วยวิธีการประมาณค่าแบบกำลังสองต่ำสุด (OLS) วิธีของแชตเทอร์จีและแมคลีช (CM) วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม (MLE_EM) วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่าเฉลี่ย (MLE_EM_MEAN) และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่ามัธยฐาน (MLE_EM_MED) เปรียบเทียบจากค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสอง โดยจำลองข้อมูลทั้งหมด 216 สถานการณ์ พบว่า เมื่อข้อมูลมีขนาดเล็ก (n=30) และมีเปอร์เซ็นต์ในการถูกตัดปลายทางขวาน้อย (r_1=10) วิธี OLS และ CM เป็นวิธีที่มีประสิทธิภาพสูงสุด แตกต่างกันตามลักษณะการกระจายตัวของตัวแปรอิสระและค่าคลาดเคลื่อน ในขณะที่หากตัวอย่างมีขนาดเล็ก (n=30) และมีเปอร์เซ็นต์ในการถูกตัดปลายทางขวามาก (r_1=30) หรือตัวอย่างมีขนาดใหญ่ (n=100) วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุด แบ่งตามช่วงการเข้ามาของข้อมูล กล่าวคือ เมื่อข้อมูลเข้ามาในช่วงต้นของการเปิดรับ วิธี MLE_EM_MED มีประสิทธิภาพสูงสุด ในขณะที่เมื่อข้อมูลเข้ามาในช่วงกลางของการเปิดรับ วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุด และเมื่อข้อมูลเข้ามาในช่วงท้ายของการเปิดรับ วิธี MLE_EM_MEAN ประสิทธิภาพสูงสุด

คำสำคัญ: ข้อมูลที่ถูกตัดปลายทางขวาแบบสุ่ม, ตัวแบบถดถอย, การประมาณค่า

Abstract

The purpose of this study is to compare the estimation methods of linear regression model for data from lognormal distribution under random right-censoring from beta distribution which are Ordinary Least Squares Method (OLS), Chatterjee and McLeish Method (CM), Maximum Likelihood Estimation using the EM algorithm Method (MLE_EM), Maximum Likelihood Estimation Method using the EM algorithm with mean-adjusted data (MLE_EM_MEAN), and Maximum Likelihood Estimation Method using the EM algorithm with median-adjusted data (MLE_EM_MED) by generating data 216 scenarios. The results present that OLS and CM are the most efficient methods if the sample size is small with small censoring proportion (r_1=10) varies by variance of independent variables and variance of error. If the sample size is small (n=30) with large censoring proportion (r_1=30) or the sample size is large (n=100) the most efficient methods are MLE_EM family methods separated by timing of data collecting point. When data is collected in early, middle, and end of time, MLE_EM_MED, MLE_EM family, and MLE_EM_MED is the most efficient method, respectively.

Keywords: RANDOMLY RIGHT-CENSORED DATA, REGRESSION MODEL, ESTIMATION

บทนำ

ในโครงการวิจัยหนึ่งต้องการติดตามระยะการอยู่รอดของคนป่วยหลังเข้ารับการรักษา เพื่อเตรียมแผนการรักษาที่มีประสิทธิภาพมากขึ้นในอนาคต โดยหาความสัมพันธ์ระหว่างระยะการอยู่รอดกับปัจจัยต่าง ๆ โดยใช้เทคนิคในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรคือการวิเคราะห์การถดถอย ซึ่งในการเปิดรับสมัครคนเข้าโครงการนั้น จะมีลักษณะการเข้ามาสมัครในรูปแบบต่าง ๆ เช่น คนสมัครกระจุกตัวในช่วงแรกหรือช่วงท้ายของการเปิดรับสมัคร ,คนสมัครเข้ามาเรื่อย ๆ อย่างสม่ำเสมอ จากลักษณะเหล่านี้จึงทำให้จุดเริ่มต้นของการเก็บข้อมูลแตกต่างกันด้วย ตัวอย่างเช่น การติดตามระยะการมีชีวิตรอดของผู้ติดเชื้อเอชไอวี โดยการหาความสัมพันธ์จากปัจจัยต่าง ๆ ที่ส่งผลต่อระยะการอยู่รอดของผู้ป่วยเอชไอวี เช่น ปัจจัยจากบุคคล (เพศ, อายุ) ปัจจัยจากโรค (ระยะอาการของโรคขณะที่ตรวจพบ (stage), ระดับ CD4) เป็นต้น โดยผู้ป่วยแต่ละคนมีการสมัครเข้าร่วมโครงการในเวลาที่แตกต่างกัน หากผู้ป่วยเสียชีวิตระหว่างโครงการ จะถูกบันทึกระยะเวลาการอยู่รอดตามความเป็นจริง แต่ถ้าหากยังมีชีวิตอยู่ไปจนถึงเวลาที่สิ้นสุดการติดตามหรือหลังจากนั้น จะถูกบันทึกระยะเวลาการอยู่รอดเป็นจุดเวลาที่สิ้นสุดโครงการ ซึ่งต่ำกว่าความเป็นจริง หรือเรียกว่าข้อมูลที่ถูกตัดปลายประเภทที่ 1 โดยจากการศึกษาพบว่า

ธนาพิพัฒน์ ทรัพย์ครองชัย (2561) ได้ศึกษาวิธีการประมาณค่าตัวแปรตามที่ข้อมูลบางค่าถูกตัดปลายทางขวาประเภทที่ 1 ซึ่งมีการกำหนดช่วงเวลาในการเริ่มเก็บข้อมูลและจุดเวลาที่เริ่มเก็บข้อมูลมีการแจกแจงแบบสม่ำเสมอ ด้วยวิธีกำลังสองต่ำสุด วิธีภาวะน่าจะเป็นสูงสุด วิธีของแชตเทอร์จีและแมคลีช วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็มเมื่อมีการปรับค่าข้อมูลก่อนคำนวณ พบว่าตัวอย่างขนาดกลางและใหญ่หรือตัวแปรตามถูกตัดปลายทางขวาปานกลางและมาก วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็มเมื่อมีการปรับค่าข้อมูลก่อนคำนวณ มีประสิทธิภาพมากที่สุด หากตัวอย่างมีขนาดเล็กหรือตัวประอิสระมีการกระจายตัวน้อยกว่าความคลาดเคลื่อน วิธีกำลังสองต่ำสุดมีประสิทธิภาพมากที่สุด แต่หากตัวแปรอิสระมีการกระจายตัวมากกว่าหรือเท่ากับความคลาดเคลื่อน วิธีของแชตเทอร์จีและแมคลีชจะมีประสิทธิภาพมากที่สุด และนอกจากนั้นยังพบว่าเมื่อขนาดตัวอย่างยิ่งใหญ่ขึ้น หรือตัวแปรตามถูกตัดปลายทางขวาน้อยลง หรือสัดส่วนช่วงเวลาที่เริ่มเก็บข้อมูลต่อระยะเวลาการเก็บข้อมูลลดลง หรือความคลาดเคลื่อนกระจายตัวน้อยกว่าตัวแปรอิสระ ยิ่งส่งผลให้ทุกวิธีมีประสิทธิภาพมากขึ้น

