บทที่ 3

วิธีนอนพาราเมตริกสำหรับการประมาณค่าฟังก์ชั่นการอยู่รอด สำหรับข้อมูลไม่สมบูรณ์

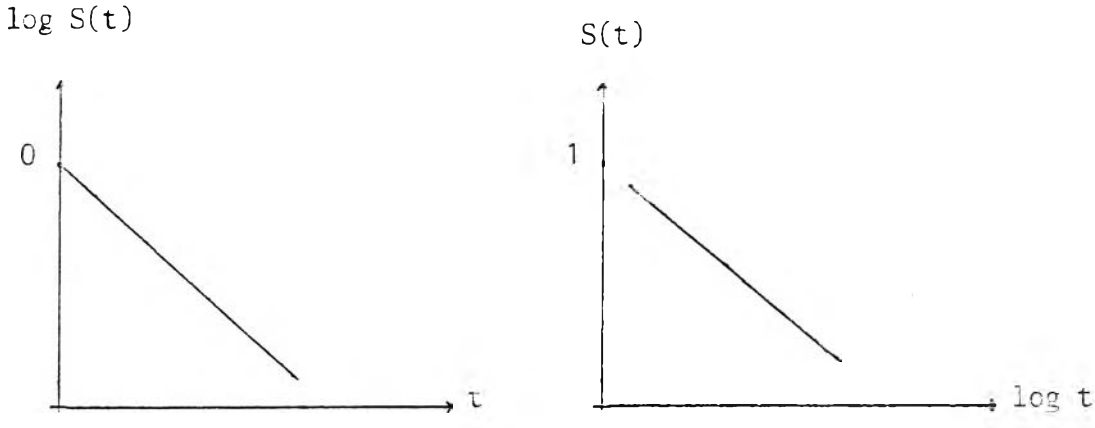
Nonparametric Methods Estimating Survival Function for Censored Data

ในการวิเคราะห์ข้อมูลโดยทั่วไปแล้ว ถ้าข้อมูลเป็นข้อมูลที่ไม่ทราบการแจกแจง หรือข้อมูลไม่เป็นไปตามข้อตกลงของวิธีพาราเมตริกแล้ว เราจะใช้วิธีนอนพาราเมตริกในการวิเคราะห์ข้อมูลนั้น ในทางปฏิบัติ วิธีนอนพาราเมตริกเป็นวิธีที่นิยมนำไปใช้กันมาก ทั้งนี้เพราะวิธีนอนพาราเมตริก สามารถเข้าใจและนำไปใช้ได้โดยง่าย นักวิจัยส่วนใหญ่จึงมักจะเสนอให้ใช้วิธีนอนพาราเมตริก ก่อนที่จะมีการพยายามในการหาทฤษฎีที่เหมาะสมกับการแจกแจง อย่างไรก็ตาม วิธีนอนพาราเมตริกจะมีประสิทธิภาพน้อยกว่าวิธีพาราเมตริก เมื่อข้อมูลมีการแจกแจงหรือมีข้อตกลงเป็นไปตามทฤษฎี และจะมีประสิทธิภาพมากกว่าเมื่อข้อมูลไม่เป็นไปตามทฤษฎีที่ทราบ ถ้ามีวัตถุประสงค์หลักในการวิเคราะห์เป็นการหารูปแบบของข้อมูล การประมาณค่าโดยวิธีนอนพาราเมตริกและวิธีกราฟ จะสามารถทำให้ทราบการแจกแจงของข้อมูลได้

วิธีนอนพาราเมตริก สำหรับการวิเคราะห์ข้อมูลการอยู่รอด ส่วนใหญ่แล้วจะพัฒนาไปใช้ในด้านการแพทย์หรือด้านชีวแพทย์ เพราะข้อมูลทางด้านนี้มักจะมีการเปลี่ยนแปลงไปตามสถานการณ์ที่แตกต่างกัน รูปแบบของการแจกแจงการอยู่รอดจึงไม่แน่นอน ส่วนการวิเคราะห์ข้อมูลโดยวิธีพาราเมตริกนั้น ส่วนใหญ่จะนำไปใช้ทางด้านวิศวกรรม และอุตสาหกรรม เพราะข้อมูลมีความถูกต้องแน่นอนกว่า รูปแบบของการแจกแจงจึงค่อนข้างจะแน่นอนด้วย

ดังนั้น สำหรับการวิเคราะห์ข้อมูลการอยู่รอด ฟังก์ชั่นการอยู่รอดหรือกราฟของฟังก์ชั่นการอยู่รอดที่เรียกว่า เส้นโค้งการอยู่รอด (survival curve) จึงเป็นสิ่งที่สำคัญและนำไปใช้ประโยชน์ได้มากที่สุด เพราะฟังก์ชั่นการอยู่รอดสามารถบอกถึงรูปแบบของการแจกแจงได้ ดัง

ตัวอย่างในรูปที่ 5 รูปที่ 5 (ก) แสดงฟังก์ชั่นการอยู่รอดของการแจกแจงแบบเอ็กโปเนนเชียล (exponential distribution) ในรูปของล็อค รูปที่ 5 (ข) แสดงฟังก์ชั่นการอยู่รอดของการแจกแจงแบบลอคนอร์มอล์ (lognormal distribution)



