Linear Regression

Petchara Pattarakijwanich

Introduction to Data Science, 9 September 2022

Goal of this week

- Simple Linear Regression
  - Least Squared Fitting
  - Significance and p-value
- Multiple Regression
- Interaction and Non-linear Terms
- Some technical details

- (Logistic Regression)

ข้อดีของ Regression

- ระบบรวมเบ็ดเสร็จ => ง่าย
- เก่าแก่ => เข้าใจดี
- ฐานของหลายๆ วิธี
  (Ridge, Lasso, Subset selection)

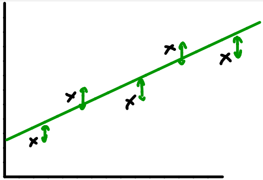
Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

noise

ค่า y เมื่อ $x = 0$

x เพิ่ม 1 หน่วย y เพิ่มเท่าไหร่



หา β_0, β_1 ที่ "ฟิตดี" $\Rightarrow$ ทำให้ $RSS = \sum_i (y - \beta_0 - \beta_1 x)^2$ น้อยที่สุด

Residual sum of Square

$$\chi^2 = \sum_i \frac{(y_{data} - y_{model})^2}{\sigma^2}$$

คล้ายกัน

$$RSS = \sum_i (y_{data} - y_{model})^2$$

$RSS = \chi^2$ ถ้า σ คงที่

Linear Least Squared Fitting

หา β_0, β_1 ที่ทำให้ $\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$ น้อยที่สุด

$$RSS = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

$$= (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + \dots + (y_n - \beta_0 - \beta_1 x_n)^2$$

$x_1, \dots, x_n$ & $y_1, \dots, y_n$ คือจุดข้อมูล = ค่าคงที่

$\Rightarrow$ RSS เป็นฟังก์ชันของ β_0, β_1

$$RSS = a\beta_0^2 + b\beta_1^2 + c\beta_0\beta_1 + d\beta_0 + e\beta_1 + f$$

$a = n$, $b = \sum (x_i)^2$

เป็นบวก

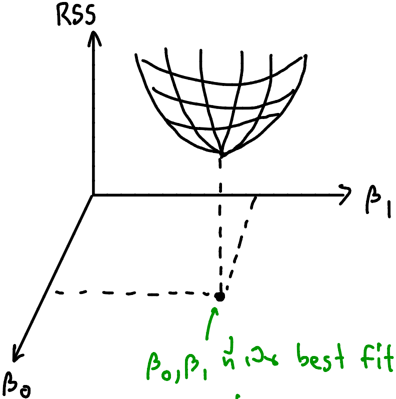
Linear Least Squared Fitting

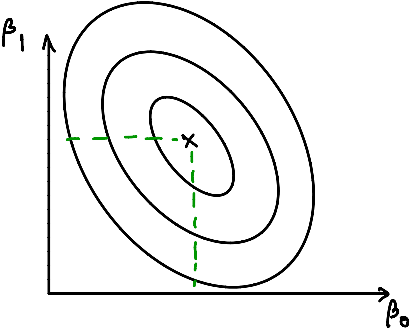
$$RSS = a\beta_0^2 + b\beta_1^2 + c\beta_0\beta_1 + d\beta_0 + e\beta_1 + f$$

$a = n$, $b = \sum (x_i)^2$

เป็นบวก

Paraboloid ใน 3 มิติ





β_0, β_1 ที่เป็น best fit

RSS น้อยสุด

Linear Least Squared Fitting

$$RSS = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

จุดต่ำสุด $\quad \dfrac{\partial(RSS)}{\partial \beta_0} = 0 \quad , \quad \dfrac{\partial(RSS)}{\partial \beta_1} = 0$

$$\frac{\partial RSS}{\partial \beta_0} = \sum_i 2(y_i - \beta_0 - \beta_1 x_i)(-1) = -2\left[\sum_i y_i - \beta_0 \underbrace{\sum_i 1}_{=n} - \beta_1 \sum x_i\right] = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = \sum_i 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2\left[\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2\right] = 0$$

$$\begin{aligned} \beta_0 n + \beta_1 \sum_i x_i &= \sum_i y_i \\ \beta_0 \sum_i x + \beta_1 \sum_i x_i^2 &= \sum_i x_i y_i \end{aligned} \Rightarrow \begin{pmatrix} n & \sum x \\ \sum x & \sum x^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum y \\ \sum xy \end{pmatrix}$$

