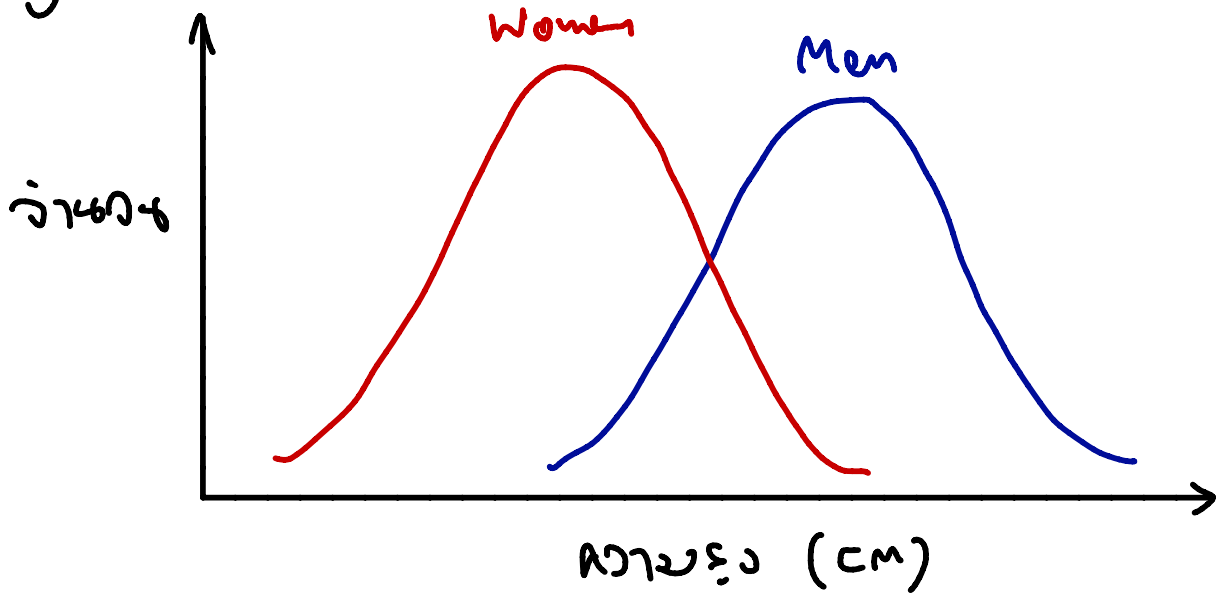
# Model Selection

Petchara Pattarakijwanich

Introduction to Data Science, 21 October 2022

HW 3




M : $x_{m1}, x_{m2}, \dots, x_{m508}$

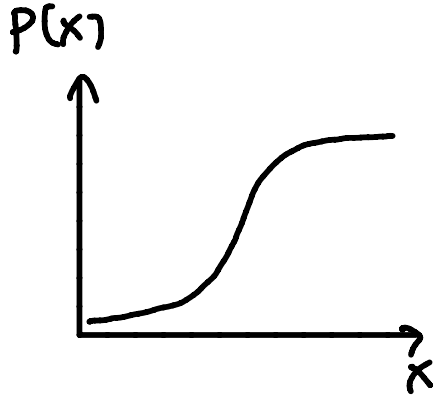
W : $x_{w1}, x_{m2}, \dots, x_{w500}$

หา $\beta_0, \beta_1$ ที่ทำให้ likelihood ได้ค่าสูงสุด!

Logistic function :

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

$$p(\text{W ที่ความสูง } x) = 1 - p(x)$$



$\mathcal{L}$ = ความน่าจะเป็นที่จะได้ข้อมูลชุดนี้ สำหรับ $\beta_0, \beta_1$ ต่างๆ

