บทที่ 2

สถิติที่ใช้ในการวิจัย

ในบทนี้กล่าวถึงทฤษฎีพื้นฐานต่างๆ ที่เกี่ยวข้อง และวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อค่าสังเกตของตัวแปรตามเป็นค่าที่ถูกตัดทิ้งทางขวาประเภทที่ 1 ด้วย 4 วิธีคือ วิธีกำลังสองต่ำสุด วิธีตัวประมาณของมิลเลอร์ วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ และวิธีการของบัคเลย์และเจมส์ ซึ่งมีรายละเอียดต่างๆ ดังนี้

2.1 ทฤษฎีพื้นฐาน

2.1.1 ประเภทของการตัดทิ้ง (Type of Censoring)

ลักษณะของข้อมูลที่ถูกตัดทิ้งบางส่วนนั้น เกิดขึ้นได้ในหลายลักษณะเช่น แบบประเภทที่ 1 แบบประเภทที่ 2 แบบสุ่ม เป็นต้น ดังรายละเอียดต่อไปนี้

ก. การตัดทิ้งประเภทที่ 1 (Type I Censoring)

การตัดทิ้งประเภทที่ 1 นี้ เกิดขึ้นเนื่องจากมีการกำหนดเวลาของการเกิดค่าที่ถูกตัดทิ้งเอาไว้ล่วงหน้าหรือมีการกำหนดค่าสูงสุดของข้อมูลล่วงหน้าด้วยค่าคงที่ T_c ซึ่งจะเรียกว่า "Fixed Censoring Time" ตัวอย่างของการตัดทิ้งประเภทนี้สำหรับข้อมูลทางด้านการประกันภัย เช่น การประกันภัยสุขภาพ บริษัทจะจ่ายเงินค่ารักษาพยาบาลให้ตามที่ผู้เอาประกันภัยจ่ายไปจริง แต่ไม่เกินผลประโยชน์สูงสุดที่กำหนดไว้ เช่น 10,000 บาท ค่า 10,000 บาทนี้จะเป็นค่า T_c ถ้าผู้เอาประกันภัยรายใดเสียค่ารักษาพยาบาลจริงไม่เกิน 10,000 บาท บริษัทจะจ่ายชดเชยให้และบันทึกการจ่ายเงินค่ารักษาพยาบาลตามจริง แต่ถ้าผู้เอาประกันภัยรายใดเสียค่ารักษาพยาบาลจริงมากกว่า 10,000 บาท บริษัทจะจ่ายชดเชยให้เพียง 10,000 บาท และบันทึกการจ่ายเงินค่ารักษาพยาบาล 10,000 บาท ข้อมูลลักษณะนี้เราจะถือว่าเป็นข้อมูลที่ถูกตัดทิ้งที่ T_c เท่ากับ 10,000 บาท

ให้ T_c เป็นเวลาที่กำหนดไว้ล่วงหน้า และให้ $T_1, T_2, \ldots, T_N$ เป็นตัวแปรสุ่มที่มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน จะได้ค่าสังเกตสุ่ม $Y_1, Y_2, \ldots, Y_N$ ซึ่ง

$$Y_i = \begin{cases} T_i \; ; \; T_i \leq T_c & \text{(ค่าสังเกตไม่ถูกตัดทิ้ง)} \\ T_c \; ; \; T_i > T_c & \text{(ค่าสังเกตถูกตัดทิ้ง)} \end{cases}$$

โดยมีฟังก์ชันภาวะน่าจะเป็น(Likelihood Function) คือ

$$L(y_i) = \begin{cases} f(y_i) & \text{(ค่าสังเกตไม่ถูกตัดทิ้ง)} \\ P(T_i > T_c) = S(T_c) & \text{(ค่าสังเกตถูกตัดทิ้ง)} \end{cases}$$

และมีฟังก์ชันภาวะน่าจะเป็นรวมดังนี้

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

เมื่อ $i \varepsilon u$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าไม่ถูกตัดทิ้ง

$i \varepsilon c$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าถูกตัดทิ้ง

ข. การตัดทิ้งประเภทที่ 2 (Type II Censoring)

ในบางกรณีไม่สามารถจะกำหนดเวลาของการเกิดค่าที่ถูกตัดทิ้งหรือค่าสูงสุดของการตัดที่เหมาะสมได้ ดังนั้นจะกำหนดจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้งแทน เช่น กำหนดเท่ากับ r นั่นคือเมื่อจำนวนของค่าสังเกตที่ไม่ถูกตัดทิ้งเกิดขึ้นจนครบแล้ว จะหยุดทำการทดลองเพื่อเป็นการประหยัดเวลาและค่าใช้จ่าย ตัวอย่างเช่น การทดสอบอายุการใช้งานของหลอดไฟ จะกำหนดจำนวนหลอดไฟที่เสื่อมสภาพไว้ล่วงหน้า เริ่มทดลองโดยเปิดหลอดไฟให้ทำงานทั้งหมดแล้วเริ่มบันทึกเวลาและนับจำนวนหลอดไฟที่เสื่อมสภาพซึ่งถือว่าเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง เมื่อได้จำนวนหลอดไฟที่เสื่อมสภาพครบแล้ว จะหยุดทำการทดสอบ

ให้ N เป็นจำนวนข้อมูลทั้งหมด และให้ n เป็นจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง โดยที่ $r \leq n$ ให้ $T_1 \leq T_2 \leq \ldots, T_n$ เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง และ $T_{n+1} \leq T_{n+2} \leq \ldots, T_N$ เป็นค่าสังเกตที่ถูกตัดทิ้ง ซึ่ง $T_i \geq T_n$; i = n+1,n+2,...,N จะไม่ทราบค่าที่แท้จริงของค่าสังเกต ดังนั้น Y_i เป็นตัวแปรสุ่มของค่าสังเกตซึ่ง

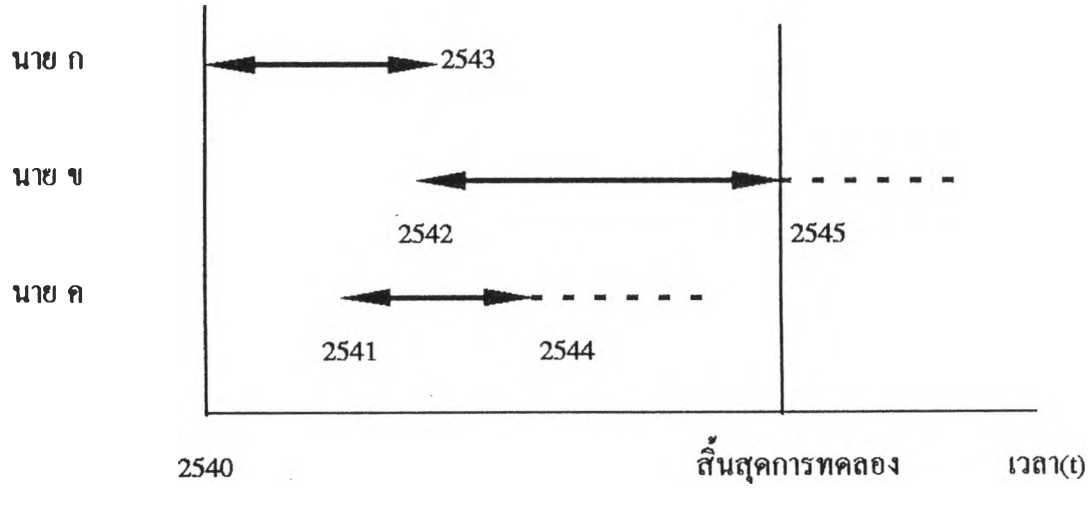
$$Y_i = \begin{cases} T_i ; & i = 1,2,\ldots,n \\ T_n ; & i = n+1,n+2,\ldots,N \end{cases}$$