สำหรับงานวิจัยนี้จะศึกษาเพื่อเปรียบเทียบการประมาณค่าจากตัวแบบการถดถอย สำหรับข้อมูลที่มีการแจกแจงแบบล็อกนอร์มอล ที่ถูกตัดปลายทางขวาแบบสุ่มที่มีการแจกแจงแบบเบตา ด้วยวิธีการประมาณค่าแบบกำลังสองต่ำสุด วิธีของแชตเทอร์จีและแมคลีช วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่าเฉลี่ย และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่ามัธยฐาน โดยเปรียบเทียบจากค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสอง

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบวิธีประมาณค่าแบบต่าง ๆ ที่เหมาะสมกับข้อมูลที่มีการแจกแจงแบบล็อกนอร์มอล ที่ถูกตัดปลายทางขวาแบบสุ่มที่มีการแจกแจงแบบเบตา

แนวคิด ทฤษฎี กรอบแนวคิด

1. ข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 (Type I censoring data)

Klein และ Moeschberger (2003) ให้ความหมายของข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 ว่าเป็นข้อมูลมีการถูกกำหนดเวลาสิ้นสุดการสังเกตที่แน่นอนไว้แล้ว หากข้อมูลที่สิ้นสุดก่อนเวลาที่กำหนดจะไม่ถือว่าเป็นข้อมูลที่ถูกตัดทิ้ง แต่ถ้าหากถึงเวลาที่กำหนดแล้วข้อมูลยังไม่มีการสิ้นสุดหรือสิ้นสุดหลังเวลาที่กำหนด จะถือเอาเวลาที่กำหนดเป็นเวลาสิ้นสุดของข้อมูลนั้นแทน

กำหนด C_r คือเวลาสิ้นสุดการสังเกต

X_i คือระยะเวลารอดชีวิตของแต่ละคน เมื่อ $i = 1,2,\dots,n$

T_i คือระยะเวลาที่สังเกตได้

δ_i คือตัวแปรที่บอกเหตุการณ์

โดย

$$T_i = \begin{cases} X_i & \text{เมื่อข้อมูลสิ้นสุดก่อนถึงเวลาสิ้นสุดการสังเกต หรือ } X_i \leq C_r \\ C_r & \text{เมื่อถึงเวลาที่สิ้นสุดการสังเกตแล้วแต่ยังมีชีวิตรอดอยู่ หรือ } X_i > C_r \end{cases}$$

$$= \min(X_i, C_r)$$

$$\delta_i = \begin{cases} 0 & \text{เมื่อเป็นข้อมูลที่ไม่ถูก censored หรือ } X_i \leq C_r \\ 1 & \text{เมื่อเป็นข้อมูลที่ถูก censored หรือ } X_i > C_r \end{cases}$$

Jöreskog (2002) กล่าวถึงฟังก์ชันภาวะน่าจะเป็นของข้อมูล ที่มีบางส่วนของข้อมูลถูกตัดปลายทางขวาประเภทที่ 1 ดังนี้

$$L(T_i) = \begin{cases} f(X_i) & \text{เมื่อ } X_i \leq C_r \text{ โดย } i = 1,2,\dots,m \\ P(X_i > C_r) = S(C_r) & \text{เมื่อ } X_i > C_r \text{ โดย } i = m+1, m+2, \dots, n \end{cases}$$

โดย $f(X_i)$ คือฟังก์ชันความหนาแน่นของ X_i และ $S(C_r)$ คือฟังก์ชันการอยู่รอดและมีฟังก์ชันภาวะน่าจะเป็นรวม ดังนี้

$$L = \prod_{i=1}^{m} f(X_i) \cdot \prod_{i=m+1}^{n} S(C_r)$$

ข้อมูลที่ถูกตัดปลายทางขวาอย่างสุ่ม (Random Right-censoring) เป็นข้อมูลมีการถูกกำหนดเวลาสิ้นสุดการสังเกตที่แน่นอนไว้แล้วอย่างสุ่ม ทำให้แต่ละข้อมูลจะมีเวลาสิ้นสุดการสังเกตที่แตกต่างกัน สามารถเขียนสัญลักษณ์ได้เป็น (T_i, δ_i)

โดย T_i บอกเวลาที่บันทึกได้จากการสังเกตของข้อมูลลำดับที่ i

δ_i บอกเหตุการณ์ว่าข้อมูลลำดับที่ i เป็นข้อมูลที่ถูกตัดปลายหรือไม่

$C_{r,i}$ เป็นเวลาที่สิ้นสุดการสังเกตของข้อมูลลำดับที่ i

2. การวิเคราะห์ความถดถอย

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระ X_{pi} และตัวแปรตาม Y_i โดยมีสัมประสิทธิ์ β_p บอกว่าตัวแปรอิสระ X_{pi} มีความสัมพันธ์กับ Y_i มากน้อยเพียงใดและมีความสัมพันธ์ไปในทิศทางใด โดยที่ความสัมพันธ์อยู่ในรูปเชิงเส้น และสามารถเขียนสมการแสดงความสัมพันธ์ได้ดังนี้

$$Y_i' = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

เมื่อ $i = 1,2,\dots,n$

หรือแสดงในรูปเมทริกซ์คือ $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

เมื่อ $\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{p1} \\ 1 & X_{12} & \dots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \dots & X_{pn} \end{bmatrix}_{n\times(p+1)}$, $\mathbf{Y}' = \begin{bmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_n' \end{bmatrix}_{n\times 1}$, $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$, $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}_{n\times 1}$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}_{(p+1)\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

โดย Y_i' คือ ค่าจริงของตัวแปรตาม ลำดับที่ i

Y_i คือ ค่าสังเกตของตัวแปรตาม ลำดับที่ i

$\hat{Y}_i$ คือ ค่าประมาณของตัวแปรตาม ลำดับที่ i

X_{pi} คือ ตัวแปรอิสระตัวที่ p ลำดับที่ i

β_0 คือ ระยะตัดแกน y ที่แท้จริง

β_p คือ สัมประสิทธิ์การถดถอยเชิงเส้นของตัวแปรอิสระที่ p

$\hat{\beta}_0$ คือ ค่าประมาณของระยะตัดแกน y ที่แท้จริง

$\hat{\beta}_p$ คือ ค่าประมาณของสัมประสิทธิ์การถดถอยเชิงเส้นของตัวแปรอิสระที่ p

ε_i คือ ความคลาดเคลื่อนอย่างสุ่ม

e_i คือ ความคลาดเคลื่อน โดย $e_i = Y_i - \hat{Y}_i$

$\hat{\beta}_p$ คือ ค่าประมาณของสัมประสิทธิ์การถดถอยเชิงเส้นของตัวแปรอิสระที่ p

3. วิธีประมาณที่ใช้ในการศึกษา

ช่วงเวลาในการศึกษาข้อมูลในงานวิจัยนี้แบ่งออกเป็น 2 ส่วน คือ

1) ช่วงเวลาที่เปิดรับข้อมูลเข้ามาเพื่อศึกษา ซึ่งคิดเป็นร้อยละ 10 และ 30 ของช่วงเวลาการศึกษาข้อมูล โดยจุดเวลาที่ข้อมูลเข้ามาในระบบมีการแจกแจงแบบเบตาที่มีการปรับค่าต่ำสุดและสูงสุดเป็นช่วงของการเปิดรับข้อมูลเพื่อเข้ามาศึกษา

