Linear Regression

Petchara Pattarakijwanich

SCPY208, 3 May 2021

<u>ตัวอย่าง</u> ข้อมูลคู่อันดับ (x, y) และความคลาดเคลื่อนมีค่าดังนี้

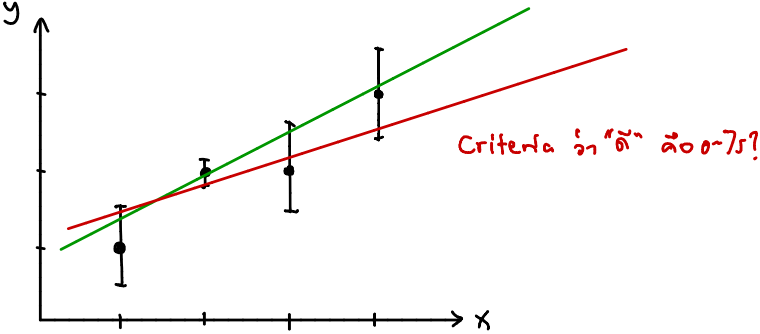
ความคลาดเคลื่อนของ y

x	1	2	3	4
y	1	2	2	3
σ_y	0.5	0.1	0.5	0.5

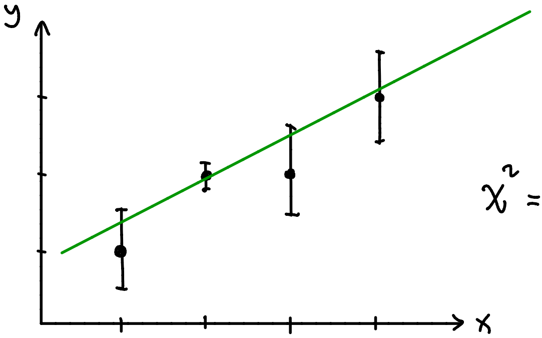
ดีกว่าจุดอื่นๆ

จงหาสมการเส้นตรงที่อธิบายข้อมูลชุดนี้ได้ดีที่สุด

เส้นไหนดีกว่า?

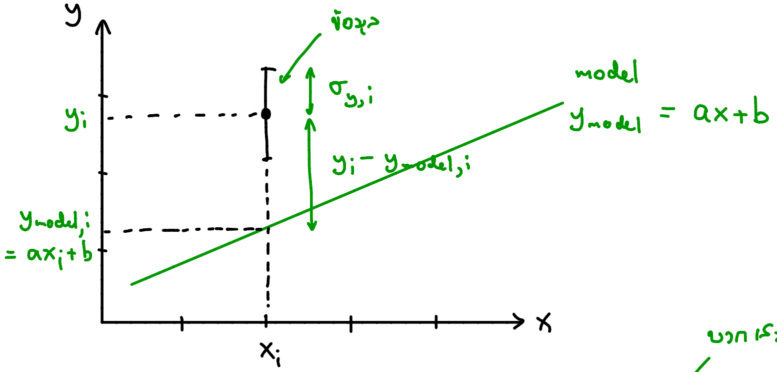


Least Squares Fitting



$$\chi^2 = \sum_i \left(\frac{y_i - y_{model,i}}{\sigma_{y,i}} \right)^2$$

เส้นฟิตข้อมูลได้ดี $\Rightarrow$ χ^2 มีค่าน้อย



บวกเสมอ

$$\left(\frac{y_i - y_{model,i}}{\sigma_{y,i}}\right)^2 = \left(\text{โมเดลกับข้อมูล ต่างกัน กี่เท่าของ Error Bar}\right)^2$$

มีค่าน้อย ⇒ โมเดลอธิบายข้อมูลได้ดี ⇒ โมเดลดี

$$\chi^2 = \sum_i \left(\frac{y_i - y_{model,i}}{\sigma_{y,i}}\right)^2 = \text{ความ "ดี" ของโมเดลแบบรวมทุกจุดข้อมูล}$$

ค่าน้อย ⇒ โมเดลฟิตข้อมูลทั้งหมดได้ดี

⇒ หาเส้นตรงที่ χ^2 น้อยสุดยังไง?

x	1	2	3	4
y	1	2	2	3
σ_y	0.5	0.1	0.5	0.5

$$\chi^2 = \sum\left(\frac{y - y_{model}}{\sigma_y}\right)^2 = \sum\left(\frac{y - ax - b}{\sigma_y}\right)^2$$

$$= \left(\frac{1 - a - b}{0.5}\right)^2 + \left(\frac{2 - 2a - b}{0.1}\right)^2 + \left(\frac{2 - 3a - b}{0.5}\right)^2 + \left(\frac{3 - 4a - b}{0.5}\right)^2$$

$$= 4(1-a-b)^2 + 100(2-2a-b)^2 + 4(2-3a-b)^2 + 4(3-4a-b)^2$$

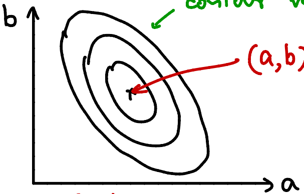
$$\chi^2 = Ca^2 + Db^2 + Eab + Fa + Gb + H$$

ตัวเลข (ค่าคงที่)

$C, D > 0$

ฟังก์ชันของ (a,b) $\Downarrow$

contour ของ χ^2



(a,b) ที่ทำให้ χ^2 น้อยที่สุด!