$\Rightarrow$ 2 สมการ 2 ตัวแปร (β_0, β_1)

Linear Least Squared Fitting

ดังนั้น

$$\beta_1 = \frac{n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}$$

$$\beta_0 = \frac{1}{n}\left[\sum_i y_i - \beta_1 \sum_i x_i\right]$$

Significance and p-value

$$y = \beta_0 + \beta_1 x + \epsilon$$

y สัมพันธ์กับ x อย่างมีนัยสำคัญมั้ย? $\Longleftrightarrow$ $\beta_1 \neq 0$ อย่างมีนัยสำคัญมั้ย?

Hypothesis test

- Null Hypothesis $\beta_1 = 0$
- Alternative hypothesis $\beta_1 \neq 0$

$\Downarrow$

หา Probability ที่จะได้ข้อมูลนี้ ถ้า $\beta_1 = 0$ [F-statistics]

$\Downarrow$

หา probability ที่จะได้ β_1 ค่านี้ ถ้า $\beta_1 = 0$ [p-value]

ยิ่งน้อย ยิ่งมีนัยสำคัญมาก

Multiple Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

ระนาบใน $n+1$ มิติ

β_0 คือค่า y เมื่อ $x_1 = x_2 = \dots = x_n = 0$

β_1 คือ Δy เมื่อ $\Delta x_1 = 1$ เมื่อ x ที่เหลือคงที่

หา $\beta_0, \beta_1, \dots, \beta_n$ ที่ทำให้ฟิตดี $\Rightarrow$ $RSS = \sum (y - \beta_0 - \beta_1 x_1 - \dots - \beta_n x_n)^2$ มีค่าน้อยสุด

$\Rightarrow \dfrac{\partial RSS}{\partial \beta_i} = 0 \Rightarrow$ $n+1$ สมการ $n+1$ ตัวแปร $[\beta_0, \beta_1, \dots \beta_n]$

$\Rightarrow$ แก้หา β ได้

Interaction Term and Non-linear Term

- Interaction term: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

$$y = \beta_0 + (\beta_1 + \beta_{12} x_2) x_1 + \beta_2 x_2$$ [ความชันของ x_1 ขึ้นกับ x_2]

$$y = \beta_0 + \beta_1 x_1 + (\beta_2 + \beta_{12} x_1) x_2$$ [ความชันของ x_2 ขึ้นกับ x_1]

- Non-linear term: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_1 x_1^2 + \alpha_2 x_2^2$

ไม่ linear ใน x แต่ยัง linear ใน β

$\Rightarrow$ หา β ที่ทำให้ RSS น้อยสุด $\Rightarrow$ ระบบสมการเชิงเส้นของ β

Technical Problems

- Parameter normalization?

  Ex. x_1 มีค่าในช่วง $\approx 10^{-5}$
  x_2 มีค่าในช่วง $\approx 10^{20}$ $\Rightarrow$ overflow, underflow, Round-off error $\}$ แก้ด้วย normalization

  Scale x_i ให้มี $\bar{x}_i = 0$, $\mathrm{var}(x_i) = 1$

- Error bars?

$$RSS = \sum (y - y_{model})^2 \quad \Rightarrow \quad \chi^2 = \sum \frac{(y - y_{model})^2}{\sigma^2}$$

- Qualitative input?