2) ช่วงเวลาการศึกษาข้อมูล โดยกำหนดจุดเริ่มต้นศึกษาข้อมูลเป็นค่าคาดหวังของการแจกแจง $\text{Beta}(\alpha, \theta)$

นิยาม random censoring ratio (r_2) คือ สัดส่วนของช่วงเวลาที่เปิดรับข้อมูลเข้ามาเพื่อศึกษาต่อช่วงเวลาที่ศึกษาข้อมูล นั่นคือ

$$r_2 = \frac{\text{ช่วงเวลาที่เปิดรับข้อมูล}}{\text{ช่วงเวลาที่ศึกษาข้อมูล}}$$

จึงทำให้ค่าที่เป็นไปได้ของ random censoring ratio คือ 0.1, 0.3

กำหนดให้ a คือ ระยะเวลาที่เปิดรับผู้ป่วย

W_i คือ ตัวแปรสุ่มที่มีการแจกแจง $\text{Beta}(\alpha, \theta)$ บนช่วง $[0,1]$

Y_c คือ เวลาสิ้นสุดในการสังเกตข้อมูลที่กำหนดไว้ล่วงหน้า

เนื่องจาก จุดเวลาที่ข้อมูลเข้ามาในระบบอยู่บนช่วง $[0, a]$ จึงต้องปรับค่าของตัวแปรสุ่ม W_i จะได้ aW_i เป็นตัวแปรสุ่มที่มีการแจกแจงเบตาบนช่วง $[0, a]$ ที่แสดงจุดเวลาที่ข้อมูลเข้ามาในระบบ จากนั้นเลื่อนจุดเวลาจากจุดที่ข้อมูลเข้ามาในระบบไปยังเวลาที่เริ่มศึกษาโดยกำหนดให้จุดเวลาที่เริ่มศึกษาข้อมูลเป็นค่าคาดหวังของ aW_i ซึ่งสามารถเขียนในรูปเมทริกซ์ได้ดังนี้

$$\mathbf{W} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}_{n\times 1}, a\mathbf{W} = \begin{bmatrix} aW_1 \\ aW_2 \\ \vdots \\ aW_n \end{bmatrix}_{n\times 1}, E(aW_i) = a\left(\frac{\alpha}{\alpha+\theta}\right),\ a\left(\frac{\alpha}{\alpha+\theta}\right)\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} - a\mathbf{W} = \begin{bmatrix} a\left(\frac{\alpha}{\alpha+\theta}\right) - aW_1 \\ a\left(\frac{\alpha}{\alpha+\theta}\right) - aW_2 \\ \vdots \\ a\left(\frac{\alpha}{\alpha+\theta}\right) - aW_n \end{bmatrix}_{n\times 1}$$

จากการเลื่อนจุดเริ่มศึกษาให้เป็นจุดเดียวกัน ทำให้เวลาสิ้นสุดการสังเกตที่กำหนดไว้ถูกเลื่อนตามไปด้วย ข้อมูลแต่ละตัวจึงมีจุดสิ้นสุดการสังเกตที่แตกต่างกัน เขียนแทนด้วย $Y_{c,i}$ ซึ่งเขียนในรูปเมทริกซ์ได้ดังนี้

$$\mathbf{Y}_{c,i} = \begin{bmatrix} Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W_1\right] \\ Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W_2\right] \\ \vdots \\ Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W_n\right] \end{bmatrix}_{n\times 1} = \begin{bmatrix} Y_{c,1} \\ Y_{c,2} \\ \vdots \\ Y_{c,n} \end{bmatrix}_{n\times 1}$$

กำหนด $Y_i^* = \min(Y_i, Y_{c,i})$ คือ เวลาที่สังเกตได้

กล่าวคือ

$$Y_i^* = \begin{cases} Y_i & , Y_i \le Y_{c,i} \\ Y_{c,i} & , Y_i > Y_{c,i} \end{cases}$$

$Y_i \le Y_{c,i}$ เป็นข้อมูลไม่ถูกตัดปลายทางขวาประเภทที่ 1 มี m ค่า โดย $i = 1,2,\dots,m$

$Y_i > Y_{c,i}$ เป็นข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 มี n-m ค่า โดย $i = m+1,\dots,n$

ให้ $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$ คือ ค่าสังเกตของตัวแปรตาม หรือเวลาที่ใช้ในการศึกษาข้อมูล ลำดับที่ i จนกว่าจะสิ้นสุดการสังเกต

เมื่อ $i = 1,2,\dots,m-1,m,m+1,\dots,n$ และ $m \le n$

σ คือ ส่วนเบี่ยงเบนมาตรฐานที่แท้จริง

$\hat{\sigma}$ คือ ค่าประมาณส่วนเบี่ยงเบนมาตรฐาน

$$\mathbf{X}^{\mathrm{NC}} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{p1} \\ 1 & X_{12} & X_{22} & \dots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1m} & X_{2m} & \dots & X_{pm} \end{bmatrix}_{m\times(p+1)}, \mathbf{X}^{\mathrm{C}} = \begin{bmatrix} 1 & X_{1(m+1)} & X_{2(m+1)} & \dots & X_{p(m+1)} \\ 1 & X_{1(m+2)} & X_{2(m+2)} & \dots & X_{p(m+2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2m} & \dots & X_{pm} \end{bmatrix}_{(n-m)\times(p+1)}$$

ดังนั้น $$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1m} & X_{2m} & \cdots & X_{pm} \\ \hline 1 & X_{1(m+1)} & X_{2(m+1)} & \cdots & X_{p(m+1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pm} \end{bmatrix}_{m\times(p+1)} \begin{matrix} \left.\begin{matrix}\\\\\\\end{matrix}\right\} \mathbf{X}^{\mathrm{NC}} \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} \mathbf{X}^{\mathrm{C}} \end{matrix}$$