ฟังก์ชันความหนาแน่นร่วมของค่าสังเกต คือ

$$\frac{N!}{(N-n)!}\prod_{i=1}^{n} f(y_i)\left[S(y_n)^{N-n}\right]$$

ค. การตัดทิ้งแบบสุ่ม (Random Censoring)

การตัดทิ้งแบบสุ่มมีลักษณะคล้ายการตัดทิ้งประเภทที่ 1 คือ มีการกำหนดระยะเวลาของการทดลองไว้ล่วงหน้าแต่การเกิดข้อมูลที่ถูกตัดทิ้งนั้น อาจเกิดขึ้นได้ก่อนสิ้นสุดการทดลองจึงเรียกว่าการตัดทิ้งแบบสุ่ม ส่วนใหญ่จะพบในการทดลองทางการแพทย์ เช่น โครงการทดลองเปลี่ยนหัวใจ คนไข้อาจถอนตัวออกจากการทดลองก่อนสิ้นสุดการทดลอง หรือคนไข้เสียชีวิตเนื่องจากสาเหตุอื่นที่ไม่เกี่ยวข้องกับสิ่งที่สนใจศึกษา หรือคนไข้มีชีวิตอยู่รอดเมื่อสิ้นสุดการทดลอง เป็นต้น จึงทำให้ไม่สามารถทราบค่าที่แน่นอนของค่าสังเกตนั้นได้ หรือการวิเคราะห์อัตราการอยู่รอดของกรมธรรม์ประกันชีวิตแบบสะสมทรัพย์ 10 ปี ชำระเบี้ยประกันภัย 10 ปี ในช่วงปี 2540-2545 และในช่วงที่ทำการศึกษานี้อาจมีผู้เอาประกันภัยบางรายเวนคืนกรมธรรม์ หรือ หยุดชำระเบี้ยประกันภัยแล้วแปลงสภาพกรมธรรม์ไปก่อนที่จะสิ้นสุดระยะเวลาที่ศึกษา เป็นต้น



รูปที่ 2.1 แสดงแผนภาพของการทดลอง

นาย ก ซื้อกรมธรรม์ประกันชีวิตแบบสะสมทรัพย์ 10 ปี ชำระเบี้ยประกันภัย 10 ปี ในปี 2535 และ ณ เวลาที่เริ่มต้นศึกษาอัตราการอยู่รอดของกรมธรรม์(ปี 2540) กรมธรรม์ของนาย ก ยังมีผลบังคับอยู่ และนาย ก เสียชีวิตในปี 2543 ค่าสังเกตนี้เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

นาย ข ซื้อกรมธรรม์ประกันชีวิตแบบสะสมทรัพย์ 10 ปี ชำระเบี้ยประกันภัย 10 ปี ในปี 2542 และกรมธรรม์ยังมีผลบังคับอยู่เมื่อสิ้นสุดระยะเวลาที่ศึกษา(ปี 2545) ค่าสังเกตนี้เป็นค่าสังเกตที่ถูกตัดทิ้ง

นาย ค ซื้อกรมธรรม์ประกันชีวิตแบบสะสมทรัพย์ 10 ปี ชำระเบี้ยประกันภัย 10 ปี ในปี 2541 และแปลงกรมธรรม์เป็นแบบตลอดชีพเมื่อปี 2544 ค่าสังเกตนี้เป็นค่าสังเกตที่ถูกตัดทิ้ง

ถ้า $T_1, ..., T_N$ เป็นตัวแปรสุ่มของค่าสังเกตที่ไม่ถูกตัดทิ้งที่มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง F และฟังก์ชันการอยู่รอด S=1-F และ $C_1, ..., C_N$ เป็นตัวแปรสุ่มของค่าสังเกตที่ถูกตัดทิ้งมีการแจกแจงที่เหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง G และฟังก์ชันการอยู่รอด 1-G

ดังนั้น T_i และ C_i ; i = 1,2,...,N เป็นอิสระกัน จากการตัดทิ้งแบบสุ่มได้นิยามให้ $Y_i = \min(T_i, C_i)$ ดังนั้น จะได้ค่าสังเกตุสุ่ม $Y_1, ..., Y_N$ ดังนี้

$$Y_i = \begin{cases} T_i \; ; \; T_i \leq C_i \\ \\ C_i \; ; \; T_i > C_i \end{cases}$$

$$\delta_i = \begin{cases} 1 \; ; \; T_i \leq C_i \\ \\ 0 \; ; \; T_i > C_i \end{cases}$$

$\delta_i = 1$ เมื่อค่าสังเกตไม่ถูกตัดทิ้ง

$\delta_i = 0$ เมื่อค่าสังเกตถูกตัดทิ้ง

จะมีฟังก์ชันภาวะน่าจะเป็นดังนี้

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

และมีฟังก์ชันภาวะน่าจะเป็นรวมดังนี้

$$L = \prod_{i \varepsilon u} f(y_i) \prod_{i \varepsilon c} S(y_i) \prod_{i \varepsilon c} g(y_i) \prod_{i \varepsilon u} (1-G(y_i))$$

เนื่องจาก $G(y_i)$ และ $g(y_i)$ ไม่เกี่ยวข้องกับพารามิเตอร์ที่สนใจ จึงละไว้โดยใช้ฟังก์ชันภาวะน่าจะเป็น ดังนี้

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

เมื่อ $i \varepsilon u$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าไม่ถูกตัดทิ้ง

$i \varepsilon c$ หมายถึง เซตของตัวแปรสุ่มที่มีค่าถูกตัดทิ้ง

2.1.2 ฟังก์ชันการอยู่รอดและฟังก์ชันการสูญเสีย (Survival Function and Hazard Function)

ให้ T เป็นตัวแปรสุ่มต่อเนื่อง

$f(t)$ เป็นฟังก์ชันความหนาแน่นของ T (Probability density function)

$F(t)$ เป็นฟังก์ชันการแจกแจงสะสมของ T (Distribution Function)

$S(t)$ เป็นฟังก์ชันการอยู่รอดของ T (Survival Function)

$h(t)$ เป็นฟังก์ชันความสูญเสียหรืออัตราการสูญเสีย (Hazard Function or Hazard Rate)

