① a) Joint PDF for Y_1, Y_2:

$$f_{Y_1 Y_2}(y_1, y_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} \cdot e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right)\right]}$$

for all y_1, y_2 : $\mu_1, \mu_2 \in R$, $\sigma_1 > 0, \sigma_2 > 0$ & $-1 \leq \rho \leq 1$

b) Bayes thm:

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1 Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}$$

The marginal distribution $f_{Y_2}(y_2)$ is given as:

$$\int_{-\infty}^{\infty} f_{Y_1, Y_2}(y_1, y_2)\, dy_1 = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}$$

placing this formula $(f_{Y_2}(y_2))$ in Bayes thm,

$$f_{Y_1|Y_2}(y_1|y_2) = \cancel{\frac{1}{e}\, e^{\frac{-1}{2(1-\rho^2)}}}$$

P.T.O.

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{\dfrac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{2}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]}}{\dfrac{1}{\sqrt{2\pi}\sigma_2} e^{-\left(\frac{(y_2-\mu_2)^2}{2\sigma_2^2}\right)}}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_1} e^{-\frac{\left(y_1 - \left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2)\right)\right)^2}{2\sigma_1^2(1-\rho^2)}}$$ — Ans.

This conditional probability is normally distributed, with mean $\mu = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)$ & variance $\sigma^2 = (1-\rho^2)\sigma_1^2$

(c) In standard regression, we treat predictors as fixed. Since predictors are fixed, it's actually like conditional probability.

→ expectation (or mean)

$E(Y_1|Y_2 = y_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(Y_2 - \mu_2)$ is a linear f^n of Y_2. This is similar to the standard form of linear regression, $y = mx + c$, where slope is m ~~is~~ & c is intercept.

$$E(y_1|y_2) = \alpha + \beta y_2$$

$$E(y_1|y_2) = E(y_1) \cancel{\alpha} + \beta\{\cancel{y_1} - E(y_2)\}$$

② a) Quadratic loss:

It's given by $L = \sum_{i=1}^{N} e_i^2$

If the second derivative of a $f^n \geq 0$, then the f^n is convex.

$$\frac{\partial L}{\partial e} = 2e$$

$$\frac{\partial^2 L}{\partial e^2} = 2$$

$$\frac{\partial^2 L}{\partial e^2} = 2 \quad \text{which is} \geq 0.$$

$\therefore$ Quadratic loss is <u>convex.</u>

b) L_1 norm. $L = \sum_{i=1}^{N} |e_i|$

$$\frac{\partial L}{\partial e} = \text{sgn}(e) \qquad \text{sgn - sign } f^n$$

$$\frac{\partial^2 L}{\partial e^2} = 0 \geq 0$$

$\therefore$ Mean absolute error is <u>convex.</u>

c) $L = \sum_{i=1}^{N} l(e_i)$, where $l(e) = \begin{cases} \frac{1}{2}e^2 & , \text{ if } |e| \leq \delta \\ \delta|e| - \frac{1}{2}\delta^2 & , \text{ if } |e| > \delta \end{cases}$

$$\frac{\partial L}{\partial e} = \begin{cases} e & , \text{ if } |e| \leq \delta \\ \delta \operatorname{sgn}(e) & , \text{ if } |e| > \delta \end{cases}$$

$$\frac{\partial^2 L}{\partial e^2} = \begin{cases} 1 & , \text{ if } |e| \leq \delta \\ 0 & , \text{ if } |e| > \delta \end{cases}$$

which is ≥ 0.

$\therefore$ Huber loss <u>IS convex.</u>

③ $Y_i = \theta_0 + \theta_1 x_1 + e_i, \quad i = 1, \dots, N$

a) Derive analytical sol^n:

hypothesis f^n: $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n$

We have to min. least squares cost:

$$J(\theta_0 \dots n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$x^{(i)}$ - i^{th} sample

$y^{(i)}$ - i^{th} expected result

m - no. of learning observations

regression coefficients θ is a vector - $(\theta_0 \; \theta_1 \dots \theta_n) \in R^{n+1}$

Since each of m i/p samples is also a column vector, we can write hypothesis as,

$$h_\theta(x) = \theta^T x$$

Now, $J(\theta)$ can be rewritten as:

$$J(\theta) = \frac{1}{2m} (X\theta - y)^T (X\theta - y)$$

Solving the eq^n,

$$J(\theta) = ((X\theta)^T - y^T)(X\theta - y) \cdot \frac{1}{2m}$$

$$J(\theta) = (X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^T y \cdot \frac{1}{2m}$$

$$J(\theta) = \theta^T X^T X\theta - 2(X\theta)^T y + y^T y \cdot \frac{1}{2m}$$

to find the min,

$$\frac{\partial J}{\partial \theta} = 0$$

$$\therefore \frac{\partial J}{\partial \theta} = 0 = \left(2X^T X\theta - 2X^T y\right) \cdot \frac{1}{2m}$$

$$X^T X\theta = X^T y$$

Multiply both sides by $(X^T X)^{-1}$

$$\therefore \theta = \cancel{X^T} (X^T X)^{-1} X^T y$$

$$\boxed{\text{Ans.} \quad \theta = (X^T X)^{-1} X^T y}$$

1) Steps of G.D. to estimate slope & intercept.

Basically, G.D. is used to minimize the cost function used for linear regression.

L.R.: $y_i = \theta_0 + \theta_1 x_i + e_i \qquad i = 1, \dots, N$

Algo. (steps) to min. loss f^n: $J(\theta_0, \theta_1)$

1) 1st, take partial derivative.

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{1}{2m} \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x_i + e_i - y^{(i)} \right)^2$$

After differentiating,

$\theta_0 \rightarrow j = 0$: $\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$

$$= \frac{\partial}{\partial \theta_0} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x_i + e_i - y^{(i)} \right)^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta x^{(i)} - y^{(i)} \right)$$

$\theta_1 \rightarrow j = 1$: $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$

$$= \frac{\partial}{\partial \theta_1} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x_i + e_i - y^{(i)} \right)^2$$

$$= \frac{1}{m} \sum^{m} \left(h_\theta x^{(i)} - y^{(i)} \right) \cdot x^{(i)}$$

2) repeat this until it converges

$$\theta_0 = \theta_0 - \frac{\overset{\nearrow \text{learning rate}}{\alpha}}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

θ_0 & θ_1 should be update simultaneously.

θ_0 & θ_1 are slope & the intercept.

c) Similarly, steps for stochastic G.D.:

1) Here, in every iteration, 1 row out of m rows of training set is considered at a time.

2) a) for $i = 1$ to no. of epochs (each iteration over dataset)
2) b) Randomly shuffle the dataset.

3) Repeat till it converges:

{ for 1, ..., m:
  { $\theta_j = \theta_j - \alpha(h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)}$ $j = 0$ to n
  }
}

(θ_0, θ_1 will be same as

- in practice a small stepsize is used.