และให้ $$\mathbf{Y}^{\mathrm{NC}^*} = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_m^* \end{bmatrix}_{m\times1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}_{m\times1}, \mathbf{Y}^{\mathrm{C}^*} = \begin{bmatrix} Y_{m+1}^* \\ Y_{m+2}^* \\ \vdots \\ Y_n^* \end{bmatrix}_{(n-m)\times1} = \begin{bmatrix} Y_{c,(m+1)} \\ Y_{c,(m+2)} \\ \vdots \\ Y_{c,n} \end{bmatrix}_{(n-m)\times1}$$

ดังนั้น $$\mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ \vdots \\ Y_m^* \\ \hline Y_{m+1}^* \\ \vdots \\ Y_n^* \end{bmatrix}_{n\times1} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \\ \hline Y_{c,(m+1)} \\ \vdots \\ Y_{c,n} \end{bmatrix}_{n\times1} \begin{matrix} \left.\begin{matrix}\\\\\\\end{matrix}\right\} \mathbf{Y}^{\mathrm{NC}^*} \\ \left.\begin{matrix}\\\\\\\end{matrix}\right\} \mathbf{Y}^{\mathrm{C}^*} \end{matrix}$$

และเมทริกซ์ของข้อมูลที่มีความสมบูรณ์ คือ $\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_n \end{bmatrix}_{n\times1}$

3.1 วิธีกำลังสองต่ำสุด (Ordinary Least Squares Method; OLS)

หาค่าประมาณ $\hat{\beta}$ จาก $\hat{Y} = X\hat{\beta}$ ที่ทำให้ผลรวมของค่าคลาดเคลื่อนกำลังสอง ซึ่งเขียนในรูปเมทริกซ์คือ $\sum_{i=1}^{n} e_i^2 = [\mathbf{Y} - \hat{\mathbf{Y}}][\mathbf{Y} - \hat{\mathbf{Y}}]^{\mathrm{T}}$ มีค่าน้อยที่สุด โดย $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

3.2 วิธีแชตเทอร์จีและแมคลีช (Chatterjee and McLeish Method; CM)

Chatterjee และ McLeish (1986) ได้เสนอวิธีที่ปรับปรุงข้อมูลที่ถูกตัดทิ้งก่อนนำไปประมาณค่าด้วยวิธีกำลังสองต่ำสุดทั้งหมด r รอบ จนกว่าค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณของ β_{p} รอบที่ r กับ $\mathrm{r}+1$ จะมีค่าไม่เกิน 0.001 โดยมีขั้นตอนดังนี้

กำหนดให้ $\hat{\beta}_p^{(r)}$ คือ ค่าประมาณของ β_p รอบที่ r โดย $\mathrm{r} = 1,2,3,\ldots$

ขั้นตอน 0 เลือกใช้เฉพาะข้อมูลที่ไม่ถูกตัดปลายทางขวา แล้วนำมาคำนวณหา $\hat{\beta}^{(0)}$ ด้วยวิธีกำลังสองต่ำสุด ได้ $\hat{\boldsymbol{\beta}}^{(0)} = ((\mathbf{X}^{\mathrm{NC}})^{\mathrm{T}}\mathbf{X}^{\mathrm{NC}})^{-1}(\mathbf{X}^{\mathrm{NC}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{NC}^*}$

ขั้นตอน a แทนค่าข้อมูลที่ถูกตัดปลายทางขวาด้วย $\mathrm{Y}_{\mathrm{c,i}}^{(\mathrm{r})}$ เมื่อ

$$Y_{c,i}^{(r)} = max\,(\hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)}X_{1i} + \hat{\beta}_2^{(r)}X_{2i} + \cdots + \hat{\beta}_p^{(r)}X_{pi}\ , Y_{c,i}^{(r-1)})$$

กำหนด $Y_{c,i}^{(0)} = \max\left(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)}X_{1i} + \hat{\beta}_2^{(0)}X_{2i} + \cdots + \hat{\beta}_p^{(0)}X_{pi}\ , Y_{c,i}^{(0)}\right)$

เมื่อ $i = m+1, m+2, \ldots, n$

ขั้นตอน b รวมข้อมูลที่ไม่ถูกตัดปลายทางขวากับข้อมูลที่ได้จากขั้นตอน a ได้เป็น

$$\mathbf{Y}^* = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \\ - - - \\ Y_{c,m+1}^{(r)} \\ \vdots \\ Y_{c,n}^{(r)} \end{bmatrix}_{n\times 1}$$ แล้วนำไปคำนวณหา $\hat{\boldsymbol{\beta}}^{(r+1)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^*$

จากนั้น ทำตามขั้นตอน a และ b วนซ้ำไปเรื่อย ๆ จนกว่า $\left|\hat{\beta}_j^{(r+1)} - \hat{\beta}_j^{(r)}\right| \le 0.001$ ทุก j = 0,1,2, … , p

3.3 วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม (Maximum Likelihood Estimation using the EM algorithm Method ; MLE_EM)

Demster, Laird และ Rubin (1977) เป็นคนคิดขั้นตอนอีเอ็ม และหลังจากนั้น Aikin (1981) นำมาใช้กับการประมาณค่าด้วยวิธีภาวะน่าจะเป็นแบบสูงสุด มีขั้นตอนดังนี้

ขั้นตอน 0 คำนวณค่า $\hat{\boldsymbol{\beta}}^{(0)}$ และ $\hat{\sigma}^{(0)}$ ด้วยวิธีภาวะน่าจะเป็นสูงสุด โดยข้อมูลที่ถูกตัดปลายจะถูกมองเสมือนเป็นข้อมูลที่ไม่ถูกตัดปลาย นั่นคือการทำให้ฟังก์ชันภาวะน่าจะเป็นรวม $L = \prod_{i=1}^{n} f(Y_i) \cdot \prod_{m+1}^{n} S(Y_{c,i})$ มีค่าสูงสุด โดย

$$L(Y_i^*) = \begin{cases} f(Y_i) & ,Y_i \le Y_{c,i} \\ P(Y_i > Y_{c,i}) = S(Y_{c,i}) & ,Y_i > Y_{c,i} \end{cases}$$

$Y_i \le Y_{c,i}$ เป็นข้อมูลไม่ถูกตัดปลายทางขวาประเภทที่ 1 โดย $i = 1,2,\ldots,m$

$Y_i > Y_{c,i}$ เป็นข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 โดย $i = m+1, \ldots, n$

ซึ่งจะได้ผลการคำนวณ $\boldsymbol{\beta}^{(0)}$ และ $\hat{\sigma}^{(0)}$ เหมือนวิธีกำลังสองต่ำสุด ฉะนั้น สามารถคำนวณ $\hat{\boldsymbol{\beta}}^{(0)}$ ได้จาก $\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^*$

ขั้นตอน E (Expectation Step : E Step)