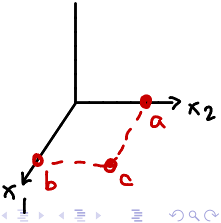
  2 กลุ่ม $\{a, b\}$

$$x = \begin{cases} 0, & a \\ 1, & b \end{cases}$$

  3 กลุ่ม $\{a, b, c\}$

$$x_1 = \begin{cases} 0, & a \\ 1, & \text{ไม่ } a \end{cases}$$

$$x_2 = \begin{cases} 0, & b \\ 1, & \text{ไม่ } b \end{cases}$$



Technical Problems

- Non-linearity

  \- เพิ่ม Non-linear หรือ Interaction terms

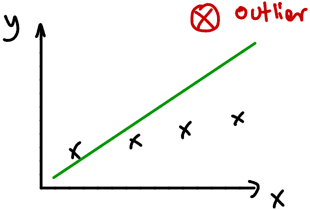
  $\{x_i^2, x_i^3, \dots\}$ $\{x_i x_j, \dots\}$

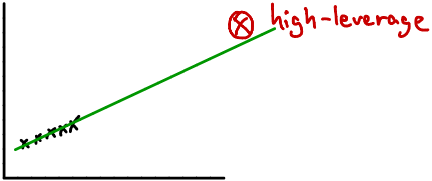
- Collinearity

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

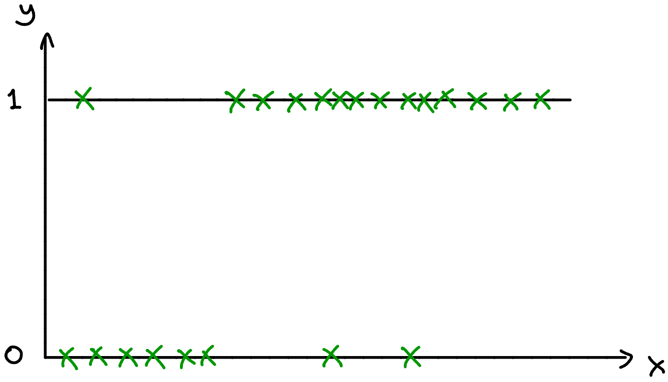
  ถ้า $x_1\, x_2$ ไม่อิสระต่อกัน ? $x_1 = k x_2 \Rightarrow$ ปริมาณ information น้อยกว่า เพราะ $x_1\, x_2$ อิสระ

