บทที่ 2

สถิติที่ใช้ในการวิจัย

ในบทนี้กล่าวถึงทฤษฎีพื้นฐานต่าง ๆ ที่เกี่ยวข้อง และวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าสังเกตของตัวแปรตามเป็นค่าที่ถูกตัดทิ้งทางขวา ด้วยวิธีการกำลังสองต่ำสุด วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด และวิธีการของบัคเลย์และเจมส์ ซึ่งมีรายละเอียดต่าง ๆ ดังนี้

2.1 ทฤษฎีพื้นฐาน

2.1.1 ประเภทของการตัดทิ้ง (Type of Censoring)

ลักษณะของข้อมูลที่ถูกตัดทิ้งบางส่วนนั้น เกิดขึ้นได้ในหลายลักษณะ คือ แบบประเภทที่ 1 แบบประเภทที่ 2 แบบสุ่ม และแบบอื่น ๆ ดังรายละเอียดต่อไปนี้

2.1.1.1 การตัดทิ้งประเภทที่ 1 (Type I Censoring)

การตัดทิ้งประเภทนี้ จะมีการกำหนดเวลาของการเกิดค่าถูกตัดทิ้งเอาไว้ล่วงหน้า เรียกว่า "Fixed Censoring Time" ตัวอย่างของการตัดทิ้งประเภทนี้ ทำการทดลองอายุการใช้งานของอุปกรณ์ไฟฟ้าจำนวนหนึ่ง กำหนดเวลาของการทดลองไว้ล่วงหน้า 50,000 ชั่วโมง เริ่มทำการทดลองโดยเปิดให้อุปกรณ์ทำงานแล้วบันทึกเวลาไว้ตั้งแต่เริ่มทำงานจนกระทั่งอุปกรณ์เสื่อมสภาพ ในระหว่างการทดลองอุปกรณ์เครื่องใดมีการเสื่อมสภาพ จะเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง (Uncensored Data หรือ Survival Time) และเมื่อสิ้นสุดการทดลอง อุปกรณ์เครื่องใดที่ยังคงอยู่ในสภาพใช้งานได้ดีจะเป็นเครื่องที่ไม่ทราบอายุการใช้งานที่แน่นอน จะบันทึกไว้ว่ามีอายุใช้งาน 50,000 ชั่วโมง ข้อมูลนี้จะเป็นค่าสังเกตที่ถูกตัดทิ้ง (Censored Data)

ให้ t_c เป็นเวลาที่กำหนดไว้ล่วงหน้า และให้ $T_1, T_2, \ldots, T_n$ เป็นตัวแปรสุ่มอายุ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน

ในกรณีนี้ จะได้ค่าสังเกตสุ่ม $Y_1, Y_2, \ldots, Y_n$ ซึ่ง

$$Y_i = \begin{cases} T_i & \text{ถ้า } T_i \leqslant t_c \quad \text{(ไม่ถูกตัด)} \\ t_c & \text{ถ้า } T_i > t_c \quad \text{(ถูกตัด)} \end{cases}$$

จะมีฟังก์ชันภาวะน่าจะเป็น (Likelihood Function)

$$L(y_i) = \begin{cases} f(y_i) & \text{ถ้าไม่ถูกตัด} \\ P(T_i > t_c) = S(t_c) & \text{ถ้าถูกตัด} \end{cases}$$

และจะมีฟังก์ชันภาวะน่าจะเป็นรวมดังนี้

$$L = \prod_{i \in u} f(y_i) \prod_{i \in c} S(t_c)$$

$i \in u$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าไม่ถูกตัดทิ้ง

$i \in c$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าถูกตัดทิ้ง

2.1.1.2 การตัดทิ้งประเภทที่ 2 (Type II Censoring)

ในบางกรณีไม่สามารถจะกำหนดเวลาของการเกิดค่าที่ถูกตัดทิ้ง (Censoring Time) ที่เหมาะสมได้ ดังนั้น จะกำหนดจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้งแทน นั่นคือ เมื่อจำนวนของค่าสังเกตที่ไม่ถูกตัดทิ้งเกิดขึ้นครบตามจำนวนที่กำหนดไว้ จะหยุดทำการทดลองเพื่อเป็นการประหยัดเวลาและค่าใช้จ่าย ตัวอย่างการตัดทิ้งประเภทนี้ เกิดได้ในทำนองเดียวกับการตัดทิ้งประเภทที่ 1 เช่น ให้ n แทนจำนวนอุปกรณ์ไฟฟ้าที่เป็นตัวอย่างทดลอง และกำหนดจำนวนค่าสังเกตไม่ถูกตัดทิ้งเท่ากับ $r \leqslant n$ ให้ $T_1 \leqslant T_2 \leqslant \ldots \leqslant T_r$ เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง และ $T_{r+1} \leqslant T_{r+2} \leqslant \ldots \leqslant T_n$ เป็นลำดับของตัวแปรสุ่มอายุการใช้งานที่ถูกตัดทิ้ง ซึ่ง $T_i \geqslant T_r$, $i = r+1, r+2, \ldots, n$ ไม่ทราบค่าที่แท้จริงของค่าสังเกต ดังนั้น สำหรับค่าสังเกต Y_i, $i = 1, \ldots, n$ ได้ว่า

$$Y_1 = T_1$$

.

$$Y_r = T_r$$

$$Y_{r+1} = T_r$$

.

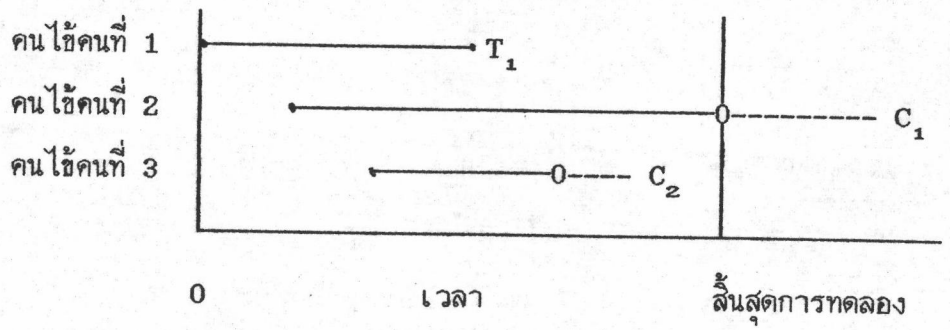
$$Y_n = T_r$$

ฟังก์ชันความหนาแน่นร่วมของค่าสังเกตที่ไม่ถูกตัดทิ้ง r ค่า คือ

$$\frac{n!}{(n-r)!} f(y_1)....f(y_r) \cdot [S(y_r)]^{n-r}$$

2.1.1.3 การตัดทิ้งแบบสุ่ม (Random Censoring)

การตัดแบบนี้มีลักษณะคล้ายแบบที่ 1 คือ มีการกำหนดระยะเวลาของการทดลอง แต่การตัดข้อมูลนั้น อาจเกิดขึ้นได้ก่อนสิ้นสุดการทดลอง จึงเรียกว่าการตัดแบบสุ่ม ส่วนใหญ่จะพบในการทดลองทางการแพทย์ เช่น คนไข้ถอนตัวจากการทดลองก่อนสิ้นสุดการทดลอง หรือคนไข้ยังมีชีวิตอยู่รอดเมื่อสิ้นสุดการทดลอง จึงทำให้ไม่สามารถบันทึกค่าที่แน่นอนของค่าสังเกตนั้นได้ รูปที่ 2.1 แสดงลักษณะความเป็นไปได้ของข้อมูลถูกตัดแบบสุ่ม



รูปที่ 2.1 แสดงแผนภาพของการทดลอง

คนไข้คนที่ 1 เข้าทำการทดลอง ณ เวลา t=0 และเสียชีวิตที่เวลา T_1 ค่าสังเกตนี้เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง T_1

คนไข้คนที่ 2 เข้าทำการทดลอง และยังมีชีวิตอยู่รอดเมื่อสิ้นสุดการทดลอง ค่าสังเกตนี้เป็นค่าสังเกตที่ถูกตัดทิ้ง C_1

คนไข้คนที่ 3 เข้าทำการทดลอง และถอนตัวออกจากการทดลองเมื่อเวลา C_2 ค่าสังเกตนี้เป็นค่าสังเกตที่ถูกตัดทิ้ง C_2

