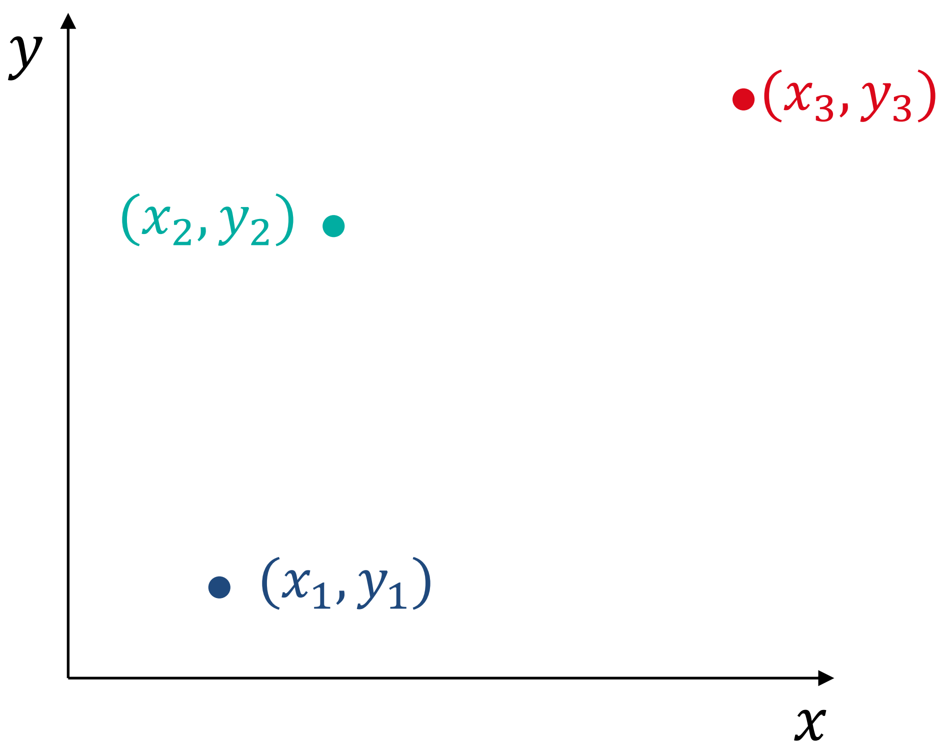
# Linear regression

Eung-Hee Kim
ehkim@sunmoon.ac.kr

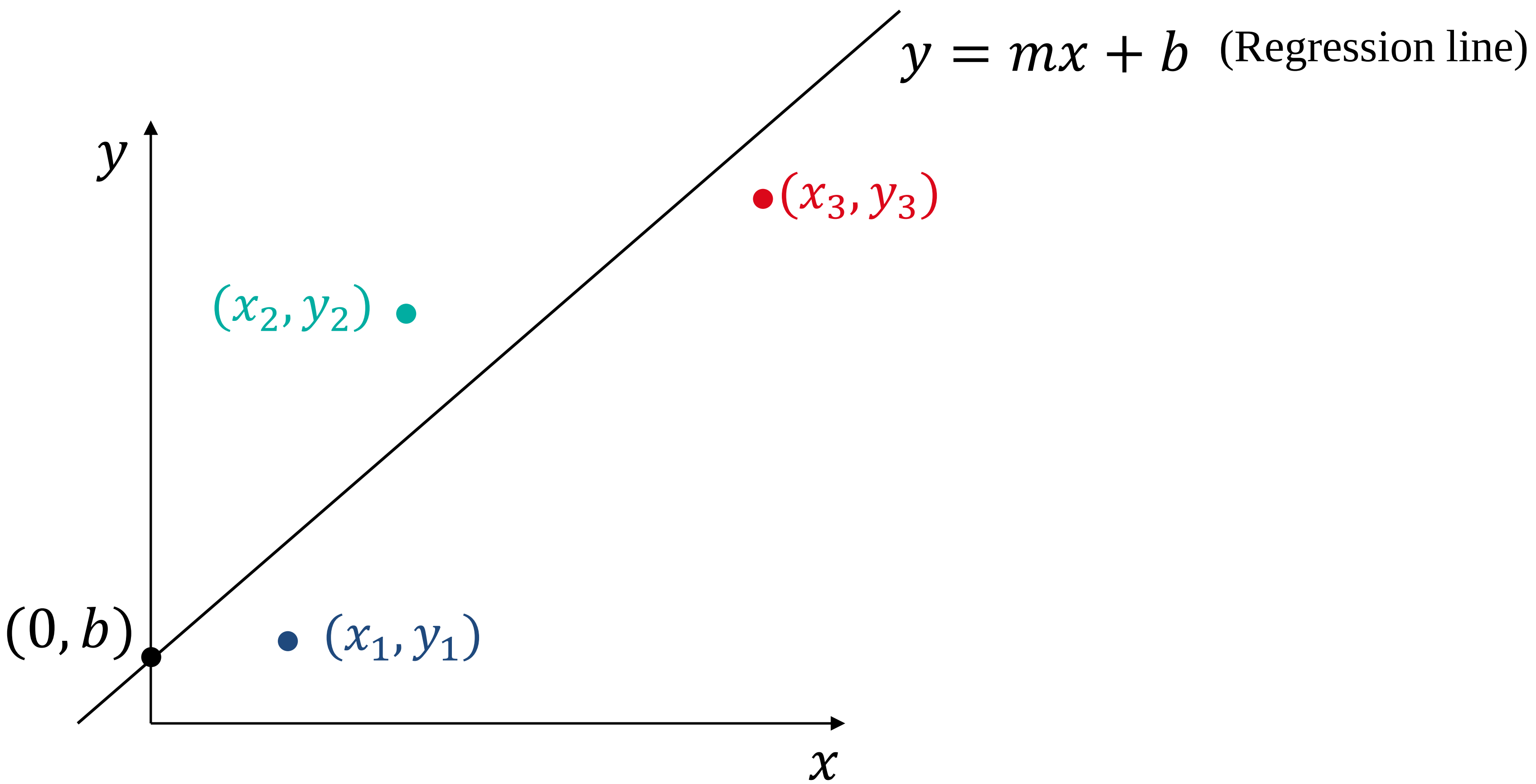
In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)



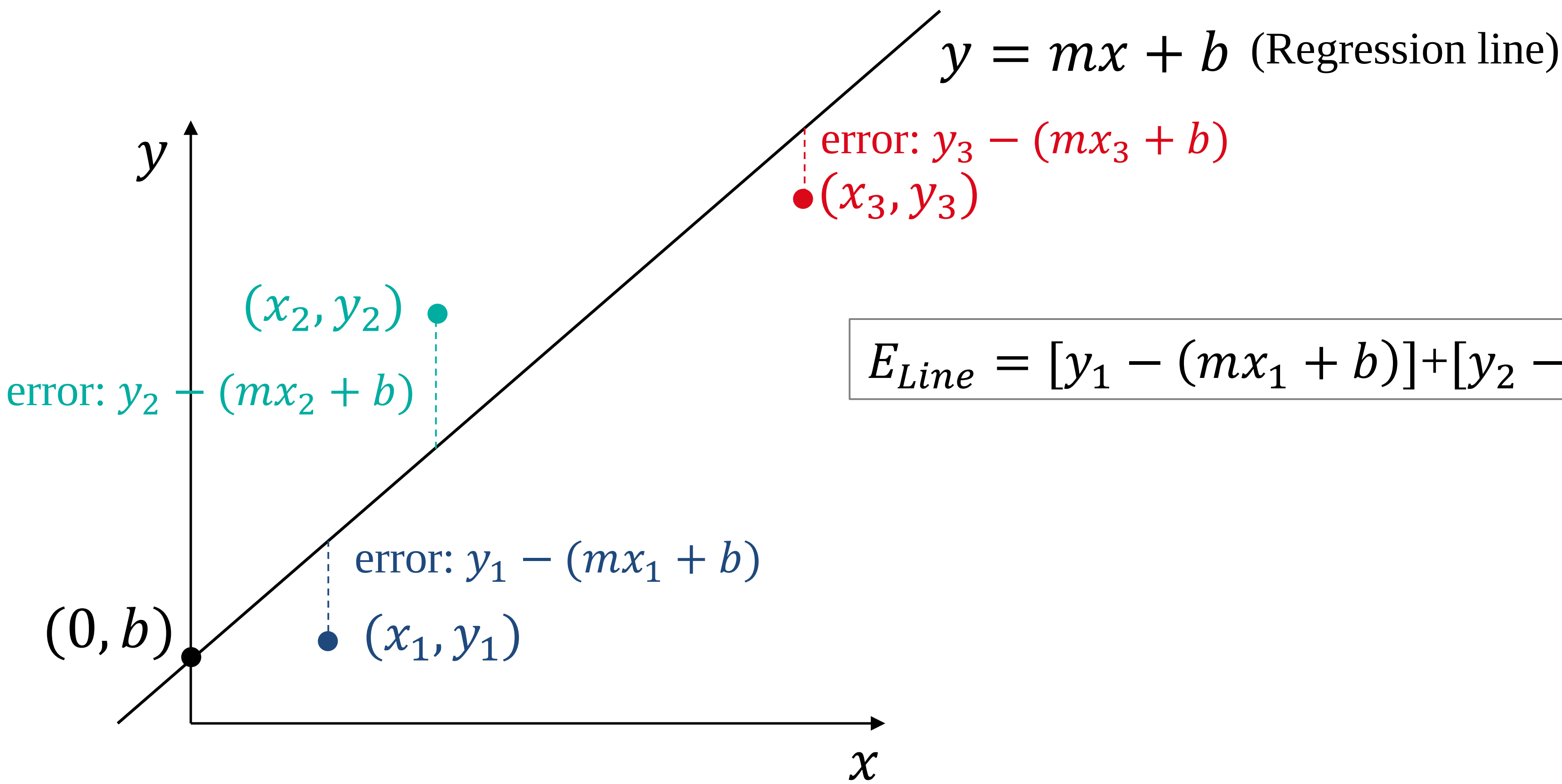
In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)