- Outliers & High-leverage points





Logistic Regression Example

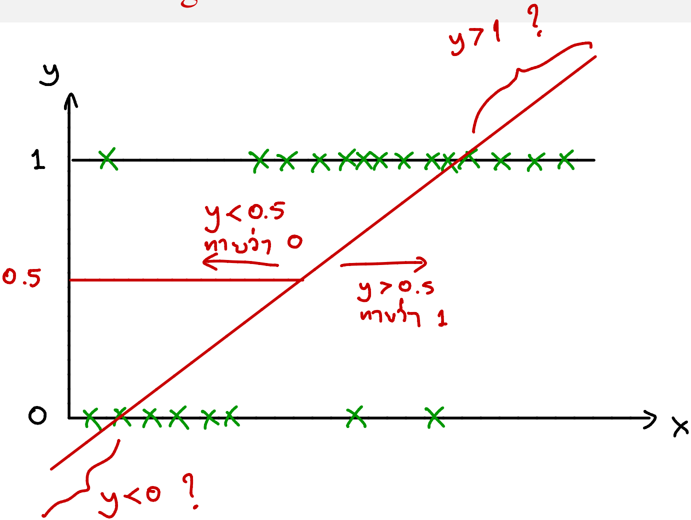


x = เกรดเฉลี่ยปี 1

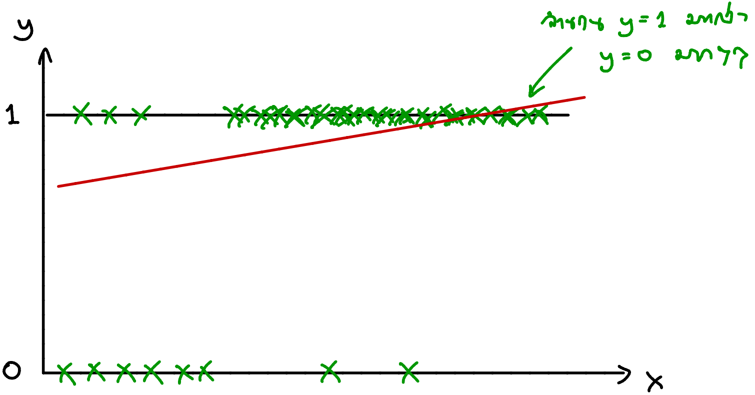
y = เรียนจบมั้ย $= \begin{cases} 1, & \text{จบ} \\ 0, & \text{ไม่จบ} \end{cases}$

ต้องการทำนาย ค่า y จาก x ?

Why not linear regression?



Why not linear regression?



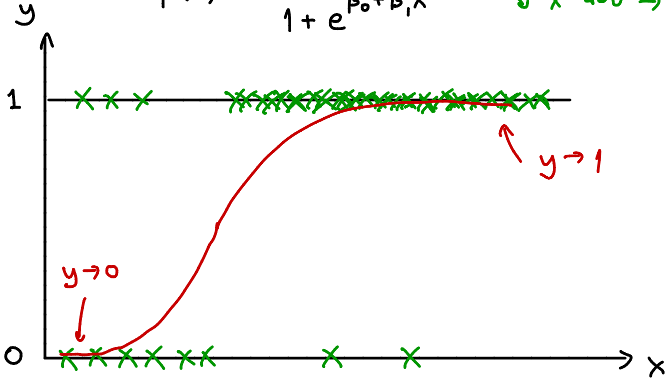
Linear regression ใช้ไม่ค่อยได้

Logistic Function

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

x มาก $\rightarrow p(x) \approx 1$

x น้อย $\rightarrow p(x) \approx 0$

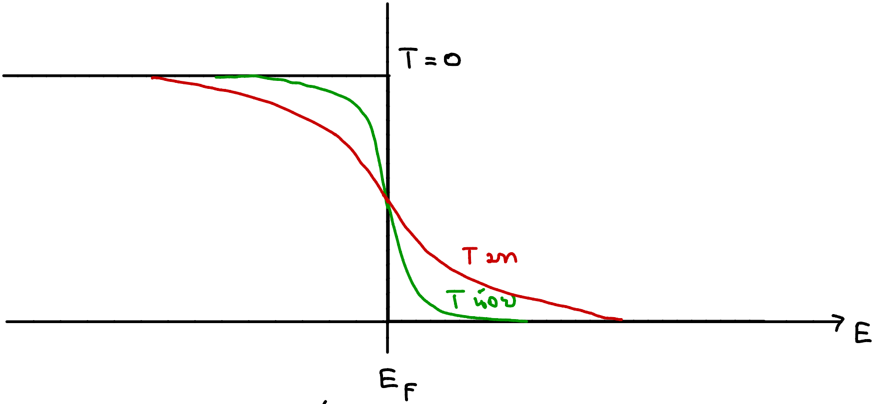


สิ่งที่ต้องการ

- ในช่วง $[0, 1]$
- monotone

Logistic Function

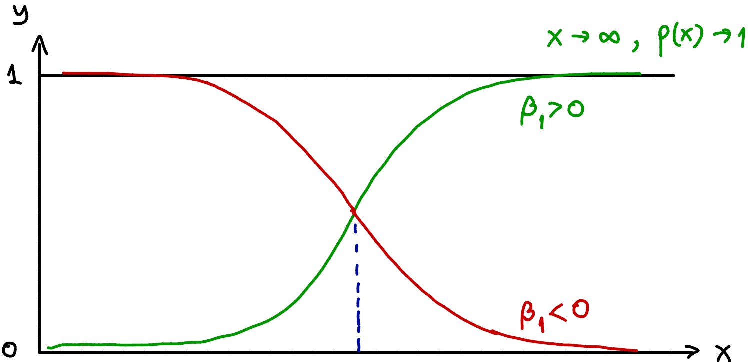
Fermi-Dirac Distribution $\quad f(E) = \dfrac{e^{-(E-E_F)/kT}}{1 + e^{-(E-E_F)/kT}}$



$$f(E) = \begin{cases} \to 0 & \text{เมื่อ } E \text{ สูง } (E \to \infty) \\ \to 1 & \text{เมื่อ } E \text{ น้อย } (E \to -\infty) \\ 0.5 & \text{เมื่อ } E = E_F \end{cases}$$

