Section 4.2 Correlation Coefficient

1 Correlation Coefficient

DEF • If $\mu_X = E(X)$ and $\mu_Y = E(Y)$, then

$$\mathrm{Cov}(X, Y) = \sigma_{XY} := E[(X-\mu_X)(Y-\mu_Y)]$$

is called the covariance of X and Y.

• If $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$ are positive, then

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

is called the correlation coefficient of X and Y.

PROP (a) $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y$.

(b) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(c) $\mathrm{Cov}(Y,X) = \mathrm{Cov}(X,Y)$.

(d) $\mathrm{Cov}(a_1X_1 + a_2X_2, Y) = a_1\mathrm{Cov}(X_1,Y) + a_2\mathrm{Cov}(X_2,Y)$

Pf) (a)

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y .\end{aligned}$$

(b) $\mathrm{Cov}(X,X) = E[(X-\mu_X)(X-\mu_X)] = E[(X-\mu_X)^2] = \mathrm{Var}(X)$.

(c) Trivial.

(d)

$$\begin{aligned}\mathrm{Cov}(a_1X_1 + a_2X_2, Y) &= E[((a_1X_1 + a_2X_2) - (a_1\mu_{X_1} + a_2\mu_{X_2}))(Y-\mu_Y)] \\ &= E[a_1(X_1-\mu_{X_1})(Y-\mu_Y) + a_2(X_2-\mu_{X_2})(Y-\mu_Y)] \\ &= a_1\mathrm{Cov}(X_1,Y) + a_2\mathrm{Cov}(X_2,Y).\end{aligned}$$