In almost every textbook dealing with the topics related to 'linear regression',

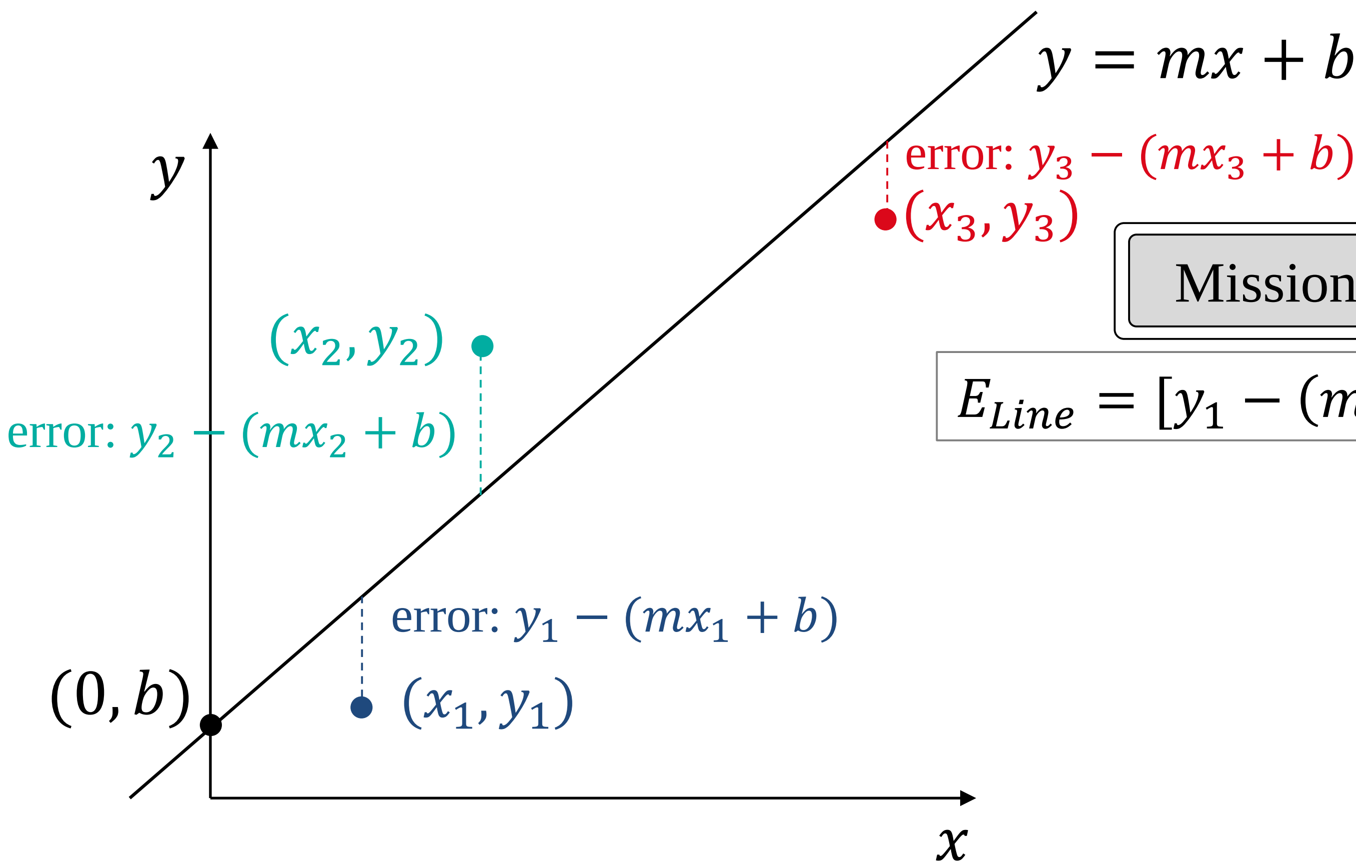
- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)




$$E_{Line} = [y_1 - (mx_1 + b)] + [y_2 - (mx_2 + b)] + [y_3 - (mx_3 + b)]$$

In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)



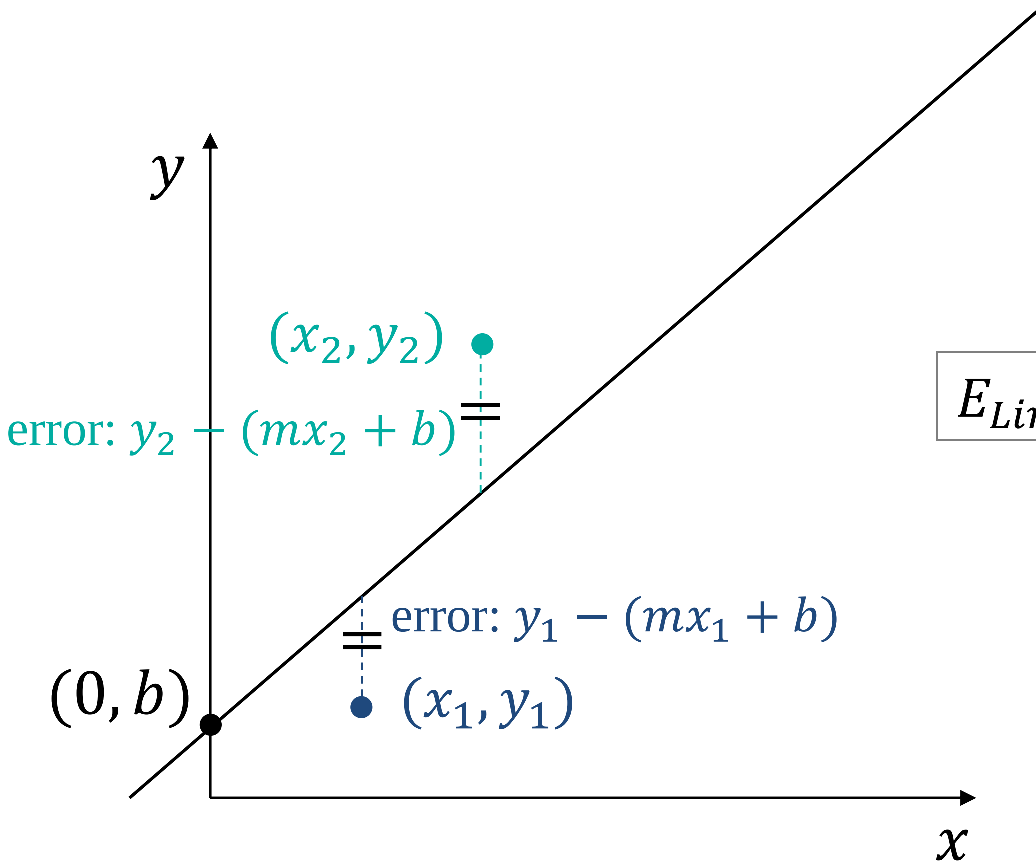

| Mission | Find $m$ and $b$ minimizing $E_{Line}$ |
|---|---|

$$E_{Line} = [y_1 - (mx_1 + b)] + [y_2 - (mx_2 + b)] + [y_3 - (mx_3 + b)]$$

# In almost every textbook dealing with the topics related to ‘linear regression’,

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)

$y = mx + b$ (Regression line)

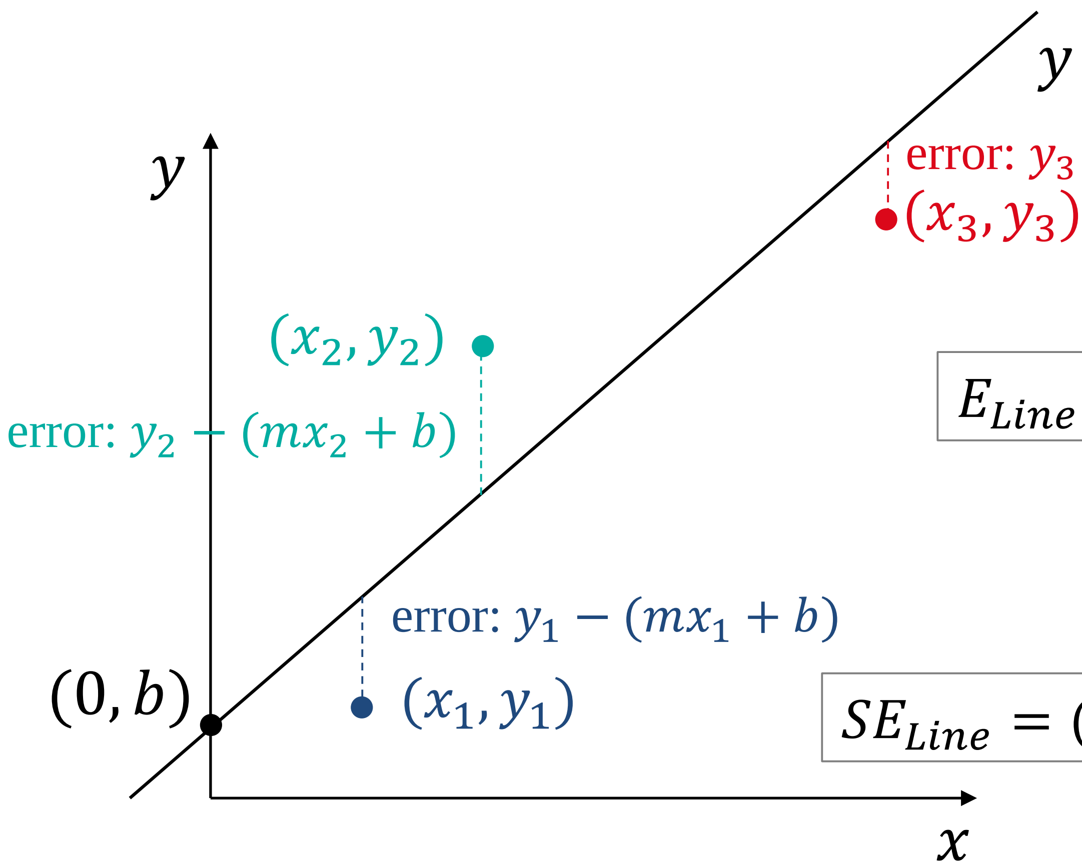



**Mission** Find $m$ and $b$ minimizing $E_{Line}$

$E_{Line} = [y_1 - (mx_1 + b)]+[y_2 - (mx_2 + b)] = \mathbf{0}$ ***Perfect line!***

In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)



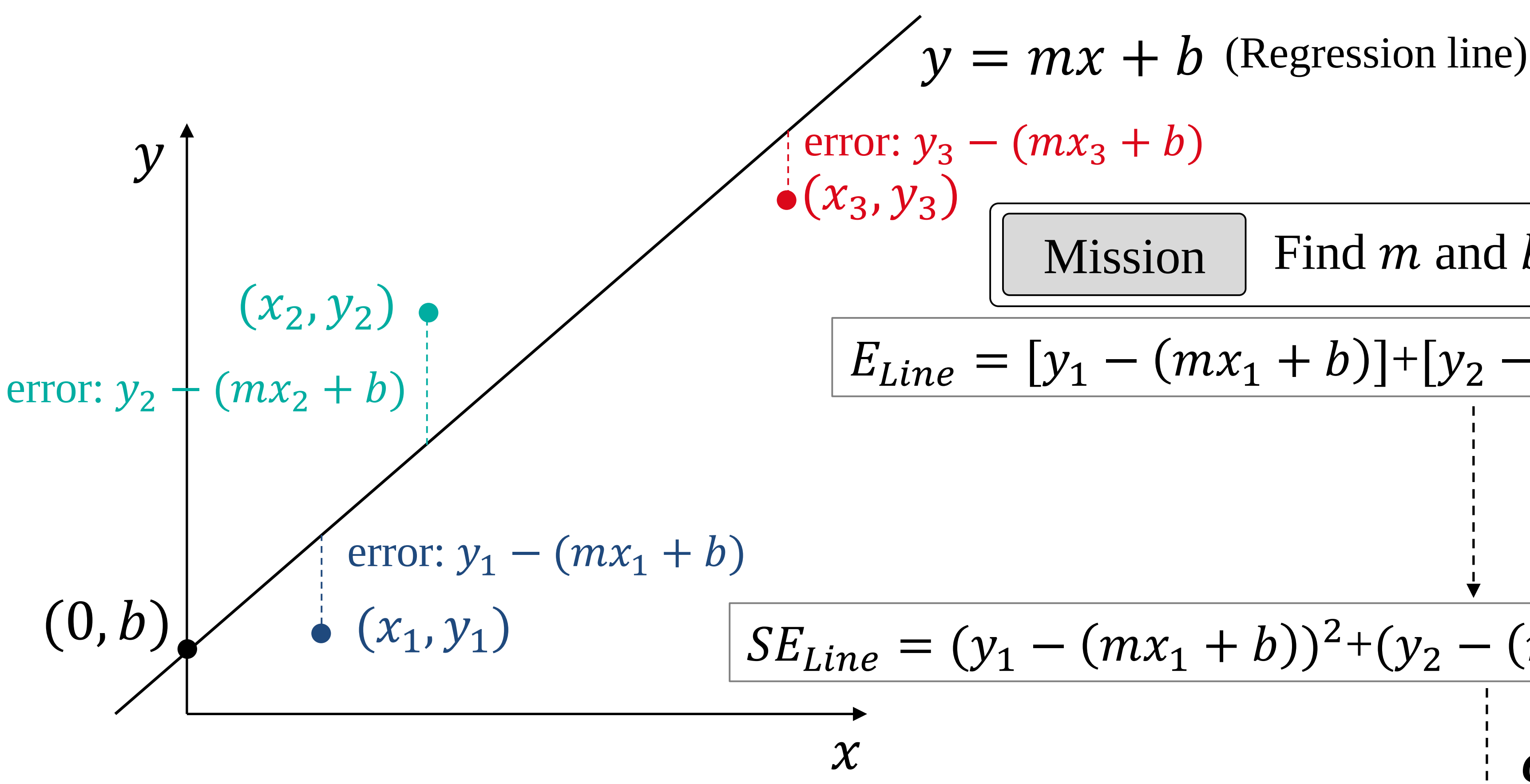

**Mission** Find $m$ and $b$ minimizing $E_{Line}$

$$E_{Line} = [y_1 - (mx_1 + b)] + [y_2 - (mx_2 + b)] + [y_3 - (mx_3 + b)]$$

$$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + (y_3 - (mx_3 + b))^2$$

In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)