(ก) เอ็กโปเนนเชียล (ข) ลอคนอร์มอล์

รูปที่ 5 แสดงฟังก์ชั่นการอยู่รอดของการแจกแจงแบบเอ็กโปเนนเชียลและลอคนอร์มอล์

ในบทนี้จะกล่าวถึง วิธีนอนพาราเมตริก สำหรับการประมาณค่าฟังก์ชั่นการอยู่รอดที่ใช้กับข้อมูลที่มีค่าสังเกตไม่สมบูรณ์ เฉพาะวิธีที่สำคัญและนำไปใช้กันอย่างกว้างขวาง อันได้แก่ วิธี Product-Limit (PL), วิธี Life-table (actuarial) และวิธี Cox's regression model

ในตอนที่ 3.1 จะเป็นวิธีการประมาณค่าฟังก์ชั่นการอยู่รอด โดยวิธี PL ซึ่งเป็นวิธีที่พัฒนาขึ้นมาใช้ โดย Kaplan & Meier (1958) วิธีนี้สามารถนำไปใช้กับตัวอย่างขนาดเล็กหรือตัวอย่างขนาดกลาง หรือตัวอย่างขนาดใหญ่ก็ได้ แต่อย่างไรก็ดี เมื่อตัวอย่างมีขนาดใหญ่มาก ๆ ก็ควรใช้วิธี Life-table ในการประมาณเพราะวิธีนี้จะกำหนดข้อมูลเป็นช่วง ทำให้

สะดวกต่อการนำมาใช้ ดูรายละเอียดในตอนที่ 3.2 วิธีการประมาณค่าฟังก์ชั่นการอยู่รอดทั้ง 2 วิธีข้างต้น จะมีวิธีการประมาณค่าเหมือนกัน แต่วิธี PL จะประมาณค่าฟังก์ชั่นของแต่ละหน่วยตัวอย่าง ส่วนวิธี Life-table จะประมาณค่าฟังก์ชั่นของแต่ละช่วง ดังนั้น ถ้าแต่ละช่วงมีค่าสังเกตเพียงค่าเดียว วิธี Life-table ก็จะเหมือนกับวิธี PL นั่นคือ วิธี PL เป็นวิธีเฉพาะของวิธี Life-table เมื่อแต่ละช่วงของข้อมูลมีค่าสังเกตเพียงค่าเดียว

ตอนที่ 3.3 จะเป็นวิธีการประมาณค่าฟังก์ชั่นการอยู่รอด โดยวิธี Cox's regression model ในบางครั้งเราอาจต้องการหาตัวแปรอิสระอื่นที่เกี่ยวข้องและมีความสัมพันธ์กับเวลาการอยู่รอดของแต่ละหน่วยตัวอย่าง เพื่อดูว่าตัวแปรอิสระเหล่านั้นมีความสัมพันธ์กับเวลาการอยู่รอดในด้านบวกหรือลบ มากน้อยเพียงไร และเพื่อใช้ในการพยากรณ์เวลาการอยู่รอดในอนาคต ซึ่งก็นับว่าเป็นสิ่งสำคัญอีกอย่างหนึ่งในการวิเคราะห์ข้อมูล ที่นักวิจัยเป็นจำนวนมากต้องการทราบ วิธี Cox's regression model จะเป็นวิธีในการหาตัวแปรอิสระที่มีความสัมพันธ์กับเวลาการอยู่รอด ดังนั้น ฟังก์ชั่นการอยู่รอดจึงขึ้นอยู่กับตัวแปรอิสระของแต่ละหน่วยตัวอย่าง ในวิธีการถดถอยนั้นตัวแปรอิสระจะต้องเป็นตัวแปรเชิงปริมาณ เช่น อายุ, ความดันโลหิต, จำนวนเม็ดเลือดขาว ฯลฯ แต่ในทางปฏิบัติมีตัวแปรอิสระหลายตัวเป็นตัวแปรเชิงคุณภาพ เช่น เพศ, กลุ่มเลือด, ชนิดยา ฯลฯ จึงต้องทำให้เป็นตัวแปรหุ่น (dummy variable) ก่อนนำไปใช้ในการวิเคราะห์

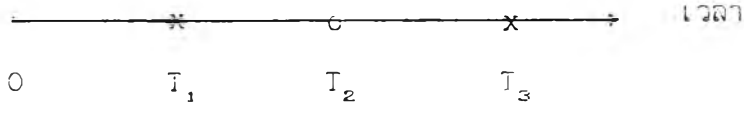
3.1 วิธี Product-Limit (PL)

วิธี PL เป็นวิธีการประมาณค่าฟังก์ชั่นการอยู่รอดสำหรับข้อมูลไม่สมบูรณ์ ที่ไม่ได้มีการจัดกลุ่ม, ไม่มีค่าสังเกตซ้ำและค่าสังเกตมีค่าไม่ต่อเนื่อง สำหรับค่าสังเกตที่มีค่าต่อเนื่องสามารถทำให้เป็นค่าไม่ต่อเนื่องได้โดยการจัดกลุ่มให้เป็นช่วง