prob ของกลุ่ม Men

$$= \left[ p(x_{m1})\, p(x_{m2})\, p(x_{m3}) \cdots p(x_{m508}) \right] \cdot$$

$$\left[ (1-p(x_{w1}))(1-p(x_{w2})) \cdots (1-p(x_{w500})) \right]$$

prob ของกลุ่ม Women

$$= \prod_{\text{men}} p(x_{mi}) \prod_{\text{women}} (1-p(x_{wi}))$$

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

$$\mathcal{L} = \mathcal{L}(\beta_0, \beta_1)$$

$\Rightarrow$ Optimize หา $\beta_0, \beta_1$ ที่ $\mathcal{L}$ มากสุด

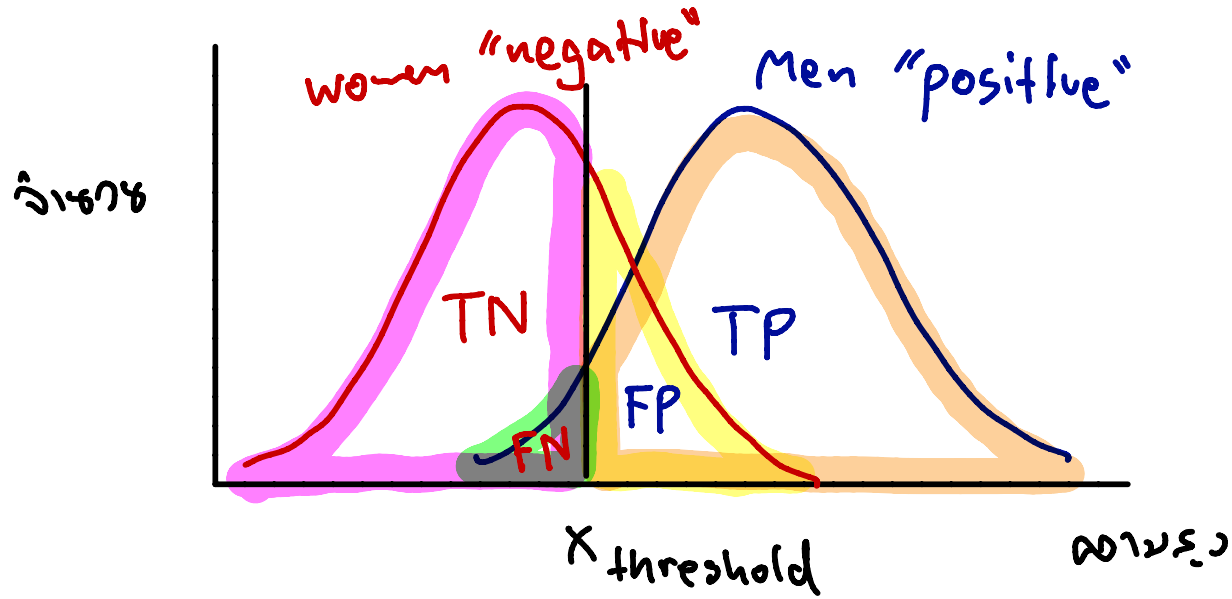
$$\mathcal{L} = \prod_{\text{men}} p(x_{mi}) \prod_{\text{women}} (1 - p(x_{wi}))$$ ← underflow ถ้าข้อมูลเยอะ

$$\ln \mathcal{L} = \sum_{\text{men}} \ln p(x_{mi}) + \sum_{\text{women}} \ln (1 - p(x_{wi}))$$ ← ข้อมูลไม่ underflow

$$-\ln \mathcal{L} = -\sum_{\text{men}} \ln p(x_{mi}) - \sum_{\text{women}} \ln (1 - p(x_{wi}))$$