**Mission** Find $m$ and $b$ minimizing $E_{Line}$

$$E_{Line} = [y_1 - (mx_1 + b)] + [y_2 - (mx_2 + b)] + [y_3 - (mx_3 + b)]$$

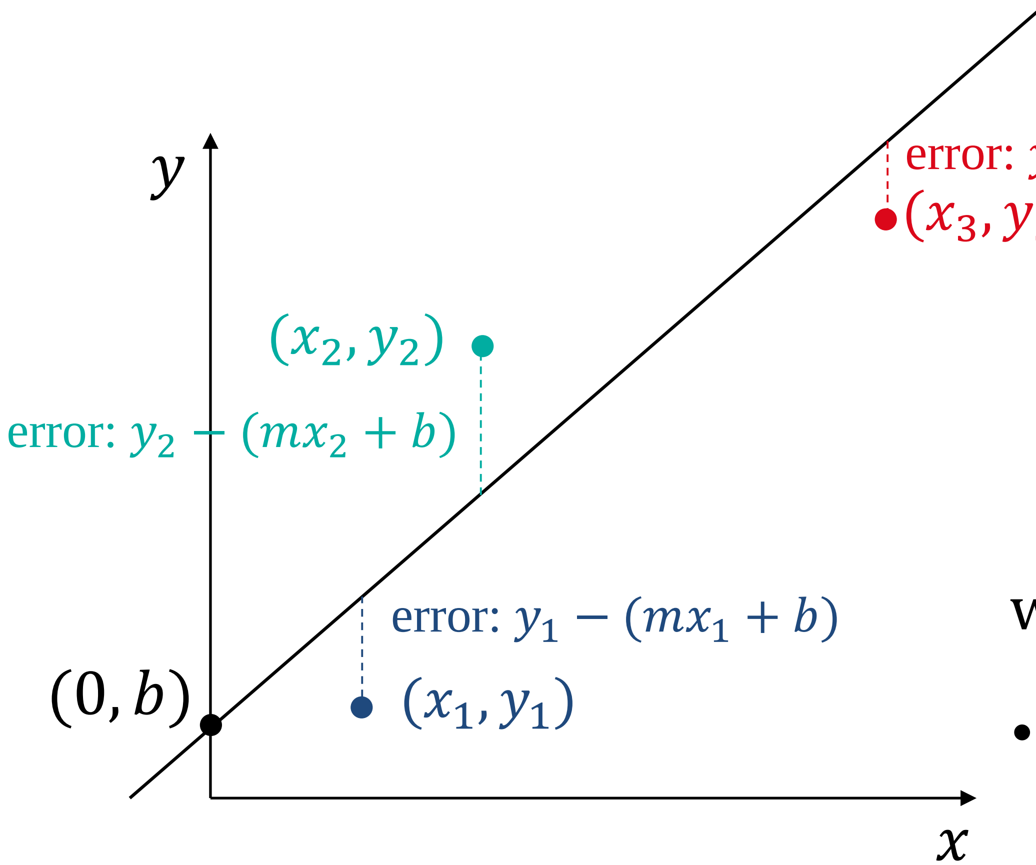
$$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + (y_3 - (mx_3 + b))^2$$

***Generalization***

$$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2$$

In almost every textbook dealing with the topics related to 'linear regression',

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)




$y = mx + b$ (Regression line)

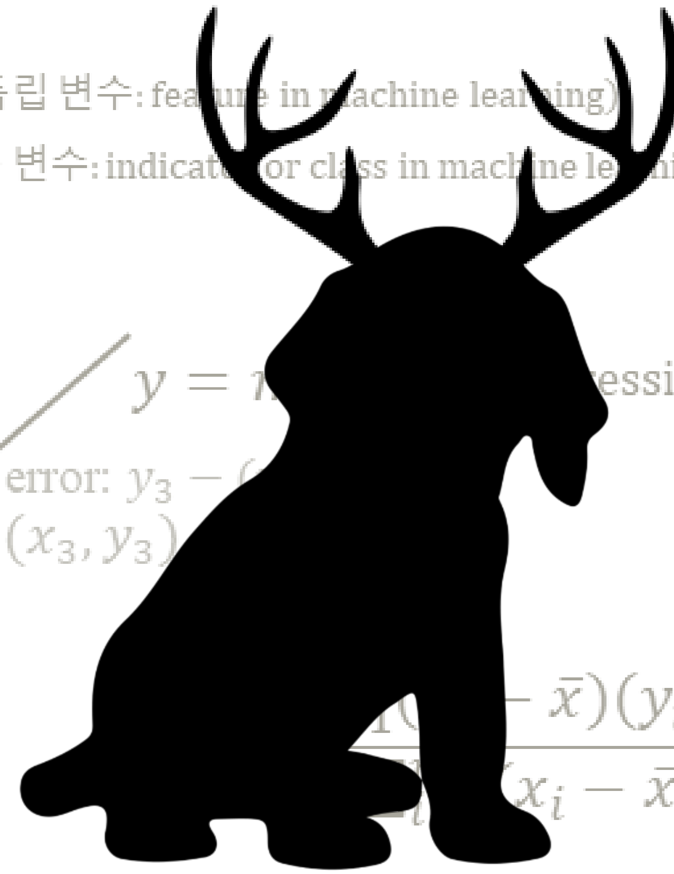
$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad b = \bar{y} - m\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the mean of $x$ and $y$ respectively.

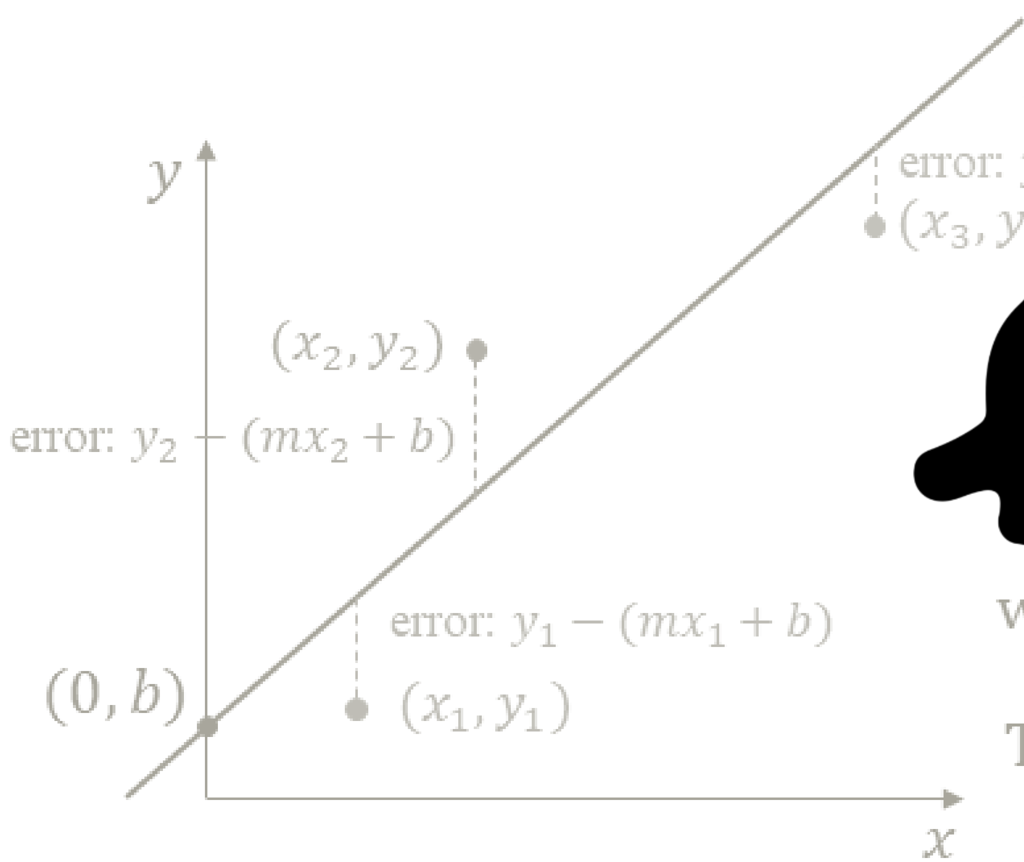
- The point $(\bar{x}, \bar{y})$ lies on the regression line.

In almost every textbook dealing with the topics related to ‘linear regression’,

- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: fea[illegible] in [illegible]achine lea[illegible]ing)
  - y: dependent variable (종속 변수: indicat[illegible] or cla[illegible]ss in mach[illegible]ine le[illegible]ning)



$y = m$ [illegible] [illegible]ession line)



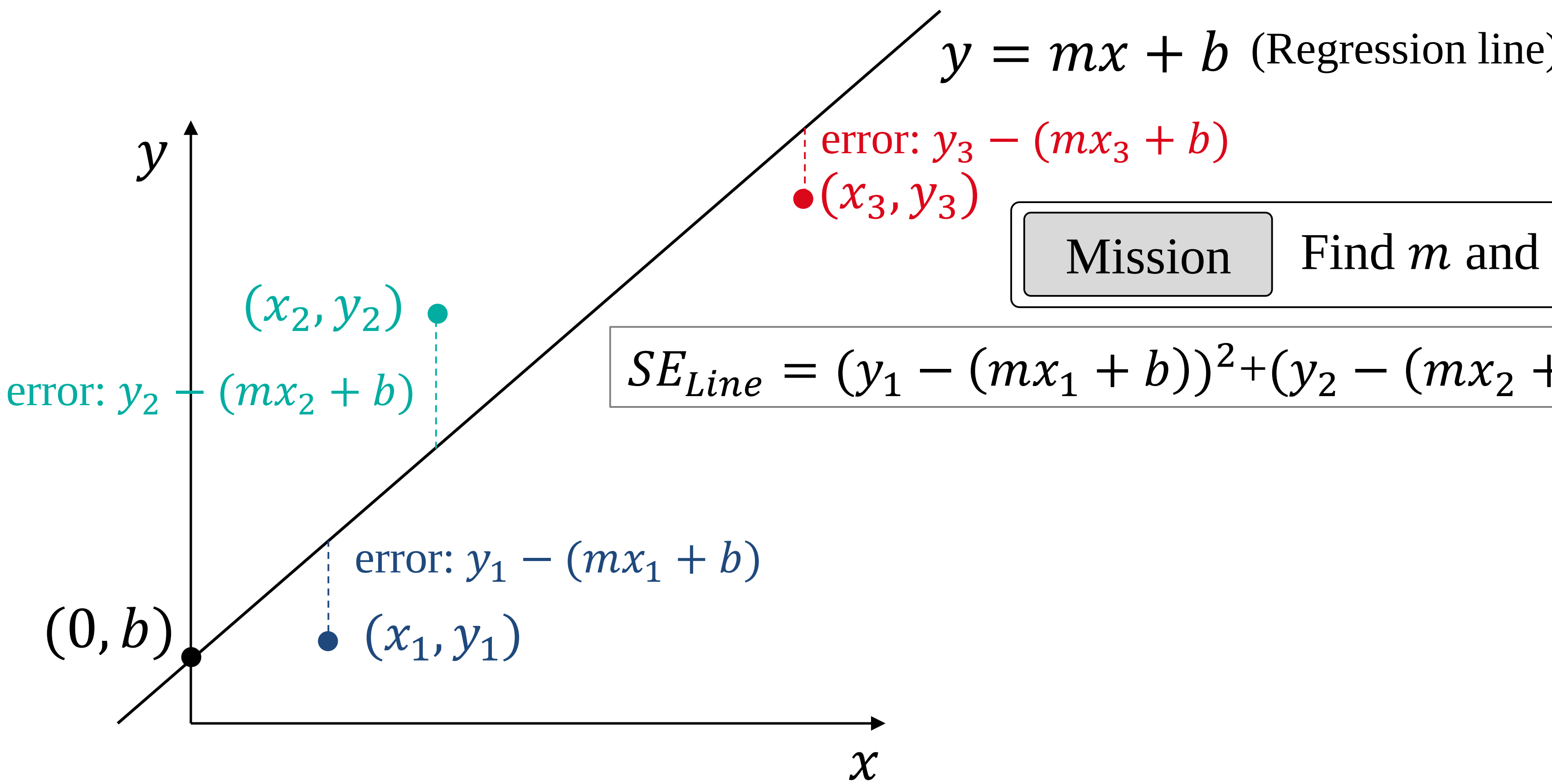

[illegible] $- \bar{x})(y_i - \bar{y})$ / [illegible] $(x_i - \bar{x})^2$ and $b = \bar{y} - m\bar{x}$