3.1.1 นิยามและวิธีการคำนวณ

การประมาณค่าฟังก์ชั่นการอยู่รอด มีวิธีการดังนี้

(ก) พิจารณาช่วงเวลาของแต่ละค่าสังเกต . $(0, T_1)$, (T_1, T_2).... ดังนี้



โดยที่เครื่องหมาย x หมายถึง ค่าสังเกตสมบูรณ์

O หมายถึง ค่าสังเกตไม่สมบูรณ์

(ข) สำหรับแต่ละช่วง (T_{j-1}, T_j), $p_j = P_j/P_{j-1}$ เป็นสัดส่วนของตัวอย่างหลังเวลา T_{j-1} ที่ยังอยู่รอดที่เวลา T_j (P_j คือ $S(t_j)$)

(ค) ถ้า t เป็นเวลาที่กำหนด $S(t)$ จะประมาณจากผลคูณของค่าประมาณ p_j ของทุกช่วงเวลาก่อนถึงเวลา t

Product-Limit ได้มาโดยการเลือกช่วงแต่ละช่วงในข้อ (ก) มาประมาณสัดส่วนในข้อ (ข) ซึ่งเป็นแบบทวินามอย่างง่าย (simple binomial) ดังนั้น ถ้าพิจารณาที่เวลา T_j และให้ $T_1 < T_2 < T_3 <...< T_k$ ค่าประมาณของ p_j คือ

$$\hat{p}_j = (n_j - d_j)/n_j \qquad (3.1)$$

เมื่อ n_j = จำนวนตัวอย่างหลังเวลา T_{j-1} ที่ยังอยู่รอดที่เวลา T_j

d_j = จำนวนที่สูญเสียที่เวลา T_j

และค่าประมาณ PL. ของ $S(t)$ คือ

$$\hat{S}(t) = \prod_{j:T_j \leq t} \hat{P}_j \qquad (3.2)$$

ถ้าเวลา T_j เป็น ค่าสังเกตไม่สมบูรณ์ นั่นคือ ยังอยู่รอดหรือยังไม่สูญเสียที่เวลา T_j ค่าของ $\hat{p}_j$ จะมีค่าเท่ากับ 1

ในทางปฏิบัติตัวประมาณ PL สามารถคำนวณได้โดยการจัดเรียงลำดับข้อมูลตัวอย่างขนาด N ที่มีทั้งค่าสังเกตสมบูรณ์และไม่สมบูรณ์ นั่นคือ

$$t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(N)}$$

กำหนดให้ $j = 1,2,\ldots,N$ และให้ $r = j$ จากสมการ (3.2) จะได้

$$\hat{S}(t) = \prod_{r:t_{(r)} \leq t} (N-r)/(N-r+1) \qquad (3.3)$$

เมื่อ $t_{(r)}$ เป็น ค่าสังเกตสมบูรณ์ และ $(N-r)/(N-r+1)$ คือสัดส่วนของตัวอย่างที่ยังอยู่รอดที่เวลา $t_{(j)}$ ถ้าค่าสังเกตทุกค่าในตัวอย่างมีค่าสมบูรณ์ ค่าประมาณจะเป็นค่าประมาณแบบทวินาม มีค่าเป็น $\hat{S}(t) = n(t)/N$ เมื่อ $n(t)$ คือ จำนวนตัวอย่างที่ยังอยู่รอดที่เวลา t และตัวประมาณ $\hat{S}(t)$ เรียกว่า ตัวประมาณ PL.

ค่าประมาณของค่ามัธยฐานของเวลาการอยู่รอดหรือเปอร์เซ็นไทล์ที่ 50 คือ ค่าของ t ที่ $\hat{S}(t) = 0.50$ จากกราฟของฟังก์ชันการอยู่รอด $\hat{S}(t)$ ดูตัวอย่างรูปที่ 2 ในตอนที่ 2.2

3.1.2 ความแปรปรวนของ $\hat{S}(t)$

พิจารณาจากสมการ (3.2) จะได้

$$\log \hat{S}(t) = \sum_{j:T_j \le t} \log \hat{p}_j$$

$$V(\hat{p}_j) \doteq p_j q_j / n_j$$

เมื่อ $q_j = 1-p_j$ และจากสูตรของความแปรปรวนของฟังก์ชัน $g(x)$;

$$V[g(x)] \doteq [g'(x)]^2 V(x)$$

ดังนั้น

$$V(\log \hat{p}_j) \doteq (1/p_j)^2 (p_j q_j / n_j) = q_j / p_j n_j$$

และ

$$V[\log \hat{S}(t)] \doteq \sum_{j:T_j \le t} (q_j / p_j n_j)$$

เพราะฉะนั้น

$$V[\hat{S}(t)] \doteq [S(t)]^2 \sum_{j:T_j \le t} (q_j / p_j n_j) \qquad (3.4)$$

ซึ่งสามารถประมาณได้โดย[1]

$$\hat{V}[\hat{S}(t)] \doteq [\hat{S}(t)]^2 \sum_{j:T_j \le t} (d_j/((n_j-d_j)n_j)) \qquad (3.5)$$

(Greenwoood's formula)

และในทางปฏิบัติถ้าใช้สมการ (3.3) สามารถประมาณค่าความแปรปรวนได้โดย

$$\hat{V}[\hat{S}(t)] \doteq [\hat{S}(t)]^2 \sum_{j:T_j \le t} (1/((N-r)(N-r+1)) \qquad (3.6)$$