ถ้า $T_1,\ldots,T_n$ เป็นตัวแปรสุ่มอายุ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง F และ $C_1,\ldots,C_n$ เป็นตัวแปรสุ่มที่ถูกตัดทิ้งมีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง G

ดังนั้น T_i และ C_i , $i = 1,\ldots,n$ เป็นอิสระกัน จากการตัดทิ้งแบบสุ่ม ได้นิยามให้ $Y_i = \min(T_i,C_i)$ ดังนั้น จะได้ค่าสังเกตสุ่ม $Y_1, Y_2,\ldots,Y_n$ ดังนี้

$$Y_i = \begin{cases} T_i & \text{ถ้า } T_i \leqslant C_i \quad \text{(ไม่ถูกตัด)} \\ C_i & \text{ถ้า } T_i > C_i \quad \text{(ถูกตัด)} \end{cases}$$

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i \leqslant C_i \\ 0 & \text{ถ้า } T_i > C_i \end{cases}$$

จะมีฟังก์ชันภาวะน่าจะเป็นดังนี้

$$L(y_i,\delta_i) = \begin{cases} f(y_i)\,(1-G(y_i)) & \text{ถ้า } \delta_i = 1 \\ g(y_i)\,S(y_i) & \text{ถ้า } \delta_i = 0 \end{cases}$$

และมีฟังก์ชันภาวะน่าจะเป็นรวมดังนี้

$$L = \left[\prod_{i\in u} f(y_i)\right]\left[\prod_{i\in c} S(y_i)\right]\left[\prod_{i\in c} g(y_i)\right]\left[\prod_{i\in u} (1-G(y_i))\right]$$

เนื่องจาก G(y) และ g(y) ไม่เกี่ยวข้องกับพารามิเตอร์ที่สนใจ จึงละไว้โดยใช้ฟังก์ชันภาวะน่าจะเป็น ดังนี้

$$L = \prod_{i\in u} f(y_i) \prod_{i\in c} S(y_i)$$

2.1.1.4 ประเภทของการตัดทิ้งแบบอื่น ๆ (Other Type of Censoring)

นอกจากประเภทของการตัดทิ้งที่เป็น การตัดทิ้งประเภทที่ 1 การตัดทิ้งประเภทที่ 2 และการตัดทิ้งแบบสุ่มแล้ว ยังมีประเภทของการตัดทิ้งแบบอื่น ๆ อีก เช่น นอกจากการตัดทิ้งทางขวาดังกล่าวไว้แล้วข้างต้น ยังมีการตัดทิ้งทางซ้าย และการตัดทิ้งทั้งทางซ้ายและทางขวา (Left and Right Censoring) การตัดทิ้งทางซ้ายเกิดขึ้นเมื่อบริษัทรับประกันภัยกำหนดจำนวนเงินขั้นต่ำที่ผู้เอาประกันต้องรับผิดชอบเอง (Deductible) ซึ่งข้อมูลค่าเสียหายส่วนที่ไม่เกินจำนวนเงินขั้นต่ำบริษัทรับประกันภัยไม่ได้บันทึกค่าไว้ จึงเป็นการตัดทิ้งทางซ้าย ส่วนการตัดทิ้งทางซ้ายและทางขวา เกิดขึ้นในกรณีที่มีการกันต่อ (Reinsurance) ความรับผิดชอบต่อค่าเสียหายของบริษัทรับประกันต่อ จะรับผิดชอบค่าเสียหายส่วนที่เกินความรับผิดชอบของบริษัทรับประกันตรงแต่ไม่เกินจำนวนเงินสูงสุดที่กำหนดไว้ ดังนั้น บริษัทรับประกันต่อมิได้บันทึกค่าเสียหายส่วนที่บริษัทรับประกันตรงรับผิดชอบ ข้อมูลส่วนนี้จะเป็นข้อมูลที่ถูกตัดทิ้งทางซ้ายของบริษัทรับประกันต่อ และไม่ได้บันทึกค่าเสียหายส่วนที่เกินจำนวนเงินสูงสุดที่บริษัทรับประกันต่อต้องรับผิดชอบ ข้อมูลส่วนนี้จะเป็นข้อมูลที่ถูกตัดทิ้งทางขวา ดังนั้น กรณีของการประกันต่อ ข้อมูลค่าเสียหายของบริษัทรับประกันต่ออาจเกิดข้อมูลที่ถูกตัดทิ้งเฉพาะทางซ้าย หรือตัดทิ้งทั้งทางซ้ายและทางขวา ส่วนค่าเสียหายของบริษัทรับประกันตรง อาจเกิดข้อมูลที่ถูกตัดทิ้งทางขวา เนื่องจากไม่ได้บันทึกค่าเสียหายส่วนที่เป็นความรับผิดชอบของบริษัทรับประกันต่อ ในการวิจัยครั้งนี้ศึกษากรณีที่มีการตัดทิ้งแบบสุ่ม และเกิดค่าถูกตัดทิ้งทางขวา ดังนั้น จึงมีฟังก์ชันภาวะน่าจะเป็นเหมือนกับข้อ 2.1.1.3

2.1.2 อัตราการสูญเสีย หรือฟังก์ชันการสูญเสีย (Failure Rate or Hazard Function)

ให้ T เป็นตัวแปรสุ่มต่อเนื่องแทนอายุการใช้งาน (Time to Failure)

f(t) แทนฟังก์ชันความหนาแน่นของ T

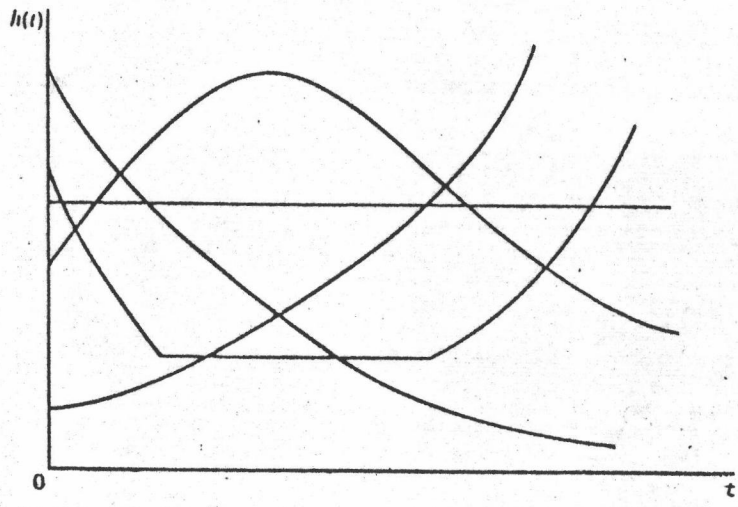
F(t) แทนฟังก์ชันการแจกแจงสะสมของ T

h(t) แทนอัตรการสูญเสีย หรือฟังก์ชันการสูญเสียของ T

นิยามฟังก์ชัน h(t) เป็นความน่าจะเป็นที่ T จะมีค่าอยู่ในช่วงเวลาสั้น ๆ $(t,t+\Delta t)$ เมื่อกำหนดค่า $T > t$:

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \left[\frac{P(t < T < t+\Delta t \mid T>t)}{\Delta t} \right] \\
&= \lim_{\Delta t \to 0} \left[\frac{F(t+\Delta t) - F(t)}{(1-F(t))\,\Delta t} \right] \\
&= \lim_{\Delta t \to 0} \left[\frac{F(t+\Delta t) - F(t)}{\Delta t} \right] \cdot \frac{1}{1-F(t)} \\
&= \frac{d\,F(t)}{dt} \cdot \frac{1}{1-F(t)} \\
&= \frac{f(t)}{1 - F(t)}, \quad t > 0
\end{aligned}
$$

ซึ่งเท่ากับฟังก์ชันความหนาแน่นของตัวแปรสุ่ม T แบบตัดปลายทางซ้าย และรูปที่ 2.2 แสดงตัวอย่างลักษณะของฟังก์ชันการสูญเสีย h(t) ในรูปแบบต่าง ๆ



รูปที่ 2.2 แสดงอัตราการสูญเสียของข้อมูลอายุ

2.1.3 ตัวประมาณพีแอล (Product Limit Estimator (PL Estimator)[1])