การแทนค่าข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 ด้วยค่าคาดหวังแบบมีเงื่อนไข คือ

$$E\left(Y_i^* \middle| Y_i^* > Y_{c,i}, \boldsymbol{\beta}, \sigma\right) = \hat{\mu}_i^{(r)} + [\hat{\sigma}^{(r)} \cdot h(\hat{z}_i)^{(r)}]$$

เมื่อ r คือ รอบที่ r ในการพารามิเตอร์นั้น ๆ โดย $r = 0,1,2,3,\ldots$

$i = m+1, m+2, \ldots, n$

โดย

$$\hat{\mu}_i^{(r)} = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)}X_{1i} + \cdots + \hat{\beta}_p^{(r)}X_{pi}$$

$$\hat{z}_i^{(r)} = \frac{Y_{c,i} - \hat{\mu}_i^{(r)}}{\hat{\sigma}^{(r)}}$$

$$f(\hat{z}_i)^{(r)} = \frac{1}{\sqrt{2\pi}} e^{(-\hat{z}_i^2/2)}$$

$$S(\hat{z}_i)^{(r)} = 1 - F(\hat{z}_i)^{(r)} = \int_{\hat{z}_i}^{\infty} f(t)\, dt$$

$$h(\hat{z}_i)^{(r)} = \frac{f(\hat{z}_i)^{(r)}}{1 - F(\hat{z}_i)^{(r)}}$$

ขั้นตอน M (Maximization Step : M Step)

รวมข้อมูลที่ถูกแทนค่าจากขั้นตอน E กับข้อมูลที่ไม่ถูกตัดปลาย จะได้ข้อมูล $\tilde{\mathbf{Y}}$ โดย

$$\tilde{Y}_i = \begin{cases} Y_i^* = Y_i & , Y_i \le Y_{c,i} \\ \hat{\mu}_i^{(r)} + \left[\hat{\sigma}^{(r)} \cdot h(\hat{z}_i)^{(r)}\right] & , Y_i > Y_{c,i} \end{cases}$$

$Y_i \le Y_{c,i}$ เป็นข้อมูลไม่ถูกตัดปลายทางขวาประเภทที่ 1 m ข้อมูล โดย $i = 1,2,\dots,m$

$Y_i > Y_{c,i}$ เป็นข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1 n-m ข้อมูล โดย $i = m+1,\dots,n$

จากนั้น นำข้อมูล $\tilde{\mathbf{Y}}$ ที่คำนวณได้ มาหาค่าประมาณของ $\tilde{\boldsymbol{\beta}}^{(r+1)}$ และ $\hat{\sigma}^{(r+1)}$ ที่ทำให้ฟังก์ชันภาวะน่าจะเป็นรวม $L = \prod_{i=1}^{n} f(\tilde{Y}_i)$ มีค่าสูงสุด นั่นคือ หาค่า $\tilde{\boldsymbol{\beta}}^{(r+1)}$ ด้วยวิธีกำลังสองต่ำสุด ฉะนั้นสามารถคำนวณได้จาก

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\tilde{\mathbf{Y}}$$

$$\hat{\sigma}^{r+1} = \sqrt{\frac{\sum_{i=1}^{m}\left(y_i - \hat{\mu}_i^{(r)}\right)^2 + (\hat{\sigma}^{(r)})^2 \sum_{m+1}^{n}\left[1 + \left(\hat{z}_j^{(r)} \cdot h\left(z_i^{(r)}\right)\right)\right]}{n}}$$

โดยที่ $\hat{\sigma}^{(0)}$ คือค่าประมาณของส่วนเบี่ยงเบนมาตรฐานในขั้นตอนที่ 0

จากนั้น ทำตามขั้นตอน E และ M วนซ้ำไปเรื่อย ๆ จนกว่า $\left|\hat{\beta}_j^{(r+1)} - \hat{\beta}_j^{(r)}\right| \le 0.001$ ทุก $j = 0,1,2,\dots,p$

3.4 วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่าเฉลี่ย (Maximum Likelihood Estimation Method using the EM algorithm with mean-adjusted data ; MLE_EM_MEAN)

ปรับค่าข้อมูลก่อนนำไปประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม โดยใช้ค่าเฉลี่ยของ $Y_{c,i}$ ซึ่ง $Y_{c,i} = Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W\right]$ และอยู่ในช่วง $\left[Y_c + a\left(\frac{\alpha}{\alpha+\theta} - 1\right), Y_c + a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$ จะได้ว่า

$$E(Y_{c,i}) = E\left(Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W\right]\right) = Y_c$$

เมื่อ $i = 1,2,3,\dots,n$

3.5 วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม เมื่อมีการปรับค่าข้อมูลก่อนคำนวณด้วยค่ามัธยฐาน (Maximum Likelihood Estimation Method using the EM algorithm with median-adjusted data ; MLE_EM_MED)

ปรับค่าข้อมูลก่อนนำไปประมาณด้วยวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนอีเอ็ม โดยใช้ค่ามัธยฐานของ $Y_{c,i}$ ซึ่ง $Y_{c,i} = Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W\right]$ และอยู่ในช่วง $\left[Y_c + a\left(\frac{\alpha}{\alpha+\theta} - 1\right), Y_c + a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$

โดยวิธี 3.4 และ 3.5 หลังจากที่เวลาสิ้นสุดการเก็บข้อมูลถูกปรับให้มีค่าเป็น Y_c แล้ว จะได้ข้อมูลที่จะนำไปประมาณด้วยวิธีภาวะน่าจะเป็นแบบสูงสุดมีลักษณะดังนี้

$$Y_i^* = \min(Y_i, Y_c) = \begin{cases} Y_i & \text{เป็นข้อมูลไม่ถูกตัดปลายทางขวาประเภทที่ 1} \\ Y_c & \text{เป็นข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1} \end{cases}$$

วิธีดำเนินการวิจัย

1. นิยามความหมายและรูปแบบของข้อมูล

แบ่งรูปแบบของการเข้ามาของข้อมูลเป็น 3 แบบ คือ ช่วงต้น, กลาง และท้าย ของช่วงเวลาที่เปิดรับข้อมูลเข้ามาเพื่อศึกษา โดยแสดงรูปแบบการแจกแจงของข้อมูลลักษณะต่าง ๆ ดังนี้

1) เมื่อข้อมูลเข้ามาในช่วงต้นของการเปิดรับ รูปแบบการแจกแจงคือ $Beta(0.3125, 0.9375)$
2) ข้อมูลเข้ามาในช่วงกลางของการเปิดรับรูปแบบการแจกแจงคือ $Beta(1,1)$
3) ข้อมูลเข้ามาในช่วงท้ายของการเปิดรับรูปแบบการแจกแจงคือ $Beta(0.9375, 0.3125)$