3.1.3 <u>การประมาณค่าเฉลี่ยและค่ามัธยฐานการอยู่รอดโดยใช้ตัวประมาณ PL</u>

(1) ค่าเฉลี่ย (μ)

$$\mu = \int_0^\infty t\, F(t)dt = -[t\, S(t)]_0^\infty + \int_0^\infty S(t)\, dt = \int_0^\infty S(t)\, dt$$

ดังนั้น ค่าประมาณของค่าเฉลี่ยของเวลาการอยู่ ก็คือ

$$\hat{\mu} = \int_0^\infty \hat{S}(t)\, dt \qquad (3.7)$$

ถ้า $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$ คือ ค่าสังเกตสมบูรณ์จำนวน r ค่า และค่าสังเกตที่ใหญ่ที่สุด คือ $t_{(r)}$ นั่นคือ $\hat{S}(t_{(r)}) = 0$

---

[1] K.L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations", <u>Journal of the American Statistical Association</u> 53, (1958), p. 466.

ดังนั้น

$$\hat{\mu} = \sum_{j=0}^{r-1} \hat{S}(t_j)\,(t_{j+1} - t_j) \qquad (3.8)$$

ถ้าค่า $\hat{S}(t_{(r)}) > 0$ ค่าเฉลี่ยจะไม่สามารถหาค่าได้

การประมาณค่าความแปรปรวนของ $\hat{\mu}$ สามารถประมาณได้โดยวิธีของ Irwin (1949)[2] ดังนี้

$$\begin{aligned} V(\hat{\mu}) &\doteq 2 \int_0^{\infty}\int_u^{\infty} S(u)\,S(v)\,U(u)\,dv\,du \\ &= 2 \int_0^{\infty} A(u)\,S(u)\,U(u)\,du \\ &= \int_0^{\infty} A^2(u)\,d\,U(u) \end{aligned}$$

เมื่อ $A(u) = \int_c^{\infty} S(v)\,dv$

และ $U(u) \doteq \sum_{r:t_{(r)}\leq u} (1/((N-r)(N-r+1))$

ซึ่งสามารถประมาณได้ โดย

$$\hat{V}(\hat{\mu}) = \sum_{i:t_i \leq t} A_i^2/((N-i)(N-i+1)) \qquad (3.9)$$

---

[2] Ibid., p. 478.

เมื่อ $A_i = \int_{t_i}^{\infty} \hat{S}(u)\,du = \sum_{j=i}^{r} \hat{S}(t_j)\,(t_{j+1} - t_j)$ ถ้าไม่มีค่าสังเกตไม่สมบูรณ์ $A_i = \sum_{j=i+1}^{N} (t_j - t_i)/N$ และค่า $\hat{V}(\hat{\mu})$ จะเท่ากับ $\Sigma\,(t_j - \bar{t})^2/N^2$

(2) ค่ามัธยฐาน (t_m)

กราฟของฟังก์ชั่นการอยู่รอด สามารถใช้หาค่ามัธยฐานได้ กล่าวคือ ค่าประมาณมัธยฐานของเวลาการอยู่รอดจะเท่ากับค่าของ t ที่ $\hat{S}(t) = 0.50$ แต่โดยทั่วไปแล้ว จะมีช่วงเวลา (t_j, t_{j+1}) ที่มี $\hat{S}(t_j) > 0.50$ และ $\hat{S}(t_{j+1}) < 0.50$ ดังนั้น ค่าประมาณมัธยฐานการอยู่รอดคือ[3]

$$\hat{t}_m = t_j + [(\hat{S}(t_j)-0.50)/((\hat{S}(t_j)-\hat{S}(t_{j+1}))]\,(t_{j+1}-t_j) \qquad (3.10)$$

หรือประมาณโดย[4] $\hat{t}_m = t_{j+1}$

และค่าประมาณของความแปรปรวนของค่าประมาณมัธยฐาน สามารถประมาณได้โดย

$$\hat{V}(\hat{t}_m) = (t_{j+1}-t_j)^2/[4n_j(\hat{S}(t_j)-\hat{S}(t_j))^2] \qquad (3.11)$$

---

[3] Lee, Statistical Methods for Survival Data Analysis, p.91.

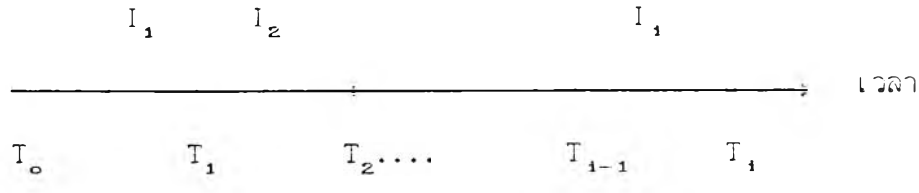
[4] W.J. Dixon, BMDP Statistical Software Mannal, (University of California Press, 1985), p.684.