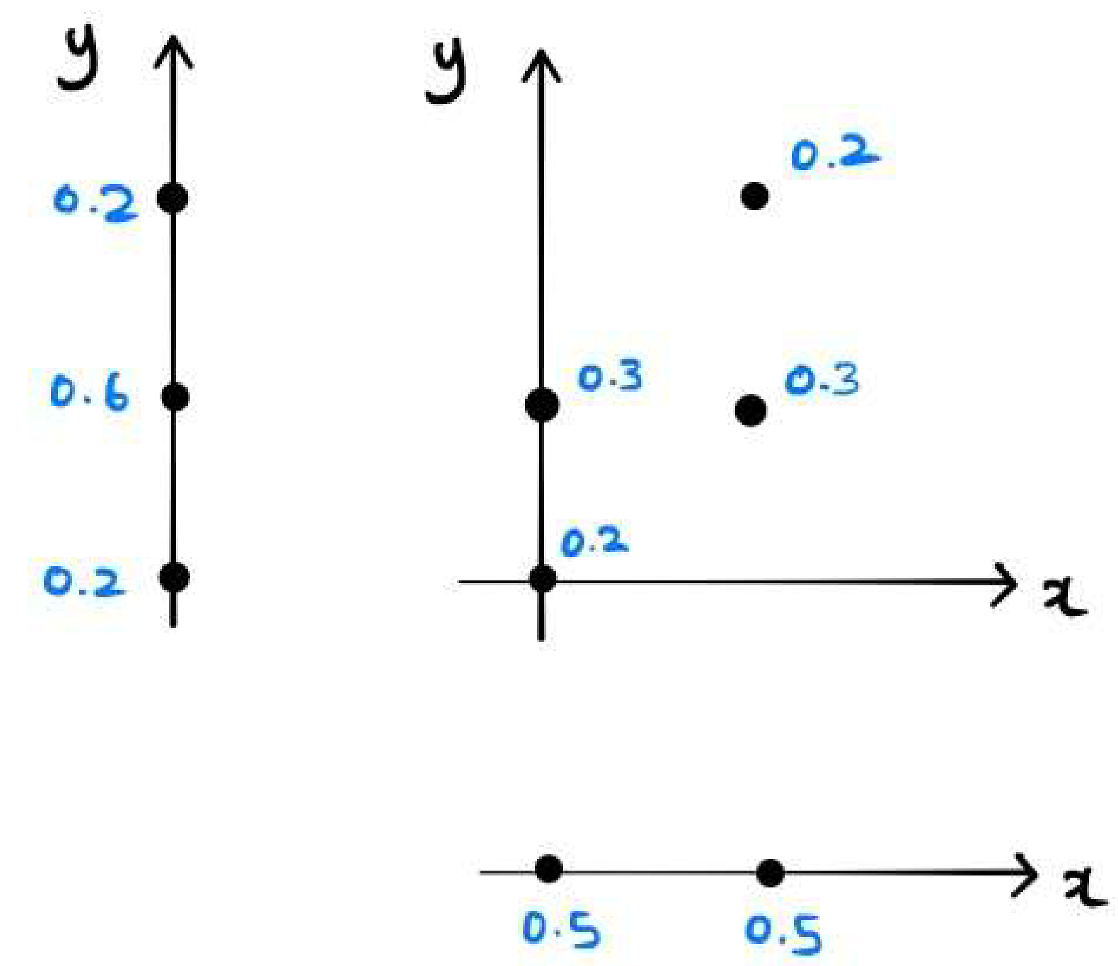
$\square$

• Remark) In general, $\mathrm{Cov}\left(\sum_{i=1}^{m} a_iX_i, \sum_{j=1}^{n} b_jY_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_ib_j\,\mathrm{Cov}(X_i,Y_j)$.

"Covariance really behaves like multiplication!"

Ex Let X and Y have the joint PMF

$$p(0,0) = 0.2, \quad p(0,1) = 0.3,$$
$$p(1,1) = 0.3, \quad p(1,2) = 0.2$$



Then

- $E(X) = (0)\cdot 0.5 + (1)\cdot 0.5 = 0.5$
- $E(Y) = (0)\cdot 0.2 + (1)\cdot 0.6 + (2)\cdot 0.2 = 1$
- $Var(X) = E(X^2) - \mu_X^2 = 0.25 \quad \Rightarrow \sigma_X = 0.5$
- $Var(Y) = E(Y^2) - \mu_Y^2 = 0.4 \quad \Rightarrow \sigma_Y = \sqrt{0.4} \approx 0.632$
- $Cov(X,Y) = E(XY) - \mu_X \mu_Y$
  $= \big((0)\cdot(0)\cdot 0.2 + (0)\cdot(1)\cdot 0.3 + (1)\cdot(1)\cdot 0.3 + (1)\cdot(2)\cdot 0.2\big) - 0.5$
  $= 0.2.$
- $\rho = Cov(X,Y) / \sigma_X \sigma_Y = \sqrt{0.4} \approx 0.632.$ $\square$