Logistic Function

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



$x \to -\infty$, $p(x) \to 0$

จุดกลาง $\beta_0 + \beta_1 x = 0$

ชันมากรอบจุดนี้

$$\Rightarrow \quad \beta_1 \left(x + \boxed{\frac{\beta_0}{\beta_1}} \right) = 0$$

ตำแหน่งกึ่งกลาง

ความกว้าง

Odds and logit function

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{\alpha}}{1+e^{\alpha}}$$

[แก้หา α ในรูปของ p]

$$(1+e^{\alpha})\,p = e^{\alpha}$$

$$p = (1-p)\,e^{\alpha}$$

$$e^{\alpha} = \frac{p}{1-p}$$

$$\alpha = \beta_0+\beta_1 x = \underbrace{\ln\left(\frac{p}{1-p}\right)}_{\text{logit function}}$$

logit(p) เป็น linear function ของ x

logit function ← logistic unit

ในรูปค่า "odd"

อัตราส่วนความน่าจะเป็น ที่ event จะเกิด ต่อ ไม่เกิด

$$\text{odd} = \frac{p}{1-p}$$

⇒ ตีความ logistic function ว่าเป็น probability ที่ $y=1$

Logistic Regression

ฟิตหา β_0, β_1 ด้วย "Maximum Likelihood"

$\mathcal{L}$ = Probability ที่จะเห็นได้ชุดข้อมูลนี้ ถ้าโมเดลมีค่า β_0, β_1

likelihood

$= P(\text{data} \mid \text{model})$

$= P(\text{จุดที่ } 1) \times P(\text{จุดที่ } 2) \times \dots \times P(\text{จุดที่ } n)$

สมมติ $y_1 = 1$ $\quad y_2 = 0$ $\quad \dots \quad y_n = 0$

$P(\text{จุดที่ } 1) = p(x_1)$ $\qquad P(\text{จุดที่ } 2) = 1 - p(x_2)$

$$\mathcal{L} = \prod_{\substack{\text{ทุก } i \text{ ที่} \\ y_i = 1}} p(x_i) \prod_{\substack{\text{ทุก } i \text{ ที่} \\ y_i = 0}} \left(1 - p(x_i)\right)$$

$\mathcal{L}$ เป็นฟังก์ชันของ β_0, β_1

$\Rightarrow$ หา β_0, β_1 ที่ $\mathcal{L}$ มากสุด

Logistic Regression

$$\mathcal{L} = \prod_{\substack{\text{ทุก } i \text{ ที่} \\ y_i = 1}} p(x_i) \prod_{\substack{\text{ทุก } i \text{ ที่} \\ y_i = 0}} \left(1 - p(x_i)\right)$$

likelihood = 1

for i = 1, ... n_{data}

if y_i == 0

likelihood *= $p(x_i)$

if y_i == 1

likelihood *= $1 - p(x_i)$

$$\mathcal{L} = \prod_{\text{ทุก } i} p(x_i)^{y_i} \left(1 - p(x_i)\right)^{1 - y_i}$$

likelihood = 1

for i = 1, ..., n_{data}

likelihood *= $p^y (1-p)^{1-y}$

ในทางปฏิบัติ ใช้ $\ln \mathcal{L}$ แทน $\mathcal{L}$

$\mathcal{L}$ มากที่สุด $\Rightarrow$ $\ln \mathcal{L}$ มากที่สุด