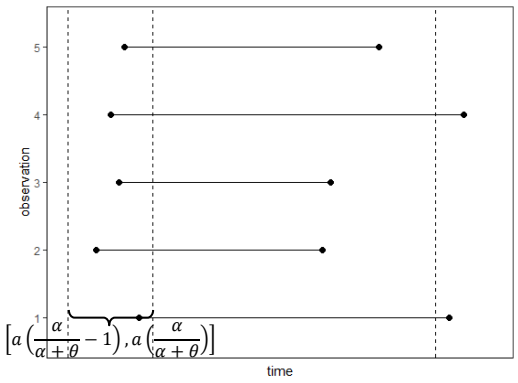
2. วิธีการจำลองข้อมูล

1) สร้างตัวแปรอิสระ $X_{1i} \overset{iid}{\sim} N(0, \sigma_{x_1}^2)$ และ $X_{2i} \overset{iid}{\sim} N(0, \sigma_{x_2}^2)$ โดยตัวแปรอิสระทั้งสองตัวอิสระต่อกัน และสร้างความคลาดเคลื่อน $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$ เมื่อ $i = 1,2,\dots,n$ โดยที่ n มีค่าเป็น 30, 100 และสัมประสิทธิ์ของสมการถดถอยมีค่าเป็น $\beta_0 = 0.3, \beta_1 = 1, \beta_2 = 1$
2) สร้างค่าจริงของตัวแปรตามที่มีความสัมพันธ์กับตัวแปรอิสระ นั่นคือ $Y_i' = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}$ ที่มีการแจกแจงแบบล็อกนอร์มอล
3) สร้างค่าสังเกตของตัวแปรตามที่มีความสัมพันธ์กับตัวแปรอิสระและความคลาดเคลื่อน นั่นคือ $Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i}$ ที่มีการแจกแจงแบบล็อกนอร์มอล

3. วิธีการสร้างตัวแปรตามให้ประกอบด้วยข้อมูลที่ไม่ถูกตัดปลายทางขวาประเภทที่ 1 และข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1

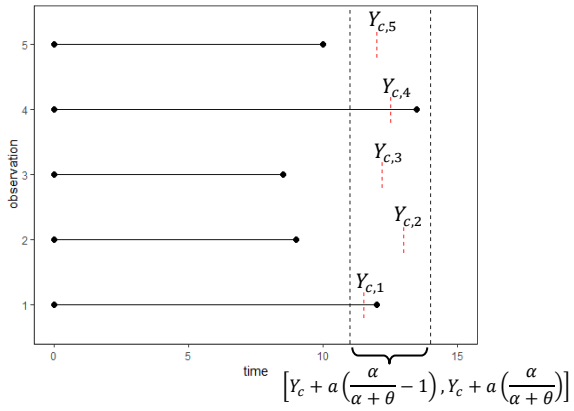
1) กำหนดค่า Y_c เป็นค่าที่เปอร์เซ็นต์ไทล์ที่ $100 - r_1$ ของตัวแปรสุ่มที่มีการแจงแจงแบบเดียวกันกับ Y_i เมื่อ $r_1 = 10\%, 30\%$
2) เนื่องจาก r_2 คือสัดส่วนของระยะเวลาเปิดรับสมัครต่อระยะเวลาติดตามผู้ป่วย จึงได้ว่า $r_2 = \frac{a}{Y_c}$ หรือนั่นคือ $a = Y_c \cdot r_2$ ดังนั้น สร้างตัวแปรสุ่ม $W_i \overset{iid}{\sim} \text{Beta}(\alpha, \theta)$ แล้วปรับค่าจากช่วง $[0,1]$ เป็นช่วง $\left[a\left(\frac{\alpha}{\alpha+\theta} - 1\right), a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$ จะได้ตัวแปรสุ่ม $U_i = \left(a\left[\frac{\alpha}{\alpha+\theta} - W_i\right]\right)$ ที่อยู่ในช่วง $\left[a\left(\frac{\alpha}{\alpha+\theta} - 1\right), a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$

a Y_c



ภาพประกอบที่ 1 ข้อมูลที่จุดเริ่มต้นเก็บข้อมูลแตกต่างกันที่อยู่ในช่วง $\left[a\left(\frac{\alpha}{\alpha+\theta}-1\right), a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$

3) นำค่า Y_c มาสร้างเป็นตัวแปรตามที่ถูกตัดปลายทางขวา $Y_{c,i}$ โดยบวกด้วยตัวแปรสุ่ม $U_i = \left(a\left[\frac{\alpha}{\alpha+\theta} - W_i\right]\right)$ จะได้ตัวแปรสุ่ม $Y_{c,i} = Y_c + a\left[\frac{\alpha}{\alpha+\theta} - W_i\right]$ และอยู่ในช่วง $\left[Y_c + a\left(\frac{\alpha}{\alpha+\theta}-1\right), Y_c + a\left(\frac{\alpha}{\alpha+\theta}\right)\right]$



ภาพประกอบที่ 2 ข้อมูลตัดปลายทางขวา $Y_{c,i}$ หลังเลื่อนจุดเริ่มต้นเก็บข้อมูล

จากข้อ 1, 2, 3

ถ้า $Y_i \leq Y_{c,i}$ แล้ว Y_i เป็นข้อมูลที่ไม่ถูกตัดปลายทางขวาประเภทที่ 1

ถ้า $Y_i > Y_{c,i}$ แล้ว Y_i เป็นข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1

และให้ $Y_i^* = \min(Y_i, Y_{c,i})$ ทำให้จะได้ตัวแปรตามที่ประกอบไปด้วยข้อมูลที่ไม่ถูกตัดปลายทางขวาประเภทที่ 1 และข้อมูลที่ถูกตัดปลายทางขวาประเภทที่ 1

4. การหาค่าประมาณ

หาค่าาประมาณของตัวแปรตาม $\hat{\boldsymbol{Y}}_i = e^{\hat{\beta}_0+\hat{\beta}_1 \boldsymbol{X_{1i}}+\hat{\beta}_2 \boldsymbol{X_{2i}}}$ ด้วยวิธี OLS, CM, MLE_EM, MLE_EM_MEAN, และ MLE_EM_MED

5. การเปรียบเทียบวิธีการประมาณค่าจากตัวแบบถดถอย

ใช้ค่าประสิทธิภาพสัมพัทธ์ของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Average of Mean Square Error; AMSE) โดยเป็นอัตราส่วน $AMSE(\hat{Y}_{MLE\_EM}) : AMSE(\hat{Y}_a)$ เมื่อ $AMSE(\hat{Y}_a) = \frac{\sum_{j=1}^{10000} MSE(\hat{Y}_a^{(j)})}{10000}$ เป็นค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแปรตามของวิธี a จากการจำลองสถานการณ์ 10,000 รอบ

$$RE(\hat{Y}_a) = \frac{AMSE(\hat{Y}_{MLE\_EM})}{AMSE(\hat{Y}_a)}$$

$MSE\left(\hat{Y}_a^{(j)}\right)$ คือ ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองของตัวแปรตาม $\hat{Y}$ ซึ่งประมาณด้วยวิธี $a =$ OLS, CM, MLE_EM, MLE_EM_MEAN, MLE_EM_MED รอบที่ $j = 1,2,\ldots,10000$ โดย $MSE\left(\hat{Y}_a^{(j)}\right) = \frac{\sum_{i=1}^{n}(\hat{Y}_{i,a}-Y_i')^2}{n}$

