

---

---

# Linear Regression

— Boston University CS 506 - Lance Galletti —

---

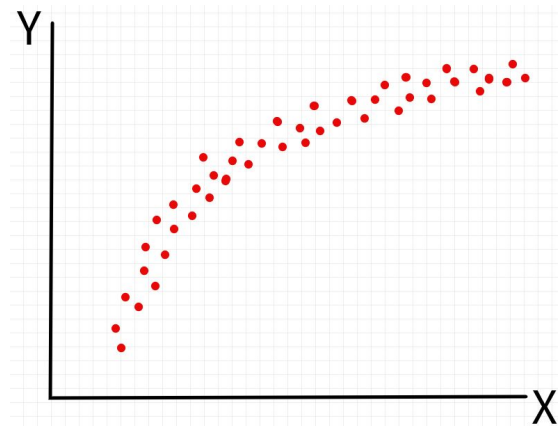
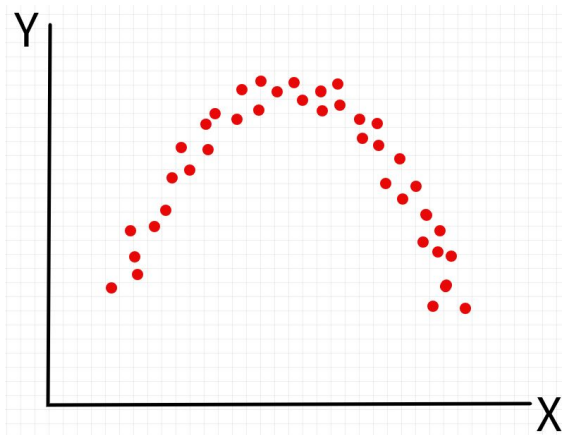
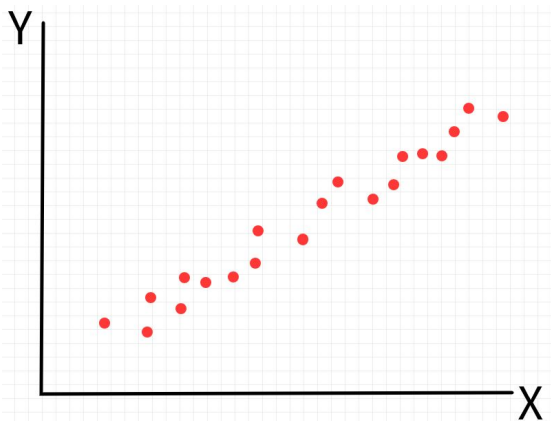
---

# Challenge for those who have LR experience

- Find the data.csv file in the regression folder of our course repo
- Challenge:
  - Every day my alarm goes off at seemingly random times...
  - I've recorded the times for the past year of so (1 - 355 days)
  - Today is day 356
  - Can you predict when my alarm will ring?

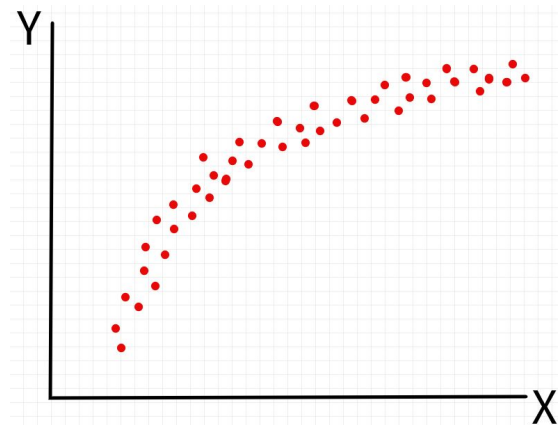
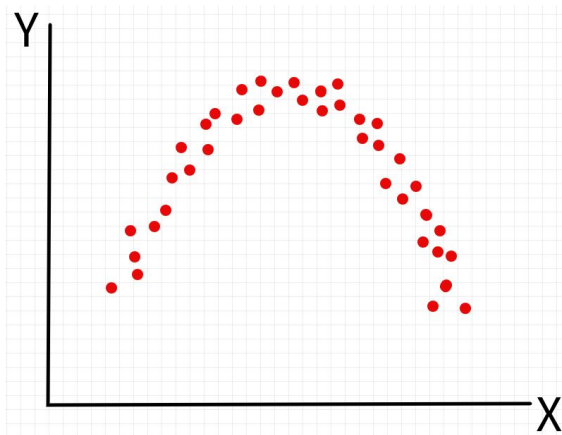
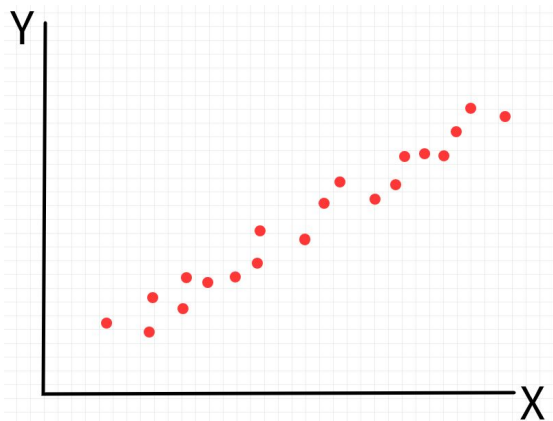
# Motivation

Given  $n$  samples / data points  $(\mathbf{y}_i, \mathbf{x}_i)$ .  $Y$  is a continuous variable (as opposed to classification).



# Motivation

Understand/explain how **y** varies as a function of **x** (i.e. find a function  **$y = h(x)$**  that best fits our data)



# Cost Function

Given our data:  $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

Suppose we are given a curve  $\mathbf{y} = \mathbf{h}(\mathbf{x})$ , how can we evaluate whether it is a good fit to our data?

Compare  $\mathbf{h}(\mathbf{x}_i)$  to  $y_i$  for all  $i$ .

# Cost Function

Given our data:  $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$

Suppose we are given a curve  $\mathbf{y} = \mathbf{h}(\mathbf{x})$ , how can we evaluate whether it is a good fit to our data?

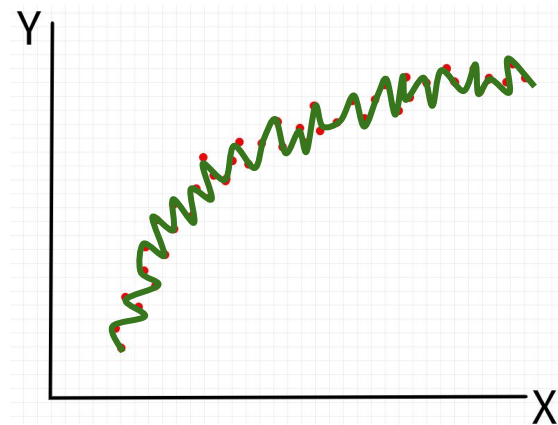
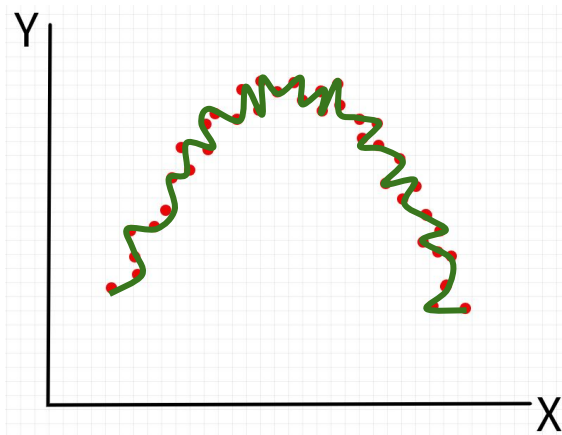
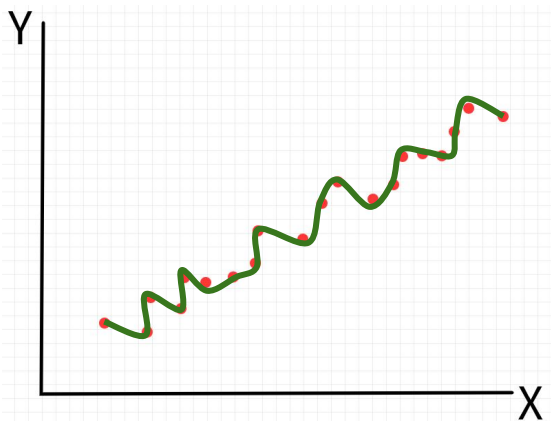
Compare  $\mathbf{h}(\mathbf{x}_i)$  to  $\mathbf{y}_i$  for all  $i$ .

