Random Variables:

(cointoss) $X \sim \text{Bernoulli}(p)$

$$X = \begin{cases} 0 & \text{with pr. } 1-p \quad \leftarrow \text{tail} \\ 1 & \text{with pr. } p \quad \leftarrow \text{Heads} \end{cases}$$

fair coin $\rightarrow p = \frac{1}{2}$

$\mathcal{X} = \{0,1\}$ (Discrete)

prob mass function of Bernoulli:

$$P(X=x): \left.\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \end{aligned}\right\} \quad P(X=x) = p^x(1-p)^{1-x}, \quad x=0,1$$

Expectation: $E(X) = \sum_{x \in \mathcal{X}} x P(X=x) = 0 \cdot P(X=0) + 1 \times P(X=1) = P(X=1) = p$

$$E(X^2) = \sum_{x \in \mathcal{X}} x^2 P(X=x) = 0^2 P(X=0) + 1^2 P(X=1) = p$$

$$E(g(X)) = g(0)P(X=0) + g(1)P(X=1) = g(0)(1-p) + g(1)p$$

Variance $\rightarrow V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$

Binomial: # success in n independent trials each with pr. p of success.

if $X_i \sim \text{Bern}(p)$ if we toss coin n times, and count # heads, then

Bernoulli (n,p) $B = \sum_{i=1}^{n} X_i$

pmf: $P(X=x) = \binom{n}{x} \underbrace{p^x(1-p)^{n-x}}$

$\mathcal{X} = \{0,1,2,\dots,n\}$

$E(X) = \sum_{x \in \mathcal{X}} x P(X=x)$

$= 0 \cdot P(X=0) + \dots + n P(X=x)$

5 coin tosses   1=head, 0=tail

1 1 0 1 0

$p \cdot p \cdot (1-p) \cdot p \cdot (1-p)$

$p^{3}(1-p)$

$P(X=3)$?

1 1 0 1 0
1 1 1 0 0
0 0 1 1 1
0 1 1 1 0

$\binom{5}{3} = \frac{5!}{3!\,2!}$

$Y_i \sim \text{Bern}(p)$

$E(A+B) = E(A) + E(B)$

$E(X) = E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p = np$

$V(X) = V\left(\sum_{i=1}^{n} Y_i\right) = \sum V(Y_i) = \sum p(1-p) = np(1-p)$

$V(X+Y) \overset{=}{\neq} V(X) + V(Y)$

Imagine we toss a coin n times, and we don't know 'p'.

data: $\boxed{X_1, X_2, X_3, \dots, X_n} \overset{iid}{\sim} \text{Bern}(p)$

goal: estimate p.

MLE: pick p that max joint prob. of the sample.

Maximum Likelihood Estimate (MLE)

Likelihood: joint prob of the sample

if A, B independent

$P(A,B) = P(A)P(B)$

$$L(p) = P(\underline{X_1, X_2, X_3, \dots, X_n} \mid p)$$
$$= \prod_{i=1}^{n} P(X_i \mid p)$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{\sum x_i}(1-p)^{n-\sum x_i} = \boxed{p^{n\bar{x}}(1-p)^{n-n\bar{x}}}$$

$2^3 \times 2^4 = 2^{3+4} = 2^7$

$$L(p) = p^{n\bar{x}}(1-p)^{n-n\bar{x}} \qquad \hat{p} = \arg\max_p L(p) = \arg\max_p \log(L(p))$$

$$\mathcal{L}(p) = \log L(p) = n\bar{x}\log p + (n-n\bar{x})\log(1-p)$$

$$\frac{\partial \mathcal{L}(p)}{\partial p} = n\bar{x}\frac{1}{p} + (n-n\bar{x})\frac{-1}{1-p} = 0$$

$$\Rightarrow \hat{p}_{MLE} = \bar{x}$$

$\boxed{X_1, X_2, \dots, X_n} \sim Bern(p)$

$p = \frac{1}{2}$

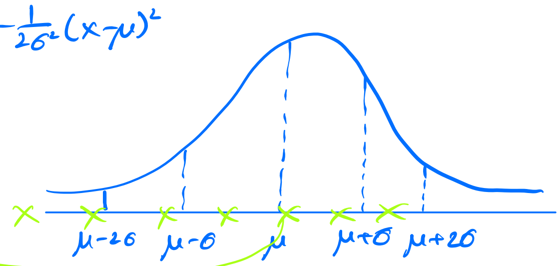
$p = 1$

$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{2}$

Continuous Distribution

$$X \sim N(\mu, \underline{\sigma^2})$$

prob density fn : $f(\underline{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$



- $X_1, X_2, \dots, X_n \overset{iid}{\sim} N(\mu, 1)$
- MLE of μ?

