บทที่ 2

สถิติที่ใช้ในการวิจัย

การวิเคราะห์การถดถอยเชิงเส้นพหุเมื่อตัวแปรตามบางค่ามีค่าขาดหาย ในการศึกษาครั้งนี้ได้ทำการศึกษาวิธีการประมาณค่า 3 วิธีคือ วิธีการของสมิธ วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด และวิธีการโมดิฟายแอคชัวเรียล ซึ่งในบทนี้จะกล่าวถึงทฤษฎีพื้นฐานและรายละเอียดของวิธีการประมาณค่าพารามิเตอร์ในแต่ละวิธีดังนี้

2.1 ทฤษฎีพื้นฐาน

2.1.1 ประเภทของการตัดทิ้ง (Type of Censoring)

2.1.1.1 การตัดทิ้งประเภทที่ 1 (Type I Censoring) เป็นการตัดทิ้งที่จะมีการกำหนดเวลาของการตัดทิ้งไว้ล่วงหน้าเรียกว่า 'Fixed Censoring Time' เช่น ศึกษาเกี่ยวกับอายุการใช้งานของเครื่องจักรชนิดหนึ่ง โดยศึกษาในระยะเวลา 8000 ชั่วโมง เริ่มทำการทดลองโดยให้เครื่องจักรนี้ทำงานแล้วบันทึกเวลาไว้ตั้งแต่เริ่มทำงานจนกระทั่งเครื่องจักรเสื่อมสภาพ ในระหว่างที่ทำการทดลอง เครื่องจักรเครื่องใดเสื่อมสภาพ จะเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง (Uncensored Data) และเมื่อสิ้นสุดการทดลองเครื่องจักรใดยังคงอยู่ในสภาพที่ใช้งานได้ ก็จะเป็นเครื่องที่ไม่ทราบอายุการใช้งานที่แน่นอน จะทำการบันทึกไว้ว่ามีอายุการใช้งาน 8000 ชั่วโมง ซึ่งข้อมูลนี้จะเป็นค่าสังเกตที่ถูกตัดทิ้ง (Censored Data)

ให้ $T_1,...,T_n$ เป็นตัวแปรสุ่มมีการแจกแจงที่เหมือนกันและเป็นอิสระกัน และ tc เป็นค่าที่กำหนดไว้ล่วงหน้า ในกรณีนี้จะได้ตัวแปรสุ่ม $Y_1,...,Y_n$ ซึ่ง

$$Y_i = \begin{cases} T_i & \text{ถ้า } T_i \le tc \quad \text{(ไม่ถูกตัดทิ้ง)} \\ tc & \text{ถ้า } T_i > tc \quad \text{(ถูกตัดทิ้ง)} \end{cases}$$

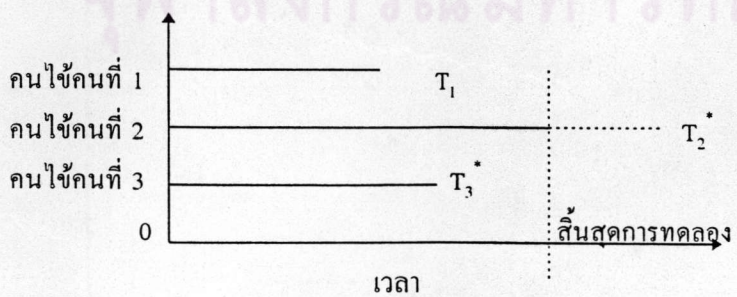
2.1.1.2 <u>การตัดทิ้งประเภทที่ 2</u> (Type II Censoring) เป็นการตัดทิ้งที่จะต้องกำหนดจำนวนค่าสังเกตที่ถูกตัดทิ้งไว้ล่วงหน้า เนื่องจากในบางกรณีไม่อาจจะกำหนดเวลาของการเกิดค่าตัดทิ้งที่เหมาะสมได้ นั่นคือเมื่อจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้งเกิดขึ้นครบตามจำนวนที่กำหนดไว้จะหยุดทำการทดลองทันทีเพื่อเป็นการประหยัดเวลาและค่าใช้จ่าย เช่น ในการทดลองเกี่ยวกับประสิทธิภาพของเครื่องใช้ไฟฟ้าชนิดหนึ่ง จำนวน 20 เครื่อง แทนที่เราจะทำการทดลองอย่างต่อเนื่องจนกระทั่งเครื่องใช้ไฟฟ้าที่นำมาทดลองทั้ง 20 เครื่องจะเสียหายหรือใช้งานไม่ได้ เราจะหยุดทำการทดลอง ณ เวลาหนึ่งของการเกิดเหตุการณ์ดังกล่าวครั้งที่ r เช่น หยุดทำการทดลองเมื่อเครื่องใช้ไฟฟ้าจะเสียเป็นเครื่องที่ 15 เป็นต้น