Goal: For a given distance function  $\mathbf{d}$ , find  $\mathbf{h}$  where  $\mathbf{L}$  is smallest.

$$L(h) = \sum_i d(h(x_i), y_i)$$

# Motivation

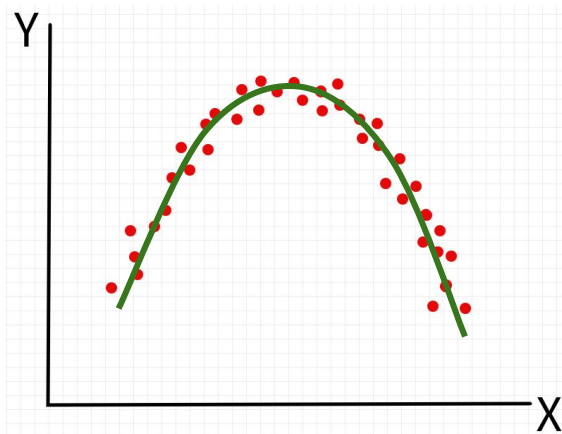
Should  $\mathbf{h}$  be the curve that goes through the most samples? I.e. do we want  $\mathbf{h}(\mathbf{x}_i) = \mathbf{y}_i$  for the maximum number of  $i$ ?



$\mathbf{h}$  may be too complex  
overfitting - may not perform well on unseen data

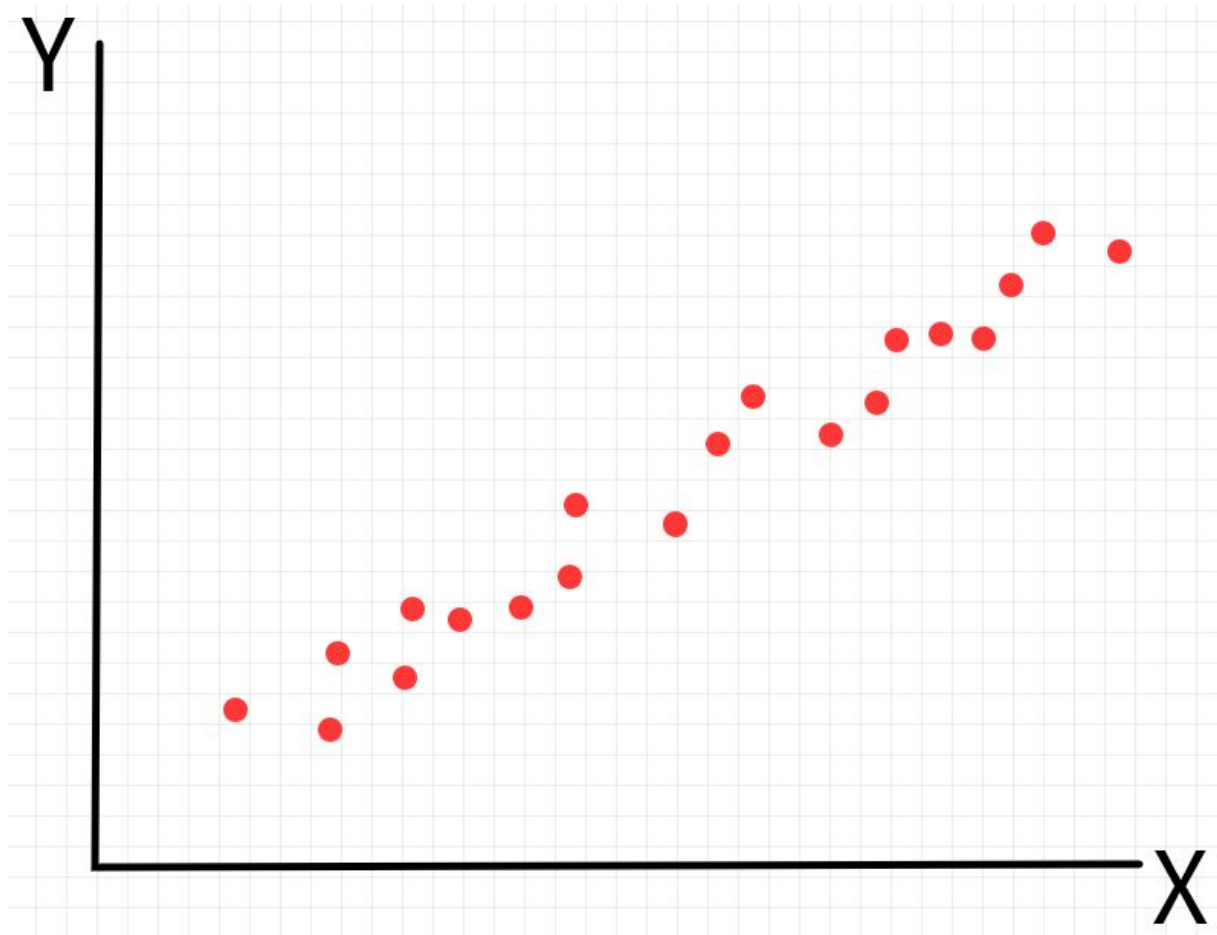
# Motivation

The following curves seem the most intuitive “best fit” to our samples. How can we define this best fit mathematically? Is it just about finding the right distance function?