↑ $-\ln \mathcal{L}$ มีค่าน้อยสุด เมื่อ $\mathcal{L}$ มีค่ามากสุด

$\Rightarrow$ หา $\beta_0, \beta_1$ ที่ minimize $-\ln \mathcal{L}$




ทายว่า Men ถ้า $x \geq x_{threshold}$

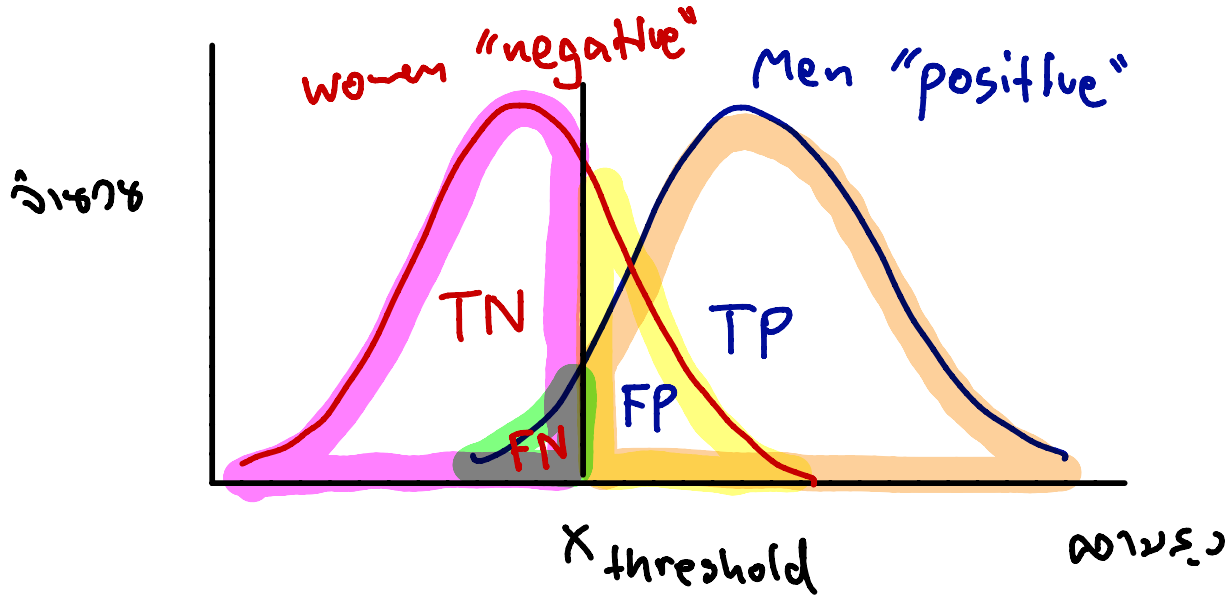
ทายว่า women ถ้า $x < x_{threshold}$

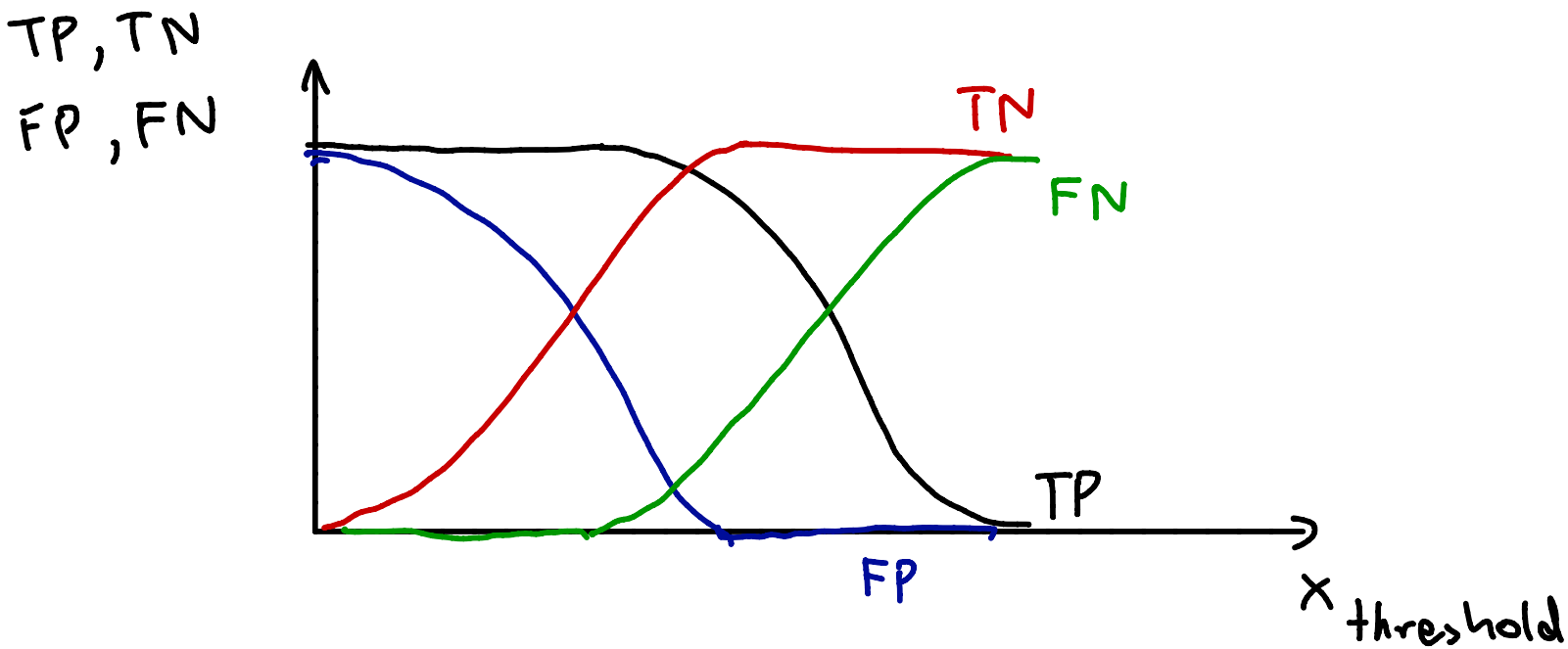
TP = จำนวนชาย ที่สูงกว่า $x_t$

TN = จำนวนหญิง เตี้ยกว่า $x_t$

FP = จำนวนหญิง สูงกว่า $x_t$

FN = จำนวนชาย เตี้ยกว่า $x_t$


จำนวน
women "negative"
Men "positive"
TN
TP
FP
FN
x threshold
ความสูง


TP, TN
FP, FN
TN
FN
TP
FP
x threshold

# Goal of this week

- Standard Linear Regression and Its Limits
- Model Selection Methods
  - Mallow's $C_p$
  - Akaike Information Criteria (AIC)
  - Bayesian Information Criteria (BIC)
  - (Reduced $\chi^2$)
- Subset Selection
  - Best Subset Selection
  - Forward/Backward Step-wise Subset Selection