where $\bar{x}$ and $\bar{y}$ are the mean of $x$ and $y$ respectively.

The point $(\bar{x}, \bar{y})$ lies on the regression line.

# Proof

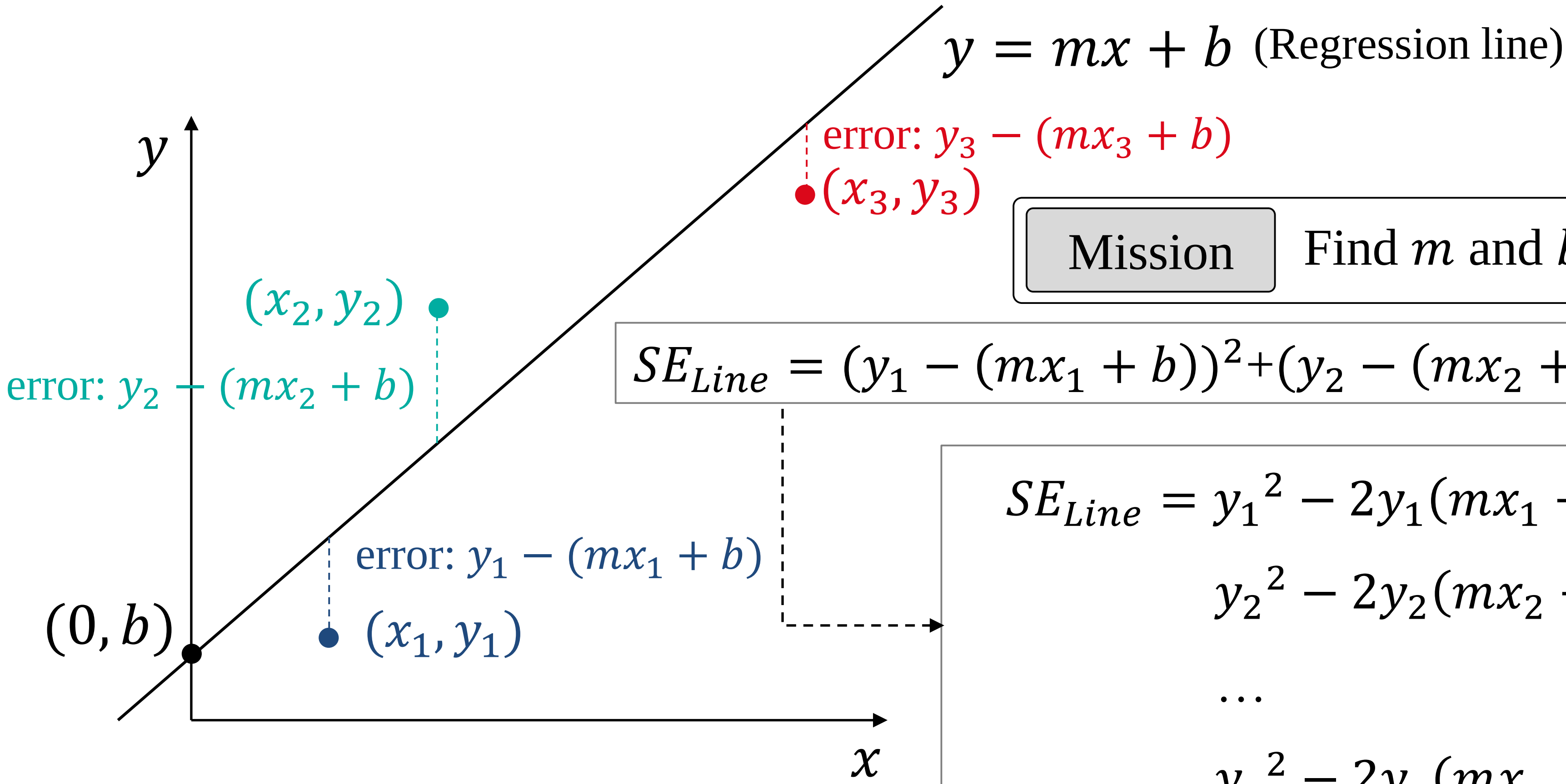
- Let $x$ and $y$ be
  - $x$: independent variable (독립 변수: feature in machine learning)
  - y: dependent variable (종속 변수: indicator or class in machine learning)




**Mission** Find $m$ and $b$ minimizing $E_{Line}$

$$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2$$

# Proof (cont.)




**Mission** Find $m$ and $b$ minimizing $E_{Line}$

$$SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2$$

$$\begin{aligned} SE_{Line} = {} & {y_1}^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 + \\ & {y_2}^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 + \\ & \ldots \\ & {y_n}^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \end{aligned}$$

# Proof (cont.)

$$\begin{aligned} SE_{Line} = {} & {y_1}^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 + \\ & {y_2}^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 + \\ & \dots \\ & {y_n}^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \end{aligned}$$

$$\begin{aligned} SE_{Line} = {} & {y_1}^2 - 2y_1mx_1 - 2y_1b + m^2{x_1}^2 + 2mx_1b + b^2 + \\ & {y_2}^2 - 2y_2mx_2 - 2y_2b + m^2{x_2}^2 + 2mx_2b + b^2 + \\ & \dots \\ & {y_n}^2 - 2y_nmx_n - 2y_nb + m^2{x_n}^2 + 2mx_nb + b^2 \end{aligned}$$

# Proof (cont.)

$$SE_{Line} = {y_1}^2 - 2y_1mx_1 - 2y_1b + m^2{x_1}^2 + 2mx_1b + b^2 +$$
$${y_2}^2 - 2y_2mx_2 - 2y_2b + m^2{x_2}^2 + 2mx_2b + b^2 +$$
$$\cdots$$
$${y_n}^2 - 2y_nmx_n - 2y_nb + m^2{x_n}^2 + 2mx_nb + b^2$$

$$SE_{Line} = ({y_1}^2 + {y_2}^2 + \ \ldots + {y_n}^2) - 2m(y_1x_1 + y_2x_2 + \cdots + y_nx_n) - 2b(y_1 + y_2 + \cdots + y_n) +$$
$$m^2({x_1}^2 + {x_2}^2 + \cdots + {x_n}^2) + 2mb(x_1 + x_2 + \cdots x_n) + nb^2$$

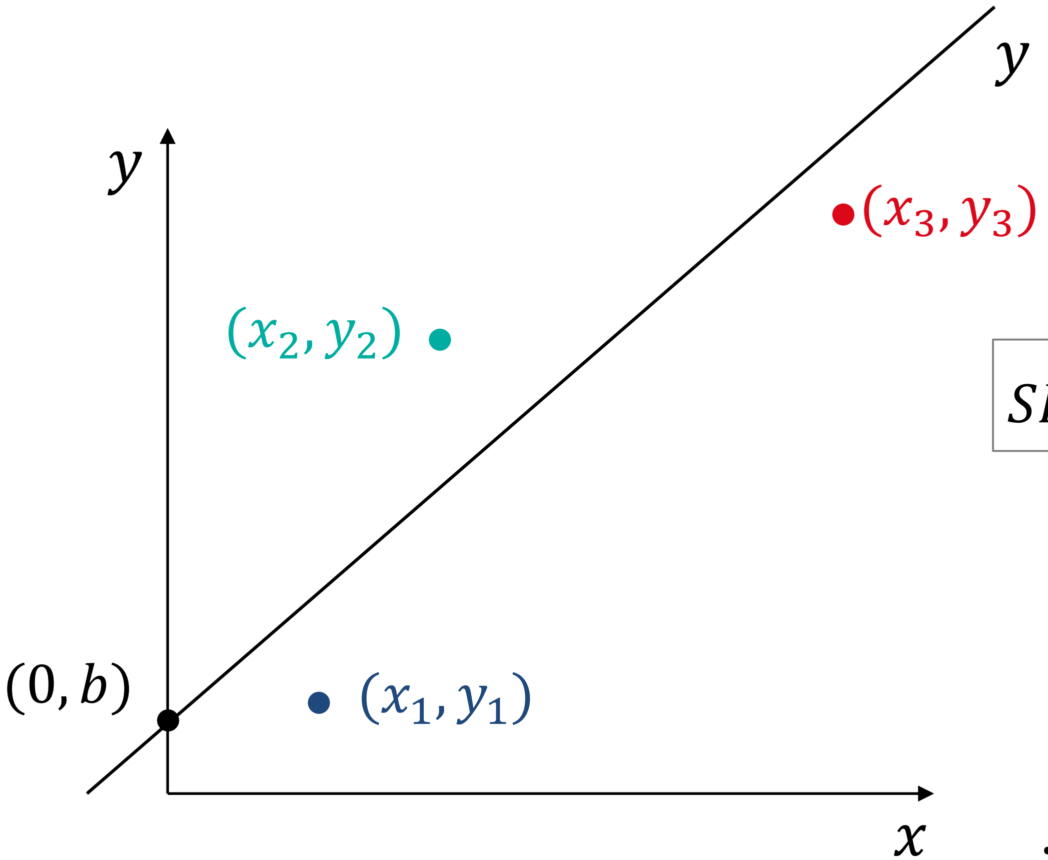
# Proof (cont.)