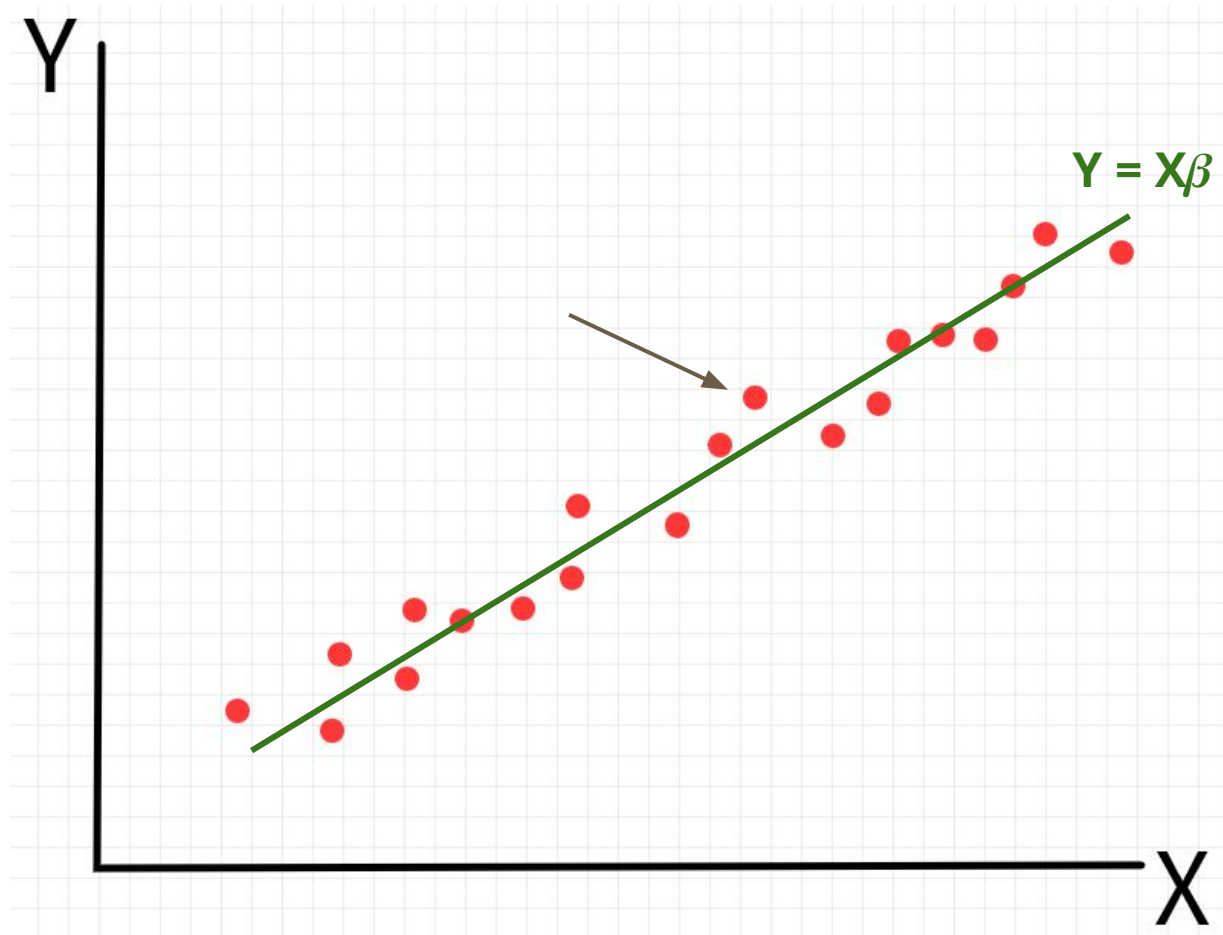
# Assumptions



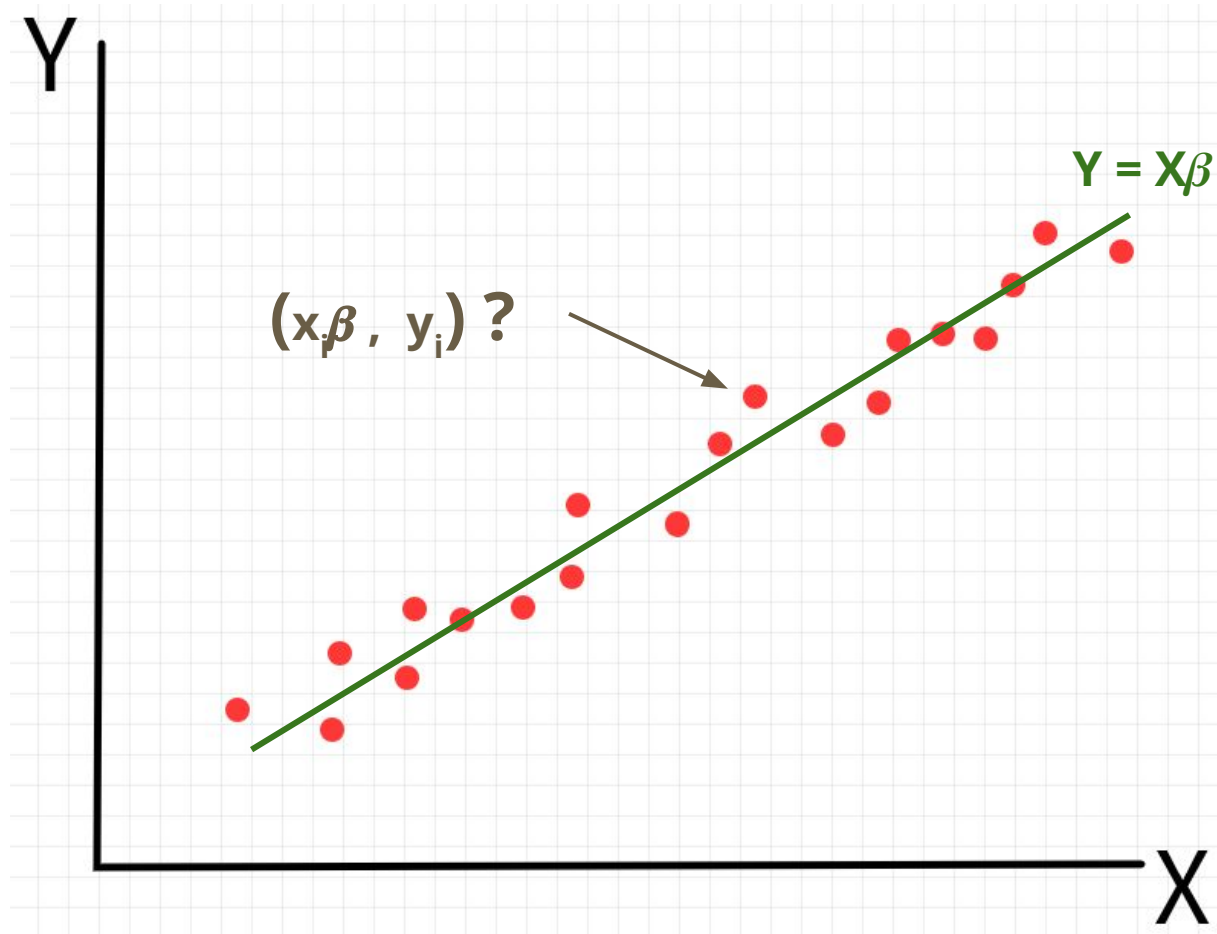
# Assumptions



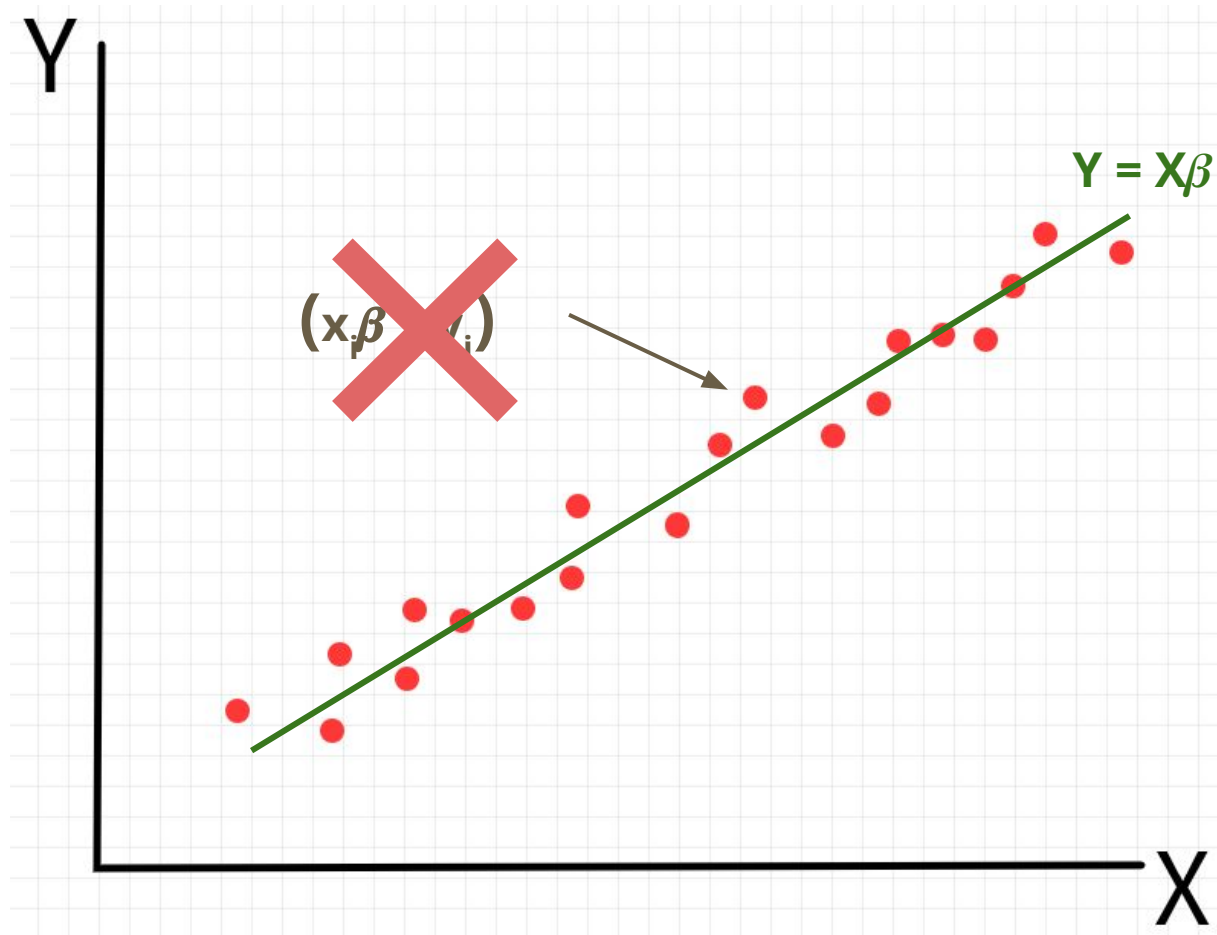
# Assumptions



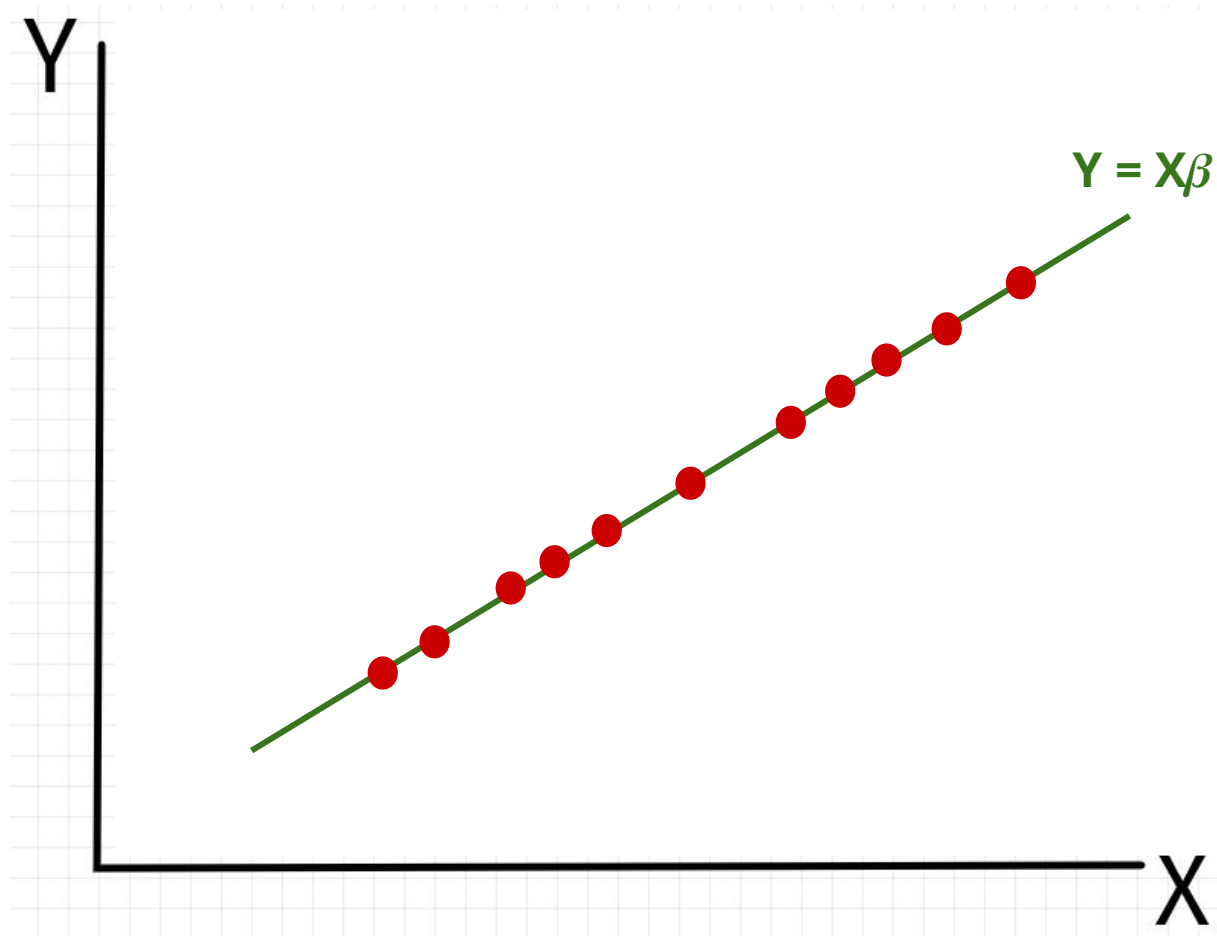
# Assumptions



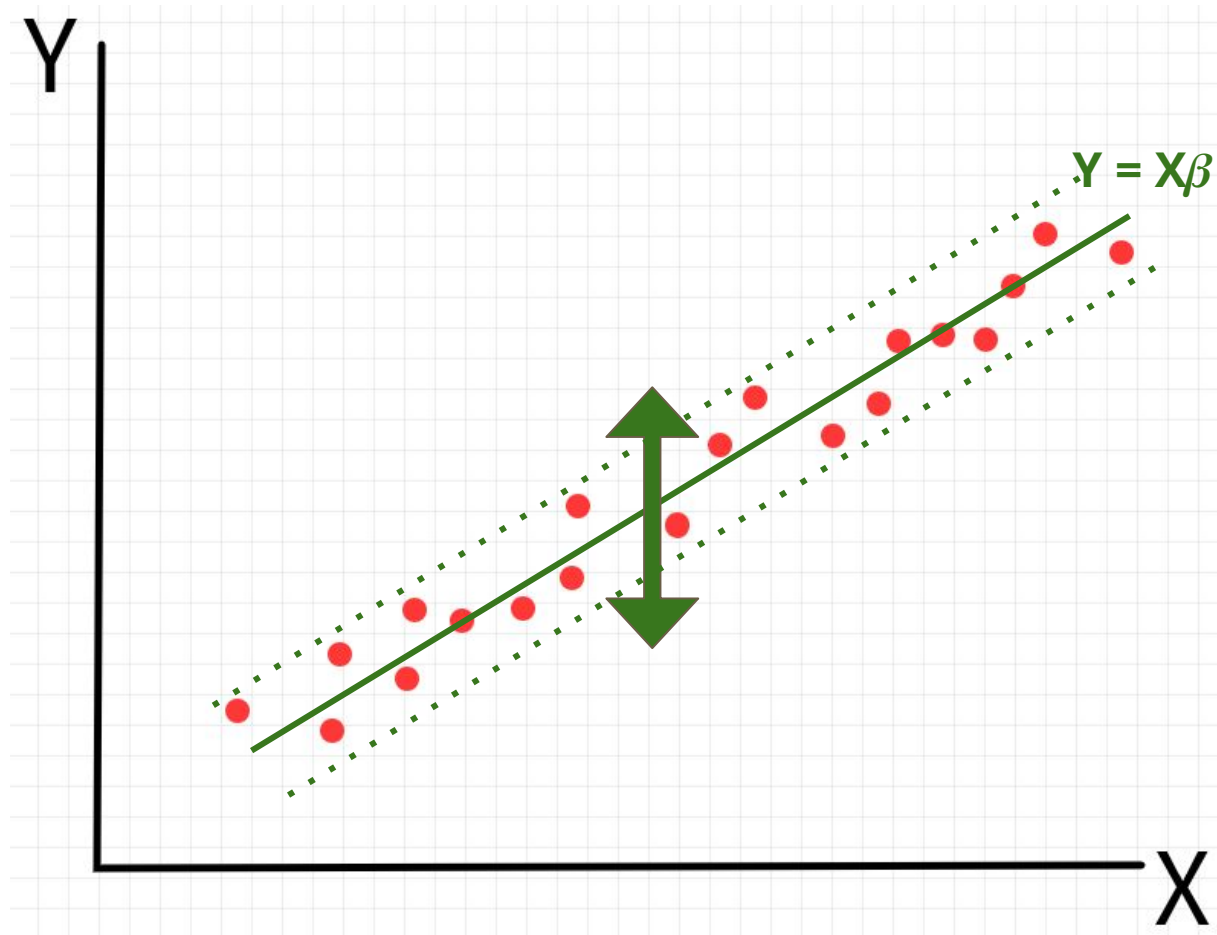
# Assumptions



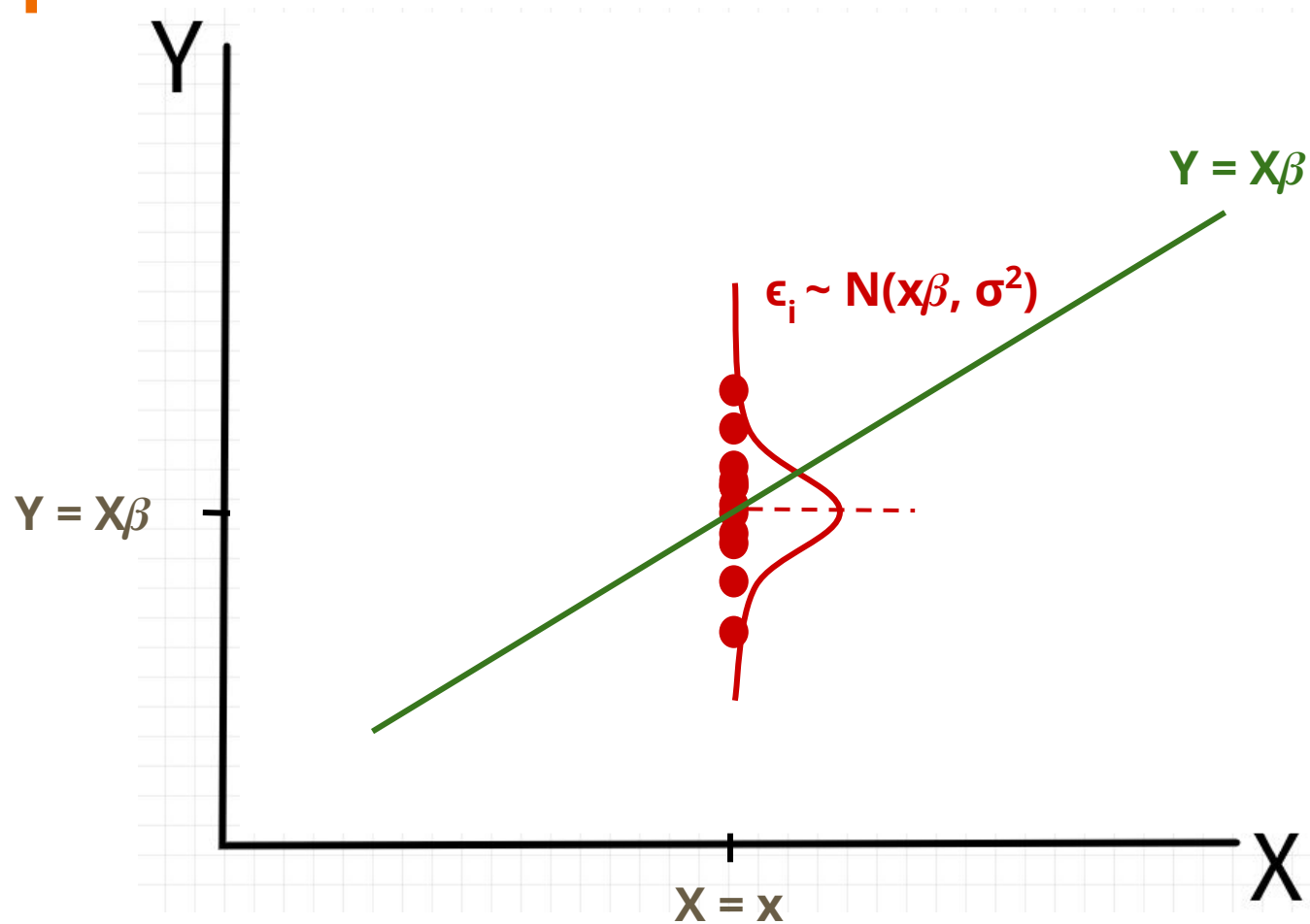