นิยามฟังก์ชัน $S(t)$ คือความน่าจะเป็นที่ตัวแปรสุ่ม T จะมีค่ามากกว่าหรือเท่ากับ t

$$S(t) = \Pr(T > t)$$
$$= 1-F(t)$$

โดยที่ $S(t)$ มีคุณสมบัติดังนี้

1. $S(t)$ เป็นฟังก์ชันไม่เพิ่ม (Nonincreasing Function)
2. $S(t)$ เป็นฟังก์ชันต่อเนื่องของ t
3. $S(t) = 1$ เมื่อ $t = 0$
4. $S(t) = 0$ เมื่อ $t = \infty$

นิยามฟังก์ชัน $h(t)$ แทนฟังก์ชันการสูญเสียมีค่าเท่ากับลิมิตของความน่าจะเป็นที่ตัวแปรสุ่ม T จะมีค่าอยู่ในช่วงเวลาสั้นๆ $(t,t+\Delta t)$ ต่อหน่วยเวลา Δt เมื่อกำหนดค่า $T > t$ และ $h(t)$ กำหนดได้ดังนี้

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < (t+\Delta t) / T \ge t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{(1-F(t))\Delta t}$$

$$= \frac{d}{dt}F(t)\frac{1}{1-F(t)}$$

$$= \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} \quad , t > 0$$

กรณีเมื่อ T เป็นตัวแปรสุ่มที่ไม่ต่อเนื่อง และมีค่าเป็น $t_1, t_2, t_3, \ldots$ โดยที่ $0 \leq t_1 < t_2 < t_3 < \ldots$ ดังนั้นฟังก์ชันความน่าจะเป็น $p(t_j)$ (Probability Function) กำหนดได้ดังนี้

$$P(t_j) = P(T = t_j) \quad ; \quad j = 1,2,3,\ldots$$

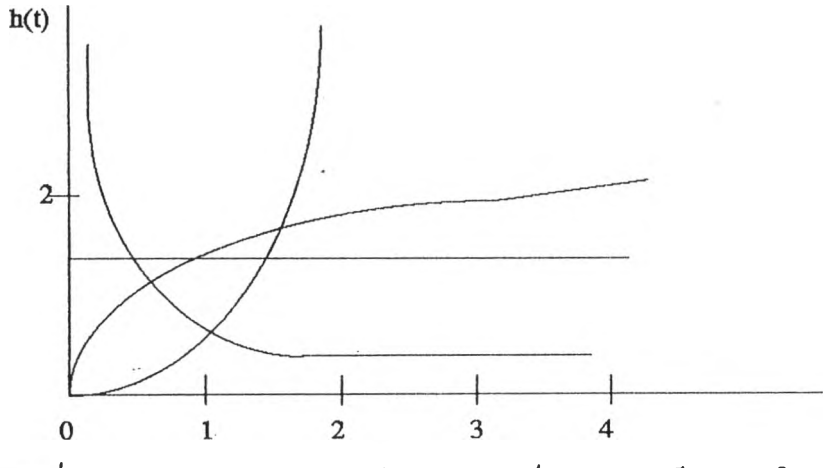
และฟังก์ชันการอยู่รอด S(t) คือ

$$S(t) = P(T \geq t_j) = \sum_{j \geq t} P(t_j)$$

ดังนั้นฟังก์ชันความสูญเสีย h(t) แสดงได้ดังนี้

$$h(t) = P(T = t_j \,/\, T \geq t_j)$$

$$= \frac{P(t_j)}{S(t_j)}$$



รูปที่ 2.2 แสดงตัวอย่างลักษณะต่างๆ ของฟังก์ชันการสูญเสีย $h(t)$ ในรูปแบบต่างๆ

สามารถเขียนความสัมพันธ์ระหว่างฟังก์ชันการอยู่รอด ฟังก์ชันความหนาแน่น ฟังก์ชันการแจกแจง และฟังก์ชันความสูญเสียได้ดังนี้

1. $f(t) = F'(t) = -S'(t)$

2. $h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$

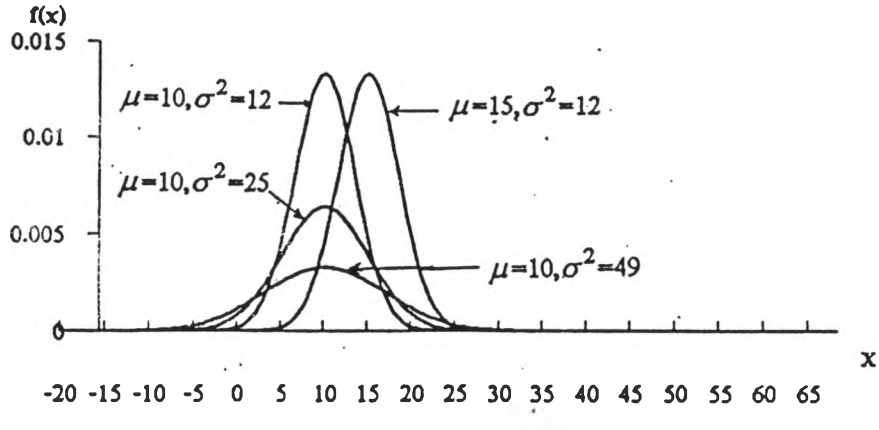
3. $H(t) = \int_0^t h(u)du = -\ln S(t)$

หรือ $S(t) = \exp(-\int_0^t h(u)du) = \exp(-H(t))$

2.1.3 ฟังก์ชันความหนาแน่น ค่าคาดหวัง และความแปรปรวน ของการแจกแจงที่ทำการศึกษาครั้งนี้

2.1.3.1 การแจกแจงปกติ (Normal Distribution)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad ; \quad -\infty < x < \infty$$

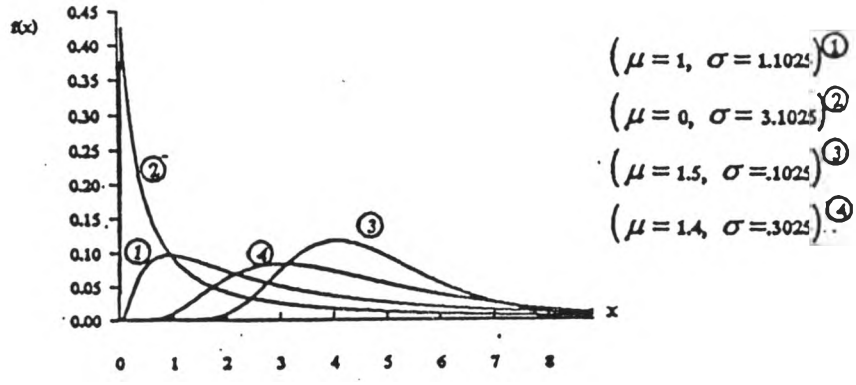


$$E(x) = \mu$$
$$V(x) = \sigma^2$$

รูปที่ 2.3 แสดงการแจกแจงแบบปกติ

2.1.3.2 การแจกแจงแบบลอกนอร์มอล (Lognormal Distribution)

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[\frac{-(\ln(x)-\mu)^2}{2\sigma^2}\right] \quad ; \quad x>0,-\infty<\mu<\infty,\sigma>0$$