หา a, b ที่ χ^2 น้อยที่สุด?

$\frac{\partial \chi^2}{\partial a} = 0$, $\frac{\partial \chi^2}{\partial b} = 0$ $\Rightarrow$ 2 สมการ 2 ตัวแปร

หา a,b ให้ผลต่ำที่สุด
(เรียกว่า)

$$\chi^2 = \sum \left(\frac{y - ax - b}{\sigma_y} \right)^2$$

1. $\frac{\partial \chi^2}{\partial a} = 0$, $\frac{\partial \chi^2}{\partial b} = 0$
2. 2 สมการ ในรูป 2 ตัวแปร (a,b)
3. แก้ 2 สมการ

(1)

$$\chi^2 = \sum \left(\frac{y - ax - b}{\sigma_y} \right)^2$$

$$\frac{\partial \chi^2}{\partial a} = \sum \left[\frac{\partial}{\partial a} \left(\frac{y - ax - b}{\sigma_y} \right)^2 \right]$$

$$= \sum \left[2 \left(\frac{y - ax - b}{\sigma_y} \right) \left(- \frac{x}{\sigma_y} \right) \right]$$

$$= -2 \sum \left[\frac{xy}{\sigma^2} - \frac{ax^2}{\sigma^2} - \frac{bx}{\sigma} \right]$$

$$\frac{\partial \chi^2}{\partial a} = -2 \left[\sum \left(\frac{xy}{\sigma^2} \right) - a \sum \left(\frac{x^2}{\sigma^2} \right) - b \sum \left(\frac{x}{\sigma^2} \right) \right]$$

$$\frac{\partial \chi^2}{\partial a} = 0 \quad \Rightarrow \quad a \sum \left(\frac{x^2}{\sigma^2} \right) + b \sum \left(\frac{x}{\sigma^2} \right) = \sum \left(\frac{xy}{\sigma^2} \right)$$

$\Sigma(\ldots) =$ ตัวเลข คำนวณได้จาก ข้อมูล

$$C_1 a + C_2 b = C_3$$

$$\chi^2 = \sum \left(\frac{y - ax - b}{\sigma_y} \right)^2$$

$$\frac{\partial \chi^2}{\partial b} = \sum \left[\frac{\partial}{\partial b} \left(\frac{y - ax - b}{\sigma} \right)^2 \right] = \sum \left[2 \left(\frac{y - ax - b}{\sigma} \right) \left(- \frac{1}{\sigma} \right) \right]$$

$$= -2 \left[\sum \left(\frac{y}{\sigma^2} \right) - a \sum \left(\frac{x}{\sigma^2} \right) - b \sum \left(\frac{1}{\sigma^2} \right) \right]$$

$$\frac{\partial \chi^2}{\partial b} = 0 \Rightarrow \quad a \sum \left(\frac{x}{\sigma^2} \right) + b \sum \left(\frac{1}{\sigma^2} \right) = \sum \left(\frac{y}{\sigma^2} \right)$$

$\sum(\ldots) =$ ตัวเลข

$$C_4 a + C_5 b = C_6$$

(2)

$$\boxed{\begin{aligned} a\sum\left(\frac{x^2}{\sigma^2}\right) + b\sum\left(\frac{x}{\sigma^2}\right) &= \sum\left(\frac{xy}{\sigma^2}\right) \quad &-(1) \\ a\sum\left(\frac{x}{\sigma^2}\right) + b\sum\left(\frac{1}{\sigma^2}\right) &= \sum\left(\frac{y}{\sigma^2}\right) \quad &-(2) \end{aligned}}$$

2 สมการ 2 ตัวแปร $\Rightarrow$ แก้หา a, b

เขียนให้อยู่ในรูป matrix

$$\boxed{\begin{bmatrix} \sum\left(\frac{x^2}{\sigma^2}\right) & \sum\left(\frac{x}{\sigma^2}\right) \\ \sum\left(\frac{x}{\sigma^2}\right) & \sum\left(\frac{1}{\sigma^2}\right) \end{bmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum\left(\frac{xy}{\sigma^2}\right) \\ \sum\left(\frac{y}{\sigma^2}\right) \end{pmatrix}}$$