# Assumptions



# Assumptions

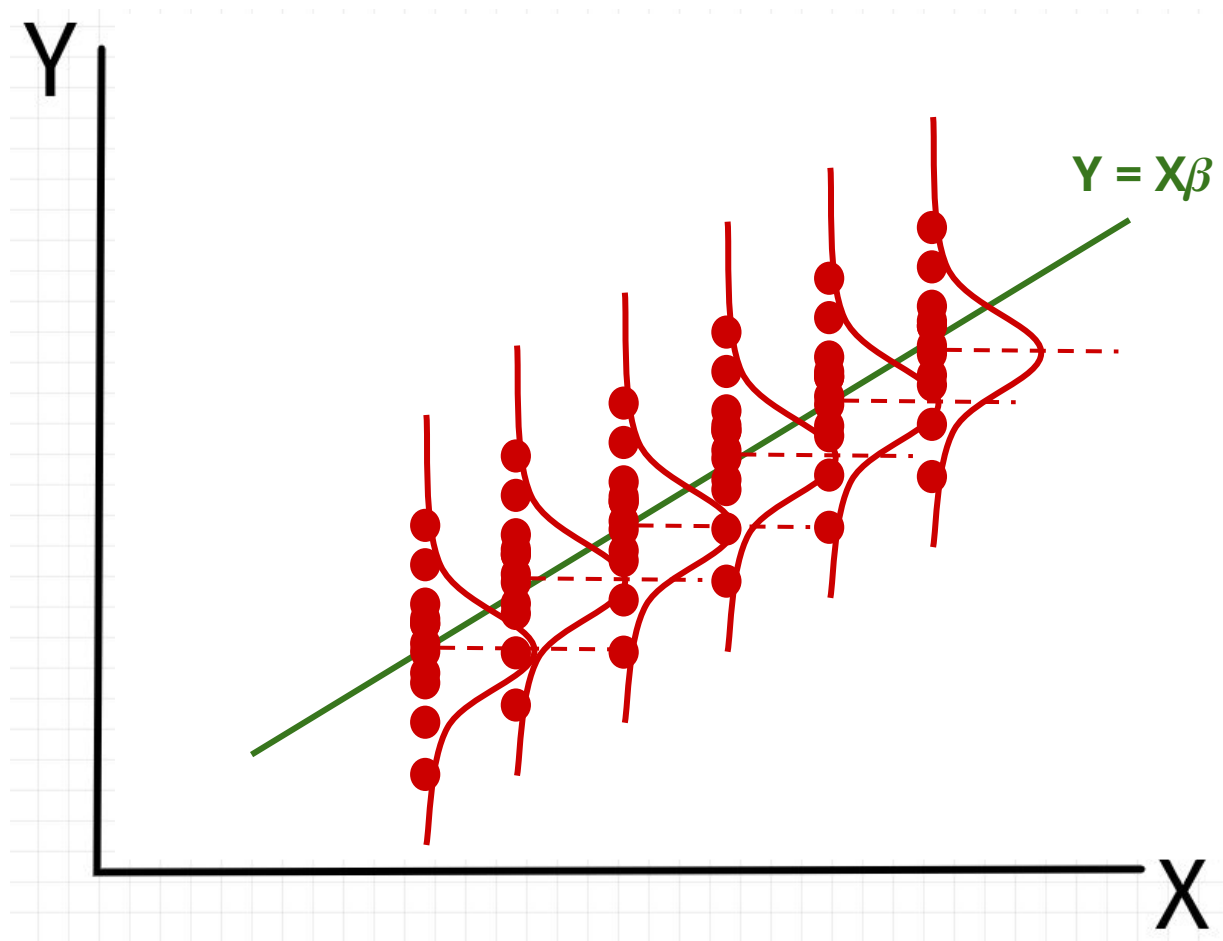


# Assumptions





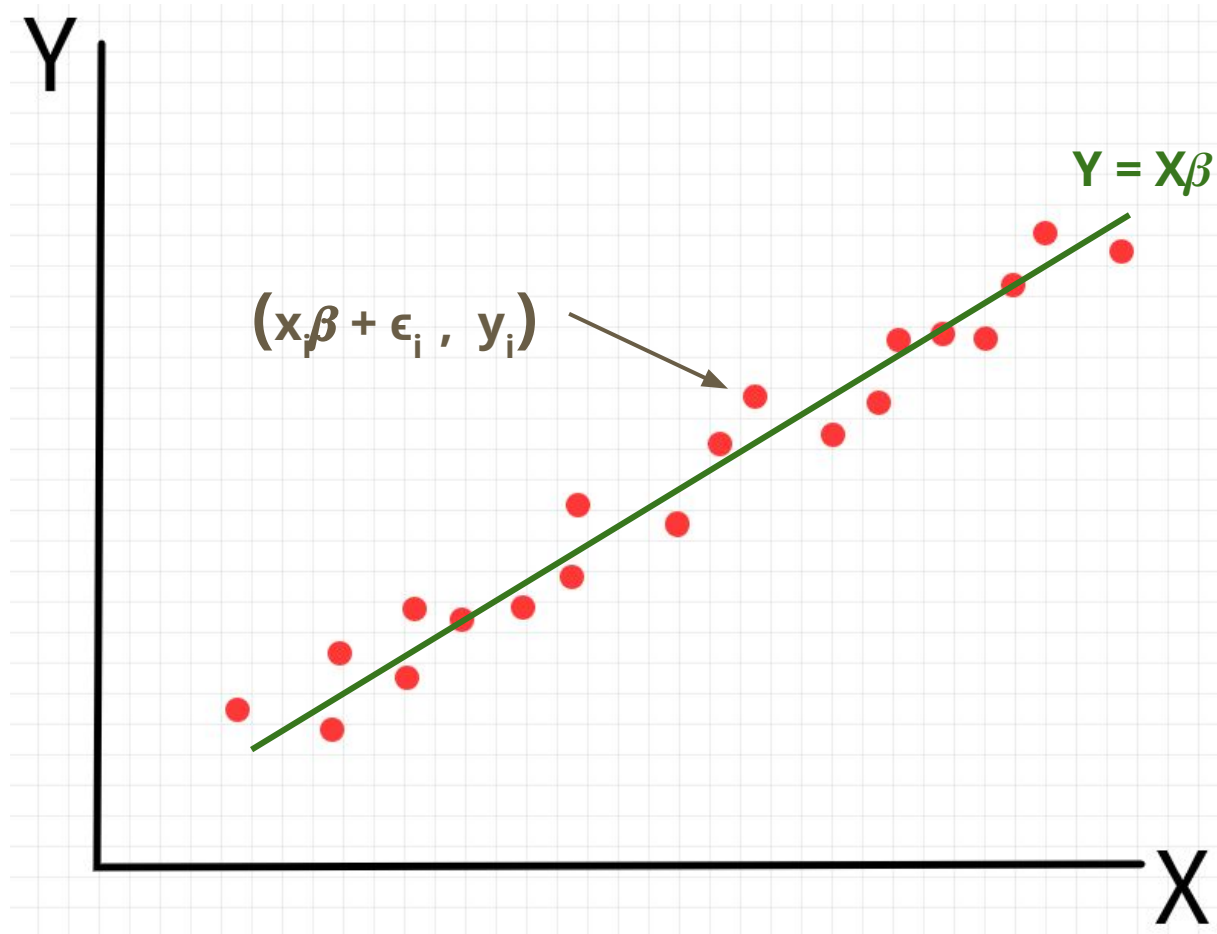
# Assumptions



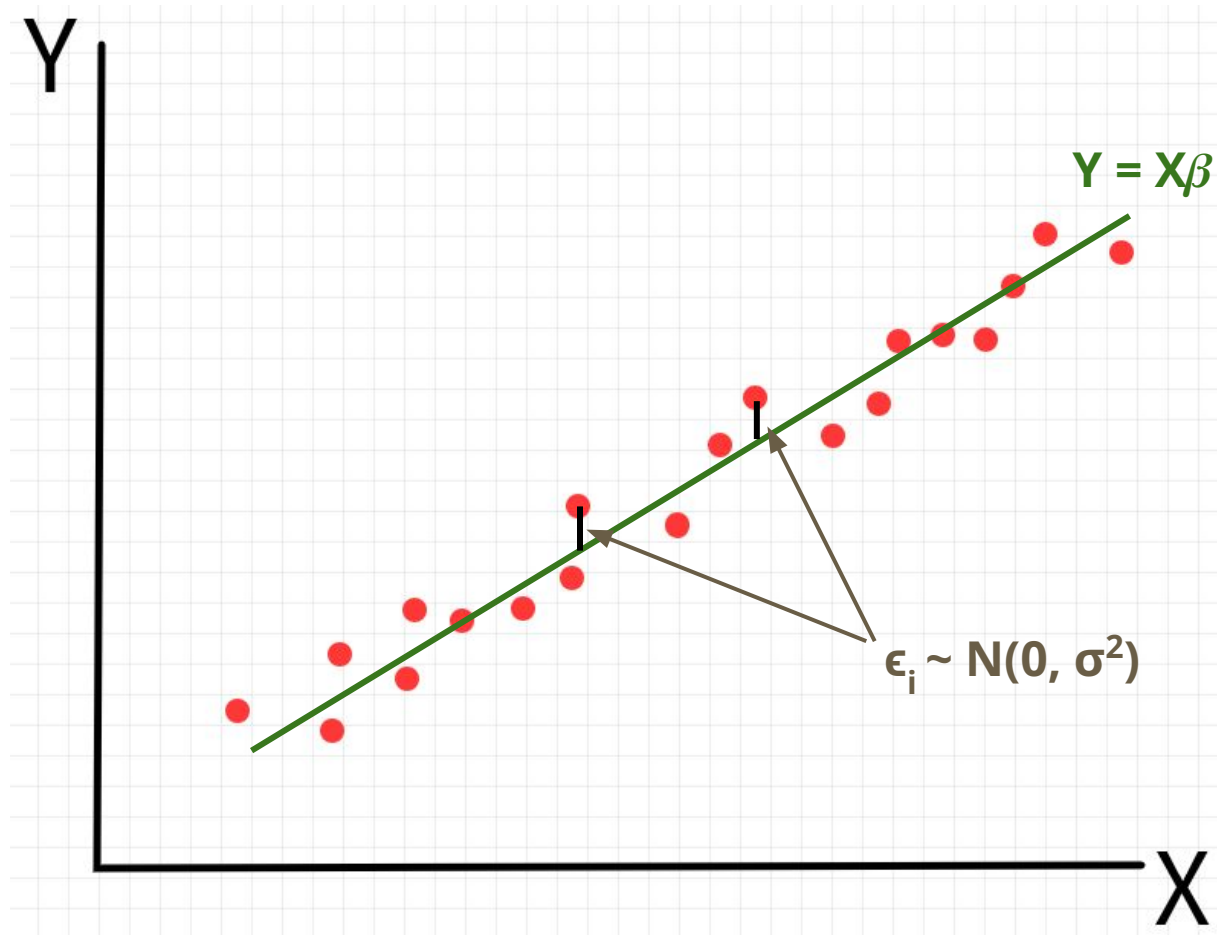
# Assumptions



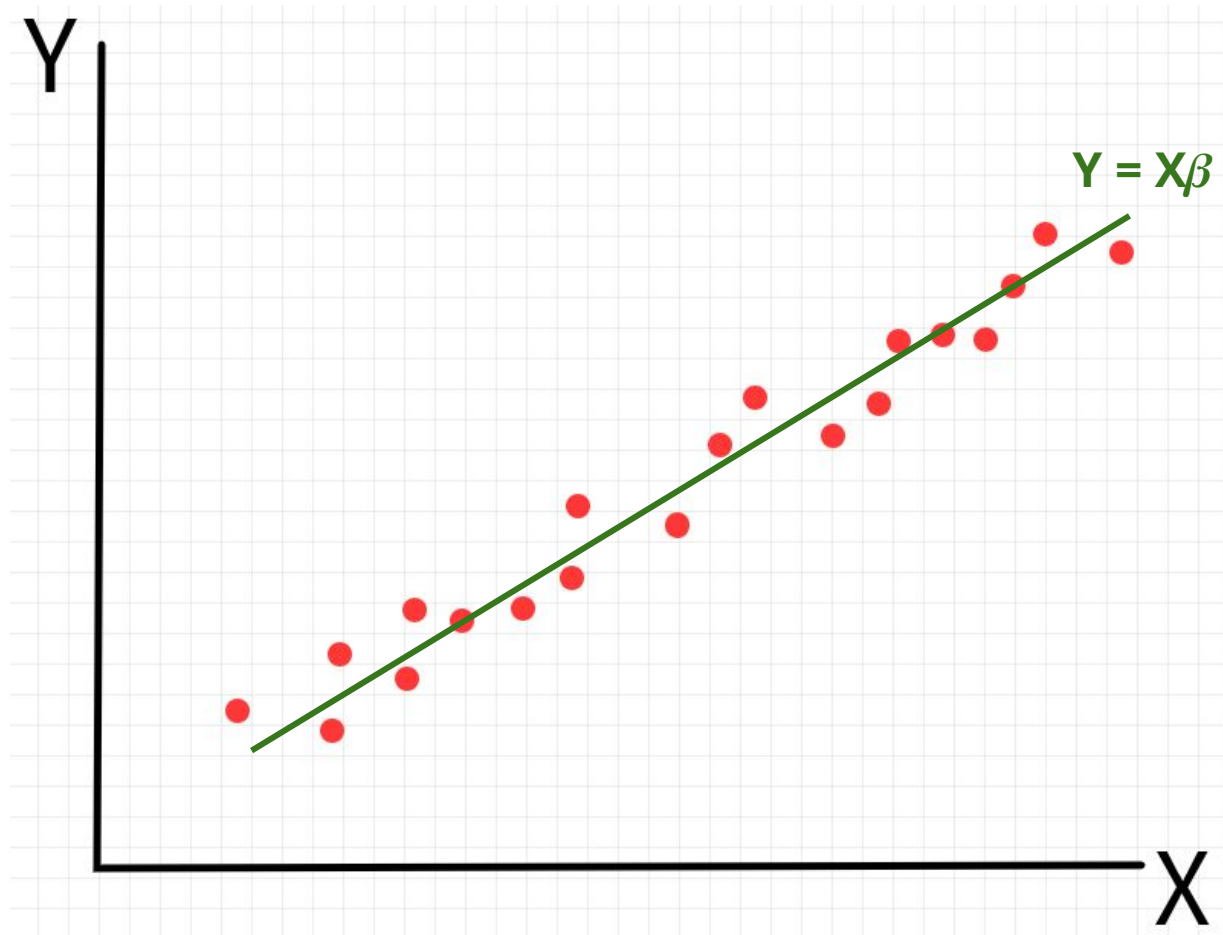
# Assumptions



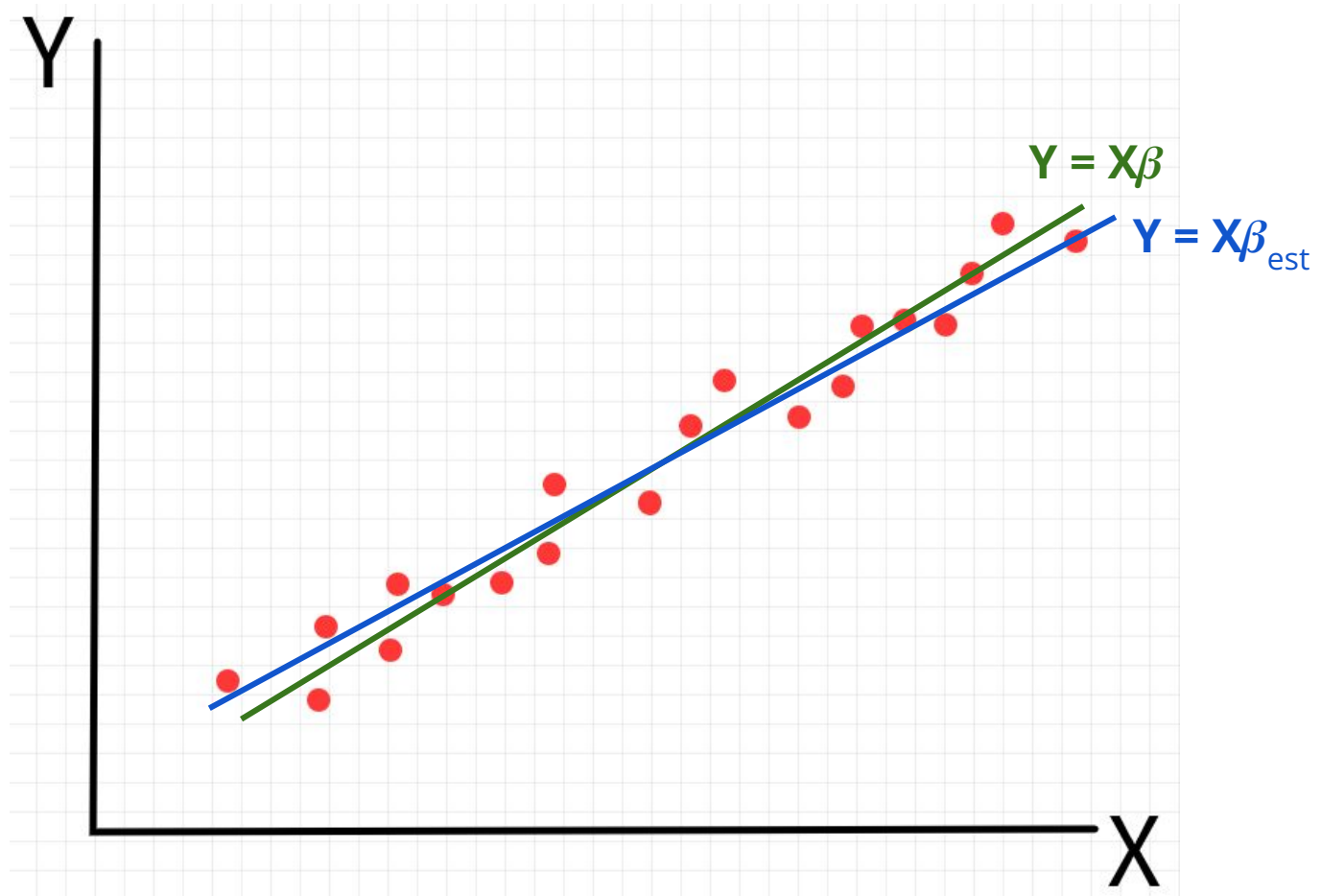
# Assumptions



# Assumptions



# Assumptions



# Assumptions

Our data was generated by a **linear function** plus some **noise**:

$$\vec{y} = h_X(\beta) + \vec{\epsilon}$$

Where **h** is linear in a parameter  **$\beta$** .

Where  **$\epsilon_i$**  are independent  **$\mathbf{N}(\mathbf{0}, \sigma^2)$**  distribution.

# Worksheet a)



# Assumptions

1. The relation between  $\mathbf{x}$  (independent variable) and  $\mathbf{y}$  (dependent variable) is linear in a parameter  $\boldsymbol{\beta}$ .
2.  $\boldsymbol{\epsilon}_i$  are independent, identically distributed random variables following a  $\mathbf{N}(\mathbf{0}, \sigma^2)$  distribution. (Note:  $\sigma$  is constant)

# Goal

Given these assumptions, let's try to minimize the cost function defined earlier!

Q: What parameter(s) are we trying to learn / estimate?

A:  $\beta$

# Least Squares

$$\beta_{LS} = \arg \min_{\beta} \sum_i d(h_{\beta}(x_i), y_i)$$



$(x_i, y_i)$  are from our dataset

# Least Squares

$$\begin{aligned}\beta_{LS} &= \arg \min_{\beta} \sum_i d(h_{\beta}(x_i), y_i) \\ &= \arg \min_{\beta} \|\vec{y} - h_{\beta}(X)\|_2^2 \\ &= \arg \min_{\beta} \|y - X\beta\|_2^2\end{aligned}$$