- Notation for average: $\bar{\alpha} = \frac{\alpha_1+\alpha_2+\cdots+\alpha_n}{n}$
- $\therefore\ \alpha_1 + \alpha_2 + \cdots + \alpha_n = n\bar{\alpha}$

$$\boxed{\begin{aligned} SE_{Line} = {} & ({y_1}^2 + {y_2}^2 + \ \ldots + {y_n}^2) - 2m(y_1x_1 + y_2x_2 + \cdots + y_nx_n) - 2b(y_1 + y_2 + \cdots + y_n) + \\ & m^2({x_1}^2 + {x_2}^2 + \cdots + {x_n}^2) + 2mb(x_1 + x_2 + \cdots x_n) + nb^2 \end{aligned}}$$

$$\boxed{SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2}$$

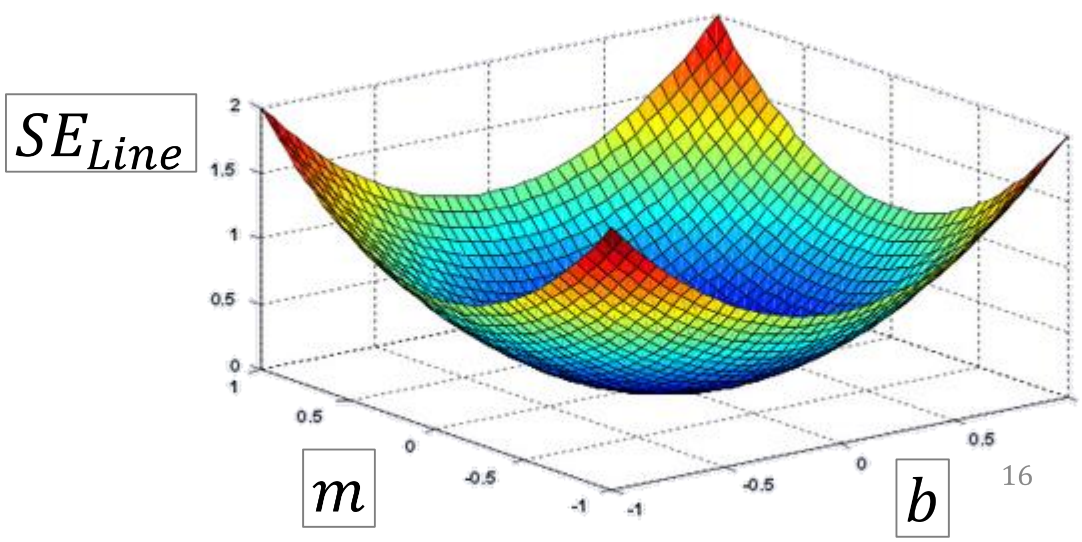
# Proof (cont.)



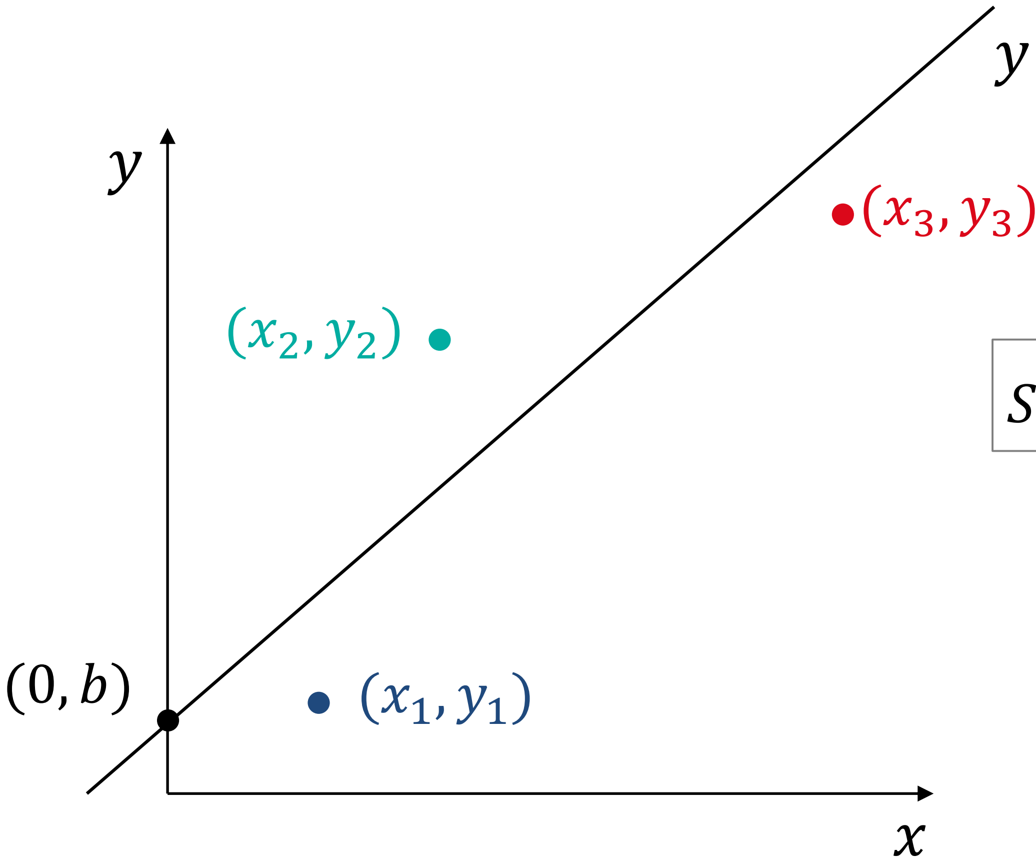

$$\boxed{SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2}$$

Brown colored ones were decided (constants) at the moment when we got our data.

$\therefore$ Only $m$ and $b$ are the **variables** of the function $SE_{Line}$.



# Proof (cont.)

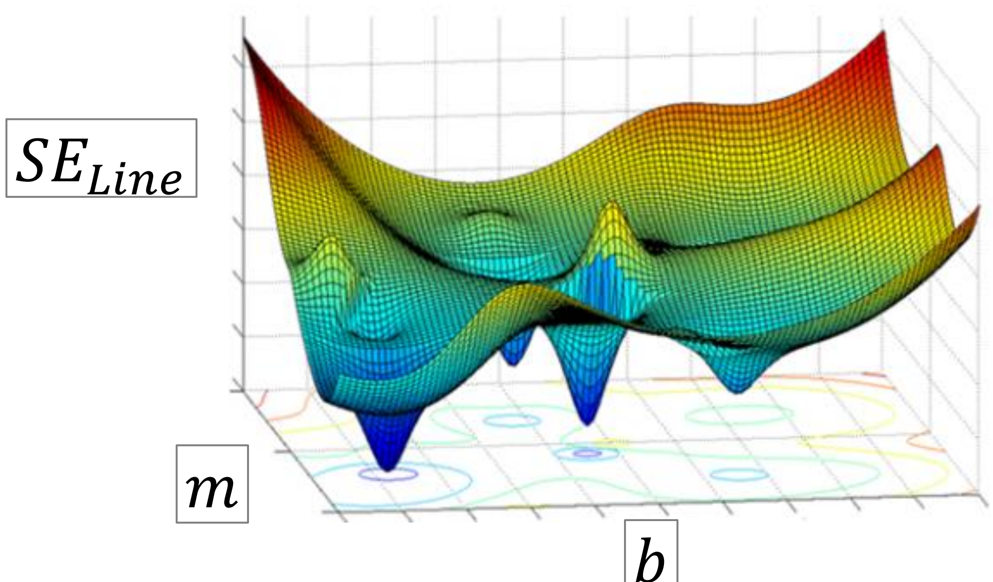



$y = mx + b$ (Regression line)

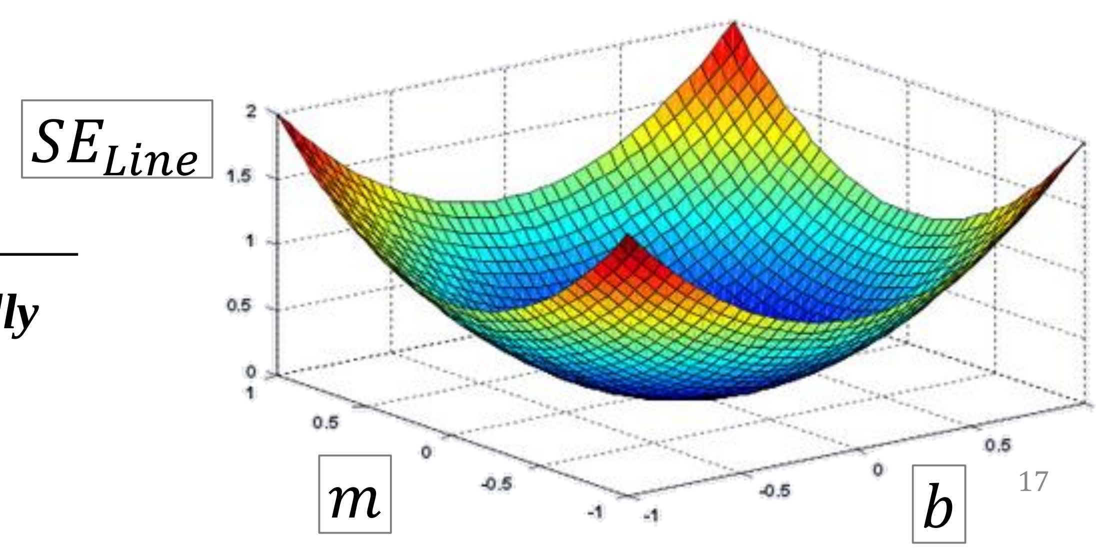
$$\boxed{SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2}$$

Brown colored ones were decided (actual numbers) at the moment when we got our data.

$\therefore$ Only $m$ and $b$ are the **variables** of the function $SE_{Line}$.




***theoretically possible***



# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

$$\frac{\partial SE_{Line}}{\partial m} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0$$

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

$$\frac{\partial SE_{Line}}{\partial m} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0$$

$$\frac{\partial SE_{Line}}{\partial m} = 0 \equiv -2n\overline{xy} + 2mn\overline{x^2} + 2bn\bar{x} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0 \equiv -2n\bar{y} + 2mn\bar{x} + 2nb = 0$$

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2$$

$$\frac{\partial SE_{Line}}{\partial m} = 0$$

$$\frac{\partial SE_{Line}}{\partial m} = 0 \equiv -2n\overline{xy} + 2mn\overline{x^2} + 2bn\bar{x} = 0$$

$$\frac{\partial SE_{Line}}{\partial m} = 0 \equiv -\overline{xy} + m\overline{x^2} + b\bar{x} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0 \equiv -2n\bar{y} + 2mn\bar{x} + 2nb = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0 \equiv -\bar{y} + m\bar{x} + b = 0$$

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

$$\frac{\partial SE_{Line}}{\partial m} = 0 \equiv -\overline{xy} + m\overline{x^2} + b\bar{x} = 0$$

$$\frac{\partial SE_{Line}}{\partial b} = 0 \equiv -\bar{y} + m\bar{x} + b = 0$$

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

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

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

$$\frac{\partial SE_{Line}}{\partial m} = 0 \equiv -\overline{xy} + m\overline{x^2} + b\bar{x} = 0$$

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

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

$$\frac{\partial SE_{Line}}{\partial b} = 0 \equiv -\bar{y} + m\bar{x} + b = 0$$

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

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

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

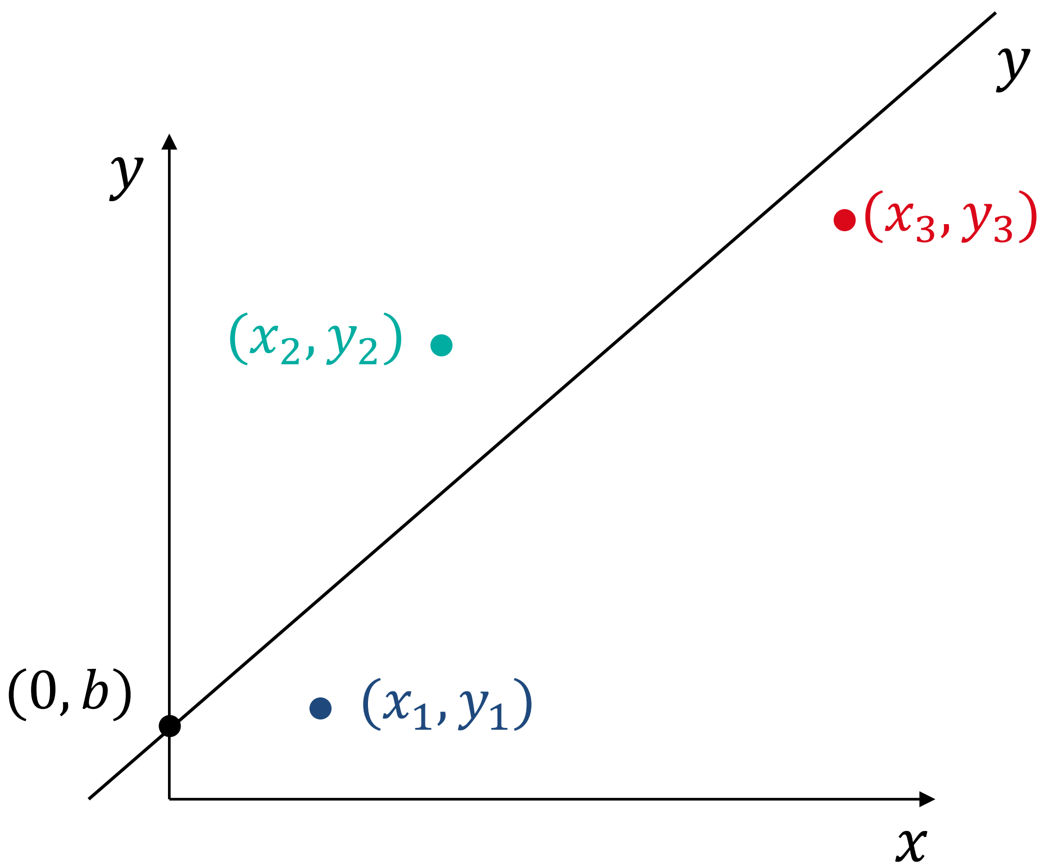
# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

$y = mx + b$ (Regression line)




① $(\bar{x}, \bar{y})$ lies on the regression line.