3.2 วิธี Life-table (actuarial)

3.2.1 การประมาณค่าฟังก์ชั่นการอยู่รอด S(t)

การประมาณค่าฟังก์ชั่นการอยู่รอด วิธีนี้มีนิยามและวิธีการเช่นเดียวกับวิธี PL แต่การประมาณโดยวิธีนี้จะขึ้นอยู่กับช่วงเวลาที่เลือกขึ้นมา จำนวนหน่วยตัวอย่างที่อยู่รอดจะพิจารณาที่จุดเริ่มต้นของช่วงเวลา และจะต้องทราบจำนวนหน่วยตัวอย่างที่สูญเสีย และจำนวนหน่วยตัวอย่างที่มีค่าไม่สมบูรณ์ในเวลานั้น ๆ

กำหนดให้ $T_o < T_1 < \ldots$ และให้ I_i แทนช่วงเวลา (T_{i-1}, T_i) นั่นคือ



ให้ n_i = จำนวนตัวอย่างที่อยู่รอด ที่เวลาเริ่มต้นของช่วง I_i

d_i = จำนวนตัวอย่างที่สูญเสียในช่วง I_i

l_i = จำนวนตัวอย่างที่ไม่สมบูรณ์หรือสูญหายในช่วง I_i

p_i = ความน่าจะเป็นของการอยู่รอดถึงเวลาสุดท้ายของช่วง I_i เมื่อการอยู่รอดเริ่มต้นที่ช่วง I_i

q_i = $1-p_i$

ตามนิยาม 3.1.1 จะได้

$$S(t_k) = \prod_{i=1}^{k} p_i$$

ถ้าช่วง I_i ไม่มีการสูญหาย p_i ประมาณได้โดย

$$\hat{p}_i = (n_i - d_i)/n_i$$

แต่ถ้าช่วงมีการสูญหายเป็นจำนวน l_i แล้วจะได้

$$\hat{p}_i = (n_i - (l_i/2) - d_i)/(n_i - (l_i/2)) \qquad (3.12)$$

ดังนั้น ค่าประมาณของฟังก์ชั่นการอยู่รอด คือ

$$\hat{S}(t_k) = \prod_{i=1}^{k} \hat{p}_i \qquad (3.13)$$

ตัวประมาณ $\hat{S}(t_k)$ เรียกว่า ตัวประมาณ Life-table และค่าประมาณของความแปรปรวนของตัวประมาณ คือ

$$\hat{V}[\hat{S}(t_k)] = [\hat{S}(t_k)]^2 \sum_{i=1}^{k} d_i/((n_i - (l_i/2) - d_i)(n_i - (l_i/2)) \qquad (3.14)$$

(Greenwood's formula)

3.2.2 <u>การประมาณค่าฟังก์ชั่นการสูญเสีย h(t)</u>

พิจารณา $h(t) = f(t)/S(t)$

ให้ T_{mi} แทนจุดกึ่งกลางของเวลาในช่วง I_i และให้ h_i แทนความกว้างของช่วง I_i, $h_i = T_{i-1} - T_i$ จากนิยาม 2.2.1 (1) ให้ $f(t_{mi})$ แทนฟังก์ชั่นความหนาแน่นของเวลา T_{mi} มีค่าเท่ากับความน่าจะเป็นของการสูญเสีย ในช่วง I_i ต่อความกว้างของช่วง (h_i) นั่นคือ

$$f(t_{mi}) = P(\text{การสูญเสียในช่วง } I_i)/h_i$$
$$= (P(T>t_{i-1}) - P(T>t_i))/h_i$$
$$= (S(t_{i-1})-S(t_i))/h_i$$

ซึ่งประมาณได้โดย

$$\hat{f}(t_{mi}) = (\hat{S}(t_{i-1})-\hat{S}(t_i))/h_i \qquad (3.15)$$

ดังนั้น ค่าประมาณของฟังก์ชันการสูญเสีย คือ

$$\hat{h}(t_{mi}) = \hat{f}(t_{mi})/(\hat{S}(t_{mi})$$
$$= [(\hat{S}(t_{i-1})-\hat{S}(t_i))/h_i]/[(\hat{S}(t_{i-1})+\hat{S}(t_i))/2]$$
$$= [2\hat{S}(t_{i-1})\ (1-\hat{p}_i)]/[h_i\ \hat{S}(t_{i-1})\ (1+\hat{p}_i)]$$
$$= (2\hat{q}_i)/(h_i(1+\hat{p}_i)) \qquad (3.16)$$

3.2.3 <u>การประมาณค่ามัธยฐานการอยู่รอดโดยใช้ตัวประมาณ Life-table</u>

มีวิธีการในการประมาณเช่นเดียวกับวิธี PL ให้ค่าประมาณของมัธยฐานการอยู่รอด แทนด้วย $\hat{t}_m$ ดังนั้น

$$\hat{t}_m = t_j + [(\hat{S}(t_j)-0.50)/(\hat{S}(t_j)-\hat{S}(t_{j+1})]\ (t_{j+1}-t_j) \qquad (3.17)$$

และค่าประมาณของความแปรปรวนประมาณได้โดย

$$\hat{V}(\hat{t}_m) = (t_{j+1} - t_j)^2/[4n_j(\hat{S}(t_j)-\hat{S}(t_{j+1}))^2] \qquad (3.18)$$