ถ้า r คือจำนวนของค่าสังเกตที่ไม่ถูกตัดทิ้งที่กำหนดไว้

$T_1,...,T_r$ คือค่าสังเกตที่ไม่ถูกตัดทิ้ง

$$Y_1 = T_1$$
$$Y_2 = T_2$$
$$\vdots$$
$$Y_r = T_r$$
$$Y_{r+1} = T_r$$
$$\vdots$$
$$Y_n = T_r$$

2.1.1.3 <u>การตัดทิ้งแบบสุ่ม</u> (Random Censoring) การตัดทิ้งแบบนี้มีลักษณะคล้ายกับการตัดทิ้งประเภทที่ 1 คือมีการกำหนดระยะเวลาของการตัดทิ้งไว้ล่วงหน้าแต่การตัดทิ้งของข้อมูลนั้นอาจจะเกิดขึ้นได้ก่อนสิ้นสุดการทดลองซึ่งส่วนใหญ่จะพบในการทดลองทางการแพทย์ เช่น คนไข้ถอนตัวออกจากการทดลองก่อนสิ้นสุดการทดลองหรือคนไข้อาจยังมีชีวิตรอดเมื่อสิ้นสุดการทดลอง ซึ่งจะทำให้ไม่สามารถได้ค่าที่แน่นอนของค่าสังเกตนั้นได้



รูปที่ 2.1 แสดงการเกิดค่าตัดทิ้งแบบสุ่ม

คนไข้คนที่ 1 เข้าทำการทดลองและตายที่เวลา T_1 ค่าสังเกตนี้เป็นค่าที่ไม่ถูกตัดทิ้ง

คนไข้คนที่ 2 เข้าทำการทดลองและเมื่อสิ้นสุดการทดลองแล้วยังมีชีวิตอยู่ ค่าสังเกตนี้เป็นค่าที่ถูกตัดทิ้ง

คนไข้คนที่ 3 เข้าทำการทดลองและถอนตัวจากการทดลองเมื่อเวลา T_3 ค่าสังเกตนี้เป็นค่าที่ถูกตัดทิ้ง

ถ้า $T_1,...,T_n$ เป็นตัวแปรสุ่มที่มีการแจกแจงเหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง F

$C_1,...,C_n$ เป็นตัวแปรสุ่มที่ถูกตัดทิ้งที่มีการแจกแจงเหมือนกันและเป็นอิสระกัน มีฟังก์ชันการแจกแจง G

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

$$Y_i = \begin{cases} Y_i & \text{เมื่อ } T_i \leq C_i \quad \text{(ไม่ถูกตัดทิ้ง)} \\ C_i & \text{เมื่อ } T_i > C_i \quad \text{(ถูกตัดทิ้ง)} \end{cases}$$

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

2.1.2 อัตราการสูญเสีย หรือฟังก์ชันการสูญเสีย (Failure Rate or Hazard Function)

ให้ T เป็นตัวแปรสุ่มต่อเนื่องแทนอายุการใช้งาน (time of failure)

f(t) แทนฟังก์ชันความหนาแน่นของ T

F(t) แทนฟังก์ชันการแจกแจงสะสมของ T

h(t) แทนอัตราการสูญเสียหรือฟังก์ชั่นการสูญเสียของ T ที่แต่ละหน่วยตัวอย่างเสียในช่วงเวลาสั้น ๆ จาก t ถึง $t+\Delta t$ ต่อหน่วยเวลา Δt เมื่อแต่ละหน่วยตัวอย่างมีอายุใช้งานมากกว่า $(T > t)$ นั่นคือ

$$h(t) = \lim_{\Delta t \to 0} \frac{[P(t<T<t+\Delta t \ /T>t]}{\Delta t}$$

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

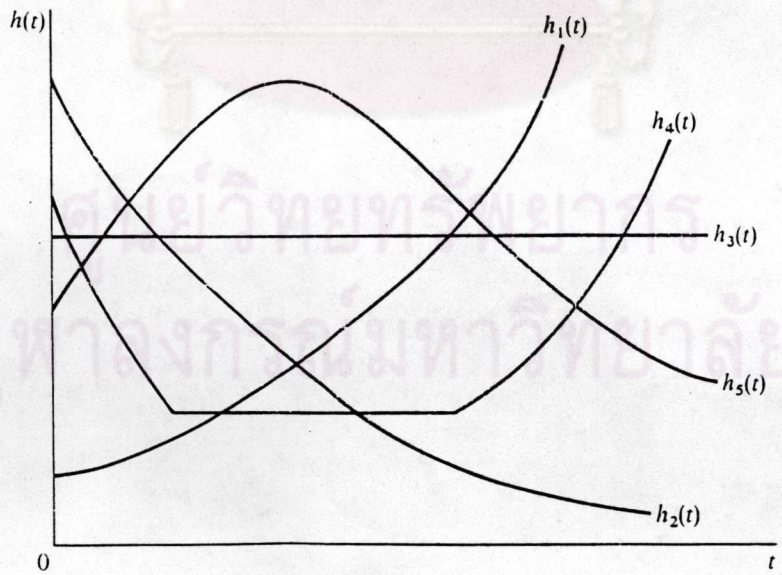
$$= \frac{d}{dt} F(t) \cdot \frac{1}{1-F(t)}$$