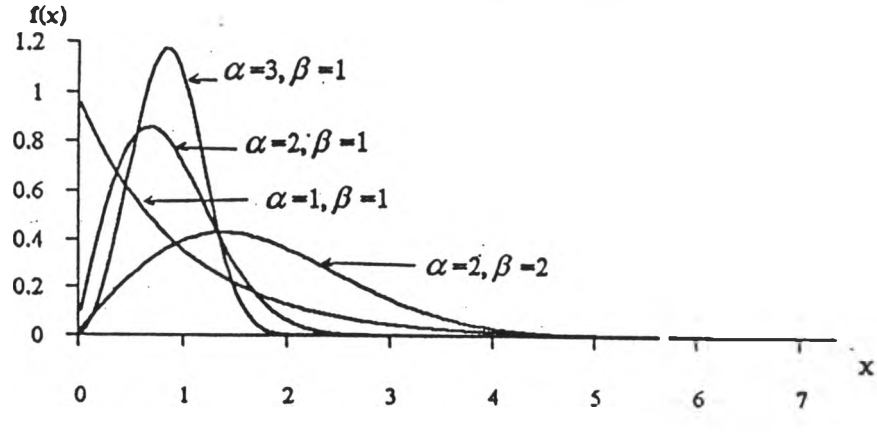
$$E(x) = \exp\left[\mu + (\sigma^2/2)\right]$$

$$V(x) = \exp(2\mu+\sigma^2)\left[\exp(\sigma^2)-1\right]$$

รูปที่ 2.4 แสดงการแจกแจงแบบลอกนอร์มอล

2.1.3.3 การแจกแจงแบบไวบูลล์ (Weibull Distribution)

$$f(x)= \quad \alpha\,\beta^{-\alpha}x^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \quad ; \quad x>0$$



$$E(x) = \frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$$

$$V(x) = \frac{\beta^2}{\alpha}\left\{2\Gamma\left(\frac{2}{\alpha}\right)-\left(\frac{1}{\alpha}\right)\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^2\right\}$$

รูปที่ 2.5 แสดงการแจกแจงแบบไวบูลล์

2.1.4 ตัวประมาณพีแอล (Product Limit Estimator:PL Estimator) [1]

ในการวิจัยครั้งนี้มีวิธีการประมาณค่าพารามิเตอร์ 3 วิธี คือ วิธีตัวประมาณของมิลเลอร์, วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ และวิธีการของบัคเลย์และเจมส์ ที่ใช้ตัวประมาณพีแอลในการประมาณฟังก์ชันการอยู่รอด และตัวประมาณพีแอลเป็นวิธีการประมาณฟังก์ชันการอยู่รอดที่เป็นแบบนอนพาราเมตริกซ์(Nonparametric) เมื่อข้อมูลบางส่วนถูกตัดทิ้ง ซึ่งตัวประมาณนี้ถูกพัฒนาขึ้นโดย แค็พแลนและไมเออร์(Kaplan and Meier, 1958) และบางครั้งอาจเรียกว่า ตัวประมาณของแค็พแลน-ไมเออร์ แนวความคิดของฟังก์ชันการอยู่รอด t_k ปี $S(t_k)$ คือความน่าจะเป็นที่ระยะเวลาของอายุที่อยู่รอด (Survival Time) มากกว่า t_k ปี ดังนั้น จะได้

$$\begin{aligned} S(t_k) &= P(T>t_k) \\ &= P(T>t_1)P(T>t_2/T>t_1)\ldots P(T>t_k/T>t_{k-1}) \\ &= p_1 p_2 \cdots p_k \end{aligned}$$

เมื่อ p_i เป็นความน่าจะเป็นที่จะอยู่รอด t_i ปี หลังจากมีชีวิตอยู่รอดมาแล้ว t_{i-1} ปี

$$p_i = P\left(T > t_i / T > t_{i-1}\right)$$

ฟังก์ชันการอยู่รอดมีค่าที่ $S(t_0) = 1$ และ $S(t_N) = 0$ และเป็นฟังก์ชันขั้นบันได

สมมติว่า ค่าสังเกตของอายุที่อยู่รอด (Survival Time) จำนวน N ตัวอย่าง มีค่าสังเกตเป็น $y_1, y_2, \ldots, y_N$ นำค่าสังเกตมาเรียงลำดับค่าจากน้อยไปมากจะได้ $y(1) < y(2) < \ldots < y(N)$
ดังนั้นหาค่าประมาณ $\hat{p}_i$ โดย

---

[1] Kaplan E.L. and Meier P., "Nonparametric Estimation from Incomplete Observation" Journal of the American Statistical Association, 53 (June 1958) : 457-481.

ให้ $\hat{q}_i = \frac{1}{n_i}$

$$\hat{p}_i = 1-\hat{q}_i = \begin{cases} 1-\left(\frac{1}{n_i}\right) & ; \delta_{(i)}=1 \\ 1 & ; \delta_{(i)}=0 \end{cases}$$

$\delta_{(i)} = 1$ เมื่อข้อมูลลำดับที่ i ไม่ถูกตัดทิ้ง

$\delta_{(i)} = 0$ เมื่อข้อมูลลำดับที่ i ถูกตัดทิ้ง

ตัวประมาณพีแอลสามารถแสดงได้ดังนี้

$$\hat{S}(t) = \prod_{y_{(i)} \le t}\left[\frac{(N-i)}{(N-i+1)}\right]^{\delta_{(i)}}$$

และตัวประมาณพีแอลเมื่อมีจำนวนค่าสังเกตที่ถูกตัด ณ y(i) เท่ากับ d_i ค่า สามารถแสดงได้ดังนี้

$$\hat{S}(t) = \prod_{y_{(i)} \le t}\left[1-\frac{d_i}{n_i}\right]^{\delta_{(i)}}$$

เมื่อ $y_{(i)}$ เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง ลำดับที่ i

i คือลำดับที่ของข้อมูล

n_i เป็นจำนวนค่าสังเกตที่อยู่รอด ณ เวลา y(i)

d_i เป็นจำนวนค่าสังเกตที่เสียชีวิต ณ เวลา y(i)

N เป็นจำนวนข้อมูลทั้งหมดที่ไม่ถูกตัดทิ้ง และที่ถูกตัดทิ้ง

ตัวอย่างการหาตัวประมาณพีแอล จากค่าสังเกตต่อไปนี้ 3.0, 4.0^+, 5.7^+, 6.5, 6.5, 8.4^+, 10.0, 10.0^+, 12.0, 15.0 สามารถแสดงการหา $\hat{S}(t)$ ได้ดังนี้

t	Rank(i)	δ_i	$[(N-i)/(N-i+1)]^{\delta_i}$	$\hat{S}(t)$
3.0	1	1	9/10	= 0.9
4.0^+	2	0	1	= 0.9
5.7^+	3	0	1	= 0.9
6.5	4	1	6/7	$\hat{S}(3)$ (6/7) = 0.771
6.5	5	1	5/6	$\hat{S}(6.5)$ (5/6) = 0.643
8.4^+	6	0	1	= 0.643
10.0	7	1	3/4	$\hat{S}(6.5)$ (3/4) = 0.482
10.0^+	8	0	-	= 0.482
12.0	9	1	1/2	$\hat{S}(10)$ (1/2) = 0.241
15.0	10	1	0	$\hat{S}(12)$ (0) = 0.0

