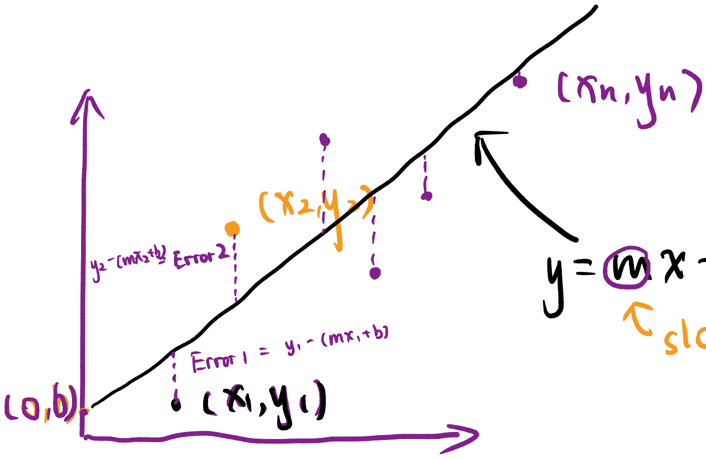


① Find a line

→ minimize the squared distance to these different points

$y = mx + b$ ← y-intercept

↑ slope

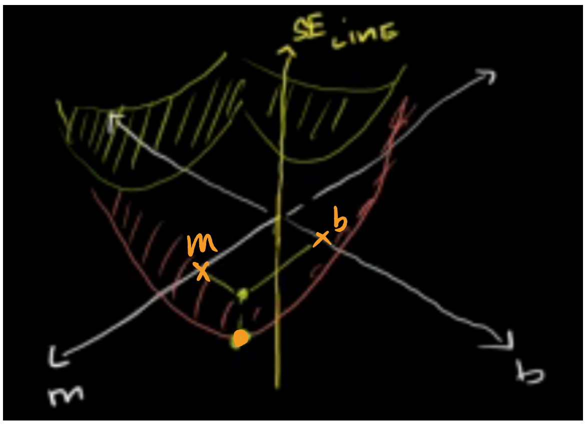
Squared Error of Line (SE_{line})

$$= (y_1-(mx_1+b))^2 + (y_2-(mx_2+b))^2 + \cdots + (y_n-(mx_n+b))^2$$

Find the $m + b$ that minimizes SE_{line}

$$= (y_1^2 + y_2^2 + \cdots + y_n)^2 - 2m(x_1y_1 + x_2y_2 + \cdots + x_ny_n) - 2b(y_1 + y_2 + \cdots + y_n) + m^2(x_1^2 + x_2^2 + \cdots + x_n^2) + 2mb(x_1 + x_2 + \cdots + x_n) + nb^2$$

$$= n\overline{y^2} - 2m \cdot n\overline{xy} - 2b \cdot n\bar{y} + m^2 n\overline{x^2} + 2mb \cdot n\bar{x} + nb^2$$



at this minimum point in three dimensions, the min point on the surface is going to occur when the slope with respect to m and the slope with respect to b

② Find the m and b values that minimize the value on the surface

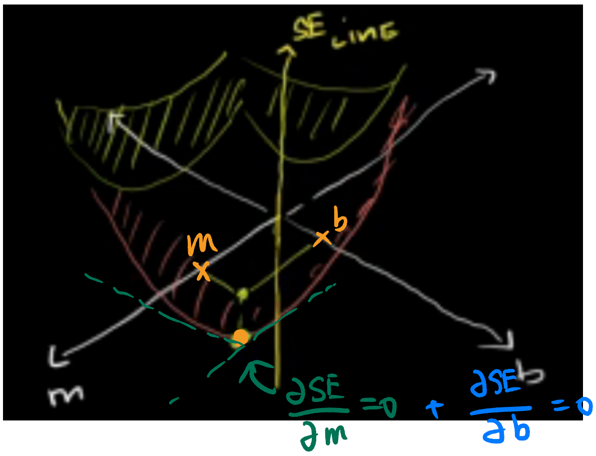
(2-1) Find the partial derivative of this w/ respect to m

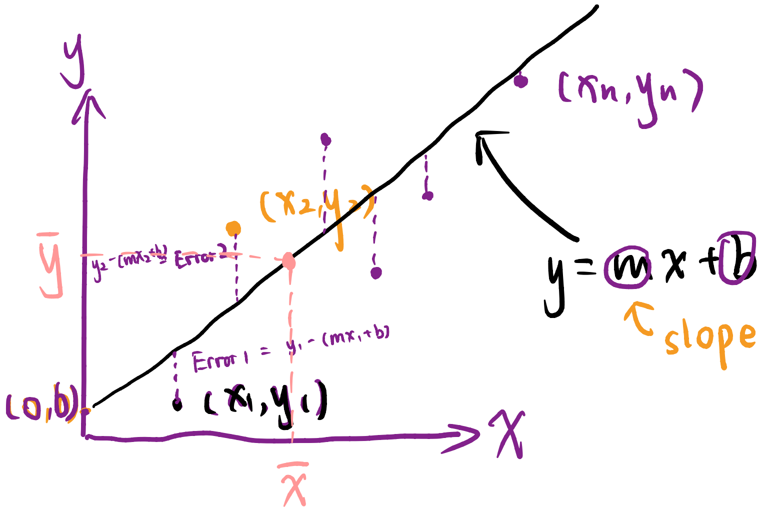
对 m 求偏导

(2-2) Find the partial derivative of this w/ respect to b

(2-3) set both of them equal to 0

$$\frac{\partial SE}{\partial m} = 0, \quad \frac{\partial SE}{\partial b} = 0$$





$$SE_{Line} = n\overline{y^2} - 2m \cdot n\overline{xy} - 2b \cdot n\bar{y} + m^2 n\overline{x^2} + 2mb \cdot n\bar{x} + nb^2$$