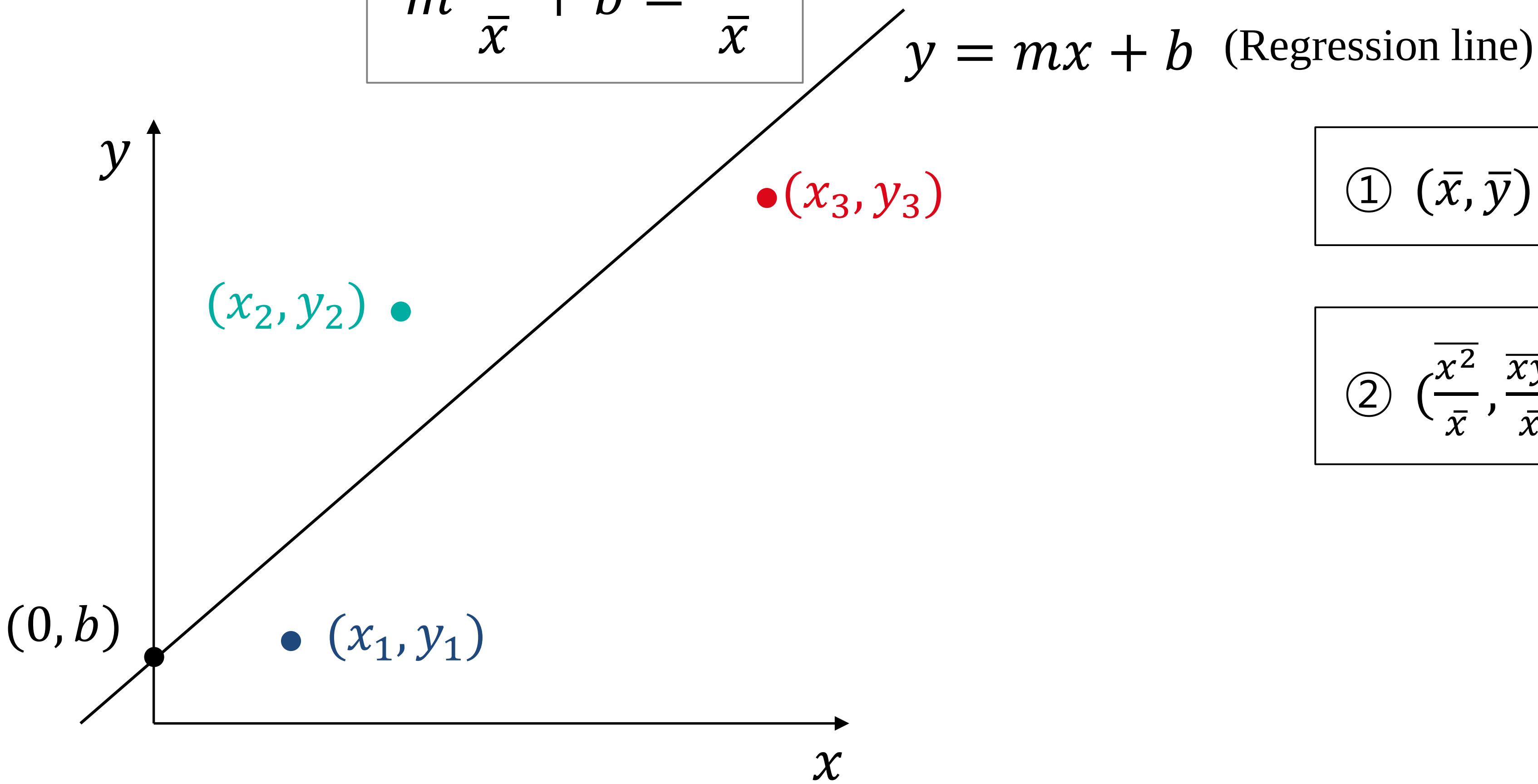
# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

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




① $(\bar{x}, \bar{y})$ lies on the regression line.

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

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

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

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

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

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

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

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

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

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

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

# Proof (cont.)

$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

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

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

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

$$\boxed{b = \bar{y} - m\bar{x}}$$

# Proof (cont.)

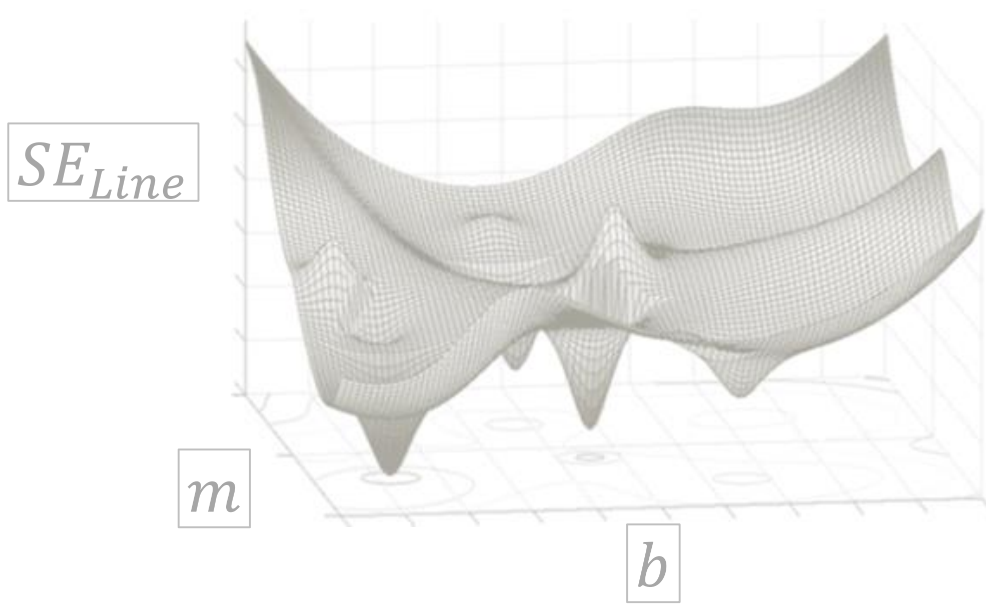
$$SE_{Line} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2n\overline{x^2} + 2mbn\bar{x} + nb^2$$

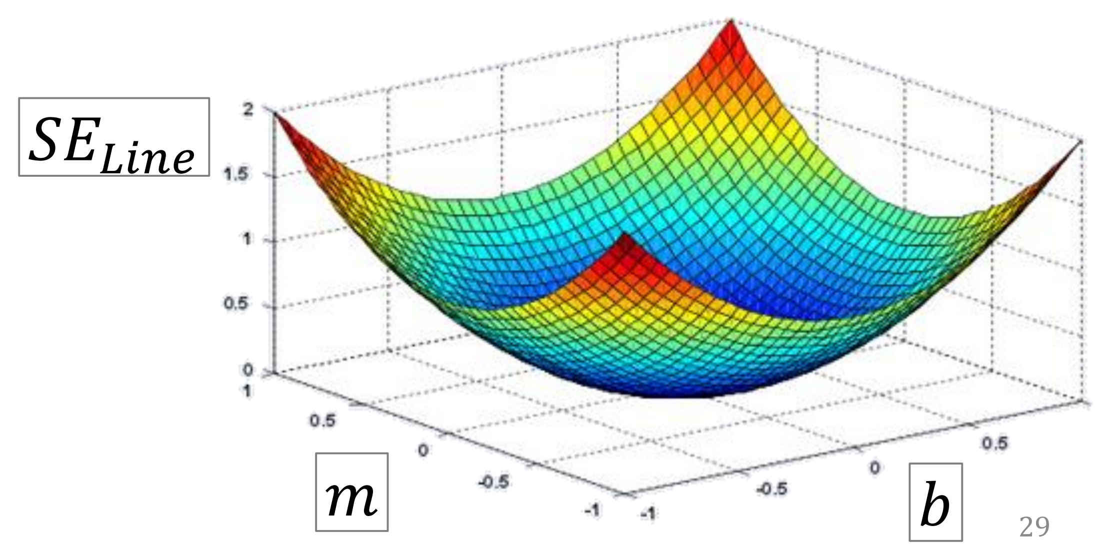
$$m\frac{\overline{x^2}}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$$

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

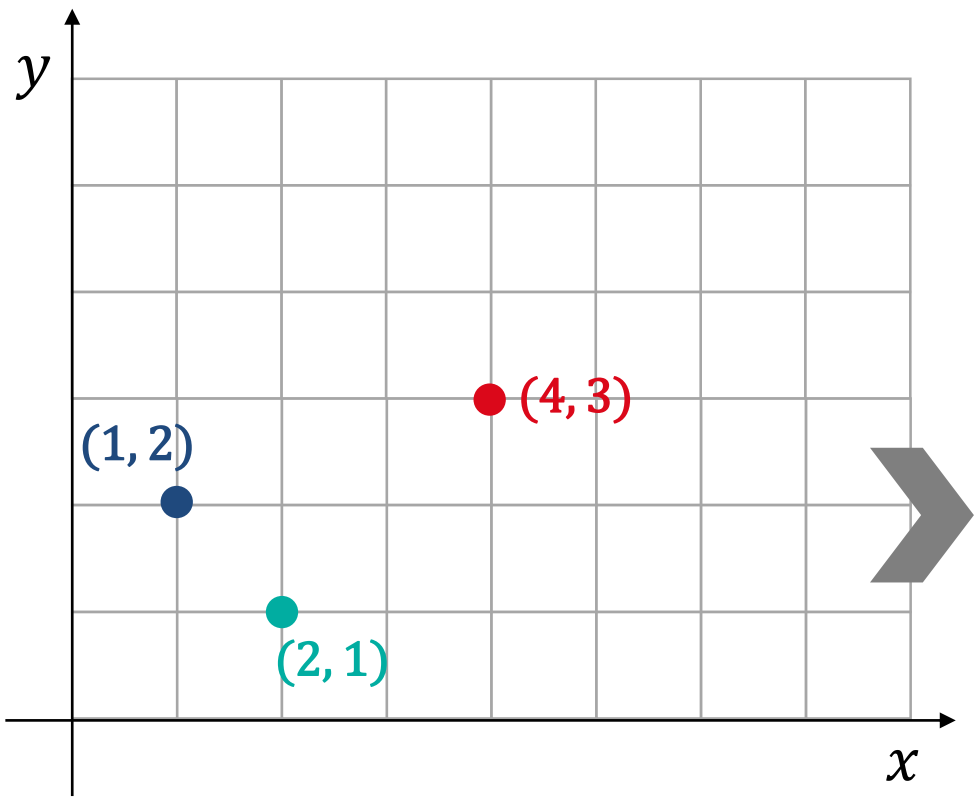
$$m\bar{x} + b = \bar{y}$$

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



# Regression line example




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

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

$$\bar{x} = \frac{1 + 2 + 4}{3} = \frac{7}{3} \qquad \bar{y} = \frac{2 + 1 + 3}{3} = 2$$