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

โดยที่ h(t) มีคุณสมบัติดังนี้

ก. $h(t) > 0$ เมื่อ $-\infty < t < \infty$

ข. $\lim_{t \to -\infty} \int_{-\infty}^{t} h(t)dt = 0$ และ $\lim_{t \to -\infty} \int_{-\infty}^{t} h(t)dt = \infty$



รูปที่ 2.2 แสดงตัวอย่างลักษณะของฟังก์ชันการสูญเสีย h(t)ในรูปแบบต่าง ๆ

2.1.3 ตัวประมาณพีแอล (Product Limit Estimator : PL Estimator) [4]

ให้ Y_1 Y_2 ... Y_n เป็นค่าสังเกตที่เรียงลำดับของข้อมูลที่ถูกตัดทิ้งและไม่ถูกตัดทิ้ง ตัวประมาณพีแอล สามารถแสดงได้ดังนี้

$$\hat{S}(Y) = \prod_{Yi \le Y} \frac{(n-1)}{(n-i+1)}$$

เมื่อ n เป็นจำนวนข้อมูลทั้งหมดที่ถูกตัดทิ้งและไม่ถูกตัดทิ้ง

i เป็นลำดับที่ของข้อมูล

Y_i เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

ตัวอย่างการหาตัวประมาณพีแอล จากค่าสังเกต

4.0 , 5.0 , 9.0 , 13.0 , 13.0^+ , 15.0 , 16.0^+ , 20.0 , 25.0^+ , 27.0

เมื่อ + หมายถึง ข้อมูลที่ถูกตัดทิ้ง จะได้ตัวประมาณพีแอล ดังนี้

Y	Rank	i	$\frac{n-i}{n-i+1}$	$\hat{S}(Y)$
4.0	1	1	9/10	0.9
5.0	2	2	8/9	$S_{(4)}x8/9 = 0.8$
9.0	3	3	7/8	$S_{(5)}x7/8 = 0.7$
13.0	4	4	6/7	$S_{(9)}x6/7 = 0.6$
13.0^+	5	-	-	-
15.0	6	6	4/5	$S_{(13)}x4/5 = 0.48$
16.0^+	7	-	-	-
20.0	8	8	2/3	$S_{(15)}x2/3 = 0.32$
25.0^+	9	-	-	-
27.0	10	10	0	$S_{(20)}x0 = 0.00$

รูป 2.3 แสดงการคำนวณฟังก์ชันการอยู่รอดโดยใช้ตัวประมาณพีแอล

---

[4] Koplan E.L and Meier P. , Nonparametric Estimation from Incomplete Observations Journal of the American Statistical Association, 53 53 (June 1958) : 475-81.

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

2.2 การประมาณค่าพารามิเตอร์

ศึกษารูปแบบสมการถดถอยพหุ ซึ่งมีรูปแบบดังนี้

$$T_i = \beta_j X_{ij} + \varepsilon_i \quad ; i=1,...,nm \; ; \; j=1,2,3$$

เมื่อ T_i เป็นตัวแปรตาม

x_{ij} เป็นตัวแปรอิสระ

β_j เป็นพารามิเตอร์ที่ไม่ทราบค่า ; j=1,2,3

ε_i เป็นค่าความคลาดเคลื่อนสุ่ม

ให้ tc เป็นค่าที่กำหนดไว้ล่วงหน้า และค่าสังเกตของตัวแปรตาม T_i, i=1,..,nm มีการแจกแจงที่เหมือนกันและเป็นอิสระกัน ดังนั้น ค่าสังเกต Y_i หาได้จาก

$$Y_i = \begin{cases} T_i & \text{ถ้า } T_i \le tc \text{ (ค่าสังเกตไม่ถูกตัดทิ้ง)} \\ tc & \text{ถ้า } T_i > tc \text{ (ค่าสังเกตที่ถูกตัดทิ้ง)} \end{cases}$$

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i \le tc \text{ (ค่าสังเกตที่ไม่ถูกตัดทิ้ง)} \\ 0 & \text{ถ้า } T_i > tc \text{ (ค่าสังเกตที่ถูกตัดทิ้ง)} \end{cases}$$