+ หมายถึงข้อมูลค่าสังเกตที่ถูกตัดทิ้ง

2.2 การประมาณค่าพารามิเตอร์

ในที่นี้จะพิจารณาการถดถอยเชิงเส้นอย่างง่าย รูปแบบ คือ

$$T_i = \alpha + \beta x_i + \varepsilon_i \quad ; i = 1,2,3,...,N$$

เมื่อ ε_i เป็นตัวแปรสุ่มของค่าคลาดเคลื่อนที่เป็นอิสระกัน และมีฟังก์ชันการแจกแจง F ที่ไม่มีรูปแบบเฉพาะที่มีค่าเฉลี่ยเป็นศูนย์และมีความแปรปรวนจำกัด และมีฟังก์ชันการอยู่รอดเป็น S = 1- F

α,β เป็นพารามิเตอร์ที่ต้องการประมาณค่า

X_i เป็นตัวแปรอิสระและเป็นค่าคงที่

Y_i เป็นตัวแปรตามและตัวแปรตามบางส่วนเป็นค่าที่ถูกตัดทิ้ง

N เป็นขนาดตัวอย่างทั้งหมด

ตัวแปรตามเป็นค่าที่ถูกตัดทิ้ง ดังนั้นจะได้ค่าสังเกตเป็น (Y_i, δ_i, X_i) ; i = 1,2,3, ... ,N

โดยที่

$$Y_i = \min(T_i, T_C) \quad ; i = 1,2,3,...,N$$

$$\delta_i = \begin{cases} 1 & ; T_i \leq T_c \\ 0 & ; T_i > T_c \end{cases}$$

$\delta_i = 1$ เมื่อข้อมูลไม่ถูกตัดทิ้ง

$\delta_i = 0$ เมื่อข้อมูลถูกตัดทิ้ง

T_c = ค่าสูงสุดของการตัดทิ้งที่กำหนดไว้ล่วงหน้า

และประมาณค่าพารามิเตอร์ 4 วิธี ดังนี้

2.2.1 วิธีกำลังสองต่ำสุด

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

1. ค่าความคลาดเคลื่อนจะต้องมีการแจกแจงแบบปกติ ที่มีค่าเฉลี่ยเป็นศูนย์และมีค่าความแปรปรวนเป็น σ^2

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

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

---

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

ดังนั้น ตัวประมาณพารามิเตอร์โดยวิธีกำลังสองต่ำสุด จะเป็นตัวประมาณเชิงเส้นที่ไม่เอนเอียงและมีความแปรความแปรปรวนต่ำสุดเรียกคุณสมบัตินี้ว่า BLUE (Best Linear Unbiased Estimator) แต่การศึกษาครั้งนี้ เนื่องจากข้อมูลที่นำมาวิเคราะห์มีค่าของตัวแปรตามถูกตัดทิ้งทางขวา ดังนั้น วิธีกำสองต่ำสุดจะทำให้ได้ตัวประมาณที่เอนเอียง และโดยเฉลี่ยการประมาณค่าจะต่ำกว่าความเป็นจริง วิธีกำลังสองต่ำสุดไม่ได้กระทำวนซ้ำ และข้อมูลที่นำมาวิเคราะห์ จะถือว่าค่าสังเกตที่ถูกตัดทิ้งเสมือนเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

การหาตัวประมาณกำลังสองต่ำสุด

จากสมการความสัมพันธ์ระหว่างตัวแปรตาม Y และตัวแปรอิสระ X คือ

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

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

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

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

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

พิจารณาผลบวกกำลังสองของค่าความคลาดเคลื่อน (Sum of Squared Error) จะพบว่า

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

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

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

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

2.2.2 วิธีตัวประมาณของมิลเลอร์ [3]

มิลเลอร์ (Miller, R.G.) ได้ทำการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นเมื่อตัวแปรตามบางค่าถูกตัดทิ้ง โดยแนะนำให้ประมาณค่าพารามิเตอร์ α,β ด้วยค่า $\hat{\alpha},\hat{\beta}$ ที่ทำให้ผลบวกกำลังสองของค่าคลาดเคลื่อนมีค่าน้อยสุด เนื่องจากค่าคลาดเคลื่อนมีการแจกแจงที่ไม่มีรูปแบบเฉพาะมีค่าเฉลี่ยเป็นศูนย์และมีความแปรปรวนจำกัด ดังนั้นจึงทำการประมาณฟังก์ชันการแจกแจงของค่าคลาดเคลื่อนด้วยตัวประมาณพีแอล $\hat{F}(e_i,\hat{\beta})$ โดยที่

$$\hat{F}(e_i,\hat{\beta}) = 1 - \prod_{i\,;e(i)\le e_i}^{N}\left[\frac{(N-i)}{(N-i+1)}\right]^{\delta_i}$$

และค่าคลาดเคลื่อน $e_i = y_i - \hat{\beta}x_i \quad ;i=1,2,3,...,N$ โดยกำหนดให้ $\hat{\alpha}$ มีค่าเป็นศูนย์ และหลังจากได้ค่าประมาณ $\hat{\beta}$ แล้วจึงทำการประมาณค่า $\hat{\alpha}$ ในกรณีที่ไม่มีค่าที่ถูกตัดทิ้ง วิธีนี้ก็คือวิธีกำลังสองต่ำสุดธรรมดา (Ordinary Least Square Method) ค่า $\hat{F}(e_i,\hat{\beta})$ จะเป็นค่าที่ไม่ต่อเนื่อง โดยจะมีค่าเปลี่ยนไปเมื่อเป็นข้อมูลที่ไม่ถูกตัดทิ้ง และถ้าค่าคลาดเคลื่อนที่มากที่สุดเป็นค่าที่ถูกตัดทิ้งจะทำให้ค่า $\hat{F}(e_i,\hat{\beta})$ ไม่ลู่เข้าสู่ค่า 1.0 ดังนั้นให้ทำการกำหนดเปลี่ยนให้ค่าคลาดเคลื่อนที่มากที่สุดนั้นเป็นค่าที่ไม่ถูกตัดทิ้ง ตัวอย่างการคำนวณหาค่าประมาณพารามิเตอร์โดยวิธีตัวประมาณของมิลเลอร์นี้แสดงในภาคผนวก ค การประมาณค่าพารามิเตอร์ $\hat{\beta}$ นี้จะใช้กระบวนการกระทำซ้ำ(Iterative) เพื่อหาค่า $\hat{\beta}$ ที่ทำให้

$$\hat{\beta} = \phi(\hat{\beta})$$

---

[3] Rupert Miller and Jerry Halpern , Regression with Censored Data. Biometrica. (1982),69,3, pp. 521-531.

$$\text{เมื่อ} \quad \phi(\hat{\beta}) = \frac{\sum_u w_i(\hat{\beta})\, y_i (x_i - \bar{x}^w)}{\sum_u w_i(\hat{\beta})(x_i - \bar{x}^w)^2}$$

โดยที่ $\sum_u$ คือ ผลรวมของข้อมูลเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