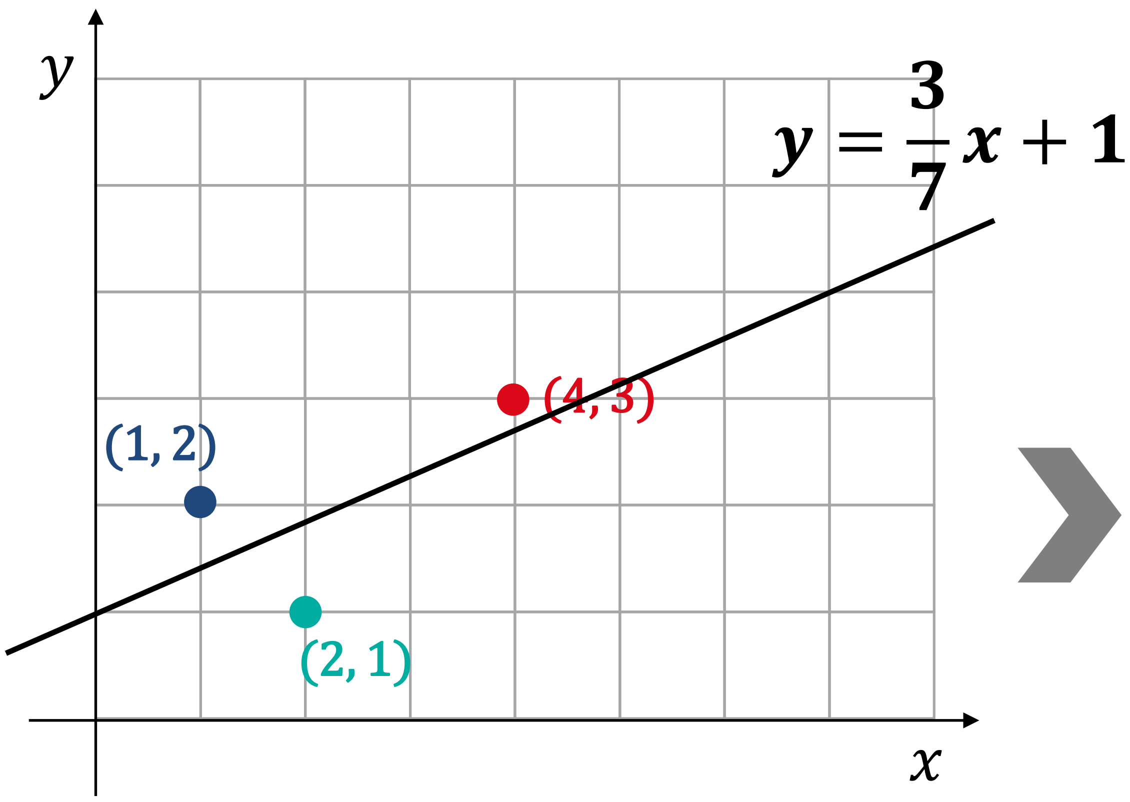
$$\overline{xy} = \frac{1 \cdot 2 + 2 \cdot 1 + 4 \cdot 3}{3} = \frac{16}{3}$$

$$\overline{x^2} = \frac{1^2 + 2^2 + 4^2}{3} = 7$$

$$m = \frac{\frac{7}{3} \cdot 2 - \frac{16}{3}}{(\frac{7}{3})^2 - 7} = \frac{\frac{14 - 16}{3}}{\frac{49 - 63}{9}} = \frac{-\frac{2}{3}}{-\frac{14}{9}} = \frac{18}{42} = \frac{3}{7}$$

$$b = 2 - \frac{3}{7} \cdot \frac{7}{3} = 1$$

# Regression line example (cont.)




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

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

$$\bar{x} = \frac{1 + 2 + 4}{3} = \frac{7}{3} \qquad \bar{y} = \frac{2 + 1 + 3}{3} = 2$$

$$\overline{xy} = \frac{1 \cdot 2 + 2 \cdot 1 + 4 \cdot 3}{3} = \frac{16}{3}$$

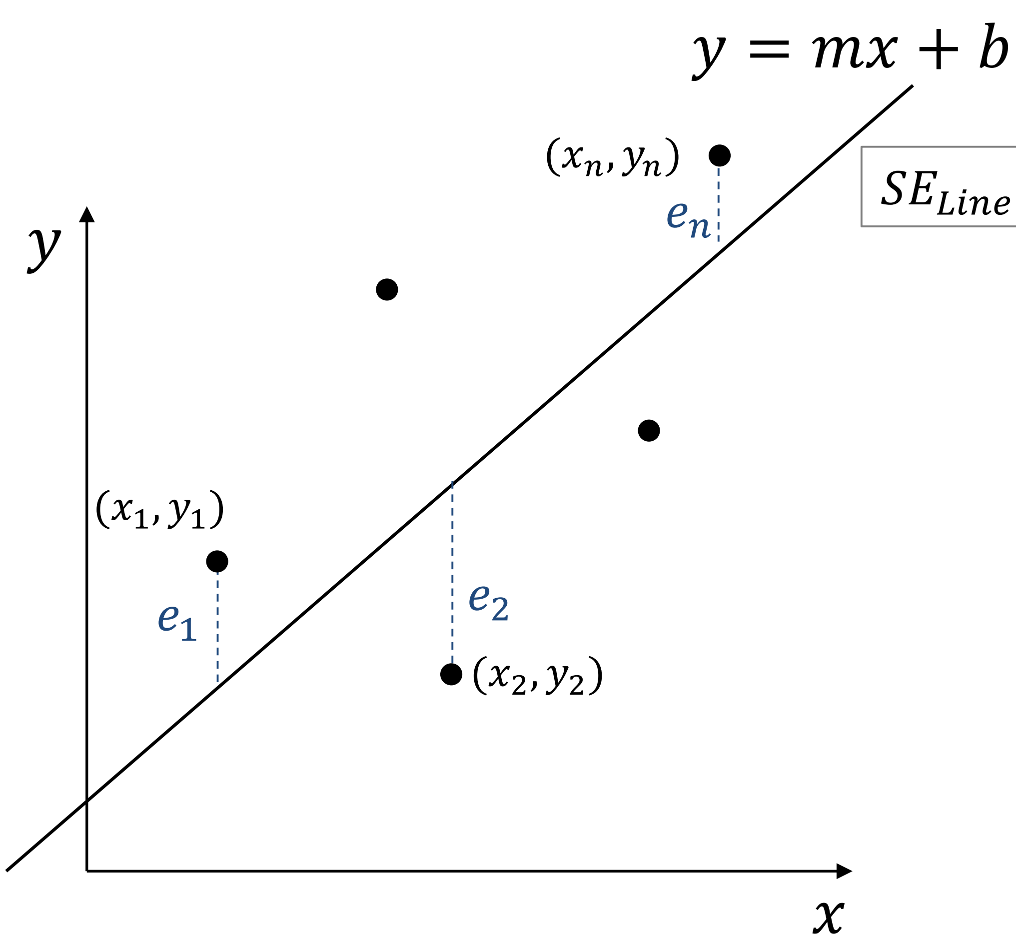
$$\overline{x^2} = \frac{1^2 + 2^2 + 4^2}{3} = 7$$

$$m = \frac{3}{7}$$

$$b = 1$$

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line (or by $x$).

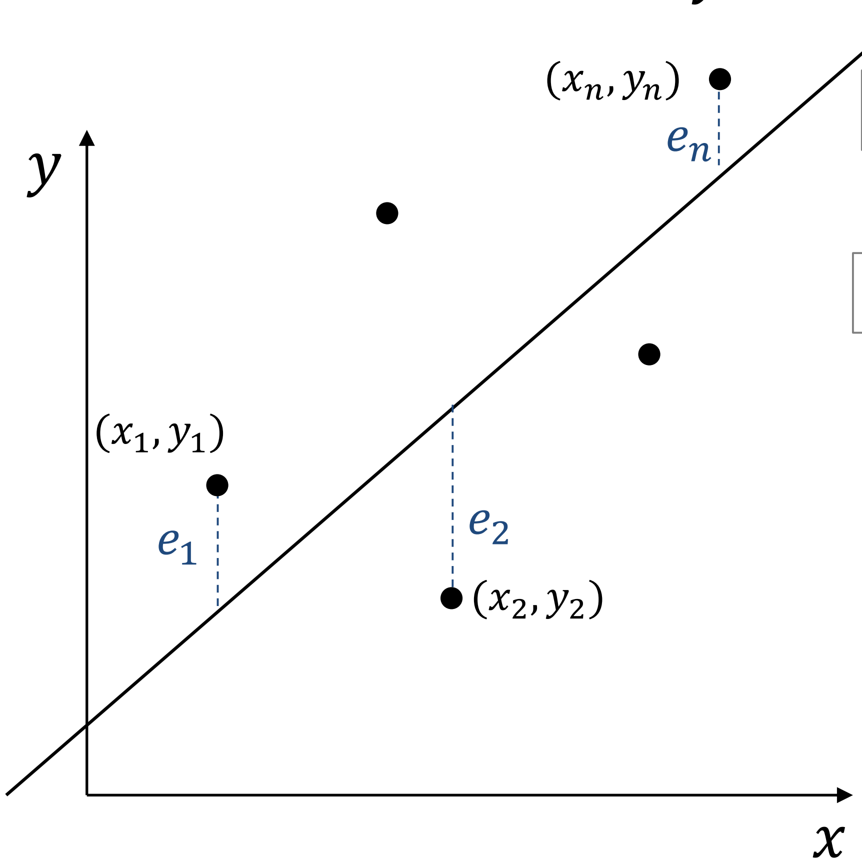



$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line (or by $x$).