3.3 วิธี Cox's regression model

วิธี Cox's regression model เป็นวิธีนอนพาราเมตริก สำหรับข้อมูลสมบูรณ์และไม่สมบูรณ์ที่มีการนำเอาไปใช้กันมากที่สุดวิธีหนึ่ง ทั้งนี้เพราะว่า นักวิจัยส่วนใหญ่นั้น มักจะอยากทราบว่ามีตัวแปรอะไรบ้างที่มีอิทธิพลต่อการอยู่รอด เพื่อนำไปใช้ในการพยากรณ์การอยู่รอดในอนาคต ดังนั้น ฟังก์ชั่นการอยู่รอดของวิธีการถดถอยขึ้นอยู่กับตัวแปรอิสระของแต่ละหน่วยตัวอย่าง

3.3.1 Cox's regression model

สมมติให้แต่ละหน่วยตัวอย่างมีการวัดค่าของตัวแปรต่าง ๆ และให้แทนด้วย $z_1, z_2, \ldots, z_p$ เรียกว่า ตัวแปรอิสระ (independent variables) สำหรับเวลาการอยู่รอด t ให้มีการแจกแจงต่อเนื่องและไม่มีค่าซ้ำ

กำหนดให้หน่วยตัวอย่างที่ i มีค่าของตัวแปร p ตัว เป็น $Z_i = (z_{1i}, z_{2i}, \ldots, z_{pi})$ ซึ่งค่าของตัวแปรเหล่านี้อาจจะเป็น เพศ, อายุ, จำนวนเม็ดเลือดขาวของคนไข้ หรืออาจจะเป็นฟังก์ชั่นของเวลา ฯลฯ ปัญหาสำคัญที่จะพิจารณาก็คือ ความสัมพันธ์ระหว่างการแจกแจงของเวลาการอยู่รอด และ Z ซึ่งจะเสนอในรูปของฟังก์ชั่นการสูญเสีย ดังนี้

$$h(t:Z_i) = h_o(t) \exp(Z_i\beta) \qquad (3.19)$$

เมื่อ β เป็นเวคเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด px1 และ $h_o(t)$ เป็นฟังก์ชั่นการสูญเสียที่ไม่ทราบค่าสำหรับ Z = 0 ในเทอมของ $Z\beta$ สามารถแทนได้โดยฟังก์ชั่นของ Z และ β ทุกฟังก์ชั่นที่ทราบค่า

ข้อตกลงของโมเดลของฟังก์ชั่นการสูญเสีย มีดังนี้

(1) ฟังก์ชั่นการสูญเสียของทุกหน่วยตัวอย่างเป็นสัดส่วนกัน เช่น ถ้ามีกลุ่มศึกษา 2 กลุ่ม ให้ตัวแปรหุ่น Z (มีค่า 1 และ 0) เป็นตัวชี้หรือแบ่งกลุ่ม อัตราการสูญเสียในกลุ่มแรกเป็น e^{β} เท่าของอีกกลุ่ม

(2) ผลของตัวแปรอิสระในฟังก์ชั่นการสูญเสียอยู่ในรูปแบบของผลคูณ

นั่นคือ $$h(t:z_{1i},z_{2i},\ldots,z_{pi}) = h_o(t)\, e^{\beta_1 z_{1i}} . e^{\beta_2 z_{2i}} \ldots e^{\beta_p z_{pi}}$$

จากสมการ (3.19) สามารถหาฟังก์ชั่นการอยู่รอดได้จาก

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

ดังนั้น $$S(t:Z_i) = \exp\left(-e^{Z_i\beta}\int_0^t h_o(u)du\right)$$

$$= S_o(t)^{e^{Z_i\beta}} \qquad (3.20)$$

เมื่อ $S_o(t) = \exp\left(-\int_0^{\infty} h_o(u)du\right)$ เป็นฟังก์ชั่นการอยู่รอดสำหรับ $Z = 0$ และสามารถประมาณได้โดย[5]

$$\hat{h}_o(t) = d_j/\left((t_j - t_{j-1}) \sum_{i\in R_i} e^{Z_i\beta}\right) \text{ ; สำหรับ } t_{j-1} < t < t_j$$

เมื่อ d_j คือ จำนวนของเวลาการอยู่รอดที่เท่ากับ t_j และ R_j คือ เซทของความเสี่ยงที่เวลา t_j และ

$$\hat{S}_o(t) = \exp\left(-\int_0^t \hat{h}_o(u)du\right) \qquad (3.21)$$

---

[5] Cox, "Regression model and Life-tables", Journal of the Royal Statistical Society, p.217.

สำหรับตัวอย่างปัญหาที่ใช้กับสมการ (3.19) ได้แก่

(1) ปัญหา 2 ตัวอย่าง สมมติให้ p = 1 นั่นคือ ตัวแปรอิสระ มีเพียงตัวเดียว เป็นตัวแปรหุ่นมีค่าดังนี้

$$z_{1i} = \begin{cases} 0 & \text{ถ้าตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 1} \\ 1 & \text{ถ้าตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 2} \end{cases}$$

ดังนั้น ฟังก์ชั่นการสูญเสียในสมการ (3.19) คือ

$$h_1(t:z_{1i}) = h_0(t) \quad ; \text{ สำหรับตัวอย่างกลุ่มย่อยที่ 1}$$

$$h_2(t:z_{1i}) = h_0(t)e^{\beta_1} \quad ; \text{ สำหรับตัวอย่างกลุ่มย่อยที่ 2}$$

