JASH MEHTA

CHAPTER 1

1.1)

consider $\mathbb{V}[X+Y]$ where X & Y are random variables.

By defination,

$$\mathbb{V}[X] = \mathbb{E}\Big[\underbrace{(X - \mathbb{E}(X))^2}_{\text{squared diff b/w } X \,\&\, \mathbb{E}(X)}\Big] \quad \ldots \text{(I)}$$

$$= \underbrace{\mathbb{E}[X^2]}_{2^{nd}\text{ moment}} - \underbrace{(\mathbb{E}[X])^2}_{(1^{st}\text{ moment})^2}$$

$$\therefore \mathbb{V}[X+Y] = \mathbb{E}\Big[\big(\underbrace{(X+Y)}_{A} - \underbrace{\mathbb{E}(X+Y)}_{B}\big)^2\Big] \quad \ldots \text{replace } X \text{ with } (X+Y) \text{ in (I)}$$

$$= \mathbb{E}\Big[\underbrace{(X+Y)^2}_{A^2} - \underbrace{2(X+Y)\,\mathbb{E}(X+Y)}_{2\cdot A\cdot B} + \underbrace{\mathbb{E}^2(X+Y)}_{B^2}\Big]$$

$$= \mathbb{E}\left[X^2 + Y^2 + 2X\cdot Y - 2(X+Y)\,\mathbb{E}(X+Y) + \mathbb{E}^2(X+Y)\right]$$

$$= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\,\mathbb{E}[XY] - 2\,\mathbb{E}^2(X+Y) + \mathbb{E}^2(X+Y)$$

$$= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\,\mathbb{E}[XY] - \mathbb{E}^2(X+Y)$$

$$\mathbb{E}^2(X+Y) \to \mathbb{E}(X+Y)\cdot\mathbb{E}(X+Y)$$

$$= [\mathbb{E}(X) + \mathbb{E}(Y)]^2$$

$$= \mathbb{E}^2(X) + \mathbb{E}^2(Y) + 2\,\mathbb{E}[X]\,\mathbb{E}[Y]$$

$$= \Big[\underbrace{\mathbb{E}[X^2] - \mathbb{E}^2(X)}_{\mathbb{V}[X]}\Big] + \Big[\underbrace{\mathbb{E}[Y^2] - \mathbb{E}^2[Y]}_{\mathbb{V}[Y]}\Big] + 2\big(\underbrace{\mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]}_{\downarrow}\big) \quad \ldots \text{(II)}$$

$$2.\,\mathrm{Cov}[X,Y]$$

Assumption : X & Y are uncorrelated

$$\Downarrow$$

$$Cov[X,Y] = 0$$

(stronger assumption is variables are independant)

From (II)

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y] + \cancel{2\,Cov[X,Y]}^{\,0}$$

$\therefore$ we ~~gor~~ can generalize this to :

$$\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathbb{V}[Y_i]$$

if s.d of Y_i is σ,

$$\boxed{\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] = n\sigma^2} \longrightarrow (1)$$

Bienaymé formula

now, consider $\mathbb{V}[aX]$

$$\mathbb{V}[aX] = \mathbb{E}[(aX)^2] - \mathbb{E}^2[(aX)]$$

2nd moment - (1st moment)^2

$$= \mathbb{E}(a^2X^2) - (a\,\mathbb{E}(X))^2$$

$$= a^2\,\mathbb{E}(X^2) - a^2[\mathbb{E}(X)]^2$$

$$= a^2\left[\underbrace{\mathbb{E}(X^2) - [\mathbb{E}(X)]^2}_{\mathbb{V}[X]}\right]$$

$$= a^2\,\mathbb{V}[X] \quad \cdots (III)$$

substitute $a = 1/n$ & $X = \sum_{i=1}^{n} Y_i$ in (III)

$$\therefore \mathbb{V}\left[n^{-1}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\,\mathbb{V}\left[\sum_{i=1}^{n} Y_i\right] = \boxed{\frac{\sigma^2}{n}} \longrightarrow (2)$$

$$s.d(x) = \sqrt{\mathbb{V}[x]}$$

from (2) :

$$s.d \text{ of } n^{-1}\sum_{i=1}^{n} Y_i = \sqrt{\mathbb{V}\left[n^{-1}\sum_{i=1}^{n} Y_i\right]}$$