$y = mx + b$

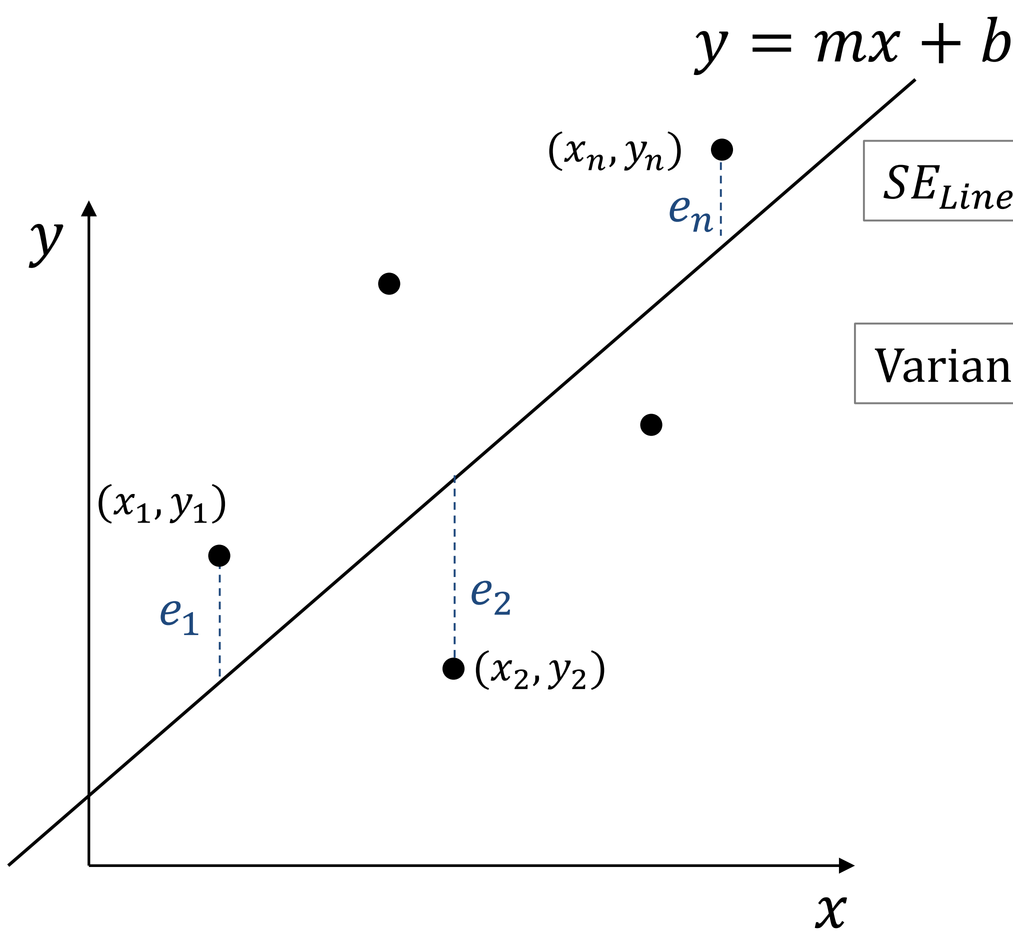



$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

$$\boxed{\text{Varianceof } y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}$$

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line (or by $x$).

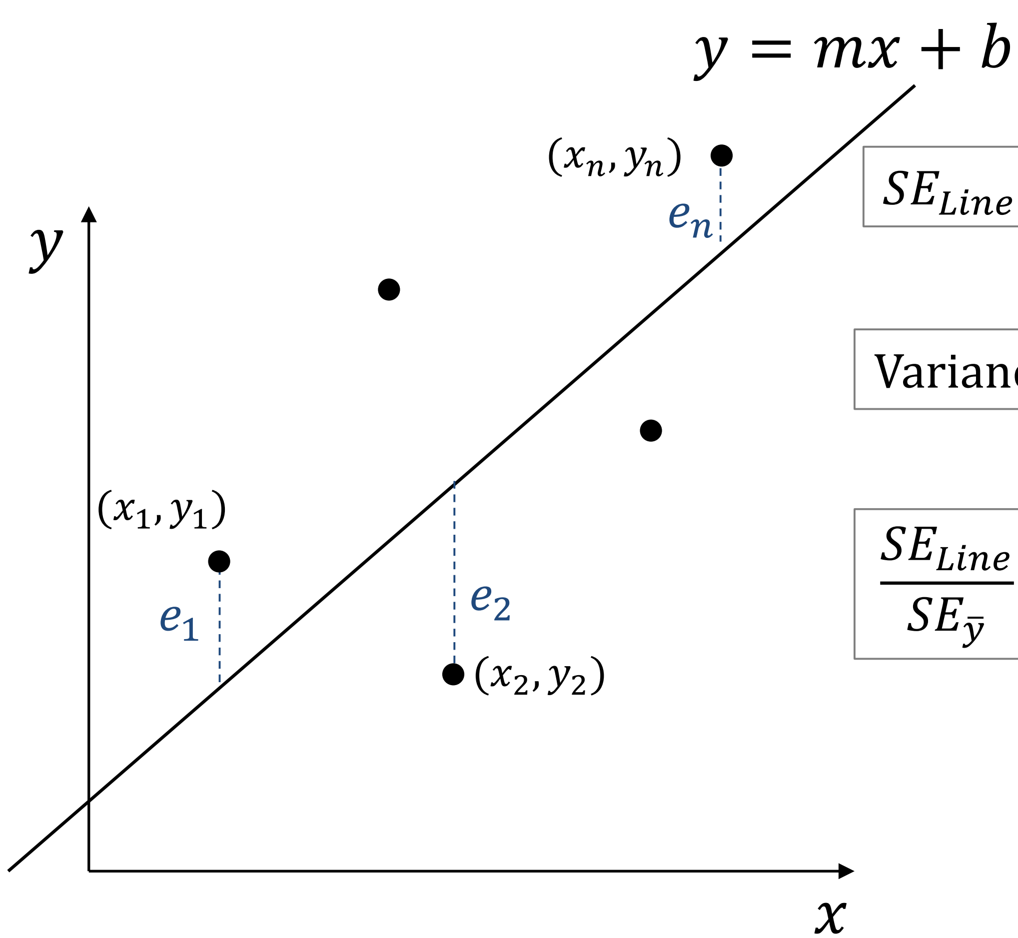



$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

$$\boxed{\text{Varianceof } y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = \boldsymbol{SE_{\bar{y}}}$$

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line.



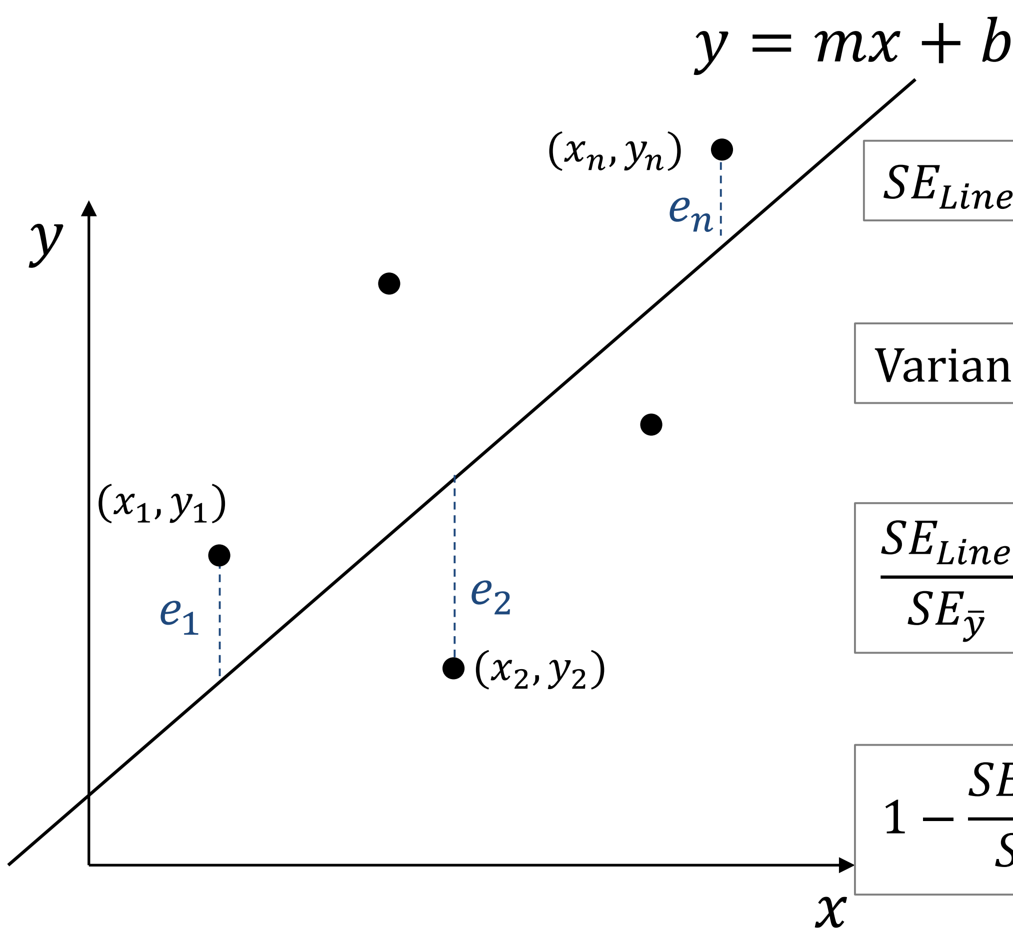

$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

$$\boxed{\text{Variance of } y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = SE_{\bar{y}}$$

$\boxed{\dfrac{SE_{Line}}{SE_{\bar{y}}}}$ : how much of the variation of $y$ is not described by the regression line?

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line?




$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

$$\boxed{\text{Variance of } y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = SE_{\bar{y}}$$

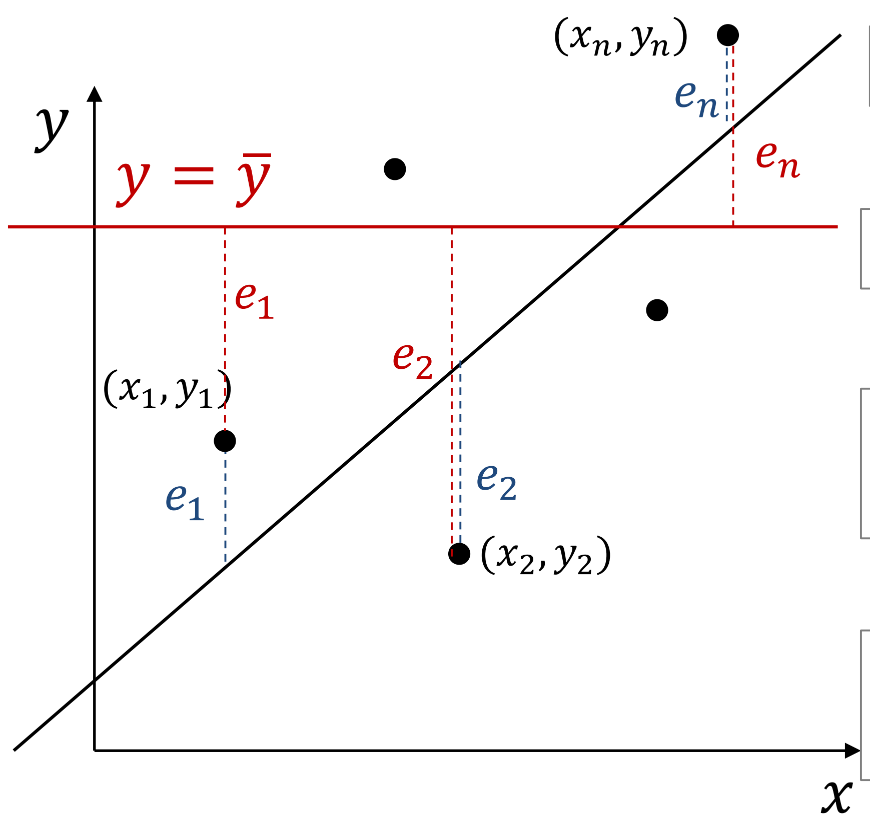
$\boxed{\dfrac{SE_{Line}}{SE_{\bar{y}}}}$ : how much of the variation of $y$ is not described by the regression line?

$\boxed{1 - \dfrac{SE_{Line}}{SE_{\bar{y}}}}$ : what % of variation in $y$ is described by the regression line?

# Coefficient of determination

- **Definition:** the proportion of the variance in the dependent variable that is predictable from independent variable(s).
  - What % of variation in $y$ is described by the regression line?

$y = mx + b$




$$\boxed{SE_{Line} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots + (y_n - (mx_n + b))^2}$$

$$\boxed{\text{Variance of } y = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = SE_{\bar{y}}$$

$\boxed{\dfrac{SE_{Line}}{SE_{\bar{y}}}}$ : how much of the variation in $y$ is not described by the regression line?

$\boxed{1 - \dfrac{SE_{Line}}{SE_{\bar{y}}}}$ : what % of variation in $y$ is described by the regression line?

# Q & A