$L(3) < L(1)$

$$L(\mu) = \prod_{i=1}^{n} f(x_i)$$

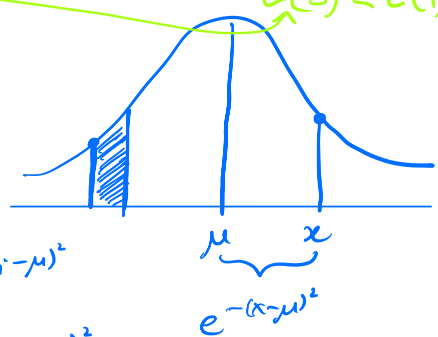
$f(1) \neq P(X=1)$

$P(X=x) = 0$

for all x

$$= \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(x_i-\mu)^2}$$

$$= (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$



$$\arg\max_\mu L(\mu) = \arg\max_\mu e^{-\frac{1}{2}\boxed{\sum(x_i-\mu)^2}} = \arg\min_\mu \sum_{i=1}^{n}(x_i-\mu)^2$$

$$\to f(\mu) = \sum(x_i-\mu)^2$$

$$f'(\mu) = 0$$

$$\hat{\mu} = \bar{x}$$

Linear Regression:

Approach 1: optimisation version (x_i, y_i) $\hat{y}_i = a + bx_i$

$$\hat{a}, \hat{b} = \underset{a,b}{\operatorname{argmin}} \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

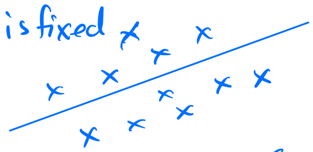
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

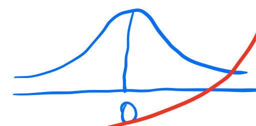
$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

Approach 2: Statistics Version

$$y_i = \underbrace{a + bx_i}_{\text{trend}} + \underbrace{\overset{N(0,\sigma^2)}{\varepsilon_i}}_{\text{error}}, \quad \boxed{\varepsilon_i \sim N(0, \sigma^2)}$$

x_i is fixed





ε_i

$$y_i \sim N(a + bx_i, \sigma^2)$$

$$y_1, y_2, \dots, y_n \overset{\text{indep}}{\sim} N(a + bx_i, \sigma^2)$$

$$Z \sim N(\mu, \sigma^2)$$

$$Z \boxed{+a} \sim N(\mu + a, \sigma^2)$$

$$Z - a \sim N(\mu - a, \sigma^2)$$

$$L(a,b) = \prod_{i=1}^{n} P(y_i | a, b)$$

$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2\right\}$$

$$\underset{a,b}{\operatorname{argmax}} L(a,b)$$

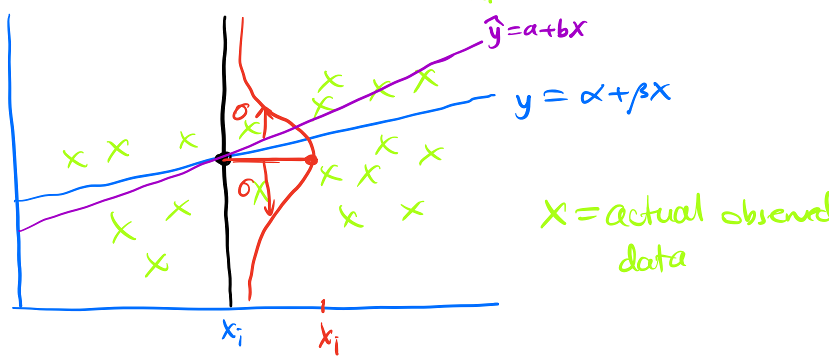
$$= \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$= \hat{a}_{MLE}, \hat{b}_{MLE} = \hat{a}_{LS}, \hat{b}_{LS}$$

$$y_i | x_i \sim N(\mu, \sigma^2) \qquad \mu_i = a + bx_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

α, β true parameters

a, b estimated parameters



A1 : optimisation : $\hat{a}_{LS}, \hat{b}_{LS}$

A2 : Statistics : $\hat{a}_{MLE}, \hat{b}_{MLE}$

$$= \hat{a}, \hat{b} = \left(\bar{y} - \hat{b}\bar{x}, \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\right)$$

under A1 : $\hat{a}, \hat{b}$ $\longrightarrow$ bias? variance? etc...