对m求偏导

$$-2n\overline{xy} + 2n\overline{x^2}m + 2bn\bar{x} = 0$$

$$-\overline{xy} + m\overline{x^2} + b\bar{x} = 0$$

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

$$m\bar{x} + b = \bar{y}$$

$$m\frac{\overline{x^2}}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$$

对b求偏导

$$-2n\bar{y} + 2mn\bar{x} + 2nb = 0$$

$$-\bar{y} + m\bar{x} + b = 0$$

$$\bar{y} = m\bar{x} + b$$

$$y = mx + b$$

the optimal fitting line

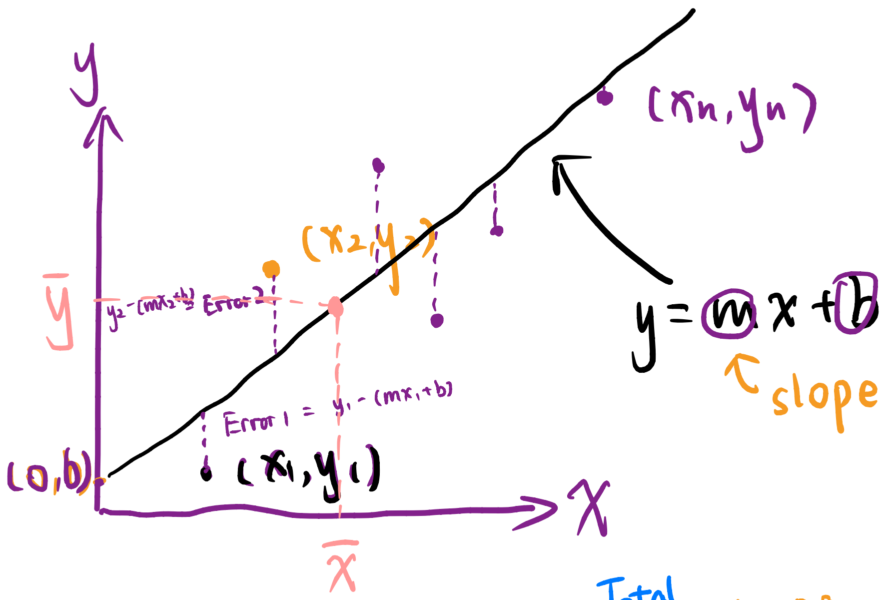
$(\bar{x}, \bar{y})$ lies on this line

$\left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ on line

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

$$m = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} \cdot \frac{(\bar{x})}{(\bar{x})}$$

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



Total 波动程度

Q: How much (what %) of the variation in y is decribed by the variation in x?

y的波动程度有多少百分比能被x的波动程度所描述

Total variation in y: $(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2 = SE_{\bar{y}}$

30%

$\boxed{\dfrac{SE_{Line}}{SE_{\bar{y}}}}$

Q: How much of total variation is not described by the regression line?

what % of total variation is not described by the variation in x (or by the regression line)

$r^2 = \boxed{1 - \dfrac{SE_{Line}\downarrow}{SE_{\bar{y}}}} =$ what % of total variation is described by the variation in x (or by the regression line)

the coefficient of determination
决定系数

SE_{Line} small ⇒ Line is a good fit 点很接近线

⇓

r^2 close to 1

covariance between 2 random variables 协方差

$$\mathrm{cov}(X,Y) = E[(X-E[X])(Y-E[Y])]$$

$$= E[XY - XE[Y] - YE[X] + E[X]E[Y]]$$

$$= E[XY] - E[X\underbrace{E[Y]}]-E[Y\underbrace{E[X]}] + E[E[X]E[Y]]$$

看作常数 如 $E[3X]=3E[X]$

$\because E[E[X]]=E[X]$

$$= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]$$

$$= E[XY] - E[Y]E[X] = \boxed{\overline{XY} - \bar{Y}\bar{X}}$$

numerator 分子 of the slope of the regression line

$E[XY] \approx \overline{XY}$ $E[Y] \approx \bar{Y}$ $E[X] \approx \bar{X}$

$$\hat{m} = \frac{\boxed{\overline{XY} - \bar{Y}\bar{X}}}{\overline{X^2} - (\bar{X})^2}$$

$\mathrm{Cov}(X,Y)$

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

$$\overline{X \cdot X} - \bar{X} \cdot \bar{X}$$

$$= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$$