$$= \sqrt{\frac{\sigma^2}{n}}$$

$$= \boxed{\frac{\sigma}{\sqrt{n}}} \longrightarrow (3)$$

now,

$$s.d \text{ of } \mu - n^{-1}\sum_{i=1}^{n} Y_i = \sqrt{\mathbb{V}\left[\mu - n^{-1}\sum_{i=1}^{n} Y_i\right]}$$

$$= \sqrt{\mathbb{V}\left[n^{-1}\sum_{i=1}^{n} Y_i\right]}$$

0 (uncorrelated)

$$\because \mathbb{V}(X-Y) = \mathbb{V}[X] + \mathbb{V}[Y] - 2\,\cancel{\mathrm{Cov}[X,Y]}$$

$$\& \ \mathbb{V}[\mu] = 0$$

$$= \sqrt{\frac{\sigma^2}{n}}$$

$$= \boxed{\frac{\sigma}{\sqrt{n}}} \longrightarrow (4)$$

1.2) $MAE(m) = \mathbb{E}[|Y-m|]$

$$\mathbb{E}[|Y-m|] = \int_{-\infty}^{\infty} p(Y)\,|Y-m|\,dY$$

$$= \int_{-\infty}^{m} p(Y)\,|Y-m|\,dY + \int_{m}^{\infty} p(Y)\,|Y-m|\,dY$$

$$|Y-m| = \begin{cases} Y-m, & m \leqslant Y \\ m-Y, & m \geqslant Y \end{cases}$$

$$\therefore = \int_{-\infty}^{m} p(Y)\,(m-Y)\,dY + \int_{m}^{\infty} p(Y)\,(Y-m)\,dY$$

In order to minimize, we will differenciate & equate it to 0.

$$\frac{dA}{dm} = \frac{d}{dm}\,\mathbb{E}[|Y-m|]$$

$$= \underbrace{\frac{d}{dm}\int_{-\infty}^{m} p(Y)\,(m-Y)\,dY}_{I} + \underbrace{\frac{d}{dm}\int_{m}^{\infty} p(Y)\,(Y-m)\,dY}_{J} \quad \ldots (I)$$

Leibnitz rule:

$$\frac{d}{dx}\left(\int_{a(x)}^{b(x)} f(x,t)\,dt\right) = f(x, b(x))\,\frac{d}{dx}b(x) - f(x, a(x))\cdot\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x,t)\,dt$$

By using Leibnitz Rule:

$$\frac{d}{dm} I = p(Y)\,\cancel{(m-m)}^{0}\,\frac{dm}{dm} + \int_{-\infty}^{m} \frac{\partial}{\partial m}\left[p(Y)\,(m-Y)\right] dY$$

$$= \int_{-\infty}^{m} p(Y)\,dY$$

$$\frac{d}{dm} J = -\,p(Y)\,\cancel{(m-m)}^{0}\,\frac{dm}{dm} + \int_{m}^{\infty} \frac{\partial}{\partial m}\left[p(Y)\,(Y-m)\right] dY$$

$$= -\int_{m}^{\infty} p(Y)\,dY$$

$$\therefore \frac{dA}{dm} = \frac{dI}{dm} + \frac{dJ}{dm} \qquad (\text{from (I)})$$

$$\Rightarrow 0 = \int_{-\infty}^{m} p(Y)\,dY - \int_{m}^{\infty} p(Y)\,dY$$

$$\Rightarrow \int_{-\infty}^{m} p(Y)\,dY = \int_{m}^{\infty} p(Y)\,dY$$

$$\Rightarrow p(Y \leq m) = p(Y \geq m)$$

we know that $P(Y \geq m) + P(Y \leq m) = 1$

$$\Rightarrow p(Y \leq m) = p(Y \geq m) = \frac{1}{2}$$

$\Rightarrow$ $\boxed{\text{m is median}}$

1.3)

Empherical MSE $= \frac{1}{n} \sum_i (y_i - b_0 - b_1 x_i)^2$

to obtain $\hat{\beta}_1$ and $\hat{\beta}_0$, we need to minimize MSE by differentiating and equating it to 0.