under A2 : $\hat{a}, \hat{b}$ $\longrightarrow$ bias

$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

$y_i \sim N(a+bx_i, \sigma^2)$

x is fixed

$E(2z) = 2E(z)$

$\text{Bias}(\hat{b}_{MLE}) = E(\hat{b}) - b$

$= E\left(\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\right) - b$

$= \frac{1}{\overline{x^2} - \bar{x}^2}\left[E(\overline{xy}) - \bar{x}E(\bar{y})\right]$

- $E(\bar{y}) = E\frac{1}{n}\sum_{i=1}^{n} y_i$

$= \frac{1}{n}\sum E(y_i)$

$= \frac{1}{n}\sum(a+bx_i)$

$= a + b\bar{x}$

$E(\overline{xy}) = E\left(\frac{1}{n}\sum x_i y_i\right) = \frac{1}{n}\sum x_i E(y_i)$

$= \frac{1}{n}\sum x_i(a+bx_i)$

$= a\bar{x} + b\overline{x^2}$

$E(\hat{b}) - b = \frac{1}{\overline{x^2} - \bar{x}^2}\left[\cancel{a\bar{x}} + b\overline{x^2} - \bar{x}\cancel{(a+b\bar{x})}\right] - b$

$= \frac{1}{\cancel{\overline{x^2} - \bar{x}^2}}\left[b\cancel{(\overline{x^2} - \bar{x}^2)}\right] - b$

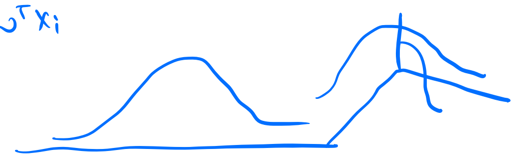
$= b - b = 0$ $\quad \text{Bias}(\hat{b}) = 0$ $\quad \hat{b}$ is unbiased $\quad \hat{a}$ is unbiased

Does this work for multivariate regression?

$y_i | x_i \sim N(\mu, \sigma^2)$ $\quad \mu_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_p x_{ip}$

$= w^T x_i$

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\underbrace{(y_i - w^T x_i)^2}}$



$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim N(Xw, \sigma^2 I)$ $\quad P(y|X) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|y - Xw\|_2^2\right\}$

$\underset{w}{\text{argmax}}\, P(y|X) \Rightarrow \underset{w}{\text{argmin}}\, \|y - Xw\|^2$

$X \in \mathbb{R}^{n\times p}$ $\quad X^T \in \mathbb{R}^{p\times n}$ $\quad \hat{w} = (X^TX)^{-1}X^Ty$

- $\hat{w} = (X^TX)^{-1}X^Ty$ $\quad X^TX \in \mathbb{R}^{p\times p}$ $\quad p\times n \;\; n\times p$

$\text{Bias}(\hat{w}) = E(\hat{w}) - w$

$= 0$

$E(\hat{w}) = E\Big(\underbrace{\underbrace{(X^TX)^{-1}}_{p\times p}\underbrace{X^T}_{C^{p\times n}}}_{p\times n}\underbrace{y}_{n\times 1}\Big) = (X^TX)^{-1}X^TE(y)$

$n\times 1$ $\quad = (X^TX)^{-1}(X^TX)w$

$= Iw = w$

$$V(\hat{b}) = \frac{\sigma^2}{\sum (x_i-\bar{x})^2}$$

$$V(Az) = AV(z)A^T$$

constant matrix

random vector

$$V(4z) = 4^2 V(z)$$

$$E(2z) = 2E(z)$$

$$E(c^Tz) = c^TE(z)$$

$$c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}$$

$$c^Tz = c_1z_1 + c_2z_2 + \dots + c_nz_n$$

$$E(c^Tz) = E(c_1z_1) + E(c_2z_2) +$$

$$c_1E(z_1) + c_2E(z_2)$$

$$= c^TE(z)$$

$$V(\hat{w}) = V(\underbrace{(X^TX)^{-1}X^T}_{A}y)$$

$$= (X^TX)^{-1}X^T \overbrace{V(y)}^{\sigma^2 I} ((X^TX)^{-1}X^T)^T$$

$$= \sigma^2 (X^TX)^{-1} \cancel{X^TX} \cancel{(X^TX)^{-1}}$$

$$= \sigma^2 (X^TX)^{-1}$$

$$E(\hat{w}) = w$$

$$V(w) = \sigma^2 (X^TX)^{-1}$$

$$y \sim N(Xw, \sigma^2 I)$$

σ^2 is variance of the measured errors

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$e_i = (y_i - \hat{y}_i)$$

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(p features)

$$E(s^2) = \sigma^2$$