2.2.1 วิธีการของสมิธ

การประมาณค่าพารามิเตอร์ด้วยวิธีการของสมิธ ซึ่งเสนอโดย ปีเตอร์ เจมส์ สมิธ[5] (Peter James Smith : 1985) ได้ดัดแปลงมาจากวิธีการของบัคเลย์และเจมส์ ซึ่งเป็น วิธีการนอนพาราเมตริกซ์ (Nonparametric Method) ที่มีข้อกำหนดคือ ค่าความคลาดเคลื่อน (ε_i) เป็นอิสระ มีการแจกแจงไม่ทราบรูปแบบที่แน่นอน มีค่าเฉลี่ยเป็น α และมีความแปรปรวนจำกัดเป็น σ^2 และมีฟังก์ชันการอยู่รอดเป็น S = 1-F โดยที่วิธีการของสมิธนี้จะศึกษาในกรณีที่เราทราบค่าของการถูกตัดทิ้ง (tc)

จากนิยามของการตัดทิ้งประเภทที่ 1 (Type I censoring) ค่าสังเกต Y_i ได้จาก

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

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i \leq tc \text{ (ค่าสังเกตไม่ถูกตัดทิ้ง)} \\ 0 & \text{ถ้า } T_i > tc \text{ (ค่าสังเกตถูกตัดทิ้ง)} \end{cases}$$

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

$$Y_i^*(\beta)^6 = Y_i\delta_i + E(Y_i / Y_i > tc)(1 - \delta_i) \quad i=1,...,nm$$

หาค่าพารามิเตอร์ β ดังนี้

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

ให้ $\beta = (\alpha, \beta)$ เป็นเวกเตอร์ของพารามิเตอร์ จะได้ว่า

$$E(Y_i^*) = \alpha + \beta X$$

---

[5] Smith, P.J. Estimation in Linear Regression with Censored Response : Proceeding of Pacific Statistical Congress,eds Francis,IF; Manly,Lam,FC.Amsterdam

[6] James,I.R. and Smith, P.J., Consistency results for Linear with censored data,Ann. Stat. 12(1984) 590-600.

แต่เนื่องจาก $E(Y_i/ Y_i > tc)$ ไม่ทราบค่า ดังนั้นจึงทำการประมาณค่าโดยใช้ตัวประมาณพีแอล โดยจะพิจารณาจากค่าความคลาดเคลื่อนบางส่วน (Partial Residual) ดังนั้นข้อมูลที่ถูกตัดทิ้ง จะถูกแทนด้วย $E(Y_i / Y_i > tc) = Y'_i(\hat{\beta})$

$$Y'_i(\hat{\beta}) = \frac{\hat{\beta} X_i + \sum_{uc} W_{(ei,\hat{\beta})}(Y-\hat{\beta} X_i)}{1-F}$$

และประมาณค่าพารามิเตอร์ตามวิธีการของสมิธ ดังนี้

$$\hat{\beta}_s = \frac{\sum (X_i - \overline{X})S_{(ei)} \rho_i(\hat{\beta})-1}{\sum (X_i - \overline{X})^2}$$

เมื่อ $S = 1-F$

$$\rho_i(\hat{\beta}) = 1 + W_{(ei,\hat{\beta})}[tc - \hat{\beta} X_i - Y'_i(\hat{\beta})]$$

ขั้นตอนในการหาค่าประมาณพารามิเตอร์ สำหรับวิธีการของสมิธมีขั้นตอนดังนี้

1. เฉพาะข้อมูลที่ไม่ตัดทิ้ง ประมาณ $\hat{\beta}_0$ เริ่มต้น โดยใช้วิธีการกำลังสองต่ำสุด

2. นำข้อมูลทั้งหมด ทั้งค่าที่ถูกตัดทิ้งและค่าไม่ถูกตัดทิ้ง มาหาค่าความคลาดเคลื่อน e_i จาก $e = Y - \hat{\beta}_0 X$ เรียงลำดับค่าความคลาดเคลื่อนจากน้อยไปหามาก ในกรณีที่ลำดับที่ของค่าความคลาดเคลื่อนของค่าที่ถูกตัดทิ้ง และค่าที่ไม่ถูกตัดทิ้งมีค่าเท่ากัน ให้ลำดับที่ของค่าที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของค่าที่ถูกตัดทิ้ง

3. เปลี่ยนลำดับของค่าที่ถูกตัดทิ้งเป็น 0 ส่วนลำดับที่ของค่าที่ไม่ถูกตัดทิ้งให้คงไว้

4. หาค่า $\hat{S}_{(e)}$ โดยใช้ตัวประมาณพีแอล

$$\hat{S}_{(e)} = \prod_{ei \le ei+1} \frac{(n-i)}{(n-i+1)}$$

เมื่อ i คือ ลำดับที่ของค่าความคลาดเคลื่อน

n คือ จำนวนข้อมูลทั้งหมด

$\hat{S}_{(e)}$ คือ ค่าของเวลาที่มีการอยู่รอด (Survival Time)