$\hat{Y}_{i,a}$ คือ ค่าประมาณของตัวแปรตามซึ่งประมาณโดยวิธี a ลำดับที่ i

Y_i' คือ ค่าจริงของตัวแปรตาม ลำดับที่ i

n คือ ขนาดตัวอย่าง

โดยถ้า

$RE(\hat{Y}_a) > 1$ หมายถึง วิธี a มีประสิทธิภาพมากกว่าวิธี MLE_EM

$RE(\hat{Y}_a) < 1$ หมายถึง วิธี a มีประสิทธิภาพด้อยกว่าวิธี MLE_EM

$RE(\hat{Y}_a) = 1$ หมายถึง วิธี a มีประสิทธิภาพเทียบเท่าวิธี MLE_EM

ผลการวิจัย

ตารางที่ 1 แสดงวิธีที่มีประสิทธิภาพสูงสุดในแต่ละสถานการณ์ เมื่อตัวอย่างมีขนาดเล็ก (n=30)

n	r_1	r_2	$\sigma^2_{x_1+x_2}:\sigma^2_\varepsilon$	$\sigma^2_{x_1}:\sigma^2_{x_2}$	$Beta(0.3125,0.9375)$	$Beta(1,1)$	$Beta(0.9375,0.3125)$
30	10	0.1, 0.3	2:1	1:1, 1:2, 1:5	CM	CM	CM
		0.1, 0.3	1:1, 1:2	1:1, 1:2, 1:5	OLS	OLS	OLS
	30	0.1	2:1	1:5	MLE_EM_MED	MLE_EM	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
			1:1	1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_MLE_EM
				1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MED	MLE_EM
		0.1, 0.3	1:2	1:1, 1:2, 1:5	CM	CM	CM
		0.3	2:1	1:2, 1:5	MLE_EM_MED	MLE_EM	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
			1:1	1:1, 1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_MLE_EM	MLE_EM_MEAN

จากตารางที่ 1 พบว่า เมื่อตัวอย่างมีขนาดเล็ก (n=30) การกระจายตัวของตัวแปรอิสระมากกว่าค่าคลาดเคลื่อน และมีเปอร์เซ็นต์ของข้อมูลที่ถูกตัดปลายทางขวาน้อย วิธี CM มีประสิทธิภาพสูงสุด ถึงแม้การกระจายตัวของตัวแปรอิสระตัวที่ 1 และ 2 จะแตกต่างกัน ไม่ได้ส่งผลอะไร และหากการกระจายตัวของตัวแปรอิสระน้อยกว่าหรือเท่ากับการกระจายตัวของค่าคลาดเคลื่อน วิธี OLS มีประสิทธิภาพที่สุด ไม่ว่าจะเริ่มเก็บข้อมูลที่ช่วงใดของการเปิดรับ แต่หากมีเปอร์เซ็นต์ของข้อมูลที่ถูกตัดปลายทางขวามาก ข้อมูลเข้ามาในช่วงต้นของการเปิดรับ วิธี MLE_EM_MED เป็นวิธีที่มีประสิทธิภาพสูงสุด ในขณะที่ข้อมูลที่เข้ามาในช่วงกลาง วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุด และเมื่อข้อมูลเข้ามาในช่วงท้าย มีแนวโน้มว่าวิธี MLE_EM_MEAN จะมีประสิทธิภาพสูงสุด

ตารางที่ 2 แสดงวิธีที่มีประสิทธิภาพสูงสุดในแต่ละสถานการณ์ เมื่อตัวอย่างมีขนาดใหญ่ (n=100)

n	r_1	r_2	$\sigma^2_{x_1+x_2}:\sigma^2_\varepsilon$	$\sigma^2_{x_1}:\sigma^2_{x_2}$	$Beta(0.3125,0.9375)$	$Beta(1,1)$	$Beta(0.9375,0.3125)$
100	10	0.1	2:1	1:5	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM	MLE_EM
			1:1	1:5	MLE_EM_MED	MLE_EM	MLE_EM
				1:1, 1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
			1:2	1:5	MLE_EM_MED	MLE_EM	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM
				1:1	CM	MLE_EM_MED	MLE_EM
		0.3	2:1	1:1, 1:5	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
			1:1	1:2, 1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
			1:2	1:5	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:1, 1:2	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
	30	0.1	2:1	1:1, 1:2, 1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
			1:1	1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MED	MLE_EM
			1:2	1:1, 1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
		0.3	2:1	1:2, 1:5	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN
				1:1	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
			1:1	1:5	MLE_EM_MED	MLE_EM	MLE_EM_MEAN
				1:1, 1:2	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
			1:2	1:5	MLE_EM_MED	MLE_EM_MEAN	MLE_EM_MEAN
				1:1, 1:2	MLE_EM_MED	MLE_EM_MED	MLE_EM_MEAN

จากตารางที่ 2 พบว่า เมื่อตัวอย่างมีขนาดใหญ่ (n=100) เมื่อข้อมูลเข้ามาในช่วงต้นของการเปิดรับ ไม่ว่าเปอร์เซ็นต์ของข้อมูลที่ถูกตัดปลาย ลักษณะการกระจายตัวของตัวแปรอิสระ และลักษณะการกระจายตัวของค่าคลาดเคลื่อนจะอยู่ในรูปแบบใด มีแนวโน้มว่าวิธี MLE_EM_MED จะมีประสิทธิภาพสูงสุด ในขณะที่เมื่อข้อมูลเข้ามาในช่วงกลาง เช่นเดียวกันกับเมื่อตัวอย่างมีขนาดเล็กและข้อมูลมีเปอร์เซ็นต์ทางขวามาก วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุด

เมื่อข้อมูลเข้ามาในช่วงท้าย ตัวอย่างมีขนาดใหญ่ (n=100) และเปอร์เซ็นต์ในการถูกตัดปลายทางขวาน้อย และมีสัดส่วนของระยะเวลาเปิดรับสมัครต่อระยะเวลาติดตามผู้ป่วย (r_2) เท่ากับ 0.1 เมื่อตัวแปรอิสระมีการกระจายตัวมากกว่าหรือเท่ากับค่าคลาดเคลื่อน วิธี MLE_EM_MEAN มีแนวโน้มที่จะมีประสิทธิภาพสูงสุด ส่วนเมื่อตัวแปรอิสระมีการกระจายตัวน้อยกว่าค่าคลาดเคลื่อน วิธี MLE_EM มีแนวโน้มที่จะมีประสิทธิภาพมากที่สุด หาก r_2 = 0.3 วิธี MLE_EM_MEAN จะมีประสิทธิภาพสูงสุด ในขณะที่ข้อมูลมีเปอร์เซ็นต์ในการถูกตัดปลายทางขวามาก มีแนวโน้มว่าวิธี MLE_EM_MEAN เป็นวิธีที่มีประสิทธิภาพสูงสุด

