# An O(n) method for linear regression

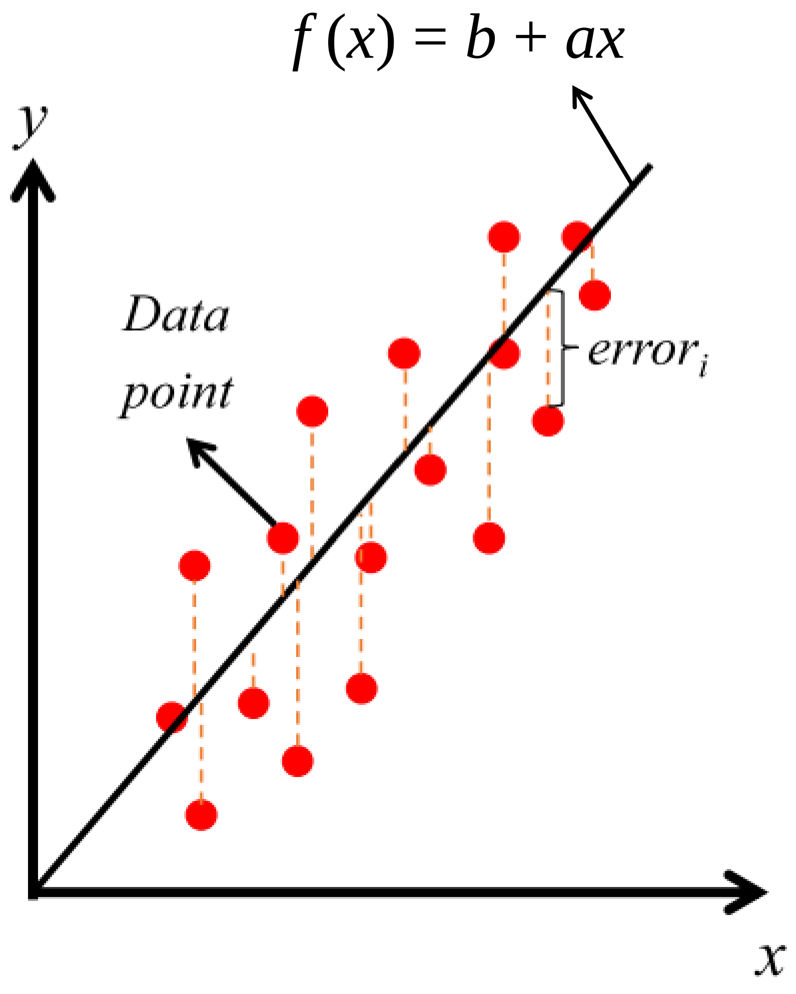
JrPhy

# Linear regression

- There are $n$ data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, $n \geq 2$ and $n \in \mathbf{Z}^+$.
- $\exists!\ y = f(x) = b + ax$ such that the square error $e$ is the least

$$E = \sum_{i=1}^{n} (y - y_i)^2 = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

- It's a parabolic eq. with concave up, so there exists a minimum value.
- Here we want to find $\boldsymbol{a}_{min}$ and $\boldsymbol{b}_{min}$ such that $E$ is minimum



# An O(n) method for linear regression

$$\begin{cases} \dfrac{\partial E}{\partial a} = 0 \rightarrow 2\sum_{i=1}^{n} x_i(ax_i + b - y_i) = 0 \\ \dfrac{\partial E}{\partial b} = 0 \rightarrow 2\sum_{i=1}^{n} (ax_i + b - y_i) = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} \left(ax_i^{\ 2} + bx_i\right) = \sum_{i=0}^{n} x_i y_i \\ \sum_{i=1}^{n} \left(ax_i + b\right) = \sum_{i=0}^{n} y_i \end{cases}$$

$$\sum_{i=1}^{n} \left(ax_i + b\right) = \sum_{i=0}^{n} y_i \rightarrow a\sum_{i=0}^{n} x_i + \sum_{i=0}^{n} b = \sum_{i=0}^{n} y_i \rightarrow a\mu_x + b = \mu_y \rightarrow b = \mu_y - a\mu_x$$

$$\sum_{i=1}^{n} \left(ax_i^{\ 2} + bx_i\right) = \sum_{i=0}^{n} x_i y_i \rightarrow a\sum_{i=0}^{n} x_i^{\ 2} + b\sum_{i=0}^{n} x_i = \sum_{i=0}^{n} x_i y_i \rightarrow a\sum_{i=0}^{n} x_i^{\ 2} + n\mu_x\mu_y - an\mu_x^{\ 2} = \sum_{i=0}^{n} x_i y_i$$

$$\rightarrow a\sum_{i=0}^{n} x_i^{\ 2} - an\mu_x^{\ 2} = \sum_{i=0}^{n} x_i y_i - n\mu_x\mu_y$$

$$\rightarrow a\left(\sum_{i=0}^{n} x_i^{\ 2} - n\mu_x^{\ 2}\right) = \sum_{i=0}^{n} x_i y_i - n\mu_x\mu_y$$

# An O(n) method for linear regression

$$\left(\sum_{i=0}^{n} x_i^{\,2} - n\mu_x^{\,2}\right) = \sum_{i=0}^{n} (x_i - \mu_x)^2$$

$$\sum_{i=0}^{n} (x_i - \mu_x)(y_i - \mu_y) = \sum_{i=0}^{n} \left(x_i y_i - x_i\mu_y - \mu_x y_i + \mu_x\mu_y\right)$$

$$= \sum_{i=0}^{n} x_i y_i - \mu_y \sum_{i=0}^{n} x_i - \mu_x \sum_{i=0}^{n} y_i + n\mu_x\mu_y$$

$$= \sum_{i=0}^{n} x_i y_i - n\mu_y\mu_x - n\mu_x\mu_y + n\mu_x\mu_y$$

$$= \sum_{i=0}^{n} x_i y_i - n\mu_y\mu_x$$

$$\boxed{a = \frac{\sum_{i=0}^{n} x_i y_i - n\mu_x\mu_y}{\sum_{i=0}^{n} x_i^{\,2} - n\mu_x^{\,2}}}$$

$$= \frac{\sum_{i=0}^{n} (x_i - \mu_x)(y_i - \mu_y)}{\sum_{i=0}^{n} (x_i - \mu_x)^2}$$

# An O(n) method for linear regression

$y = ax + b$

$\quad = ax + (\mu_y - a\mu_x) \quad \Rightarrow \quad y - \mu_y = a(x - \mu_x)$

$\quad = \mu_y + a(x - \mu_x)$

- The regression line pass through the average of *x* and *y*

# An O(n) method for linear regression

```
for(i=0; i<n; i++)
{
    xavg += x[i]
    yavg += y[i]
    xiyi += x[i]*y[i]
    xixi += x[i]*x[i]
}
a = (xiyi – n* xavg* yavg)/(xixi – n* xavg* xavg)
y = a*(x – xavg) – yavg
```