และ

$$S_1(t:z_{1i}) = S_0(t) \quad ; \text{ สำหรับตัวอย่างย่อยกลุ่มที่ 1}$$

$$S_2(t:z_{1i}) = S_0(t)^{e^{\beta_1}} \quad ; \text{ สำหรับตัวอย่างย่อยกลุ่มที่ 2}$$

สำหรับการทดสอบ 2 ตัวอย่างที่พัฒนาจากสมการ (3.19) คือ การทดสอบ Cox-Mantel (ดูในภาคผนวก ก) และในปัญหา 2 ตัวอย่างนี้สามารถขยายไปใช้ในปัญหา k ตัวอย่างได้

(2) ปัญหา 2 ตัวอย่างที่มีตัวแปรอิสระหลายตัว นอกจากจะมีตัวแปรอิสระ z_1 เช่นเดียวกับปัญหา (1) แล้ว ในสมการ (3.19) ยังมีตัวแปรอิสระอื่นอีกอาจจะเป็น 2 ตัว หรือมากกว่า

(3) ปัญหา 2 ตัวอย่างที่มีตัวแปรอิสระเป็นฟังก์ชั่นของเวลา ในสมการ (3.19) ตัวแปรอิสระ z สามารถเป็นตัวแปรที่เป็นฟังก์ชั่นของเวลา t เช่น ถ้าให้ $z_2 = t z_1$ เมื่อ z_1 คือ ตัวแปรแบ่งกลุ่มเช่นเดียวกับปัญหา (1) นั่นคือ

$$Z_{1i} = \begin{cases} 0 & \text{เมื่อตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 1} \\ 1 & \text{เมื่อตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 2} \end{cases}$$

$$Z_{2i} = \begin{cases} 0 & \text{เมื่อตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 1} \\ t & \text{เมื่อตัวอย่างที่ i อยู่ในตัวอย่างย่อยที่ 2} \end{cases}$$

ดังนั้น

$$h_1(t:z_{1i},z_{2i}) = h_o(t) \quad ; \text{ สำหรับตัวอย่างย่อยที่ 1}$$

$$h_2(t:z_{1i},z_{2i}) = h_o(t)e^{\beta_1+\beta_2 t} \quad ; \text{ สำหรับตัวอย่างย่อยที่ 1}$$

(4) ปัญหาการถดถอยจากสมการ (3.19)

$$h(t:Z_i) = h_o(t)e^{Z_i\beta} = h_o(t)\, e^{\beta_1 z_{1i} + \beta_2 z_{2i} + \cdots + \beta_p z_{pi}}$$

$$\log\ (h(t:Z_i)/h_o(t)) = Z_i\beta = \beta_1 z_{1i} + \beta_2 z_{2i} + \ldots + \beta_p z_{pi}$$

ซึ่งเทอมทางด้านขวา คือ ฟังก์ชั่นของการสูญเสียสำหรับตัวอย่างที่ i และเทอมทางด้านซ้าย คือ การรวมเชิงเส้นของตัวแปร $z_{1i}, z_{2i}, \ldots z_{pi}$ ด้วยสัมประสิทธิ์ $\beta_1, \beta_2, \ldots \beta_p$ ถ้าให้ $y_i = \log [h(t:Z_i)/h_o(t)]$ จะได้

$$y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \ldots + \beta_p z_{pi}$$

ซึ่งก็คือ สมการการถดถอยพหุมาตรฐาน ที่มีฟังก์ชั่นของการสูญเสียเป็นตัวแปรตาม และมีตัวแปร z เป็นตัวแปรอิสระ

3.3.2 <u>การประมาณค่าพารามิเตอร์ β โดยวิธี Maximum-Likelihood (ML)</u>

สมมติให้ $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ เป็นค่าสังเกตสมบูรณ์จำนวน k ค่า และให้ $R(t_{(i)})$ เป็นเซทของความเสี่ยงที่เวลา $t_{(i)}$ ซึ่ง $R(t_{(i)})$ ประกอบด้วยหน่วยตัวอย่างทั้งหมดที่มีการอยู่รอดอย่างน้อยที่สุด $t_{(i)}$ และให้กำหนด $h_o(t)$ ขึ้นเอง

สำหรับเวลา $t_{(i)}$ ในเงื่อนไข $R(t_{(i)})$ ความน่าจะเป็นที่ตัวอย่าง i สูญเสียที่เวลา $t_{(i)}$ เมื่อเกิดการสูญเสีย 1 ค่าที่เวลา $t_{(i)}$ สามารถประมาณได้โดย

$$P(\text{ตัวอย่าง } i \text{ สูญเสียที่เวลา } t_{(i)} \mid \text{เกิดการสูญเสีย 1 ค่าที่เวลา } t_{(i)}) = e^{Z_i\beta} \Big/ \sum_{j\in R(t_{(i)})} e^{Z_j\beta} \qquad (3.22)$$

ซึ่งเป็นความน่าจะเป็นแบบมีเงื่อนไข (conditional probability) สำหรับแต่ละเวลาการอยู่รอดที่มีค่าสมบูรณ์ ดังนั้น conditional log-likelihood คือ

$$L(\beta) = \sum_{i=1}^{k} Z_i\beta - \sum_{i=1}^{k} \log\left[\sum_{j\in R(t_{(i)})} \exp(Z_j\beta)\right]$$