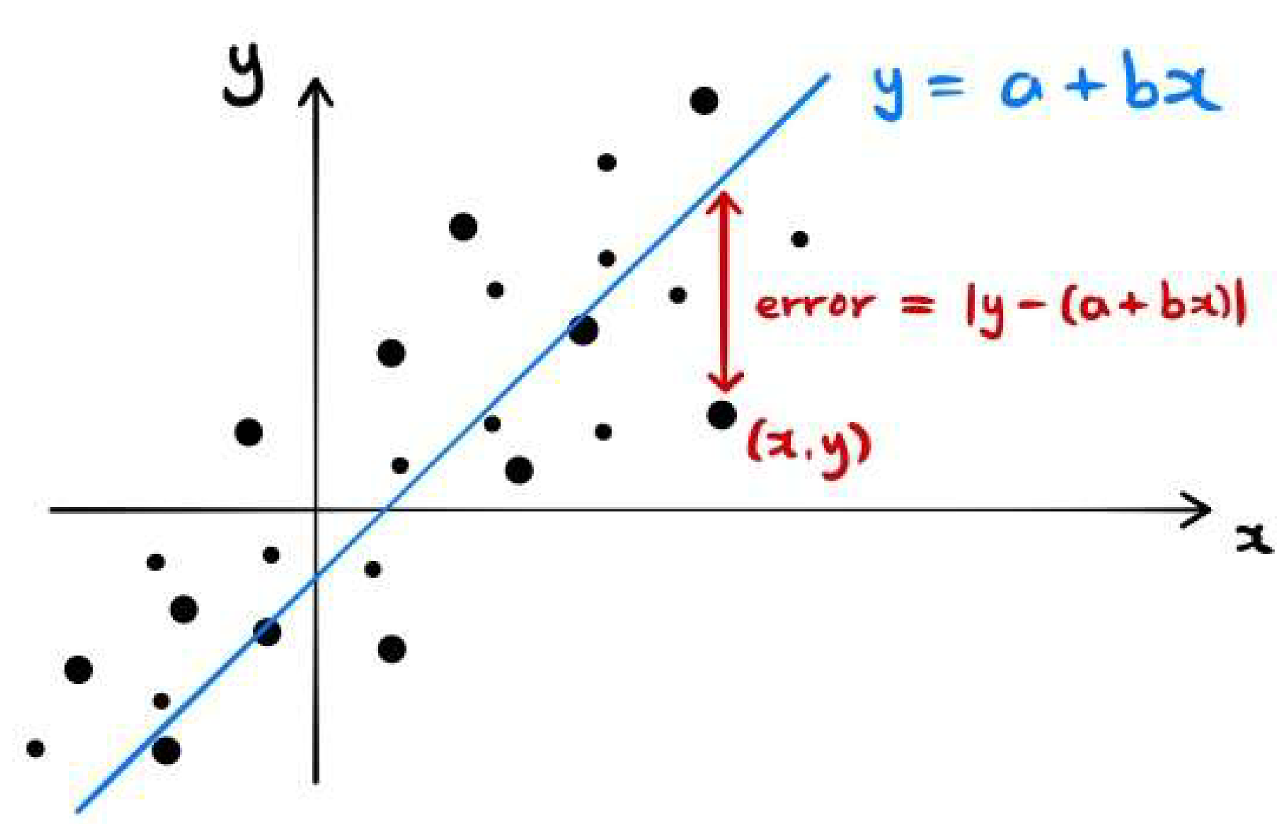
2 Least Square Regression Line

Q What is "the best" approximation of Y in a linear function of X?

- Consider the mean-square error

$$K = K(a,b) = E\left[(Y - (a+bX))^2\right].$$

Want to find a and b that minimizes $K(a,b)$.



- Such min. point can be obtained by finding the critical point of K, i.e., solving

$$\frac{\partial K}{\partial a} = 0 \quad \text{and} \quad \frac{\partial K}{\partial b} = 0.$$

- Expanding K,

$$\begin{aligned} K &= E[Y^2 - 2Y(a+bX) + (a^2 + 2abX + b^2X^2)] \\ &= E(Y^2) - 2a\mu_Y - 2bE(XY) + a^2 + 2ab\mu_X + b^2E(X^2), \end{aligned}$$

Differentiating w.r.t. a and b gives:

$$\frac{\partial K}{\partial a} = -2\mu_Y + 2a + 2b\mu_X,$$

$$\frac{\partial K}{\partial b} = -2E(XY) + 2a\mu_X + 2bE(X^2).$$

Equating both derivatives with zero, we obtain:

$$\begin{cases} a + \mu_X b = \mu_Y \\ \mu_X a + E(X^2) b = E(XY) \end{cases}$$

Solving this system of equations give:

$$b = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_Y}{\sigma_X}\rho, \qquad a = \mu_Y - b\mu_X.$$

- In other words,

$$\begin{aligned} y &= a + bx = \mu_Y + b(x - \mu_X) \\ &= \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho\,(x - \mu) \end{aligned}$$

may be considered as the "best linear relationship between X and Y", called the least square regression line.

- Additionally, the min. value of K is

Check by yourself!

$$\min_{a,b} K(a,b) = E\left[\left((Y-\mu_X) - \frac{\sigma_Y}{\sigma_X}\rho\,(X-\mu_X)\right)^2\right] = \sigma_Y^2(1-\rho^2)$$

This leads to :

THM (1) $-1 \le \rho \le 1$ always holds.

(2) $\rho = \pm 1 \iff Y = a + bX$ for some constants a and b.

Pf) (1) With a and b as before,

$$E[\underbrace{(Y-(a+bX))^2}_{\geq 0}] = \sigma_Y^2(1-\rho^2).$$

So $\sigma_Y^2(1-\rho^2) \geq 0$ always holds, which then implies $-1 \le \rho \le 1$.

(2) $Y = a+bX$ for some a & b

$\iff$ $K(a,b) = 0$ for some a & b

$\iff$ min. of K is zero

$\iff$ $\rho^2 = 1$.

$\iff$ $\rho = \pm 1$. $\square$

3 Independence and Correlation

THM If X and Y are independent, then

$$E[u(X)v(Y)] = E[u(X)]E[v(Y)]$$

for any functions $u(x)$ & $v(y)$, provided all the expectations exist.

<u>Pf)</u> We will only prove this when X & Y are discrete. Then

$$
\begin{aligned}
E[u(X)v(Y)] &= \sum_{x,y} u(x)v(y)\, p_{X,Y}(x,y) \\
&= \sum_{x,y} u(x)v(y)\, p_X(x)\, p_Y(y) \\
&= \left(\sum_x u(x)\, p_X(x)\right)\left(\sum_y v(y)\, p_Y(y)\right) = E[u(X)]\,E[v(Y)]
\end{aligned}
$$

Independence!

□

- As a crucial consequence:

COR X and Y independent $\Longrightarrow$ $\mathrm{Cov}(X,Y) = 0$.

indep. is used here

<u>Pf)</u> $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0.$ □

- But the converse is not true in general.