$$\frac{\partial}{\partial b_0} MSE = \frac{\partial}{\partial b_0} \sum (y_i - (b_0 + b_1 x_i))^2 \times \frac{1}{n}$$

$$= \sum \frac{\partial}{\partial b_0} (y_i - (b_0 + b_1 x_i))^2 \times \frac{1}{n}$$

$$= \sum 2 (y_i - (b_0 + b_1 x_i))(-1) \times \frac{1}{n}$$

$$= \frac{-2}{n} \sum (y_i - (b_0 + b_1 x_i)) \quad \ldots (\mathrm{I})$$

$$\frac{\partial}{\partial b_1} MSE = \frac{\partial}{\partial b_1} \sum (y_i - (b_0 + b_1 x_i))^2 \times \frac{1}{n}$$

$$= \sum \frac{\partial}{\partial b_1} (y_i - (b_0 + b_1 x_i))^2 \times \frac{1}{n}$$

$$= \sum 2 (y_i - (b_0 + b_1 x_i))(-x_i) \times \frac{1}{n}$$

$$= \frac{-2}{n} \sum x_i (y_i - (b_0 + b_1 x_i)) \quad \ldots (\mathrm{II})$$

Equating partial derivatives to 0:

$$-\cancel{\frac{2}{n}} \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0 \longrightarrow 1$$

$$-\cancel{\frac{2}{n}} \sum x_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0 \longrightarrow 2$$

Solving for $\hat{\beta_0}$:

$$\sum \left(y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)\right) = 0$$

$$\sum y_i - \sum \hat{\beta_0} - \sum \hat{\beta_1} x_i = 0$$

$$\sum y_i - n\hat{\beta_0} - \hat{\beta_1} \sum x_i = 0$$

$$n\hat{\beta_0} = \sum y_i - \hat{\beta_1} \sum x_i$$

$$\boxed{\hat{\beta_0} = \overline{Y} - \hat{\beta_1}\,\overline{x}}$$

Solving for $\hat{\beta_1}$:

$$\sum x_i \left(y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)\right) = 0$$

substitute $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$:

$$\sum x_i \left(y_i - (\bar{y} - \hat{\beta_1}\bar{x} + \hat{\beta_1} x_i)\right) = 0$$

$$\sum x_i \left(y_i - \bar{y} - \hat{\beta_1}(x_i - \bar{x})\right) = 0$$

$$\sum x_i (y_i - \bar{y}) - \sum \hat{\beta_1} x_i (x_i - \bar{x}) = 0$$

$$\sum x_i (y_i - \bar{y}) = \sum \hat{\beta_1} x_i (x_i - \bar{x})$$

$$\hat{\beta_1} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} \quad \ldots (\text{III})$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i (y_i - \bar{y}) - \sum \bar{x}(y_i - \bar{y})$$

$$= \sum x_i (y_i - \bar{y}) - \bar{x} \sum \cancel{(y_i - \bar{y})} \; 0$$

$$\sum(y_i - \bar{y}) = \sum y_i - n\bar{y} = n\bar{y} - n\bar{y}$$

$$\left.\begin{aligned} &= \sum x_i (y_i - \bar{y}) \\ &= \sum y_i (x_i - \bar{x}) \end{aligned}\right\} \ldots (\text{IV})$$

substitute (IV) in (III)

$$\hat{\beta}_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

$$\therefore \boxed{\hat{\beta}_1 = \frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\mathbb{V}[x]}}$$

1.4)

For linear smoothers:

$$\hat{\mu}(x) = \sum_i y_i \, \hat{w}(x_i, x)$$

For nearest neighbor regression,

$$\hat{w}(x_i, x) = \begin{cases} 1, & x_i \text{ is nearest neighbor of } x \\ 0, & \text{otherwise} \end{cases}$$

For Gaussian kernel in 1 D, the kernel function is

$$K(x_i, x) \propto e^{-(x_i - x)^2/2h^2} \quad \dots \text{(I)}$$

If x_i is the nearest neighbor of x, & $|x_i - x| = L$ where $L \gg h$ then $\because h \to 0$

$$K(x_i, x) \propto e^{-L^2/2h^2}$$

But if there is some point x_j which is farther away from x,

$$K(x_j, x) \propto e^{-(x_j - x)^2/2h^2} \quad \dots \text{(II)}$$

also,

$$(x_j - x)^2 = (\underbrace{(x_j - x_i)}_{A} + \underbrace{(x_i - x)}_{B})^2$$

$$= [\underbrace{(x_j - x_i)^2}_{A^2} + \underbrace{(x_i - x)^2}_{B^2} + \underbrace{2(x_j - x_i)(x_i - x)}_{2AB}]$$