ในหัวข้อนี้กล่าวถึงวิธีการประมาณฟังก์ชันการอยู่รอด S(t) สมมติค่าสังเกตของอายุที่อยู่รอด (Survival Time) จำนวน n คน มีค่าเป็น $t_1, t_2, \ldots, t_n$ นำอายุที่อยู่รอดมาเรียงลำดับจากน้อยไปหามากสมมติเป็น $t_1 < t_2 < \ldots < t_n$ ฟังก์ชันการอยู่รอด ณ เวลา t_i ประมาณได้ด้วย

$$S(t_i) = \frac{[\text{จำนวนหน่วยตัวอย่างที่มีอายุอยู่รอดมากกว่าเวลา } t_i]}{[\text{จำนวนหน่วยตัวอย่างทั้งหมด}]}$$

$$= \frac{n - i}{n} = 1 - \frac{i}{n}$$

$S(t_0) = 1$ และ $S(t_n) = 0$ ฟังก์ชันการอยู่รอด เป็นฟังก์ชันขั้นบันได (Step Function) เริ่มต้นที่ 1 และลดลงไปทีละขั้น 1/n ไปยังค่า 0

จากวิธีข้างต้นจะใช้ได้ในกรณีที่ทราบอายุที่อยู่รอดของคนไข้ทั้งหมดเท่านั้นถ้าคนไข้บางคนยังมีชีวิตอยู่เมื่อสิ้นสุดการทดลอง ก็จะไม่ทราบอายุที่มีชีวิตอยู่รอดของคนไข้นั้น ดังนั้นวิธีการประมาณฟังก์ชันการอยู่รอด จะใช้วิธีการที่เป็นนอนพาราเมตริก (Nonparametric) โดยใช้ตัวประมาณพีแอล ประมาณฟังก์ชันการอยู่รอด ซึ่งตัวประมาณพีแอล พัฒนาโดย แค็พแลนและไมเออร์ (Kaplan and Meier 1958)

แนวคิดของฟังก์ชันการอยู่รอด คือ คนไข้ที่มีชีวิตอยู่รอด 2 ปี หมายถึงคนไข้ที่มีชีวิตอยู่รอดในปีแรก และเมื่อมีชีวิตอยู่รอดมากกว่า 1 ปี

---

[1] ดูรายละเอียดเพิ่มเติมได้จาก

Rupert G. Miller. Survival Analysis. (New York : John Wiley 1981), pp. 46-50.

Elisa T. Lee. Statistical Methods for Survival Data Analysis. (California : Wadsworth, 1980), pp. 75-82.

ดังนั้น ความน่าจะเป็นของการมีชีวิตอยู่รอด 2 ปี คือ ความน่าจะเป็นของการมีชีวิตอยู่รอดในปีแรก และเมื่อมีชีวิตอยู่รอดมากกว่า 1 ปี แค็พแลน และไมเออร์ ประมาณ S(2) ได้จากสมการ

$$S(2) = P(\text{มีชีวิตอยู่รอดในปีแรก และมีชีวิตอยู่รอดมากกว่า 1 ปี})$$
$$= P(\text{มีชีวิตอยู่รอด 2 ปี เมื่อมีชีวิตอยู่รอดมาแล้ว 1 ปี}) \times P(\text{มีชีวิตอยู่รอด 1 ปี})$$

จึงเขียนเป็นกฎทั่ว ๆ ไป ได้ดังนี้ ความน่าจะเป็นที่จะมีชีวิตอยู่รอด $k \geq 2$ หรือจำนวนปีที่มีชีวิตอยู่รอดได้มากกว่าปีที่เริ่มต้นศึกษา เป็นผลคูณของอัตราที่มีชีวิตอยู่รอดของ k ค่า

$$S(k) = P_1 \times P_2 \times P_3 \times \ldots \times P_k$$

P_1 = สัดส่วนของคนที่มีชีวิตอยู่รอดอย่างน้อย 1 ปี นับจากปีเริ่มต้น

P_2 = สัดส่วนของคนที่มีชีวิตอยู่รอด 2 ปี หลังจากมีชีวิตอยู่รอดมาแล้ว 1 ปี นับจากปีเริ่มต้น

P_3 = สัดส่วนของคนที่มีชีวิตอยู่รอด 3 ปี หลังจากมีชีวิตอยู่รอดมาแล้ว 2 ปี นับจากปีเริ่มต้น

.

P_k = สัดส่วนของคนที่มีชีวิตอยู่รอด k ปี หลังจากมีชีวิตอยู่รอดมาแล้ว k-1 ปี นับจากปีเริ่มต้น

ค่าสังเกตของอายุที่อยู่รอดของคนไข้ n คน นำมาเรียงลำดับจากน้อยไปหามาก $t_1 < t_2 < \ldots < t_n$ ฟังก์ชันการอยู่รอด หาได้จากสูตรดังนี้

$$S(t) = \prod_{t_i < t} \left[\frac{n-i}{n-i+1}\right]$$

t_i = เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

i เป็นลำดับที่ของข้อมูล

n เป็นจำนวนข้อมูลทั้งหมดที่ไม่ถูกตัดทิ้ง และถูกตัดทิ้ง

ตัวอย่างการหาตัวประมาณพีแอล มีค่าสังเกตตัวอย่างดังต่อไปนี้ 3.0, 4.0^+, 5.7^+, 6.5, 6.5, 8.4^+, 10.0, 10.0^+, 12.0, 15.0 แสดงการหา S(t) ได้ดังรูปที่ 2.3

t	Rank(i)	i	$\left[\frac{n-i}{n-i+1}\right]$	S(t)
3.0	1	1	9/10	= 0.900
4.0^{+}	2	-	-	-
5.7^{+}	3	-	-	-
6.5	4	4	6/7	S(3)x(6/7) = 0.771
6.5	5	5	5/6	S(6.5)x(5/6)= 0.643
8.4^{+}	6	-	-	-
10.0	7	7	3/4	S(6.5)x(3/4)= 0.482
10.0^{+}	8	-	-	-
12.0	9	9	1/2	S(10)x(1/2) = 0.241
15.0	10	10	0	S(12)x(0) = 0.000

\+ หมายถึงข้อมูลที่ถูกตัดทิ้ง

\- หมายถึง ไม่กำหนดอันดับที่

รูปที่ 2.3 แสดงการคำนวณฟังก์ชันการอยู่รอด โดยใช้ตัวประมาณพีแอล

2.1.4 อีเอ็ม อัลกอริทึม (EM Algorithm)

เนื่องจากข้อมูลที่นำมาวิเคราะห์มีลักษณะเป็นข้อมูลที่ถูกตัดทิ้งบางส่วน หรือเป็นข้อมูลที่ไม่สมบูรณ์ (Incomplete Data) ซึ่งสถิติที่เพียงพอ (Sufficient Statistics) เป็นสถิติที่เพียงพอสำหรับพารามิเตอร์ของข้อมูลที่สมบูรณ์ (Complete Data) เท่านั้น ดังนั้นจึงควรประมาณค่าข้อมูลที่ถูกตัดทิ้ง จะได้ค่าประมาณของข้อมูลที่ถูกตัดทิ้ง m ค่า รวมทั้งค่าข้อมูลที่ไม่ถูกตัดทิ้ง n ค่า ทั้งหมดจะถือเป็นข้อมูลที่สมบูรณ์ n+m ค่า การประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าของตัวแปรตามมีค่าที่ถูกตัดทิ้ง ด้วยวิธีการประมาณด้วยภาวะน่า

จะเป็นสูงสุด และวิธีการของบัคเลย์และเจมส์ ซึ่งทั้งสองวิธีได้ใช้การกระทำวนซ้ำ และใช้ทฤษฎี อีเอ็ม อัลกอริทึม เสนอโดย เด็มสเตอร์ ลายด์และรูบิน (Dempster, Laird and Rubin 1977) ทฤษฎีดังกล่าวได้ทำการประมาณค่าเป็น 2 ขั้น คือ ขั้นการประมาณค่าข้อมูลที่ถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไข (Conditional Expectation) คือ $E(Y_i|Y_i>C_i,\beta,\sigma)$ ซึ่งขั้นการประมาณค่าสังเกตที่ถูกตัดทิ้ง ด้วยค่าคาดหวังที่มีเงื่อนไขนี้ เรียกว่า ขั้นหาค่าความคาดหวัง (Expectation Step : E Step) เมื่อประมาณค่าที่ถูกตัดทิ้งได้แล้วจะทำให้ได้ข้อมูลที่สมบูรณ์ Y_i^* แสดงได้ดังนี้

$$Y_i^* = Y_i\sigma_i + E(Y_i|Y_i>C_i,\beta,\sigma)(1-\sigma_i)$$

ขั้นต่อมานำข้อมูลที่สมบูรณ์ Y^* มาประมาณค่าพารามิเตอร์ ขั้นนี้เรียกว่า การประมาณค่าสูงสุดของพารามิเตอร์ (Maximization Step : M Step) จากขั้นนี้จะได้สถิติที่เพียงพอสำหรับพารามิเตอร์ของข้อมูลที่สมบูรณ์