$w_i(\hat{\beta})$ คือ ฟังก์ชันความน่าจะเป็นของค่าคลาดเคลื่อน ซึ่งเท่ากับ

$$\hat{F}_i(e_i, \hat{\beta}) - \hat{F}_{i-1}(e_{i-1}, \hat{\beta})$$

$$\bar{x}^w = \sum_u w_i(\hat{\beta})\, x_i$$

การประมาณค่าพารามิเตอร์โดยวิธีนี้อาจจะต้องกระทำซ้ำมากกว่า 1 ครั้ง หรือค่าพารามิเตอร์อาจแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นๆ เป็นค่าประมาณพารามิเตอร์ หรือบางครั้งค่าพารามิเตอร์ในรอบปัจจุบันกับรอบที่ผ่านมาอาจมีค่าต่างกันไม่เกิน 0.001 ในกรณีให้ใช้ค่าประมาณพารามิเตอร์ในรอบปัจจุบัน หรือบางครั้งอาจจะไม่มีคำตอบก็ได้ หลังจากได้ค่าประมาณพารามิเตอร์ $\hat{\beta}$ แล้ว จะทำการประมาณค่า $\hat{\alpha}$ จาก

$$\hat{\alpha} = \sum_u w_i(\hat{\beta})(y_i - \hat{\beta} x_i)$$

ขั้นตอนการหาค่าประมาณพารามิเตอร์ สำหรับวิธีตัวประมาณของมิลเลอร์

ขั้นที่ 1 หาค่าประมาณ $\hat{\beta}$ เริ่มต้น โดยการใช้วิธีกำลังสองต่ำสุดกับเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

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

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

ขั้นที่ 2 จากข้อมูลทั้งหมดคือรวมทั้งข้อมูลที่ถูกตัดทิ้งและข้อมูลที่ไม่ถูกตัดทิ้ง ให้หาความคลาดเคลื่อน e_i โดยกำหนดให้ $\hat{\alpha}$ เป็นศูนย์ จะได้ว่า

$$e_i = y_i - \hat{\beta} x_i \quad ; i = 1,2,3,...,N$$

ให้เรียงลำดับค่าคลาดเคลื่อนจากน้อยไปหามาก จะได้ $e(1) < e(2) < ... < e(N)$ และในการเรียงลำดับให้นำ y_i , x_i , และ δ_i ที่สอดคล้องกับ e_i ตามไปด้วย ในกรณีที่ลำดับที่ของค่าคลาดเคลื่อนของข้อมูลที่ถูกตัดทิ้งและที่ไม่ถูกตัดทิ้งมีค่าเท่ากัน ให้ลำดับที่ของข้อมูลที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของข้อมูลที่ถูกตัดทิ้ง

ขั้นที่ 3 ถ้าค่าคลาดเคลื่อนที่มากที่สุด เป็นข้อมูลที่ถูกตัดทิ้ง ให้ทำการเปลี่ยนให้ค่าคลาดเคลื่อนที่มากที่สุดนั้นเป็นข้อมูลที่ไม่ถูกตัดทิ้งเพื่อทำการปรับให้ $\hat{F}(e_i, \hat{\beta})$ ของค่าคลาดเคลื่อนที่มากที่สุด(e(N)) มีค่าเท่ากับ 1.0

ขั้นที่ 4 หาค่าฟังก์ชันการอยู่รอดด้วยตัวประมาณพีแอล(PL Estimator) จาก

$$\hat{S}_i(e_i, \hat{\beta}) = \prod_{i\,;\,e(i) \le e_i} \left[\frac{(N-i)}{(N-i+1)} \right]^{\delta_i}$$

เมื่อ i คือ ลำดับที่ของความคลาดเคลื่อน

N คือ จำนวนข้อมูลทั้งหมด

$\hat{S}_i(e_i, \hat{\beta})$ คือ ตัวประมาณพีแอล หรือ ตัวประมาณแค็พแลน-ไมเออร์

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

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

ขั้นที่ 6 คำนวณหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$ จาก

$$w_1(\hat{\beta}) = \hat{F}_1(e_1, \hat{\beta})$$
$$w_i(\hat{\beta}) = \hat{F}_i(e_i, \hat{\beta}) - \hat{F}_{i-1}(e_{i-1}, \hat{\beta}) \quad ; \quad i = 2,3,...,N$$

ขั้นที่ 7 นำค่า $w_i(\hat{\beta})$ มาหาค่า $\bar{x}^w$ จาก

$$\bar{x}^w = \Sigma_u w_i(\hat{\beta})\, x_i$$

ขั้นที่ 8 คำนวณค่า $\hat{\beta}$ จาก

$$\hat{\beta} = \frac{\sum_u w_i(\hat{\beta})\, y_i (x_i - \bar{x}^w)}{\sum_u w_i(\hat{\beta})\, (x_i - \bar{x}^w)^2}$$

ขั้นที่ 9 แทนค่า $\hat{\beta}$ จากขั้นที่ 8 ลงในขั้นที่ 2 แล้วทำการคำนวณวนซ้ำจากขั้นตอนที่ 2 จนถึงขั้นตอนที่ 8 จนกระทั่งค่า $\hat{\beta}$ ที่ได้เท่ากับค่า $\hat{\beta}$ ในรอบที่ผ่านมา หรือมีผลต่างกันไม่เกิน 0.001 จะหยุด ก็จะได้ค่าประมาณของ $\hat{\beta}$ ในบางครั้งค่าประมาณ $\hat{\beta}$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณ $\hat{\beta}$

ขั้นที่ 10 คำนวณ $\hat{\alpha}$ จาก

$$\hat{\alpha} = \Sigma_u w_i(\hat{\beta})\, (y_i - \hat{\beta} x_i)$$

ขั้นที่ 11 นำค่า $\hat{\alpha}$ และ $\hat{\beta}$ จากขั้นตอนที่ 10 และขั้นตอนที่ 9 ตามลำดับ มาหาค่าประมาณของตัวแปรตาม

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

2.2.3 วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ [4]

การประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นเมื่อตัวแปรตามบางค่าถูกตัดทิ้งทางขวาด้วยวิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ เสนอโดย มิลเลอร์ (Miller , R.G.) เป็นวิธีที่ใช้ตัวประมาณของแค็พแลน-ไมเออร์ เป็นค่าถ่วงน้ำหนัก และเป็นวิธีที่ทำให้ผลบวกกำลังสองของการถ่วงน้ำหนักมีค่าน้อยที่สุด การหาคำตอบจะใช้วิธีการกระทำซ้ำ การประมาณค่าพารามิเตอร์เริ่มต้นมีหลายวิธี มิลเลอร์ได้แนะนำให้ใช้วิธีกำลังสองต่ำสุดกับข้อมูลที่ไม่ถูกตัดทิ้งดังนี้

---

[4] Rupert G. Miller, Least Square Regression with Censored Data. Biometrica(1976), 63,3, pp. 449-464

$$\hat{\beta}_0 = \frac{\sum_u y_i (x_i - \bar{x}^u)}{\sum_u (x_i - \bar{x}^u)^2}$$

เมื่อ $\sum_u$ คือ ผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือ ค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