③

$$a\sum\left(\frac{x^2}{\sigma^2}\right) + b\sum\left(\frac{x}{\sigma^2}\right) = \sum\left(\frac{xy}{\sigma^2}\right)$$

$$a\sum\left(\frac{x}{\sigma^2}\right) + b\sum\left(\frac{1}{\sigma^2}\right) = \sum\left(\frac{y}{\sigma^2}\right)$$

แก้หา a,b ได้

- ลองทดหน่วย ของแต่ละพจน์

$[b] = [y]$

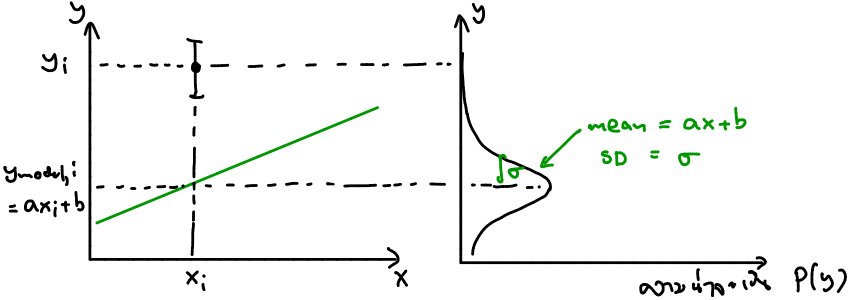
$[a] = \frac{[y]}{[x]}$

$$a = \frac{\sum\left(\frac{1}{\sigma^2}\right)\sum\left(\frac{xy}{\sigma^2}\right) - \sum\left(\frac{x}{\sigma^2}\right)\sum\left(\frac{y}{\sigma^2}\right)}{\sum\left(\frac{1}{\sigma^2}\right)\sum\left(\frac{x^2}{\sigma^2}\right) - \left[\sum\left(\frac{x}{\sigma^2}\right)\right]^2}$$

$$b = \frac{\sum\left(\frac{x^2}{\sigma^2}\right)\sum\left(\frac{y}{\sigma^2}\right) - \sum\left(\frac{x}{\sigma^2}\right)\sum\left(\frac{xy}{\sigma^2}\right)}{\sum\left(\frac{1}{\sigma^2}\right)\sum\left(\frac{x^2}{\sigma^2}\right) - \left[\sum\left(\frac{x}{\sigma^2}\right)\right]^2}$$

determinant ของ matrix

ความหมายทางสถิติของ Least Squares Fitting



สมมติว่า $y = ax + b$ เป็น "ความจริง" ⇒ หาความน่าจะเป็นที่จะวัด y ที่ x_i ได้ y_i

[y เป็น random variable

mean = $ax+b$

variance = σ^2]

Normal Distribution

หาความน่าจะเป็นที่จะ "วัด" ได้ y_i จาก Normal Distribution

mean $= ax_i + b$, SD $= \sigma_{y,i}$

$$\underbrace{P(y_i \mid y = ax + b)} = \frac{1}{\sqrt{2\pi}\,\sigma_{y,i}} \exp\left[-\frac{(y_i - ax_i - b)^2}{2\sigma_{y,i}^2}\right]$$

ความน่าจะเป็นที่จะวัดได้ข้อมูล 1 จุด
ถ้าเส้นตรงที่ฟิตเป็นความจริง

ความน่าจะเป็นที่จะวัดได้ข้อมูลทั้งชุด (n จุด) ถ้าเส้นตรงเป็นความจริง

$$P(\text{ชุดข้อมูล} \mid y = ax+b) = P(\text{จุดที่ } 1) \times P(\text{จุดที่ } 2) \times \ldots \times P(\text{จุดที่ } n)$$

independent

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\,\sigma_{y,i}} \exp\left[-\frac{(y_i - ax_i - b)^2}{2\sigma_{y,i}^2} \right] \right]$$

$$= \frac{1}{(\sqrt{2\pi})^n (\sigma_{y,1} \cdot \sigma_{y,2} \cdots \sigma_{y,n})} \exp\left[\sum_{i=1}^{n} \left(-\frac{(y_i - ax_i - b)^2}{2\sigma_{y,i}^2} \right) \right]$$

$$= \frac{1}{(\sqrt{2\pi})^n (\sigma_{y,1} \cdot \sigma_{y,2} \cdots \sigma_{y,n})} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{(y_i - ax_i - b)^2}{\sigma_{y,i}^2} \right) \right]$$

$$P(\text{ชุดข้อมูล} \mid y = ax+b) = \frac{1}{(\sqrt{2\pi})^n (\sigma_{y,1} \cdot \sigma_{y,2} \cdots \sigma_{y,n})} \exp\left(-\frac{1}{2} \chi^2 \right)$$

$$P(\text{ชุดข้อมูล} \mid y = ax+b) = \frac{1}{(\sqrt{2\pi})^n (\sigma_{y,1} \cdot \sigma_{y,2} \cdots \sigma_{y,n})} \exp\left(-\frac{1}{2}\chi^2\right)$$