2.2 การประมาณค่าพารามิเตอร์

ศึกษารูปแบบสมการถดถอยอย่างง่าย มีรูปแบบดังนี้

$$T_i = \beta_o + \beta_1 X_i + \varepsilon_i \quad , \ i = 1,2,\ldots,n+m$$

T_i เป็นตัวแปรตาม

X_i เป็นตัวแปรอิสระ

β_i เป็นพารามิเตอร์ที่ไม่ทราบค่า , $i = 0,1$

ε_i เป็นค่าความคลาดเคลื่อนสุ่ม

ค่าสังเกตของตัวแปรตาม T_i , $i = 1,\ldots,n+m$ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง F

ค่าสังเกตของตัวแปรตาม C_i , $i = 1,\ldots,n+m$ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง G

ดังนั้น T_i และ C_i เป็นอิสระกัน จากนิยามของการตัดทิ้งแบบสุ่ม ค่าสังเกต Y_i ได้จาก

$$Y_i = \min(T_i, C_i)$$

$$\sigma_i = \begin{cases} 1 & \text{ถ้า } T_i \leq C_i \quad \text{(เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง)} \\ 0 & \text{ถ้า } T_i > C_i \quad \text{(เป็นค่าสังเกตที่ถูกตัดทิ้ง)} \end{cases}$$

ค่าสังเกตที่ถูกตัดทิ้งประมาณได้ด้วย ค่าคาดหวังที่มีเงื่อนไข $E(Y_i | Y_i > C_i, \beta X_i)$ ค่าสังเกต $Y_i^*(\beta)$ ได้จาก

$$Y_i^*(\beta) = Y_i\sigma_i + E(Y_i | Y_i > C_i, \beta X_i)(1-\sigma_i)$$

การประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta}$ ได้ดังนี้

$$\underset{\sim}{\beta} = (X'X)^{-1} X'\underset{\sim}{Y}^*(\beta)$$

การประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta}$ ใช้วิธีกำลังสองต่ำสุด วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด และวิธีการของบัคเลย์และเจมส์

2.2.1 วิธีกำลังสองต่ำสุด

วิธีการหาตัวประมาณของพารามิเตอร์วิธีนี้ เป็นวิธีที่มีรากฐานมาจากทฤษฎีการประมาณเชิงเส้น (Theory of Linear Estimation) เป็นวิธีที่คิดขึ้นโดย คาร์ล เฟรดริก เกาส์ (Karl Freidrich Gauss 1777-1855) และ อังเดร แอนดรีวิช มาร์คอฟ (Andrie Andreevich Markov 1856-1922)[2] โดยมีหลักเกณฑ์ว่าหาค่าประมาณของพารามิเตอร์ ที่ทำให้ผลบวกกำลังสองของผลต่างระหว่างค่าสังเกตกับค่าประมาณมีค่าต่ำสุด ในกรณีที่ข้อมูลเป็นไปตามข้อตกลงเบื้องต้นของการวิเคราะห์ความถดถอย คือ

---

[2]ประชุม สุวัตถี,ดร., ทฤษฎีการอนุมานเชิงสถิติ. (กรุงเทพมหานคร:2527),หน้า 158

1. ค่าความคลาดเคลื่อนจะต้องมีการแจกแจงเป็นแบบปกติ ที่มีค่าเฉลี่ยเป็น 0 และมีค่าความแปรปรวนเป็น σ^2

2. ค่าความคลาดเคลื่อนจะต้องเป็นอิสระต่อกัน คือ ε_i และ ε_j จะต้องไม่มีความสัมพันธ์ต่อกัน เมื่อ $i \neq j$, $i=1,\ldots,n$ $j=1,\ldots,n$, n คือ ขนาดตัวอย่าง

3. ค่าความคลาดเคลื่อน ε_i จะต้องเป็นอิสระกับตัวแปรอิสระ X หรือ $Cov(\varepsilon_i, X_i)$ เท่ากับ 0 , $i = 1,\ldots,n$ เมื่อ n คือขนาดตัวอย่าง

ดังนั้น ตัวประมาณพารามิเตอร์โดยวิธีกำลังสองต่ำสุด จะเป็นตัวประมาณเชิงเส้นที่ไม่เอนเอียงและมีความแปรปรวนต่ำสุด เรียกคุณสมบัตินี้ว่า BLUE (Best Linear Unbiased Estimator) แต่การศึกษาครั้งนี้ เนื่องจากข้อมูลที่นำมาวิเคราะห์มีค่าของตัวแปรตามถูกตัดทิ้งทางขวา ดังนั้น วิธีกำลังสองต่ำสุดจะทำให้ได้ตัวประมาณที่เอนเอียง และโดยเฉลี่ยการประมาณค่าจะต่ำกว่าความเป็นจริง วิธีกำลังสองต่ำสุดไม่ได้กระทำวนซ้ำ และข้อมูลที่นำมาวิเคราะห์ จะถือว่าค่าสังเกตที่ถูกตัดทิ้งเสมือนเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

การหาตัวประมาณกำลังสองต่ำสุด

จากสมการความสัมพันธ์ระหว่างตัวแปรตาม Y และตัวแปรอิสระ X คือ

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad \text{เมื่อ } \underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$$

ให้ $\underset{\sim}{\hat{\beta}} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$ เป็นเวกเตอร์ของตัวประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta}$ จะได้ความสัมพันธ์คาดหมาย คือ

$$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}} + \underset{\sim}{e}$$

เมื่อ $\underset{\sim}{e}$ คือ ค่าความคลาดเคลื่อนระหว่างค่าสังเกตของ $\underset{\sim}{Y}$ กับค่าประมาณ $\underset{\sim}{\hat{Y}}$ ดังนั้น

$$\underset{\sim}{e} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}$$

พิจารณาผลบวกกำลังสองของค่าความคลาดเคลื่อน (Sum of Squared Error) จะพบว่า

$$\begin{aligned}
\underset{\sim}{e}'\underset{\sim}{e} &= (\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}})' (\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}) \\
&= (\underset{\sim}{Y}' - X'\underset{\sim}{\hat{\beta}}') (\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}) \\
&= (\underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\hat{\beta}}'X'\underset{\sim}{Y} + \underset{\sim}{\hat{\beta}}'X'X\underset{\sim}{\hat{\beta}})
\end{aligned}$$

ตัวประมาณกำลังสองต่ำสุด คือ ตัวประมาณที่ได้จากการทำให้ผลบวกกำลังสองของความคลาดเคลื่อน หรือ $\underset{\sim}{e}'\underset{\sim}{e}$ มีค่าต่ำสุด การหาค่าต่ำสุดของผลบวกกำลังสองของความคลาดเคลื่อน ทำได้โดยหาอนุพันธ์ (Differentiate) เทียบกับ $\underset{\sim}{\hat{\beta}}$ แล้วกำหนดให้เท่ากับ $\underset{\sim}{0}$ ดังนั้น

$$\frac{\partial}{\partial \underset{\sim}{\hat{\beta}}} (\underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\hat{\beta}}'\underset{\sim}{X}'\underset{\sim}{Y} + \underset{\sim}{\hat{\beta}}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\hat{\beta}}) = \underset{\sim}{0}$$

$$2\underset{\sim}{X}'\underset{\sim}{Y} + 2\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\hat{\beta}}) = \underset{\sim}{0}$$

$$(\underset{\sim}{X}'\underset{\sim}{X})\underset{\sim}{\hat{\beta}} = \underset{\sim}{X}'\underset{\sim}{Y}$$

$$\underset{\sim}{\hat{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$$

2.2.2 วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด

วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด เป็นการประมาณค่าพารามิเตอร์ที่ไม่ใช้ฟังก์ชันการเสี่ยงในการคัดเลือกตัวประมาณที่เหมาะสม แต่ใช้การวิเคราะห์จากสภาพความเป็นจริง ผู้ที่ค้นพบวิธีนี้เป็นคนแรกชื่อ เกาส์ ซี เอฟ (Gauss C.F. 1821) ซึ่งเป็นนักคณิตศาสตร์ชาวเยอรมัน ต่อมานักสถิติชาวอังกฤษชื่อ อาร์ เอ ฟิชเชอร์ (R.A. Fisher 1922) ได้ปรับปรุงวิธีการและตรวจสอบคุณสมบัติต่าง ๆ วิธีการนี้จะใช้ได้เมื่อตัวอย่างสุ่มมีการแจกแจงแบบมีพารามิเตอร์ (Parametric Distribution)[3]

สำหรับการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อมีการตัดค่าทางขวาของตัวแปรตาม ด้วยวิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด เสนอโดยเมอร์เรย์ ไอท์เคน (Murray Aitkin 1981)[4] นั้น การประมาณค่าพารามิเตอร์ ด้วยวิธีการ

---

[3] ธีระพร วีระถาวร, ดร. การอนุมานเชิงสถิติชั้นกลาง: โครงสร้างและความหมาย (พิทักษ์การพิมพ์ : 2531) หน้า 99.

[4] Murray Aitkin. A Note on the Regression Analysis of Censored Data. Technometrics. 1981,23,2, pp. 161-163.

ประมาณด้วยภาวะน่าจะเป็นสูงสุดได้ใช้ อีเอ็ม อัลกอริทึม เสนอโดย เด็มสเตอร์ ลายด์และรูบิน วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด ใช้วิธีการกระทำวนซ้ำ และข้อมูลที่นำมาวิเคราะห์จะถือว่าค่าสังเกตที่ถูกตัดทิ้งเสมือนเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

จากสมการ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ เมื่อ $\varepsilon_i \sim N(0,\sigma^2)$

ให้ $\mu_i = \sum_{j=0}^{1} \beta_j x_{ij}$, $x_{i0} = 1$

$z_i = (y_i - \mu_i)/\sigma$

$f(z) = (\sqrt{2\pi})^{-1} \exp(-z^2/2)$, $z \sim N(0,1)$

$S(z) = 1-F(z) = \int_z^{\infty} f(t)\, dt$

$h(z) = f(z) / (1-F(z))$

$\phi(y) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2} (y-\mu)\right)^2$

ฟังก์ชันภาวะน่าจะเป็นของการตัดทิ้งแบบสุ่ม และเกิดค่าถูกตัดปลายทางขวา

$$L = \prod_{i=1}^{n} \phi(y_i) \prod_{n+1}^{n+m} S(y_i)$$

$$= \frac{1}{\sigma^n} \prod_{i=1}^{n} f(z_i) \prod_{n+1}^{n+m} S(z_i)$$

($i=1,\ldots,n$ ค่าสังเกตที่ไม่ถูกตัดทิ้ง , $i=n+1,\ldots,n+m$ ค่าสังเกตที่ถูกตัดทิ้ง)

$$L = \frac{1}{\sigma^n} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2) \prod_{n+1}^{n+m} \int_z \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\, dt$$

$$= \frac{1}{\sigma^n} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2) \prod_{n+1}^{n+m} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2)$$

$$= \frac{1}{\sigma^n} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2} (y_i-\mu_i)^2\right) \prod_{n+1}^{n+m} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2} (y_i-\mu_i)^2\right)$$

$$= \frac{1}{\sigma^n} \frac{1}{(2\pi)^{(n+m)/2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n+m} (y_i-\mu_i)^2\right)$$

$$\ln L = -n\ln\sigma - \frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}(y_i-\mu_i)^2$$

$$= -n\ln\sigma - \frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}(y_i^2 - 2y_i\mu_i + \mu_i^2)$$

$$= -n\ln\sigma - \frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n+m}(y_i^2 - 2y_i\sum_{j=0}^{1}\beta_j x_{ij} + (\sum_{j=0}^{1}\beta_j x_{ij})^2)$$

ประมาณค่าพารามิเตอร์ β_j ได้โดยหาอนุพันธ์บางส่วน (Partial derivatives) ของล็อกของภาวะน่าจะเป็น (log-likelihood) เทียบกับ β_j และให้สมการอนุพันธ์บางส่วนเท่ากับ 0

$$\frac{\partial \ln L}{\partial \beta_j} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n+m}(-2y_i x_{ij} + 2\sum_{j=0}^{1}\beta_j x_{ij}^2)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n+m}(y_i - \sum_{j=0}^{1}\beta_j x_{ij})x_{ij}$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n+m}(w_i - \mu_i)x_{ij}$$

โดยที่ $$w_i = \begin{cases} y_i & , \ (i=1,2,\ldots,n) \\ \mu_i + \sigma\, h(z_i) & , \ (i=n+1,\ldots,n+m) \end{cases}$$

ประมาณค่าพารามิเตอร์ σ ได้โดยหาอนุพันธ์บางส่วน (Partial derivatives) ของล็อกของภาวะน่าจะเป็น เทียบกับ σ และให้สมการอนุพันธ์บางส่วนเท่ากับ 0

$$L = \frac{1}{\sigma^n}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-1}{2\sigma^2}(y_i-\mu_i)^2\right)\prod_{n+1}^{n+m}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-1}{2\sigma^2}(y_i-\mu_i)^2\right)$$

$$\ln L = -n\ln\sigma - \frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 - \frac{1}{2\sigma^2}\sum_{n+1}^{n+m}(y_i-\mu_i)^2$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu_i)^2 + \frac{1}{\sigma^3}\sum_{n+1}^{n+m}(y_i-\mu_i)^2$$

$$= -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 + \sum_{n+1}^{n+m}\frac{(y_i-\mu_i)}{\sigma}\,\frac{(y_i-\mu_i)}{\sigma}$$

$$= -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 + \sum_{n+1}^{n+m} z_i\,\frac{(\mu_i+\sigma\, h(z_i)-\mu_i)}{\sigma}$$

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 + \sum_{n+1}^{n+m} z_i\, h(z_i) = 0$$

$$\sigma^2 = \sum_{i=1}^{n}(y_i-\mu_i)^2 \,/\, \left(n - \sum_{n+1}^{n+m} z_i\, h(z_i) \right)$$

ในทางปฏิบัติการประมาณค่าสูงสุดของพารามิเตอร์ (Maximization Step : M Step) ในรอบที่ (k+1) จะประมาณค่าพารามิเตอร์ $\beta^{(k+1)}$ โดยวิธีกำลังสองต่ำสุด และประมาณ $\sigma^2_{(k+1)}$ ได้ดังต่อไปนี้

(n+m) $$\hat{\sigma}^2_{(k+1)} = \sum_{i=1}^{n+m}(y_i-\hat{\mu}_i^{(k)})^2$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \sum_{n+1}^{n+m}(y_i-\hat{\mu}_i^{(k)})^2$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \sum_{n+1}^{n+m}(y_i^2 - 2y_i\hat{\mu}_i^{(k)} + \hat{\mu}_i^{2(k)}) \quad .....(1)$$

เนื่องจากสถิติที่เพียงพอของพารามิเตอร์ สำหรับข้อมูลที่สมบูรณ์ $\sum_{i=1}^{n+m} x_i y_i$ และ $\sum_{i=1}^{n+m} y_i^2$ แสดงได้ดังนี้

$$E\left(\sum_{i=1}^{n+m} x_i y_i\right) = \sum_{i=1}^{n} x_i y_i + \sum_{n+1}^{n+m} x_i\, E(Y_i/Y_i > c_i,\beta,\sigma^2)$$

$$E\left(\sum_{i=1}^{n+m} y_i^2\right) = \sum_{i=1}^{n} y_i^2 + \sum_{n+1}^{n+m} E(Y_i^2/Y_i > c_i,\beta,\sigma^2)$$

โดย $$E(Y_i|Y_i > c_i,\beta,\sigma^2) = \hat{\mu}_i + \hat{\sigma}\, h(z_i)$$

$$E(Y_i^2|Y_i > c_i,\beta,\sigma^2) = \hat{\mu}_i^2 + \hat{\sigma}^2 + \hat{\sigma}(y_i+\hat{\mu}_i)\, h(z_i)$$