Substituting this in (II):

$$K(x_j, x) \propto e^{-(x_j - x_i)^2 - (x_i - x)^2 - 2(x_j - x_i)(x_i - x)/2h^2}$$

$$\propto e^{-(x_j - x_i)^2/2h} \cdot e^{-(x_i - x)^2/2h^2} \cdot e^{-2(x_j - x_i)(x_i - x)/2h^2}$$

put $|x_i - x| = L$

$$K(x_j, x) \propto e^{-L^2/2h^2} e^{-(x_j - x_i)L/2h^2} e^{-(x_j - x_i)^2/2h^2}$$

$\therefore K(x_j, x) << K(x_i, x)$

and if $h \to 0$, we can say that

$K(x_i, x) \approx 0$ if x_i is not the nearest neighbour

because $K(x_i, x) \longrightarrow 0$

$\therefore$ As $h \to 0$, Gaussian Kernel Regression approaches nearest neighbour regression.

1.5)

$$W = X(X^TX)^{-1}X^T$$

To prove that W is idempotent, we need to prove

$$W^2 = W$$

$$\therefore W^2 = \left(X(X^TX)^{-1}X^T\right)\left(X(X^TX)^{-1}X^T\right)$$

$$= X\left((X^TX)^{-1}(X^TX)(X^TX)^{-1}\right)X^T$$

$$= X(X^TX)^{-1}X^T$$

$$= W$$

$\therefore$ W is <u>idempotent</u>

1.6) $$\mathbb{V}[\hat{\mu}(x)] = \sigma^2 \sum_{j=1}^{n} w^2(x_j, x) \quad \text{(eqn 1.61)} \quad \ldots \text{(I)}$$

For ordinary linear regression,

$$\hat{\mu}(x) = \sum_{i=1}^{n} \frac{1}{n}\left(1 + \frac{(x-\bar{x})(x_i-\bar{x})}{\hat{\sigma}_x^2}\right) y_i$$

$$\therefore W = \frac{1}{n}\left[1 + \frac{(x-\bar{x})(x_i-\bar{x})}{\hat{\sigma}_x^2}\right]$$

Substitute this w in (I):

$$\mathbb{V}[\hat{\mu}(x)] = \sigma^2 \sum_{j=1}^{n}\left(\frac{1}{n}\left[1 + \frac{(x-\bar{x})(x_i-\bar{x})}{\hat{\sigma}_x^2}\right]\right)^2$$

$$= \frac{\sigma^2}{n^2} \sum_{i=1}^{n}\left[1 + \frac{(x-\bar{x})(x_i-\bar{x})}{S_x^2}\right]^2$$

$$= \frac{\sigma^2}{n^2}\left[\sum_{i=1}^{n}\left(1 + \frac{2(x-\bar{x})(x_i-\bar{x})}{S_x^2} + \frac{(x-\bar{x})^2(x_i-\bar{x})^2}{S_x^4}\right)\right]$$

$$= \frac{\sigma^2}{n^2}\left[n + 2\,\cancel{\frac{\sum (x-\bar{x})\overset{0}{(x_i-\bar{x})}}{S_x^2}} + \frac{(x-\bar{x})^2 \times n S_x^2}{S_x^4}\right]$$

b/c $\sum(x_i-\bar{x})$

$= \sum x_i - n\bar{x}$

$= n\bar{x} - n\bar{x}$

$= 0$

$$= \frac{\sigma^2}{n^2}\left[n + \frac{n(x-\bar{x})^2}{S_x^2}\right]$$

$$= \frac{\sigma^2}{n}\left[1 + \frac{(x-\bar{x})^2}{S_x^2}\right]$$

Hence proved.