5. หาค่าฟังก์ชัน $\hat{F}_{(ei,\hat{\beta})}$ จาก

$$\hat{F}_{(ei,\hat{\beta})} = 1- \hat{S}_{(e)}$$

6. หาค่าถ่วงน้ำหนัก $W_{(ei,\hat{\beta})}$ จาก

$$W_{(e1,\hat{\beta})} = \hat{F}_1$$

$$W_{(e2,\hat{\beta})} = \hat{F}_2 - \hat{F}_1$$

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

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

$$W^{*}_{(ei,\hat{\beta})} = \frac{W_{(ei,\hat{\beta})}}{\sum_{uc} W_{(ei,\hat{\beta})}}$$

7. เปลี่ยนลำดับที่ของค่าที่ไม่ถูกตัดทิ้งเป็น 0 ส่วนลำดับที่ของค่าที่ถูกตัดทิ้งให้คงไว้

8. หาค่า $\hat{S}_{(c)}$ โดยใช้ตัวประมาณพีแอล

$$\hat{S}_{(c)} = \prod_{ei \le ei+1} \frac{(n-i)}{(n-i+1)}$$

$\hat{S}_{(c)}$ คือ ค่าของเวลาที่ถูกตัดทิ้ง (Censoring Time)

9. หาค่าฟังก์ชัน $\hat{F}_{(ci,\hat{\beta})}$ จาก

$$\hat{F}_{(ci,\hat{\beta})} = 1 - \hat{S}_{(c)}$$

10. จากขั้นที่ 6 และ 9 จะได้ค่า $W_{(ei,\hat{\beta})}$ และ $\hat{F}_{(ci,\hat{\beta})}$ นำมาหาค่าประมาณของค่าสังเกตที่ถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไข ดังนี้

$$E(Y_i / Y_i > tc) = Y'_i(\hat{\beta})$$

$$Y'_i(\hat{\beta}) = \frac{\hat{\beta} X_i + \sum_{uc} W_{(ei,\hat{\beta})}(Y - \hat{\beta} X_i)}{1 - F}$$

11. หาค่า $\rho_i(\hat{\beta})$ จาก

$$\rho_i(\hat{\beta}) = 1 + W_{(ei,\hat{\beta})}[tc - \hat{\beta} X_i - Y'_i(\hat{\beta})]$$

เมื่อ tc เป็นค่าเวลาที่กำหนดไว้ล่วงหน้า

12. ประมาณค่าพารามิเตอร์ $\hat{\beta}_s$ ดังนี้

$$\hat{\beta}_s = \frac{\sum (X_i - \overline{X}) S_{(ei)} \rho_i(\hat{\beta}) - 1}{\sum (X_i - \overline{X})^2}$$

13. แทนค่า $\hat{\beta}_s$ จากขั้นที่ 12 ลงในขั้นที่ 2 แล้วทำการวนซ้ำจากขั้นที่ 2 ถึงขั้นที่ 12 ทำไปจนกระทั่งค่าของ $\hat{\beta}_s$ ของรอบปัจจุบัน ได้เท่ากับค่าของ $\hat{\beta}_s$ ในรอบที่แล้วจึงหยุด ในบางครั้งค่าของ $\hat{\beta}_s$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้น เป็นค่าประมาณของ $\hat{\beta}_s$

14. นำค่าประมาณ $\hat{\beta}_s$ หาค่าประมาณของตัวแปรตาม $\hat{Y}_i = \sum_{j=0}^{3} \hat{\beta}_{sj} X_{ij}$ จากนั้น หาค่าความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริงในรูปของค่ารากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (RMSE)

$$MSE = \frac{\sum_{uc} (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

เมื่อ n คือจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

Y_i คือค่าจริงของค่าสังเกตที่ไม่ถูกตัดทิ้ง

2.2.2 วิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด

การประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นตรงพหุ เมื่อตัวแปรตามมีค่าถูกตัดทิ้งทางขวา ด้วยวิธีการประมาณด้วยภาวะน่าจะเป็นสูงสุด ได้ใช้อีเอ็ม อัลกอริทึม ซึ่งเสนอโดย เด็มสเตอร์ ลายด์แและรูบิน โดยจะใช้วิธีการกระทำวนซ้ำ และข้อมูลที่นำมาวิเคราะห์นั้นจะถือว่า ค่าสังเกตที่ถูกตัดทิ้งเสมือนเป็นค่าที่ไม่ถูกตัดทิ้ง

จากสมการ $Y = \beta X + \varepsilon$ เมื่อ $\varepsilon \sim N(0, \sigma^2)$

ให้ $\mu_i = \sum_{j=0}^{3} \beta_j X_{ij}$; $X_{i0} = 1$, $\beta_0 = 1$

$$Z_i = (Y_i - \mu_i) / \sigma$$

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

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

$$h(Z) = \frac{f(Z)}{(1 - F(Z))}$$

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

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

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

i=1,...,n ค่าสังเกตที่ไม่ถูกตัดทิ้ง

i=n+1,...,n+m ค่าสังเกตที่ถูกตัดทิ้ง

$$= \frac{1}{\sigma^n}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-Z_i^2}{2}\right)\cdot\prod_{n+1}^{n+m}\int_{z}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)dt$$