ค่าถ่วงน้ำหนัก $w_i(\hat{\beta}_0)$ คำนวณโดยใช้ตัวประมาณของแค็พแลน-ไมเออร์ กระทำกับข้อมูลค่าคลาดเคลื่อน $e_i = y_i - \hat{\beta}_0 x_i \quad ; i = 1,2,3,...,N$ ตัวประมาณพารามิเตอร์สำหรับวิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ คือค่า $\hat{\alpha}, \hat{\beta}$ ที่ทำให้สมการข้างล่างนี้มีค่าน้อยที่สุด

$$\Sigma_u w_i(\hat{\beta}_0)(y_i - \hat{\beta}_0 x_i)^2$$

ในกรณีที่ค่าคลาดเคลื่อนที่มากที่สุดเป็นข้อมูลที่ถูกตัดทิ้งให้ปรับค่าถ่วงน้ำหนักเป็น

$$w_i'(\hat{\beta}_0) = \frac{w_i(\hat{\beta}_0)}{\sum_u w_j(\hat{\beta}_0)}$$

และค่าประมาณพารามิเตอร์ $\hat{\beta}$ รอบต่อไปคือ

$$\hat{\beta} = \frac{\sum_u w_i'(\hat{\beta}_0)\, y_i (x_i - \bar{x}^k)}{\sum_u w_i'(\hat{\beta}_0)\, (x_i - \bar{x}^k)^2}$$

การประมาณค่าพารามิเตอร์ $\hat{\beta}$ จะใช้การกระทำซ้ำ (Iteration) จนกระทั่งค่าพารามิเตอร์ที่ได้เท่ากับค่าในรอบที่ผ่านมาจึงจะหยุด ในบางครั้งค่าประมาณพารามิเตอร์จะแกว่งระหว่าง 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณ $\hat{\beta}$ ตัวอย่างการคำนวณหาค่าประมาณพารามิเตอร์โดยวิธีนี้อยู่ในภาคผนวก ค หลังจากได้ค่าประมาณ $\hat{\beta}$ แล้วจะประมาณค่า $\hat{\alpha}$ จากสมการ

$$\hat{\alpha} = \Sigma_u w_i'(\hat{\beta})(y_i - \hat{\beta} x_i)$$

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^k$ คือค่าเฉลี่ยแบบถ่วงน้ำหนักของค่าสังเกต x ซึ่งเท่ากับ $\sum_u w_i'(\hat{\beta}_0) x_i$

ขั้นตอนการหาค่าประมาณพารามิเตอร์ สำหรับวิธีกำลังสองต่ำสุดแบบตัดแปลงแค็พแลน-ไมเออร์

ขั้นที่ 1 หาค่าประมาณ $\hat{\beta}$ เริ่มต้น โดยการใช้วิธีกำลังสองต่ำสุดกับเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

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

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

ขั้นที่ 2 จากข้อมูลทั้งหมดคือรวมทั้งข้อมูลที่ถูกตัดทิ้งและข้อมูลที่ไม่ถูกตัดทิ้ง ให้หาความคลาดเคลื่อน e_i โดยกำหนดให้ $\hat{\alpha}$ เป็นศูนย์ จะได้ว่า

$$e_i = y_i - \hat{\beta} x_i \quad ; i = 1,2,3,...,N$$

ให้เรียงลำดับค่าคลาดเคลื่อนจากน้อยไปหามาก จะได้ $e(1) < e(2) < ... < e(N)$ และในการเรียงลำดับให้นำ y_i , x_i , และ δ_i ที่สอดคล้องกับ e_i ตามไปด้วย ในกรณีที่ลำดับที่ของค่าคลาดเคลื่อนของข้อมูลที่ถูกตัดทิ้งและที่ไม่ถูกตัดทิ้งมีค่าเท่ากัน ให้ลำดับที่ของข้อมูลที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของข้อมูลที่ถูกตัดทิ้ง

ขั้นที่ 3 หาค่าฟังก์ชันการอยู่รอดด้วยตัวประมาณพีแอล(PL Estimator) จาก

$$\hat{S}_i(e_i, \hat{\beta}) = \prod_{i\,;\,e(i) \le e_i} \left[\frac{(N-i)}{(N-i+1)} \right]^{\delta_i}$$

เมื่อ i คือ ลำดับที่ของความคลาดเคลื่อน

N คือ จำนวนข้อมูลทั้งหมด

$\hat{S}_i(e_i, \hat{\beta})$ คือ ตัวประมาณพีแอล หรือ ตัวประมาณแค็พแลน-ไมเออร์

ขั้นที่ 4 คำนวณหาฟังก์ชัน $\hat{F}_i(e_i, \hat{\beta})$ จาก

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

ขั้นที่ 5 คำนวณหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$ จาก

$$w_1(\hat{\beta}) = \hat{F}_1(e_1, \hat{\beta})$$
$$w_i(\hat{\beta}) = \hat{F}_i(e_i, \hat{\beta}) - \hat{F}_{i-1}(e_{i-1}, \hat{\beta}) \quad ; \quad i = 2,3,\ldots,N$$

ในกรณีที่ค่าคลาดเคลื่อนที่มากที่สุดเป็นข้อมูลที่ถูกตัดทิ้งให้ปรับค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$ เป็น

$$w_i'(\hat{\beta}) = \frac{w_i(\hat{\beta})}{\sum_u w_j(\hat{\beta})}$$

ขั้นที่ 6 นำค่า $w_i'(\hat{\beta})$ มาหาค่า $\bar{x}^k$ จาก

$$\bar{x}^k = \sum_u w'_i(\hat{\beta})\, x_i$$

ขั้นที่ 7 คำนวณค่า $\hat{\beta}$ จาก

$$\hat{\beta} = \frac{\sum_u w_i'(\hat{\beta})\, y_i (x_i - \bar{x}^k)}{\sum_u w_i'(\hat{\beta})\, (x_i - \bar{x}^k)^2}$$

ขั้นที่ 8 แทนค่า $\hat{\beta}$ จากขั้นที่ 7 ลงในขั้นที่ 2 แล้วทำการคำนวณวนซ้ำจากขั้นตอนที่ 2 จนถึงขั้นตอนที่ 7 จนกระทั่งค่า $\hat{\beta}$ ที่ได้เท่ากับค่า $\hat{\beta}$ ในรอบที่ผ่านมา หรือมีผลต่างกันไม่เกิน 0.001 จะหยุด ก็จะได้ค่าประมาณของ $\hat{\beta}$ ในบางครั้งค่าประมาณ $\hat{\beta}$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณ $\hat{\beta}$

ขั้นที่ 9 คำนวณ $\hat{\alpha}$ จาก

$$\hat{\alpha} = \sum_u w'_i(\hat{\beta})\, (y_i - \hat{\beta} x_i)$$

ขั้นที่ 10 นำค่า $\hat{\alpha}$ และ $\hat{\beta}$ จากขั้นตอนที่ 9 และขั้นตอนที่ 8 ตามลำดับ มาหาค่าประมาณของตัวแปรตาม

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

2.2.4 วิธีการของบัคเลย์และเจมส์ [5]

วิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นหลายวิธีมาจากรากฐานของ Least Square Normal Equations รวมทั้งวิธีการของบัคเลย์และเจมส์ซึ่งเสนอโดยโจนาธานบัคเลย์ และเอียนเจมส์ (Jonathan Buckley and Ian James) วิธีการของบัคเลย์และเจมส์เป็นวิธีประมาณค่าพารามิเตอร์แบบนอนพาราเมตริกซ์ในสมการถดถอยเชิงเส้น เมื่อตัวแปรตามบางส่วนมีค่าถูกตัดทิ้ง โดยวิธีนี้มีการประมาณค่าของข้อมูลที่ถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไขทำการประมาณค่า $E[y_i / y_i > T_c, x_i, \hat{\beta}]$ เมื่อ T_c เป็นค่าสูงสุดที่ถูกตัดทิ้ง แต่เนื่องจากไม่ทราบฟังก์ชันการแจกแจง F จึงไม่สามารถหาค่า $E[y_i / y_i > T_c, x_i, \hat{\beta}]$ ได้จึงประมาณค่า F ด้วยตัวประมาณพีแอล ตัวอย่างการคำนวณหาค่าประมาณพารามิเตอร์โดยวิธีการของบัคเลย์และเจมส์นี้แสดงในภาคผนวก ค

ดังนั้นข้อมูลที่ถูกตัดทิ้งจะถูกแทนด้วย

$$\bar{Y}_i(\hat{\beta}) = \hat{\beta} x_i + \frac{\sum_{k \in u} w_k(\hat{\beta})(y_k - \hat{\beta} x_k)}{1 - \hat{F}(T_c - \hat{\beta} x_i)}$$

และประมาณค่าพารามิเตอร์ $\hat{\beta}$ ด้วยการกระทำซ้ำตามสมการ

$$\hat{\beta} = \phi(\hat{\beta})$$

โดยที่

$$\phi(\hat{\beta}) = \frac{\sum_u y_i(x_i - \bar{x}) + \sum_C \bar{y}_i(\hat{\beta})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

---

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

เมื่อ $\sum_u$ คือ ผลรวมเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

$\sum_c$ คือ ผลรวมเฉพาะค่าสังเกตที่ถูกตัดทิ้ง

หลังจากที่ได้ค่า $\hat{\beta}$ แล้วจะประมาณค่า $\hat{\alpha}$ จาก

$$\hat{\alpha} = \left\{\frac{\left[\sum_u y_i + \sum_c \bar{y}_i(\hat{\beta})\right]}{N}\right\} - \hat{\beta}\,\bar{x}$$

ขั้นตอนการหาค่าประมาณพารามิเตอร์ สำหรับวิธีการของบัคเลย์และเจมส์

ขั้นที่ 1 หาค่าประมาณ $\hat{\beta}$ เริ่มต้น โดยการใช้วิธีกำลังสองต่ำสุดกับเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

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

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

ขั้นที่ 2 จากข้อมูลทั้งหมดคือรวมทั้งข้อมูลที่ถูกตัดทิ้งและข้อมูลที่ไม่ถูกตัดทิ้ง ให้หาความคลาดเคลื่อน e_i โดยกำหนดให้ $\hat{\alpha}$ เป็นศูนย์ จะได้ว่า

$$e_i = y_i - \hat{\beta}\,x_i \quad ; i = 1,2,3,\ldots,N$$

ให้เรียงลำดับค่าคลาดเคลื่อนจากน้อยไปหามาก จะได้ $e(1) < e(2) < \ldots < e(N)$ และในการเรียงลำดับให้นำ y_i , x_i , และ δ_i ที่สอดคล้องกับ e_i ตามไปด้วย ในกรณีที่ลำดับที่ของค่าคลาดเคลื่อนของข้อมูลที่ถูกตัดทิ้งและที่ไม่ถูกตัดทิ้งมีค่าเท่ากัน ให้ลำดับที่ของข้อมูลที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของข้อมูลที่ถูกตัดทิ้ง

ขั้นที่ 3 หาค่า PL Estimator S จาก

$$\hat{S}_i(e_i,\hat{\beta}) \quad = \quad \prod_{i\,;\,e(i)\le e_i}^{N}\left[\frac{(N-i)}{(N-i+1)}\right]^{\delta_i}$$

เมื่อ i คือลำดับที่ของความคลาดเคลื่อน

N คือจำนวนข้อมูลทั้งหมด

$\hat{S}_i(e_i,\hat{\beta})$ คือ ตัวประมาณพีแอล หรือ ตัวประมาณแค็พแลน-ไมเออร์

ขั้นที่ 4 คำนวณหาฟังก์ชัน $\hat{F}_i(e_i,\hat{\beta})$ จาก

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

ขั้นที่ 5 คำนวณหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$จาก

$$w_1(\hat{\beta}) \quad = \quad \hat{F}_1(e_1,\hat{\beta})$$

$$w_i(\hat{\beta}) \quad = \quad \hat{F}_i(e_i,\hat{\beta})-\hat{F}_{i-1}(e_{i-1},\hat{\beta}) \quad ; \quad i \;=2,3,\ldots,N$$

ในกรณีที่ค่าคลาดเคลื่อนที่มากที่สุดเป็นของข้อมูลที่ถูกตัดทิ้งให้ปรับค่าถ่วงน้ำหนัก$w_i(\hat{\beta})$เป็น

$$w_i'(\hat{\beta}) \quad = \quad \frac{w_i(\hat{\beta})}{\sum_u w_j(\hat{\beta})}$$

ขั้นที่ 6 ประมาณค่าที่ถูกตัดทิ้งด้วย

$$\bar{Y}_i(\hat{\beta}) \quad = \quad \hat{\beta}\,x_i+\frac{\sum_{k\in u} w_k'(\hat{\beta})(y_k-\hat{\beta}\,x_k)}{1-\hat{F}(T_c-\hat{\beta}\,x_i)}$$

ขั้นที่ 7 คำนวณค่า $\hat{\beta}$ จาก

$$\hat{\beta} \quad = \quad \frac{\sum_u y_i(x_i-\bar{x})+\sum_C \bar{y}_i(\hat{\beta})(x_i-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

ขั้นที่ 8 แทนค่า $\hat{\beta}$ จากขั้นที่ 7 ลงในขั้นที่ 2 แล้วทำการคำนวณวนซ้ำจากขั้นตอนที่ 2 จนถึงขั้นตอนที่ 7 จนกระทั่งค่า $\hat{\beta}$ ที่ได้เท่ากับค่า $\hat{\beta}$ ในรอบที่ผ่านมา หรือมีผลต่างกันไม่เกิน 0.001 จะหยุด ก็จะได้ค่าประมาณของ $\hat{\beta}$ ในบางครั้งค่าประมาณ $\hat{\beta}$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณ $\hat{\beta}$

ขั้นที่ 9 คำนวณ $\hat{\alpha}$ จาก

$$\hat{\alpha} = \left\{ \frac{\left[\sum_u y_i + \sum_c \bar{y}_i (\hat{\beta}) \right]}{N} \right\} - \hat{\beta}\, \bar{x}$$

ขั้นที่ 10 นำค่า $\hat{\alpha}$ และ $\hat{\beta}$ จากขั้นตอนที่ 9 และขั้นตอนที่ 8 ตามลำดับ มาหาค่าประมาณของตัวแปรตาม

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i$$