# Standard Linear Regression and Its Limits

input : $\{x_1, x_2, \ldots, x_p\}$ output : $y$

Linear Regression :

$$y_{model} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

(+ interaction terms + non-linear terms)

$\Rightarrow$ หา $\beta_0, \beta_1, \ldots, \beta_p$ ที่ทำให้ $RSS = \sum (y - y_{model})^2$ น้อยสุด

[ เช่น RSS , $\frac{\partial RSS}{\partial \beta_0} = 0$, $\frac{\partial RSS}{\partial \beta_1} = 0$, ... $\Rightarrow$ เหมาะสม $\Rightarrow$ เหมาะสม ]

$\Rightarrow$ หา statistical significance ของ $\beta_i$ ได้

# Standard Linear Regression and Its Limits

รูปแบบที่ว่า ตัวแปร $y$ ขึ้นกับ $x_1$ อย่างเดียว (ไม่ขึ้นกับ $x_i$)

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

$\Rightarrow$ ควรจะฟิตได้ $\beta_2, \beta_3, \ldots \beta_p \approx 0$ } โมเดลควรจะบอก
แต่ไม่มี statistical significance } โดยอัตโนมัติ

ปัญหา :

$n \approx p$ ฟิตได้ แต่ variance เยอะ

$n < p$ ฟิตไม่ได้

เท่าที่คำนวณมา $n \gg p$
(ข้อมูลเยอะๆ)

# Standard Linear Regression and Its Limits

ข้อมูลที่ $n < p$ precision medicine (personalized)

Genome sequence $\Rightarrow$ วัด Gene ได้ $\approx 500{,}000$ ตัว

(binary $\approx 500{,}000$ ตัว) input

ตัวอย่างงาน : โอกาสเป็นโรคมะเร็ง output

ตัวอย่างข้อมูล : กลุ่มตัวอย่าง $\approx 1000$ คน

$n \approx 1000$ , $p \approx 500{,}000$

# Standard Linear Regression and Its Limits

วิธีที่ดีกว่า Linear Regression

① ดีกว่าแบบ Linear Regression

- Subset Selection
- Dimensionality Reduction
- Regularization (Ridge/Lasso)

② ใช้วิธีอื่นแทน (non-linear)

- Tree-based
- Support vector Machine
- Neural Network

# Model Selection

ฟังก์ชันประกอบด้วย 3 โมเดล

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$ ← ขึ้นกับ $x_2, x_3$

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$y_3 = \beta_0 + \beta_1 x_1$$

} ไม่ขึ้นกับ $x_2, x_3$

**Model Selection**

จำนวนตัวแปร / ความซับซ้อน ต่างกัน

⇒ โมเดลไหน "ฟังก์ชัน" ที่ดี ?

$MSE_{training}$ บอกไม่ได้ โมเดลที่ซับซ้อน overfit

- ทดสอบกับ test data ⇒ $MSE_{test}$ น้อยสุด คือ "ฟังก์ชัน" ที่ดี
- Cross Validation (เหมือนกับ test data โดยประมาณ)
- พยายาม Estimate $MSE_{test}$ จาก $MSE_{training}$ ด้วยทฤษฎีทางสถิติ

# Model Selection

พยายาม Estimate $MSE_{test}$ จาก $MSE_{training}$ ด้วยวิธีการต่างๆ

$$MSE_{test} = MSE_{training} + \text{correction}$$

$MSE_{test} > MSE_{training}$ เพราะ overfit

วิธีต่างๆ correction

Likelihood
- Mallow's $C_p$
- Akaike Information Criteria (AIC)

Posterior
- Bayesian Information Criteria (BIC)