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

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

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

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

$$= -n\ln\sigma - \frac{(n+m)}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\Sigma(Y_i^2 - 2Y_i\mu_i + \mu_i)^2$$

$$= -n \ln \sigma - \frac{(n+m)}{2} \ln (2\pi) - \frac{1}{2\sigma^2} \Sigma(Y_i^2 - 2Y_i \sum_{j=0}^{3} \beta_j X_{ij} + (\sum_{j=0}^{3} \beta_j X_{ij})^2)$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{-1}{2\sigma^2} \sum_{i=1}^{n+m} (-2Y_i X_{ij} + 2 \sum_{j=0}^{3} \beta_j X_{ij}^2)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n+m} (Y_i - \sum_{j=0}^{3} \beta_j X_{ij}) X_{ij}$$

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

โดยที่ $W_i = \begin{cases} Y_i & ; i=1,..,n \\ \mu_i + \sigma h(Z_i) & ; i=n+1,...,n+m \end{cases}$

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

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

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

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

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

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

$$= -n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2 + \sum_{n+1}^{n+m} Z_i \cdot h(Z_i) = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{n - \sum_{n+1}^{n+m} Z_i \cdot h(Z_i)}$$

ในทางปฏิบัติ การประมาณค่าสูงสุดของพารามิเตอร์ (Maximization Step : M.step) ในรอบที่ K+1 จะประมาณค่าพารามิเตอร์ β^{k+1} โดยวิธีกำลังสองต่ำสุด และประมาณ σ^2_{k+1} ได้ดังต่อไปนี้

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

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

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

เนื่องจากสถิติที่เพียงพอของพารามิเตอร์ สำหรับข้อมูลที่สมบูรณ์ $\sum_{i=1}^{n+m} X_iY_i$ และ $\sum_{i=1}^{n+m} Y_i^2$ แสดงดังนี้

$$E\left(\sum_{i=1}^{n+m} X_iY_i\right) = \sum_{i=1}^{n} X_iY_i + \sum_{n+1}^{n+m} X_i\, E\,(Y_i / Y_i > t)$$

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

โดยที่

$$E(Y_i / Y_i > t) = \hat{\mu}_i + \hat{\sigma}\ h(Z_i)$$

$$E(Y_i^2 / Y_i > t) = \hat{\mu}_i^2 + \hat{\sigma}^2 + \hat{\sigma}(Y_i + \hat{\mu}_i)\ h(Z_i)$$

ดังนั้น Y_i, Y_i^2 , i=n+1,...,n+m มีค่าประมาณดังต่อไปนี้

$$Y_i = E(Y_i / Y_i > t) = \hat{\mu}_i + \hat{\sigma}\ h(Z_i)$$

$$Y_i^2 = E(Y_i^2 / Y_i > t) = \hat{\mu}_i^2 + \hat{\sigma}^2 + \hat{\sigma}\ (Y_i + \hat{\mu}_i)\ h(Z_i)$$

นำค่าประมาณ Y_i, Y_i^2 , i=n+1,...,n+m แทนค่าในสมการ (1) จะได้

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

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

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

$$= \sum_{i=1}^{n} (Y_i - \hat{\mu}_i^{(k)})^2 + \hat{\sigma}_{(k)}^2 \cdot \sum_{n+1}^{n+m} (1 + \frac{[Y_i - \hat{\mu}_i^{(k)}]}{\hat{\sigma}_{(k)}} \cdot h(Z_i)^{(k)})$$

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

$$\hat{\sigma}_{(k+1)}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_i^{(k)})^2 + \hat{\sigma}_{(k)}^2 \cdot \sum_{n+1}^{n+m} (1 + Z_i^{(k)} \cdot h(Z_i)^{(k)})}{n+m}$$

ขั้นตอนในการหาค่าประมาณพารามิเตอร์สำหรับวิธีการภาวะน่าจะเป็นสูงสุด มีขั้นตอนดังนี้

1. ข้อมูลที่นำมาวิเคราะห์ จะถือว่าค่าสังเกตที่ถูกตัดทิ้ง เสมือนเป็นค่าที่ไม่ถูกตัดทิ้ง นำข้อมูลทั้งหมดประมาณค่าพารามิเตอร์ $\hat{\beta}_0, \hat{\sigma}_0$ เริ่มต้น ด้วยวิธีการกำลังสองต่ำสุด