1.7)

$$df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$$

By defination

$$df(\hat{y}) = \frac{1}{\sigma^2} tr\left(Cov(\hat{y}, y)\right)$$

now, for global mean as linear smoother :-

$$\hat{y}^{mean} = (\bar{y}, \dots, \bar{y}) \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\therefore df(\hat{y}^{mean}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov(\bar{y}, y_i)$$

$\because$ all y_i are $\bar{y}$, $Cov(y, y) = \sigma_y^2$

$$\Rightarrow df(\hat{y}^{mean}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{\sigma^2}{n}$$

$$= \frac{1}{\sigma^2} \times \frac{\sigma^2}{n} \times n$$

$$= \boxed{1}$$

influence matrix

$$W = \begin{bmatrix} 1/n & 1/n & 1/n \\ 1/n & 1/n & 1/n \\ 1/n & \vdots & 1/n \\ \vdots & \vdots & \ddots \end{bmatrix}_{n\times n}$$

all diagonal elements are $1/n$

1.8) For linear smoothers :

$$\hat{y}_i^{linearsm} = \hat{r}^{linearsm}(x) = \sum_{j=1}^{n} w(x, x_j)\, y_i$$

$$\therefore \hat{y}^{linearsm} = Sy \quad \text{where } S \in \mathbb{R}^{n\times n}$$

$$s_{ij} = w(x_i, x_j)$$

$$df(\hat{y}^{linear\,sm}) = \frac{1}{\sigma^2} tr\left(Cov(Sy, y)\right)$$

... by defination

$$= \frac{1}{\sigma^2} tr(S \, Cov(y,y))$$

$$= \frac{1}{\cancel{\sigma^2}} \times \cancel{\sigma^2} \, tr(S)$$

$$= \sum_{i=1}^{n} w(x_i, x_i) \quad \ldots \; (I)$$

For K-nearest neighbours:

$$w(x, x_j) = \begin{cases} 1/K, & \text{if } x_j \text{ is one of the K closest points to } x \\ 0, & \text{otherwise} \end{cases}$$

$$\therefore \; w(x_i, x_i) = 1/K$$

$$\Rightarrow df(\hat{y}^{knn}) = \sum_{i=1}^{n} \frac{1}{K} \quad (\text{from } (I))$$

$$= \boxed{\frac{n}{K}}$$

influence matrix:

$$W = \begin{bmatrix} 1/K & & & \\ & 1/K & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix}_{n \times n}$$

all diagonal elements are $1/K$

Scanned by CamScanner

1.9)

$$\mathbb{V}[\hat{\mu}_i] = \sigma^2 \sum_{j=1}^{n} w_{ij}^2$$ for a linear smoother

$$\therefore \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}[\hat{\mu}_i] = \frac{1}{n}\sum_{i=1}^{n}\left[\sigma^2 \sum_{j=1}^{n} w_{ij}^2\right]$$

$$= \frac{\sigma^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 \quad \ldots (\mathrm{I})$$

consider A to be real matrix

$\therefore$ j^{th} column of A is A_j

$(i,j)^{th}$ entry of A^TA is $A_i^T A_j$

Similarly $(j,j)^{th}$ entry is $A_j^T A_j = \|A_j\|^2$

$$= \sum_{i=1}^{n} a_{ij}^2$$

$\therefore$ $(j,j)^{th}$ element of A^TA measures $\|A_j\|^2$

$$\therefore tr(A^TA) = \sum_{j=1}^{m} \|A_j\|^2 = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ij}^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2$$

property of Gramiam matrix

$\therefore$ using this result in (I):

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{V}[\hat{\mu}_i] = \frac{\sigma^2}{n} tr(ww^T)$$

Hence proved.

1.10)

eqn 1.71 :

$$df(\hat{\mu}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} Cov[Y_i, \hat{\mu}(x_i)]$$

modified eqn :

$$df(\hat{\mu}) = \sum_{i=1}^{n} \frac{Cov[Y_i, \hat{\mu}_i]}{\sigma_i^2} \quad \ldots (I)$$

$$Cov[Y_i, \hat{\mu}(x_i)] = Cov\left[Y_i, \sum_{j=1}^{n} w_{ij} Y_i\right]$$

$$= \sum_{j=1}^{n} w_{ij} Cov[Y_i, Y_j]$$

$$= w_{ii} \mathbb{V}[Y_i]$$

$$= w_{ii} \sigma_i^2$$

Substitute in (I):

$$df(\hat{\mu}) = \sum_{i=1}^{n} \frac{Cov[Y_i, \hat{\mu}_i]}{\sigma_i^2}$$

$$= \sum_{i=1}^{n} \frac{w_{ii} \cancel{\sigma_i^2}}{\cancel{\sigma_i^2}}$$

$$= \underline{\underline{tr(W)}}$$

Hence proved.