สรุปและอภิปรายผล

จากผลการศึกษา พบว่าขนาดของตัวอย่าง ลักษณะการกระจายตัวของตัวแปรอิสระกับค่าคลาดเคลื่อน มีผลต่อประสิทธิภาพของวิธีการประมาณค่า เมื่อตัวอย่างมีขนาดเล็ก และการกระจายตัวของตัวแปรอิสระน้อยกว่าหรือเท่ากับค่าคลาดเคลื่อน วิธี OLS มีประสิทธิภาพสูงสุด ถึงแม้ว่าวิธี OLS จะเป็นวิธีที่จะให้ค่าประมาณที่ต่ำกว่าความเป็นจริง เนื่องจากวิธี OLS ไม่ได้พิจารณาว่าข้อมูลที่นำมาใช้เพื่อประมาณค่าเป็นข้อมูลที่ถูกตัดปลายทางขวา ทำให้เมื่อถ้าข้อมูลที่ถูกตัดปลายทางขวามากขึ้น แต่ขนาดตัวอย่างยังเป็นขนาดเล็ก วิธี OLS จึงมีประสิทธิภาพลดลง โดนวิธี CM จะมีประสิทธิภาพสูงสุดแทน เนื่องจากวิธี CM มีขั้นตอนการประมาณค่าข้อมูลที่ถูกตัดปลายเพิ่มขึ้นมา ทำให้ค่าประมาณตัวแปรตามเคลื่อนน้อยลง แต่หากตัวอย่างมีขนาดใหญ่ขึ้น วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุด เนื่องจากวิธีในกลุ่ม MLE_EM มีการนำข้อมูลที่ถูกตัดปลายมาแทนที่โดยใช้ขั้นตอนการแทนค่าคาดหวังและการหาค่าสูงสุด แล้วจึงนำมาหาค่าประมาณของตัวแปรตาม

โดยเมื่อพิจารณาจากการเข้ามาของข้อมูลในช่วงของการเปิดรับ เมื่อข้อมูลเข้ามาในช่วงต้นและตัวอย่างมีขนาดเล็กและเปอร์เซ็นต์ของข้อมูลที่ถูกตัดปลายทางขวามาก หรือเมื่อตัวอย่างมีขนาดใหญ่ พบว่าวิธี MLE_EM_MED มีประสิทธิภาพสูงสุด เนื่องด้วยหลังจากมีการปรับข้อมูลให้เป็นข้อมูลที่ถูกตัดปลายทางขวา ณ จุดเวลาเดียวกัน ทำให้ข้อมูลมีลักษณะเบ้ซ้าย เมื่อปรับข้อมูลไปที่ค่ามัธยฐานก่อนนำไปประมาณด้วยวิธี MLE_EM จึงมีประสิทธิภาพสูงสุด ในทางกลับกันเมื่อข้อมูลเข้ามาในช่วงท้ายของการเปิดรับ หลักจากปรับข้อมูลให้เป็นข้อมูลที่ถูกตัดปลายทางขวา ทำให้ข้อมูลมีลักษณะเบ้ขวา การปรับข้อมูลไปที่ค่าเฉลี่ยก่อนนำไปประมาณด้วยวิธี MLE_EM วิธี MLE_EM_MEAN จึงมีประสิทธิภาพสูงสุด และเมื่อข้อมูลเข้ามาในช่วงกลางของการเปิดรับ หรืออีกนัยหนึ่งคือการแจกแจงแบบยูนิฟอร์ม วิธีในกลุ่ม MLE_EM จะมีประสิทธิภาพสูงสุดปะปนกัน ซึ่งสอดคล้องกับงานวิจัยก่อนหน้าโดยธนาพิพัฒน์ ที่เมื่อขนาดตัวอย่างใหญ่ขึ้นวิธีในกลุ่ม MLE_EM จะเป็นวิธีที่มีประสิทธิภาพสูงสุดและใกล้เคียงกัน

เอกสารอ้างอิง

Aitkin, M. (1981). A Note on the Regression Analysis of Censored Data. *Technometrics, 23*(2), 161-163. doi:10.1080/00401706.1981.10486259

Chatterjee, S. and D. L. McLeish (1986). Fitting linear regression models to censored data by least squares and maximum likelihood methods. *Communications in Statistics - Theory and Methods 15(11)*: 3227-3243.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38. http://www.jstor.org/stable/2984875

Jöreskog, K.G. (2002). Censored Variables and Censored Regression. [Online]. Retrieved April 1, 2020, from : https://www.ssicentral.com/wp-content/uploads/2021/04/lis_censor.pdf

Klein, J.P., and Moeschberger, M.L. 2003. *Survival Analysis Techniques for Censored and Truncated Data*. Second Edition. New York : Springer-Verlag New York, Inc.

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications*: Cengage Learning. Seventh Edition. : 177,194

จำเนียน จำนงรักษ์. (2539). การพยากรณ์ในสมการถดถอยเชิงเส้นพหุเมื่อค่าตัวแปรตามถูกตัดปลายทางขวา (วิทยานิพนธ์ปริญญามหาบัณฑิต). สาขาวิชาประกันภัย ภาควิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

บังอร กุมพล. (2539). การวิเคราะห์การถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามมีค่าถูกตัดทิ้งทางขวากรณีค่าตัดทิ้งประเภทที่ 1 (วิทยานิพนธ์ปริญญามหาบัณฑิต). ภาควิชาสถิติ สาขาวิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ศิวพร ทิพย์พันธุ์. (2561). การศึกษาเปรียบเทียบการประมาณค่าจากตัวแบบถดถอยสำหรับข้อมูลที่ถูกตัดปลายทางขวาแบบที่ 1 ที่มีการแจกแจงแบบล็อกนอร์มอล (วิทยานิพนธ์ปริญญามหาบัณฑิต). ภาควิชาสถิติ สาขาวิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

ธนาพิพัฒน์ ทรัพย์ครองชัย. (2561). การศึกษาเปรียบเทียบการประมาณค่าจากตัวแบบการถดถอยสำหรับข้อมูลที่มีการแจกแจงแบบล็อกนอร์มอล ที่ถูกตัดปลายทางขวาแบบสุ่มที่มีการแจกแจงแบบยูนิฟอร์ม (วิทยานิพนธ์ปริญญามหาบัณฑิต). ภาควิชาสถิติ สาขาวิชาสถิติ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.