2. เฉพาะข้อมูลที่ถูกตัดทิ้ง หาค่า

$$\hat{\mu}_{i0} = \sum_{j=0}^{3} \hat{\beta}_j X_{ij}$$

$$\hat{Z}_{io} = (tc - \hat{\mu}_{i0}) / \hat{\sigma}_0$$

$$f(Z_i)_0 = \frac{1}{\sqrt{2\pi}} \exp \frac{(-Z_i^2)}{2}$$

$$S(Z_i)_0 = 1-F(Z_i)_0 = \int_z^{\infty} f(t)dt$$

$$h(Z_i)_0 = \frac{f(Z_i)_0}{1 - F(Z_i)_0} \qquad i=n+1,..,n+m$$

tc คือช่วงเวลาที่กำหนดไว้ล่วงหน้า i=1,..,m ; m คือจำนวนข้อมูลที่ถูกตัดทิ้ง

3. ประมาณค่าถูกตัดทิ้งด้วยค่าคาดหวัง ที่มีเงื่อนไข $E(Y_i / Y_i > tc) = w_i$

$$w_{i0} = \hat{\mu}_{i0} + \hat{\sigma}_0 \; h(Z_i)_0 \qquad ; i=n+1,...,n+m$$

ค่าสังเกต $Y'_i = Y_i \delta_i + E(Y_i/Y_i>tc) (1- \delta_i)$

4. นำค่าสังเกต Y'_i มาหาค่าประมาณพารามิเตอร์ $\hat{\beta}_k$ ด้วยวิธีกำลังสองต่ำสุด และหาค่า $\hat{\sigma}^2_{MLE(K)}$ จาก

$$\hat{\sigma}^2_{MLE(K)} = \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_{i0})^2 + \hat{\sigma}_0^2 \sum_{n+1}^{n+m} (1+Z_{i0}h(Z_i)_0)}{n+m}$$

5. เปรียบเทียบค่าประมาณพารามิเตอร์จาก 1 และ 4 คือ $\hat{\beta}_0, \hat{\sigma}^2_0$ กับ $\hat{\beta}_{MLE}, \hat{\sigma}^2_{MLE}$ ถ้ายังไม่เท่ากันให้ทำขั้นต่อไป

6. เฉพาะข้อมูลตัดทิ้ง หาค่า

$$\hat{\mu}_{i(k)} = \sum_{j=0}^{3} \hat{\beta}_{j(k)} X_{i(k)}$$

$$\hat{Z}_{i(k)} = \frac{tc - \hat{\mu}_{i(k)}}{\hat{\sigma}_{MLE(K)}}$$

$$f(Z_i)_k = \frac{1}{\sqrt{2\pi}} \exp \frac{(-Z_i^2)}{2}$$

$$S(Z_i)_k = 1\text{-}F(Z_i)_k = \int_z^{\infty} f(t)dt$$

$$H(Z_i)_k = \frac{F(Z_i)_k}{1 - F(Z_i)k} \qquad i=n+1,..,n+m$$

7. ประมาณค่าที่ถูกตัดทิ้งด้วยค่าคาดหวังที่มีเงื่อนไข $E(Y_i / Y_i > t) = W_i$

$$w_{i(k)} = \hat{\mu}_{ik} + \hat{\sigma}_{MLE(k)} \; h(Z_i)_k \qquad ; i=n+1,...,n+m$$

ดังนั้นค่าสังเกต $Y'_i = Y_i \, \delta_i + E(Y_i/Y_i>tc) \, (1- \delta_i)$

8. นำค่าสังกต Y'_i ไปหาค่าประมาณพารามิเตอร์ $\hat{\beta}_{(k+1)}$ ด้วยวิธีการกำลังสองต่ำสุด และหาค่า $\hat{\sigma}^2_{MLE(K+1)}$ จาก

$$\hat{\sigma}^2_{MLE(K+1)} = \frac{\sum_{i=1}^{n} E(Y_i - \hat{\mu}_{ik})^2 + \hat{\sigma}^2_{MLE(k)} \sum_{n+1}^{n+m} (1+Z_{i(k)}.h(Z_i)_k)}{n+m}$$

9. เปรียบเทียบค่าพารามิเตอร์จากขั้นที่ 4 และ 8 คือ $\hat{\beta}_k$, $\hat{\sigma}^2_{MLE(k)}$ กับ $\hat{\beta}_{k+1}$, $\hat{\sigma}^2_{MLE(k+1)}$ ถ้าหากว่าประมาณในรอบที่ K+1 เท่ากับในรอบที่ K จึงหยุดการกระทำวนซ้ำ จะได้ค่าประมาณ $\hat{\beta}_{MLE}$, $\hat{\sigma}^2_{MLE}$ ในกรณีที่ค่าประมาณในรอบที่ K+1 ไม่เท่ากับค่าประมาณในรอบที่ K ให้นำค่า $\hat{\beta}_{k+1}$, $\hat{\sigma}^2_{k+1}$ แทนค่าในขั้นที่ 6 แล้วกระทำซ้ำจากขั้นที่ 6 ถึง 8 ทำจนกระทั่งค่าประมาณพารามิเตอร์ของรอบปัจจุบัน เท่ากับค่าประมาณพารามิเตอร์รอบที่แล้วจึงหยุด