ดังนั้น y_i , y_i^2 , $i = n+1,\ldots,n+m$ มีค่าประมาณดังต่อไปนี้

$$y_i = E(Y_i|Y_i > c_i,\beta,\sigma^2) = \hat{\mu}_i + \hat{\sigma}\, h(z_i)$$

$$y_i^2 = E(Y_i^2|Y_i > c_i,\beta,\sigma^2) = \hat{\mu}_i^2 + \hat{\sigma}^2 + \hat{\sigma}(y_i+\hat{\mu}_i)h(z_i)$$

นำค่าประมาณ y_i, y_i^2 , $i = n=1,\ldots,n+m$ แทนค่าในสมการ (1) จะได้

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \sum_{n+1}^{n+m}[\,\hat{\sigma}^2_{(k)} + \hat{\sigma}^2_{(k)}h(z_i)^{2(k)}\,]$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{(k)}\sum_{n+1}^{n+m}[1+ h(z_i)^{2(k)}\,]$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{(k)}\sum_{n+1}^{n+m}[1+ h(z_i)^{(k)}\, h(z_i)^{(k)}\,]$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{(k)}\sum_{n+1}^{n+m}\left[1+ \left[\frac{y_i-\hat{\mu}_i^{(k)}}{\hat{\sigma}_{(k)}}\right] h(z_i)^{(k)}\,\right]$$

$$= \sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{(k)}\sum_{n+1}^{n+m}[1+ z_i^{(k)}\, h(z_i)^{(k)}\,]$$

$$\hat{\sigma}^2_{(k+1)} = \{\sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{(k)}\sum_{n+1}^{n+m}[1+ z_i^{(k)}\, h(z_i)^{(k)}]\,\}/(n+m)$$

ขั้นตอนในการหาค่าประมาณพารามิเตอร์ $\hat{\beta}$ **และ** $\hat{\sigma}^2$ สำหรับวิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด มีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 ข้อมูลที่นำมาวิเคราะห์ จะถือว่าค่าสังเกตที่ถูกตัดทิ้งเสมือนเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง จากข้อมูลทั้งหมดประมาณค่าพารามิเตอร์เริ่มต้น $\hat{\beta}_o^{(o)}$,$\hat{\beta}_1^{(o)}$ และ $\hat{\sigma}^2_{LS}$ ด้วยวิธีกำลังสองต่ำสุด

ขั้นที่ 2 เฉพาะข้อมูลที่ถูกตัดทิ้ง หาค่า

$$\hat{\mu}_i^{(o)} = \hat{\beta}_o^{(o)} + \hat{\beta}_1^{(o)}X_i$$

$$\hat{z}_i^{(o)} = (\,c_i^{(o)} - \hat{\mu}_i^{(o)}\,) / \hat{\sigma}_{LS}$$

$$f(z_i)^{(o)} = (\sqrt{2\pi})^{-1} \exp(-z_i^2/2)$$

$$S(z_i)^{(o)} = 1-F(z_i)^{(o)} = \int_{z_i}^{\infty} f(t)\,dt$$

$$h(z_i)^{(o)} = f(z_i)^{(o)} / (1-F(z_i)^{(o)}) \quad , \quad i=n+1,\ldots,n+m$$

$c_i^{(o)}$ คือ ค่าสังเกตที่ถูกตัดทิ้ง $i=1,\ldots,m$, m คือ จำนวนข้อมูลที่ถูกตัดทิ้ง

ขั้นที่ 3 ประมาณค่าถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไข $E(Y_i|Y_i>c_i,\beta,\sigma) = w_i$

$$w_i^{(o)} = \hat{\mu}_i^{(o)} + \hat{\sigma}_{LS}\, h(z_i)^{(o)} \quad , \quad i = n+1,\ldots,n+m$$

ดังนั้นได้ค่าสังเกต $y_i^*(\beta) = y_i\delta_i + E(Y_i|Y_i > c_i,\beta,\sigma)(1-\delta_i)$

ขั้นที่ 4 นำค่าสังเกต $y_i^*(\beta)$ จากขั้นที่ 3 หาค่าประมาณพารามิเตอร์ $\hat{\beta}_o^{(k)}$, $\hat{\beta}_1^{(k)}$ ด้วยวิธีกำลังสองต่ำสุด และหาค่า $\hat{\sigma}^2_{MLE(k)}$ จากสูตร

$$\hat{\sigma}^2_{MLE(k)} = \{ \sum_{i=1}^{n} (y_i - \hat{\mu}_i^{(o)})^2 + \hat{\sigma}^2_{LS} \sum_{n+1}^{n+m} [1 + z_i^{(o)}\, h(z_i)^{(o)}] \} / (n+m)$$

ขั้นที่ 5 เปรียบเทียบค่าประมาณพารามิเตอร์จากขั้นที่ 1 และขั้นที่ 4 คือ $\hat{\beta}_o^{(o)}$, $\hat{\beta}_1^{(o)}$ และ $\hat{\sigma}_{LS}$ เทียบกับ $\hat{\beta}_o^{(k)}$, $\hat{\beta}_1^{(k)}$ และ $\hat{\sigma}_{MLE(k)}$ ถ้าค่ายังไม่เท่ากันให้ทำขั้นต่อไป

ขั้นที่ 6 เฉพาะข้อมูลที่ถูกตัดทิ้ง หาค่า

$$\hat{\mu}_i^{(k)} = \hat{\beta}_o^{(k)} + \hat{\beta}_1^{(k)} x_i$$

$$\hat{z}_i^{(k)} = (c_i^{(o)} - \hat{\mu}_i^{(k)}) / \hat{\sigma}_{MLE(k)}$$

$$f(z_i)^{(k)} = (\sqrt{2\pi})^{-1} \exp(-z_i^2/2)$$

$$S(z_i)^{(k)} = 1-F(z_i)^{(k)} = \int_{z_i}^{\infty} f(t)\,dt$$

$$h(z_i)^{(k)} = f(z_i)^{(k)} / (1-F(z_i)^{(k)}) \quad , \quad i=n+1,\ldots,n+m$$

$c_i^{(o)}$ คือ ค่าสังเกตที่ถูกตัดทิ้ง $i=1,\ldots,m$, m คือ จำนวนข้อมูลที่ถูกตัดทิ้ง

ขั้นที่ 7 ประมาณค่าถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไข $E(Y_i|Y_i>c_i,\beta,\sigma) = w_i$

$$w_i^{(k)} = \hat{\mu}_i^{(k)} + \hat{\sigma}_{MLE(k)}\, h(z_i)^{(k)} \quad , i= n+1,\ldots,n+m$$

ดังนั้นค่าสังเกต $y_i^*(\beta) = y_i\sigma_i + E(Y_i|Y_i > c_i,\beta,\sigma)(1-\sigma_i)$

ขั้นที่ 8 นำค่าสังเกต $y_i^*(\beta)$ จากขั้นที่ 7 หาค่าประมาณพารามิเตอร์ $\hat{\beta}_o^{(k+1)}$, $\hat{\beta}_1^{(k+1)}$ ด้วยวิธีกำลังสองต่ำสุด และหาค่า $\hat{\sigma}_{MLE(k+1)}$ จากสูตร

$$\hat{\sigma}^2_{MLE(k+1)} = \{\sum_{i=1}^{n}(y_i-\hat{\mu}_i^{(k)})^2 + \hat{\sigma}^2_{MLE(k)}\sum_{n+1}^{n+m}[1+ z_i^{(k)}\, h(z_i)^{(k)}]\} / (n+m)$$