Ad hoc
- Reduced $\chi^2$

# Mallow's $C_p$

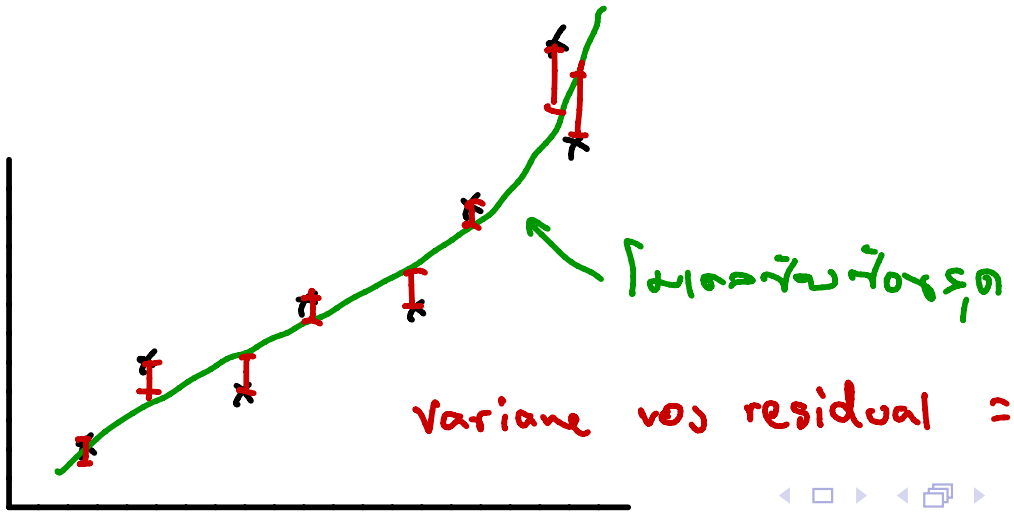
Residual Sum of Squares
$= \sum (y - y_{model})^2$
(ของ training data)

variance ของ residual
(ค่าความแปรปรวนที่ขับข้อมูล)
ของ noise ของข้อมูล

$$C_p = \frac{1}{n}\left(RSS + 2d\sigma^2\right)$$

จำนวนจุดข้อมูล

จำนวน free parameter
ของโมเดล



# Mallow's $C_p$

$$C_p = \frac{1}{n}\left(RSS + 2d\sigma^2\right)$$

$$= \frac{RSS}{n} + \frac{2d\sigma^2}{n}$$

$$C_p = MSE_{training} + \underbrace{\frac{2d\sigma^2}{n}}_{\text{correction}}$$

$C_p$ $\leftarrow$ $MSE_{test}$

(ราคาของ free parameter)

$C_p$ ต่ำๆ $\Rightarrow$ โมเดลดูว่า "ฟิตดี"

$C_p$ น้อย ถ้า
- โมเดลฟิตดี (MSE น้อย)
- ใช้ free parameter เท่าที่จำเป็น ($d$ น้อย)

# Mallow's $C_p$

ฮังก์ชันของ $C_p \Rightarrow \quad C_p' = \frac{RSS}{\sigma^2} + 2d - n$

$\Rightarrow$ ถ้า $C_p$ น้อยสุด $C_p'$ น้อยสุดด้วย

# Akaike Information Criteria (AIC)

$$AIC = \frac{1}{n\sigma^2}\left(RSS + 2d\sigma^2\right)$$

$$= \frac{1}{\sigma^2}\left[\frac{RSS}{n} + \frac{2d\sigma^2}{n}\right]$$

$$AIC = \frac{1}{\sigma^2} C_p$$

จำนวน free parameter

$n \gg p$ ไม่จำเป็น

$n \approx p$ จำเป็น

AIC น้อย ⇒ โมเดลน้อย "ฟิตดี"

# Bayesian Information Criteria (BIC)

$$BIC = \frac{1}{n\sigma^2}\left(RSS + d\sigma^2 \ln n\right)$$

$$BIC = \frac{1}{\sigma^2}\left(\frac{RSS}{n} + d\sigma^2 \frac{\ln n}{n}\right)$$

$$= \frac{1}{\sigma^2}\left(MSE + d\sigma^2 \frac{\ln n}{n}\right)$$

$$\left[AIC = \frac{1}{\sigma^2}\left(MSE + \frac{2d\sigma^2}{n}\right)\right]$$

$BIC > AIC$ ถ้า $n \gtrsim 7-8$

$\Rightarrow$ BIC ตัวแปรแพงกว่า AIC

$\Rightarrow$ ถ้าใช้ BIC จะมีแนวโน้มได้โมเดลที่ซับซ้อนน้อย

# Reduced $\chi^2$

$$\chi^2 = \sum \left( \frac{y - y_{model}}{\Delta y} \right)^2 \approx RSS$$

$$\text{Reduced } \chi^2 = \frac{\chi^2}{dof} = \frac{\chi^2}{n - d}$$