10. นำค่าประมาณ $\hat{\beta}_{MLE}$ จากขั้นที่ 9 หาค่าประมาณของตัวแปรตาม $\hat{Y}_i = \sum_{j=0}^{3} \hat{\beta}_{MLEj} X_{ij}$ จากนั้นหาค่าความคลาดเคลื่อน ระหว่างค่าประมาณของตัวแปรตามกับค่าจริงในรูปของค่ารากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (RMSE)

$$MSE = \frac{\sum_{uc} (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

เมื่อ n คือ จำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

Y_i คือ ค่าจริงของค่าสังเกตที่ไม่ถูกตัดทิ้ง

2.2.3 วิธีการโมดิไฟแอคชัวเรียล

การประมาณค่าพารามิเตอร์ ด้วยวิธีการโมดิไฟแอคชัวเรียลนี้ เป็นวิธีที่ผู้วิจัยได้ดัดแปลงขึ้นมาจากวิธีการกำลังสองน้อยที่สุด และใช้ตัวถ่วงน้ำหนักจากวิธีการแอคชัวเรียล โดยมีหลักการในการประมาณค่าดังนี้

1. เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง ประมาณค่าพารามิเตอร์ $\hat{\beta}_0$ โดยวิธีการกำลังสองน้อยที่สุด

2. นำข้อมูลทั้งหมดทั้งค่าที่ถูกตัดทิ้งและค่าที่ไม่ถูกตัดทิ้ง มาหาค่าความคลาดเคลื่อน e_i จาก $e_i = Y_i - \hat{\beta}_0 X_i$

3. เรียงลำดับที่ของค่าความคลาดเคลื่อนจากน้อยไปหามาก $e_1, e_2, ..., e_{nm}$ ในกรณีที่ลำดับที่ของค่าความคลาดเคลื่อนของค่าที่ถูกตัดทิ้งและค่าที่ไม่ถูกตัดทิ้ง มีค่าเท่ากัน ให้ลำดับที่ของค่าที่ไม่ถูกตัดทิ้งนำหน้าลำดับที่ของค่าที่ถูกตัดทิ้ง

4. ในช่วง (e_1, e_{nm}) หาค่าถ่วงน้ำหนัก (W_i) จาก

$$W_i = \frac{d_i}{(e_{i+1} - e_i)(n - 0.5\, l_i - 0.5 d_i)}$$

เมื่อ d_i คือ จำนวนค่า Partial Residual ของค่าสังเกตที่ไม่ถูกตัดทิ้ง

l_i คือ จำนวนค่า Partial Residual ของค่าสังเกตที่ถูกตัดทิ้ง

n คือ จำนวนค่า Partial Residual ของค่าสังเกตทั้งหมด

5. ทำการประมาณค่าพารามิเตอร์ $\hat{\beta}_{MAM}$ จาก

$$\hat{\beta}_{MAM} = \frac{\sum_{\mu c} W_i Y_i (X_i - \overline{X})}{\sum_{\mu c} W_i (Xi - \overline{X})^2} \quad \text{เมื่อ} \quad \overline{X} = \sum_{\mu c} W_i X_i$$

6. แทนค่า $\hat{\beta}_{MAM}$ ในขั้นที่ 5 ลงในขั้นที่ 2 แล้วทำการกระทำวนซ้ำจากขั้นที่ 2 ถึงขั้นที่ 5 ทำไปจนกระทั่งค่าของ $\hat{\beta}_{MAM}$ ในรอบปัจจุบันเท่ากับค่าในรอบที่แล้ว จึงหยุด ในบางกรณี ค่า $\hat{\beta}_{MAM}$ จะแกว่งอยู่ระหว่างค่า 2 ค่า ในกรณีนี้จะใช้ค่าเฉลี่ยของทั้ง 2 ค่านั้นเป็นค่าประมาณของ $\hat{\beta}_{MAM}$

7. นำค่าประมาณ $\hat{\beta}_{MAM}$ จากขั้นที่ 6 หาค่าประมาณของตัวแปรตาม $\hat{Y}_i = \sum_{j=0}^{3} \hat{\beta}_{MAMj} X_{ij}$ จากนั้นหาค่าความคลาดเคลื่อน ระหว่างค่าประมาณของตัวแปรตามกับค่าจริงในรูปของค่ารากที่สองของค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (RMSE)

$$MSE = \frac{\sum_{uc} (Y_i - \hat{Y}_i)^2}{n}$$

$$RMSE = \sqrt{MSE}$$

เมื่อ n คือ จำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

Y_i คือ ค่าจริงของค่าสังเกตที่ไม่ถูกตัดทิ้ง