ขั้นที่ 9 เปรียบเทียบค่าพารามิเตอร์จากขั้นที่ 4 และขั้นที่ 8 คือ $\hat{\beta}_o^{(k)}, \hat{\beta}_1^{(k)}$ และ $\hat{\sigma}_{MLE(k)}$ เทียบกับ $\hat{\beta}_o^{(k+1)}$, $\hat{\beta}_1^{(k+1)}$ และ $\hat{\sigma}_{MLE(k+1)}$ ถ้าหากว่าค่าประมาณพารามิเตอร์ของรอบที่ k+1 เท่ากับค่าประมาณพารามิเตอร์ของรอบที่ k จึงหยุดการกระทำวนซ้ำ และจะได้ค่าประมาณของ $\hat{\beta}_o$, $\hat{\beta}_1$ และ $\hat{\sigma}_{MLE}$ ถ้าหากค่าประมาณของพารามิเตอร์ รอบที่ k+1 ไม่เท่ากับรอบที่ k ให้นำค่าประมาณพารามิเตอร์ $\hat{\beta}_o^{(k+1)}, \hat{\beta}_1^{(k+1)}$ และ $\hat{\sigma}_{MLE(k+1)}$ แทนค่าในขั้นตอนที่ 6 แล้วกระทำวนซ้ำจากขั้นตอนที่ 6 ถึงขั้นตอนที่ 8 ทำจนกระทั่งค่าประมาณพารามิเตอร์ของรอบปัจจุบัน เท่ากับค่าประมาณพารามิเตอร์รอบที่แล้วจึงหยุด

ขั้นที่ 10 นำค่าประมาณ $\hat{\beta}_o$ และ $\hat{\beta}_1$ จากขั้นที่ 9 หาค่าประมาณของตัวแปรตาม $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$ และหาค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตาม กับค่าจริงในรูปของค่ารากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (RMSE)

$$MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n$$

$$RMSE = \surd MSE$$

n คือ จำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

y_i คือ ค่าจริงของค่าสังเกตที่ไม่ถูกตัดทิ้ง , i=1,...,n

2.2.3 วิธีการของบัคเลย์และเจมส์

การประมาณค่าพารามิเตอร์ ด้วยวิธีการของบัคเลย์และเจมส์ ซึ่งเสนอโดยโจนาทาน บัคเลย์ และเอียน เจมส์ (Jonathan Buckley and Ian James 1979)[5] วิธีการของบัคเลย์และเจมส์ เป็นวิธีการที่เป็นนอนพาราเมตริก (Nonparametric Method) ซึ่งวิธีนี้ได้ใช้อีเอ็ม อัลกอริทึม ซึ่งเสนอโดย เด็มสเตอร์ ลายด์และรูบิน เช่นเดียวกับวิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด ในการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้น วิธีการของบัคเลย์และเจมส์ มีข้อสมมติว่า ค่าความคลาดเคลื่อน ε_i เป็นอิสระ มีฟังก์ชันการแจกแจง F ที่ไม่มีรูปแบบเฉพาะ มีค่าเฉลี่ยเป็น α มีค่าความแปรปรวนเป็น σ^2 และมีฟังก์ชันการอยู่รอดเป็น $S = 1 - F$ ตัวแบบเชิงเส้นแสดงดังนี้

$$T_i = \alpha + \beta X_i + \varepsilon_i \ , \ i = 1,2,\ldots,n$$

n คือ จำนวนข้อมูลทั้งหมด

$$E(T_i) = \alpha + \beta X_i$$

ค่าสังเกตของตัวแปรตาม T_i , $i = 1,\ldots,n$ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง F

ค่าสังเกตของตัวแปรตาม C_i , $i = 1,\ldots,n$ มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง G

---

[5] Jonathan Buckley And Ian James. Linear Regression with Censored Data. Biometrika. (1979),66,3, pp. 429-436.

Rupert G. Miller. Survival Analysis. (New York : John Willey, 1981), pp. 150-154

ดังนั้น T_i และ C_i เป็นอิสระกัน จากนิยามของการตัดทิ้งแบบสุ่ม ค่าสังเกตุ Y_i ได้จาก

$$Y_i = \min(T_i, C_i)$$

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i \leqslant C_i \\ 0 & \text{ถ้า } T_i > C_i \end{cases}$$

เนื่องจากว่า $E(Y_i) \neq \alpha + \beta X_i$ บัคเลย์และเจมส์ จึงนิยามค่าสังเกตของตัวแปรสุ่ม $Y_i^*(\beta)$ เพื่อใช้ในการคำนวณค่าพารามิเตอร์ ดังนี้คือ

$$Y_i^*(\beta) = Y_i\delta_i + E(Y_i | Y_i > C_i, \beta X_i)\ (1-\delta_i)$$

ดังนั้น นำค่าสังเกต $Y_i^*(\beta)$, $i = 1,\ldots,n$, n คือ จำนวนข้อมูลทั้งหมด นำมาหาค่าพารามิเตอร์ $\underset{\sim}{\beta}$ ได้ดังนี้

$$\underset{\sim}{\beta} = (X'X)^{-1} X'\underset{\sim}{Y}^*(\beta)$$

ให้ $\underset{\sim}{\beta} = (\alpha, \beta)'$ เป็นเวกเตอร์ของพารามิเตอร์ จะได้ว่า

$$E(Y_i^*)^{6} = \alpha + \beta X_i$$

เนื่องจากค่าสังเกตที่ถูกตัดทิ้งประมาณได้ด้วย ค่าคาดหวังที่มีเงื่อนไข และในรูปแบบปกติ (Normal Model) เมื่อ $F = \Phi$ จะได้

$$E(Y_i | Y_i > c_i, \beta x_i) = \beta x_i + \sigma\, h[(c_i - \beta x_i)/\sigma]$$

โดยที่ $h(u) = \phi(u) / (1 - \Phi(u))$ ซึ่ง $h(u)$ เป็นอัตราการสูญเสีย (Failure Rate) ของการแจกแจงปกติ การหาค่า $E(Y_i | Y_i > c_i, \beta x_i)$ โดยแทนค่าประมาณ β และ σ แต่บัคเลย์และเจมส์ (1979) พิจารณาว่ากรณีที่ไม่ทราบฟังก์ชันการแจกแจง F จึงไม่สามารถหาค่า $E(Y_i | Y_i > c_i, \beta x_i)$ ได้ จึงต้องประมาณการแจกแจง F โดยใช้ตัวประมาณพีแอล

---

[6] Ruper G. Miller. Survival Analysis. (New York : John Wiley & Sons, 1981) pp. 151.

$$\hat{F}(e_i,\hat{\beta}) = 1 - \prod_{i:e(\hat{\beta})<e} \left[\frac{n-i}{n-i+1}\right]$$

$$= 1 - \hat{S}(e_i,\hat{\beta})$$

ซึ่งได้มาจากการหาลำดับที่ของความคลาดเคลื่อนบางส่วน (Partial Residuals) โดยคำนวณที่จุด $\hat{\alpha} = 0$, $\hat{e}_i(0,\hat{\beta}) = y_i - \hat{\beta}x_i$, $\hat{e}_1 \leqslant e_1$, $\hat{e}_1 < \hat{e}_2 , \ldots, < \hat{e}_n$ โดยค่าความคลาดเคลื่อนบางส่วนเป็นอิสระและมีฟังก์ชันการแจกแจง F ที่ไม่มีรูปแบบเฉพาะ มีค่าเฉลี่ยเป็น α และค่าความแปรปรวน σ^2 มีฟังก์ชันการอยู่รอด $\hat{S} = 1 - \hat{F}$

ดังนั้น ข้อมูลที่ถูกตัดทิ้งจะถูกแทนด้วย $E(Y_i|Y_i > c_i,\beta x_i) = \bar{y}_i(\hat{\beta})$

$$\bar{y}_i(\hat{\beta}) = \hat{\beta}x_i + \Sigma_{uc} \frac{W_{ik}(e,\hat{\beta})(y_k - \hat{\beta}x_k)}{\{1 - \hat{F}(c_i - \hat{\beta}x_i)\}}$$

$i = 1,\ldots,n$ $k = 1,\ldots,n$ และประมาณค่าพารามิเตอร์ ตามวิธีการของบัคเลย์และเจมส์ ได้ดังนี้

$$\hat{\beta}_{BJ} = \frac{\{\Sigma_{uc} y_i(x_i - \bar{x}) + \Sigma_c \bar{y}_i(\hat{\beta})(x_i - \bar{x})\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha}_{BJ} = n^{-1}\{\Sigma_{uc} y_i + \Sigma_c \bar{y}_i(\hat{\beta})\} - \hat{\beta}_{BJ}\bar{x}$$