ดังนั้น ตัวประมาณ Maximum-Likelihood (ML) โดยปกติสามารถประมาณได้จาก

$$U_m(\beta) = (\partial L(\beta)/\partial\beta_m) = \sum_{i=1}^{k} [z_{mi} - A_{mi}(\beta)] = 0 \; ; \; m = 1,2,..,p \qquad (3.23)$$

เมื่อ $$A_{mi}(\beta) = \sum_{j\in R(t_{(i)})} z_{mj}\exp(Z_i\beta) \Big/ \sum_{j\in R(t_{(i)})} \exp(Z_i\beta) \qquad (3.24)$$

นั่นคือ $A_{mi}(\beta)$ เป็นค่าเฉลี่ยของ z_m ภายใต้ $R(t_{(i)})$

การประมาณค่าพารามิเตอร์ในสมการ (3.23) ถ้าทำด้วยมือจะยากมาก จึงจำเป็นต้องใช้คอมพิวเตอร์ โดยใช้วิธี Newton-Raphson (1967)[6] (ดูวิธีการในภาคผนวก ข) ซึ่งจะใช้สมการ (3.23) และ

$$I_{mn}(\beta) = -(\partial^2 L(\beta))/(\partial\beta_m \partial\beta_n) = \sum_{i=1}^{k} C_{mni}(\beta) \qquad (3.25)$$

เมื่อ $C_{mni}(\beta) = [(\sum_{j\epsilon R(t_{(i)})} z_{mi} z_{nj} \exp(Z_i\beta)) /$

$$(\sum_{j\epsilon R(t_{(i)})} \exp(Z_j\beta))] - (A_{mi}(\beta)\, A_{ni}(\beta))$$

สมมติให้ n_i เป็นจำนวนตัวอย่างที่ยังอยู่รอดที่เวลา $t_{(i)}$ และ d_i เป็นจำนวน ที่สูญเสียที่เวลา $t_{(i)}$ ภายใต้เงื่อนไข $R(t_{(i)})$ สมการ (3.23) และ (3.25) สามารถประมาณได้โดย

$$U_m(\beta) = \sum_{i=1}^{k} [z_{n,i} - d_i A_{mni}(\beta)] \qquad (3.26)$$

และ
$$I_{mn}(\beta) = \sum_{i=1}^{k} (d_i(n_i-d_i)/(n_i-1)\, C_{mni}(\beta) \qquad (3.27)$$

สำหรับการทดสอบสมมติฐานหลัก $\beta = 0$ จะถือว่า $U_m(0)$ มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็นเวคเตอร์ 0 และความแปรปรวนร่วมเป็นเมทริกซ์ $I(0)$ นั่นคือ ภายใต้สมมติฐานหลัก

$$[U(0)]^T [U(0)]/I(0) \qquad (3.28)$$

---

[6] Lee, Statistical Methods for Survival Data Analysis. p.310.

จะมีการแจกแจงแบบไค-สแควร์ ด้วยองศาแห่งความอิสระ p และจากสมการ (3.23) และ (3.25) ค่าของ U(0) และ I(0) คือ

$$U_m(0) = \sum_{i=1}^{k} [z_{mi} - A_{mi}(0)] \qquad (3.29)$$

และ

$$I_{mn}(0) = \sum_{i=1}^{k} C_{mni}(0) \qquad (3.30)$$

3.3.3 <u>ปัญหา 2 ตัวอย่าง</u> (two-sample problem)

ในตอนนี้จะพิจารณาเฉพาะปัญหา 2 ตัวอย่าง ซึ่งมีฟังก์ชั่นการสูญเสีย ดังแสดงในตัวอย่างปัญหา (1) ตอนที่ 3.3.1 และมี p = 1 ดังนั้น พารามิเตอร์ที่จะประมาณค่าจึงมีเพียงตัวเดียว จากสมการ (3.29) และ (3.30) จะได้

$$U(0) = \sum_{i=1}^{k} (z_i - d_i A_i) = n_{(1)} - \sum_{i=1}^{k} d_i A_i \qquad (3.31)$$

และ $$I(0) = \sum_{i=1}^{k} (d_i(n_i-d_i))/(n_i-1)\, A_i(1-A_i)$$

เมื่อ A_i เป็นสัดส่วนของ n_i เมื่อ z = 1 หรืออยู่ในตัวอย่างย่อยที่ 2 และ $n_{(1)}$ คือ จำนวนค่าสังเกตรวมในตัวอย่างย่อยที่ 2

สำหรับการทดสอบสมมติฐาน ตัวสถิติ

$$U(0) / \sqrt{I(0)}$$

จะมีการแจกแจงแบบปกติมาตรฐาน ภายใต้สมมติฐานหลัก

ในทางปฏิบัติ ถ้า β มีค่าน้อย ตัวประมาณ ML ประมาณได้โดย

$$\hat{\beta} = [n_{(1)} - \sum_{i=1}^{k} A_i] / [\sum_{i=1}^{k} A_i (1-A_i)]$$

วิธีการที่กล่าวข้างต้นใช้สำหรับข้อมูลที่จัดเรียงลำดับแล้วเท่านั้น