ความน่าจะเป็นที่จะวัดได้ชุดข้อมูล
เมื่อสมมติ เรียกว่า ⇒ "Likelihood" $\mathcal{L}$

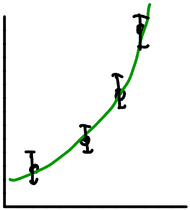
$$\ln \mathcal{L} = \underbrace{\ln\left[\frac{1}{(\sqrt{2\pi})^n (\sigma_{y,1} \cdot \sigma_{y,2} \cdots \sigma_{y,n})}\right]}_{\text{ค่าคงที่ ไม่สำคัญ}} - \frac{1}{2}\chi^2$$

⇒ ถ้า χ^2 น้อยสุด ⇒ $\mathcal{L}$ มากสุด

(a,b ที่ให้ χ^2 น้อย, ฟิตดี) ⇐ (สมมติ a,b ความน่าจะเป็นที่ได้ชุดข้อมูล มากสุด)

Least Squares Fitting ⟺ Maximum Likelihood Method

Least Square Fitting ด้วย Quadratic Function



$$y_{model} = ax^2 + bx + c$$

$$\chi^2 = \sum \left(\frac{y - ax^2 - bx - c}{\sigma_y^2} \right)$$

- หา a, b, c ที่ทำให้ χ^2 น้อยสุด
- $\frac{\partial \chi^2}{\partial a} = 0$, $\frac{\partial \chi^2}{\partial b} = 0$, $\frac{\partial \chi^2}{\partial c} = 0$
- 3 สมการ (เชิงเส้น) 3 ตัวแปร (a, b, c) $\Rightarrow$ แก้ได้

Least Square Fitting ด้วย Polynomial Function

$$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

$\Rightarrow$ $n+1$ สมการ $n+1$ ตัวแปร

$$\begin{pmatrix} \sum\left(\frac{x^{2n}}{\sigma^2}\right) & \longrightarrow & \sum\left(\frac{x^{n}}{\sigma^2}\right) \\ \downarrow & & \downarrow \\ \sum\left(\frac{x^{n}}{\sigma^2}\right) & \longrightarrow & \sum\left(\frac{1}{\sigma^2}\right) \end{pmatrix} \begin{pmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} \sum\left(\frac{x^n y}{\sigma^2}\right) \\ \vdots \\ \sum\left(\frac{y}{\sigma^2}\right) \end{pmatrix}$$

กำลังของ x ลดช่องละ 1

Least Square Fitting ด้วย “Linear” Function

สมมติ $y_{model} = \underline{\underline{a}}\,x + \underline{\underline{b}}\left(\frac{1}{x}\right) + \underline{\underline{c}}\,e^{x} + \underline{\underline{d}}\sin x$

“Linear” function ในรูปของ a, b, c, d

- เขียน χ^2
- $\frac{\partial \chi^2}{\partial a} = 0$, ...
- 4 สมการเชิงเส้น 4 ตัวแปร (a, b, c, d)

Least Square Fitting ด้วย Exponential Function

$$y_{model} = ae^{bx}$$ หา a,b ที่ฟิตข้อมูลได้ดี ยังไง?

วิธีแรก หาตรงๆ

$$\chi^2 = \sum \left(\frac{y - ae^{bx}}{\sigma_y} \right)^2$$

Error bar ของ y

- หา a,b ที่ χ^2 น้อยที่สุด (หาด้วย numerical methods)

วิธีที่ 2 เปลี่ยนเป็น linear

$$\ln y = bx + \ln a$$

$$\tilde{\chi}^2 = \sum \left(\frac{\ln y - bx - \ln a}{\sigma_{\ln y}} \right)^2$$

Error bar ของ $\ln y = \frac{\sigma_y}{y}$

$$\tilde{\chi}^2 = \sum \frac{y^2}{\sigma_y^2} (\ln y - bx - \ln a)^2$$

พิสูจน์ได้ว่า ถ้า $\tilde{\chi}^2$ น้อยที่สุด

$\Rightarrow$ χ^2 น้อยที่สุด (ถ้า σ น้อยๆ)

Least Square Fitting ด้วย Power-Law Function

$$y_{model} = a x^b$$

$$\Rightarrow \ln y = b \ln x + \ln a$$

$\Rightarrow$ ใช้ linear regression แต่ Error bar ต้องเป็นของ $\ln y$