ขั้นตอนการดำเนินงานสำหรับวิธีการของบัคเลย์และเจมส์ มีดังต่อไปนี้

ขั้นที่ 1 เฉพาะข้อมูลที่ไม่ถูกตัดทิ้งประมาณค่าพารามิเตอร์เริ่มต้น $\hat{\alpha}$, $\hat{\beta}$ โดยใช้วิธีกำลังสองต่ำสุด

$$\hat{\beta} = \frac{\Sigma_{uc} y_i (x_i - \bar{x}^{uc})}{\Sigma_{uc}(x_i - \bar{x}^{uc})^2}$$

$$\hat{\alpha} = \bar{y}^{uc} - \hat{\beta}\bar{x}^{uc}$$

Σ_{uc} คือ ผลรวมของค่าสังเกตเฉพาะที่ไม่ถูกตัดทิ้ง

$\bar{x}^{uc}$ เป็นค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

$\bar{y}^{uc}$ เป็นค่าเฉลี่ยของ y_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

ขั้นที่ 2 ค่าที่ถูกตัดทิ้งเสมือนค่าที่ไม่ถูกตัดทิ้งนำข้อมูลทั้งหมดหาค่าความคลาดเคลื่อนบางส่วน โดยคำนวณที่ $\hat{\alpha} = 0$ ได้ดังนี้

$$\hat{e}_i(0,\hat{\beta}) = y_i - \hat{\beta}x_i \quad , \quad i = 1,\ldots,n$$

ให้เรียงลำดับ (Rank) ค่าความคลาดเคลื่อนบางส่วน $\hat{e}_i(0,\hat{\beta})$ จากน้อยไปหามาก จะได้ $\hat{e}_1 < \hat{e}_2, \ldots, < \hat{e}_n$ ในกรณีที่ลำดับที่ของค่าความคลาดเคลื่อนของค่าถูกตัดทิ้ง และค่าที่ไม่ถูกตัดทิ้งมีค่าเท่ากัน ให้ลำดับที่ของค่าที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของค่าถูกตัดทิ้ง

ขั้นที่ 3 ให้เปลี่ยนค่าลำดับที่ของ <u>ค่าถูกตัดทิ้งเป็น 0</u> ส่วนลำดับที่ของค่าที่ไม่ถูกตัดทิ้งให้คงไว้

ขั้นที่ 4 หาค่า $\hat{S}$ โดยใช้ตัวประมาณพีแอล

$$\hat{S}(e_i,\hat{\beta}) = \prod_{i:e(\hat{\beta})<e} \left[\frac{n-i}{n-i+1}\right]$$

i คือ ลำดับที่ของค่าความคลาดเคลื่อน , n คือ จำนวนข้อมูลทั้งหมด ค่า $\hat{S}$ นี้คือ ค่าของเวลาที่มีการอยู่รอด (Survival Time)

ขั้นที่ 5 หาค่าฟังก์ชัน $\hat{F}(e_i,\hat{\beta})$ จาก

$$\hat{F}(e_i,\hat{\beta}) = 1 - \hat{S}(e_i,\hat{\beta})$$

ขั้นที่ 6 หาค่าถ่วงน้ำหนัก $W(e_i,\hat{\beta})$ ได้จาก

$$W(e_1,\hat{\beta}) = \hat{F}_1$$

$$W(e_2,\hat{\beta}) = \hat{F}_2 - \hat{F}_1$$

$$\vdots$$

$$W(e_n,\hat{\beta}) = \hat{F}_n - \hat{F}_{n-1}$$

ในกรณีที่ลำดับที่สูงสุดของความคลาดเคลื่อนเป็นลำดับที่ของค่าที่ถูกตัดทิ้ง ให้ปรับค่าถ่วงน้ำหนักเป็น $W^*(e_i,\hat{\beta})$ ดังต่อไปนี้

$$W^*(e_i,\hat{\beta}) = \frac{W(e_i,\hat{\beta})}{\Sigma_{uc}\, W(e_j,\hat{\beta})}$$

Σ_{uc} คือผลรวมของค่าสังเกตเฉพาะที่ไม่ถูกตัดทิ้ง

ขั้นที่ 7 ให้เปลี่ยนค่าลำดับที่ของ ค่าที่ไม่ถูกตัดทิ้งเป็น 0 ส่วนลำดับที่ของค่าที่ถูกตัดทิ้งให้คงไว้

ขั้นที่ 8 หาค่า $\hat{S}$ โดยใช้ตัวประมาณพีแอล

$$\hat{S}(e_i,\hat{\beta}) = \prod_{i:e(\hat{\beta})<e} \left[\frac{n-i}{n-i+1}\right]$$

i คือ ลำดับที่ของค่าความคลาดเคลื่อน , n คือ จำนวนข้อมูลทั้งหมด ค่า $\hat{S}$ นี้คือ ค่าของเวลาที่ถูกตัดทิ้ง (Censored Time)

ขั้นที่ 9 หาค่าฟังก์ชัน $\hat{F}(e_i,\hat{\beta})$ หรือ $\hat{F}(c_i-\hat{\beta}x_i)$ จาก

$$\hat{F}(c_i,\hat{\beta}x_i) = 1 - \hat{S}(e_i,\hat{\beta})$$

ขั้นที่ 10 จากขั้นที่ 6 และ 9 ได้ค่า $W(e_i,\hat{\beta})$ และ $\hat{F}(c_i,\hat{\beta}x_i)$ นำมาหาค่าประมาณของค่าสังเกตที่ถูกตัดทิ้งได้ด้วยค่าคาดหวังที่มีเงื่อนไข ดังต่อไปนี้

$$E(Y_i|Y_i > c_i,\beta x_i) = \bar{y}_i(\hat{\beta})$$

$$\bar{y}_i(\hat{\beta}) = \hat{\beta}x_i + \Sigma_{uc} \frac{W_{ik}(e,\hat{\beta})(y_k - \hat{\beta}x_k)}{\{1 - \hat{F}(c_i-\hat{\beta}x_i)\}}$$

$i = 1,\ldots,n \qquad k = 1,\ldots,n$

ขั้นที่ 11 ประมาณค่าพารามิเตอร์ $\hat{\beta}_{BJ}$ ได้จาก

$$\hat{\beta}_{BJ} = \frac{\{ \Sigma_{uc}\, y_i(x_i - \bar{x}) + \Sigma_c\, \bar{y}_i(\hat{\beta})\ (x_i - \bar{x}) \}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Σ_{uc} คือ ผลรวมของค่าสังเกตเฉพาะที่ไม่ถูกตัดทิ้ง

Σ_c คือ ผลรวมของค่าสังเกตเฉพาะที่ถูกตัดทิ้ง

$\bar{x} = \sum_{i=1}^{n} x_i / n$, n คือจำนวนข้อมูลทั้งหมด

ขั้นที่ 12 แทนค่า $\hat{\beta}_{BJ}$ จากขั้นที่ 11 ลงในขั้นที่ 2 แล้วทำการวนซ้ำจากขั้นที่ 2 จนถึงขั้นที่ 11 ทำไปจนกระทั่งค่าของ $\hat{\beta}_{BJ}$ ของรอบปัจจุบันได้เท่ากับค่าของ $\hat{\beta}_{BJ}$ ในรอบที่แล้วจึงหยุดและจะได้ค่าประมาณของ $\hat{\beta}_{BJ}$ ในบางครั้งค่าของ $\hat{\beta}_{BJ}$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้น เป็นค่าประมาณของ $\hat{\beta}_{BJ}$

ขั้นที่ 13 คำนวณค่า $\hat{\alpha}_{BJ}$ จาก

$$\hat{\alpha}_{BJ} = n^{-1} \{ \Sigma_{uc}\, y_i + \Sigma_c\, \bar{y}_i(\hat{\beta}) \} - \hat{\beta}_{BJ}\bar{x}$$

ขั้นที่ 14 นำค่าประมาณ $\hat{\alpha}_{BJ}$ และ $\hat{\beta}_{BJ}$ จากขั้นที่ 13 และ 12 หาค่าประมาณของตัวแปรตาม $\hat{y}_i = \hat{\alpha}_{BJ} + \hat{\beta}_{BJ}x_i$ และหาค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตาม กับค่าจริง ในรูปของค่ารากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (RMSE)

$$MSE = \sum_{i=1}^{n_{uc}} (y_i - \hat{y}_i)^2 / n_{uc}$$

$$RMSE = \surd MSE$$

n_{uc} คือ จำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

y_i คือ ค่าจริงของค่าสังเกตที่ไม่ถูกตัดทิ้ง , $i = 1,\ldots,n_{uc}$