# Least Squares

$$\frac{\partial}{\partial \beta} (y - X\beta)^T (y - X\beta) = 0$$

$$\frac{\partial}{\partial \beta} (y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta) = 0$$

$$\frac{\partial}{\partial \beta} (y^T y - 2\beta^T X^T y - \beta^T X^T X\beta) = 0$$

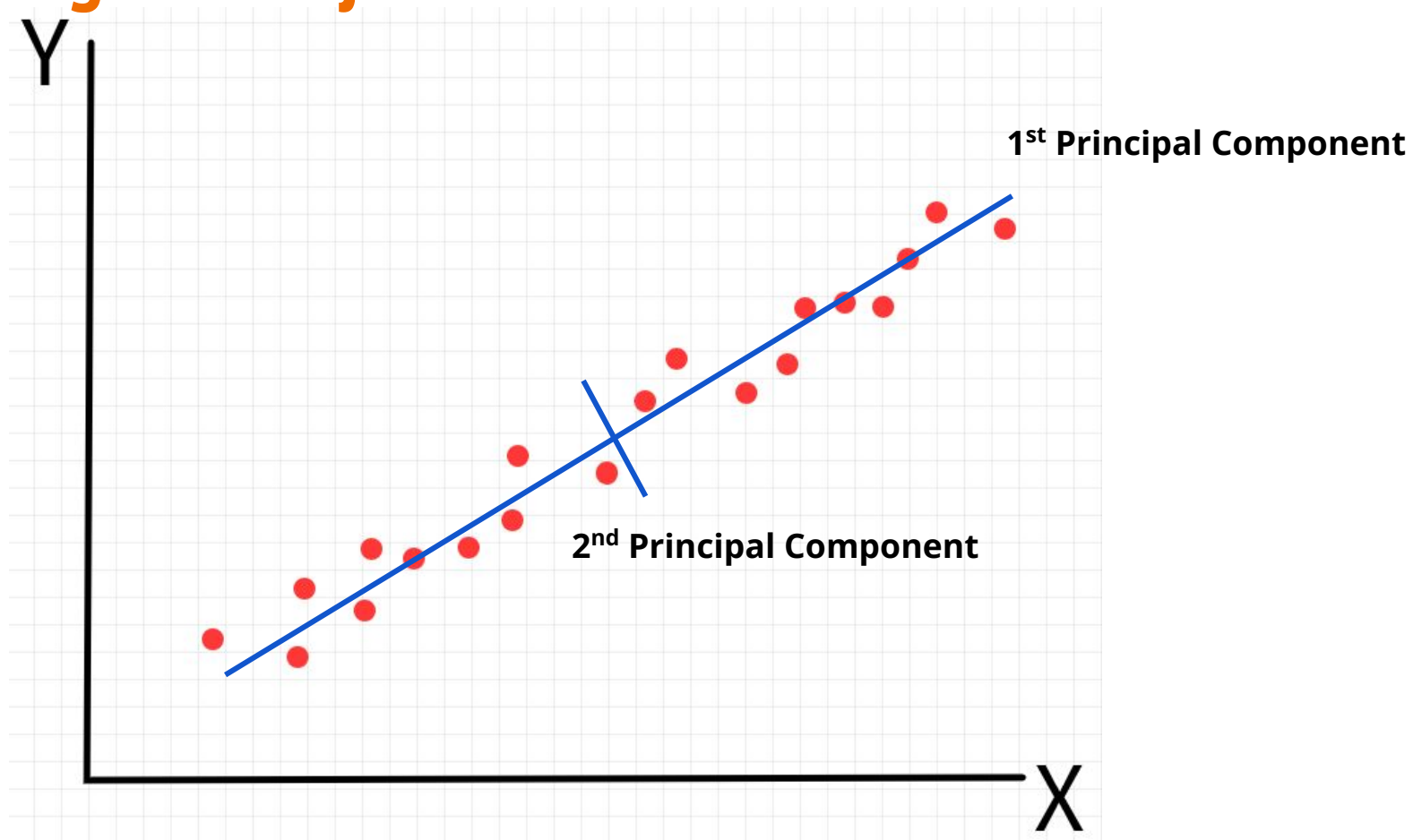
$$-2X^T y - X^T X\beta = 0$$

$$X^T X\beta = X^T y$$

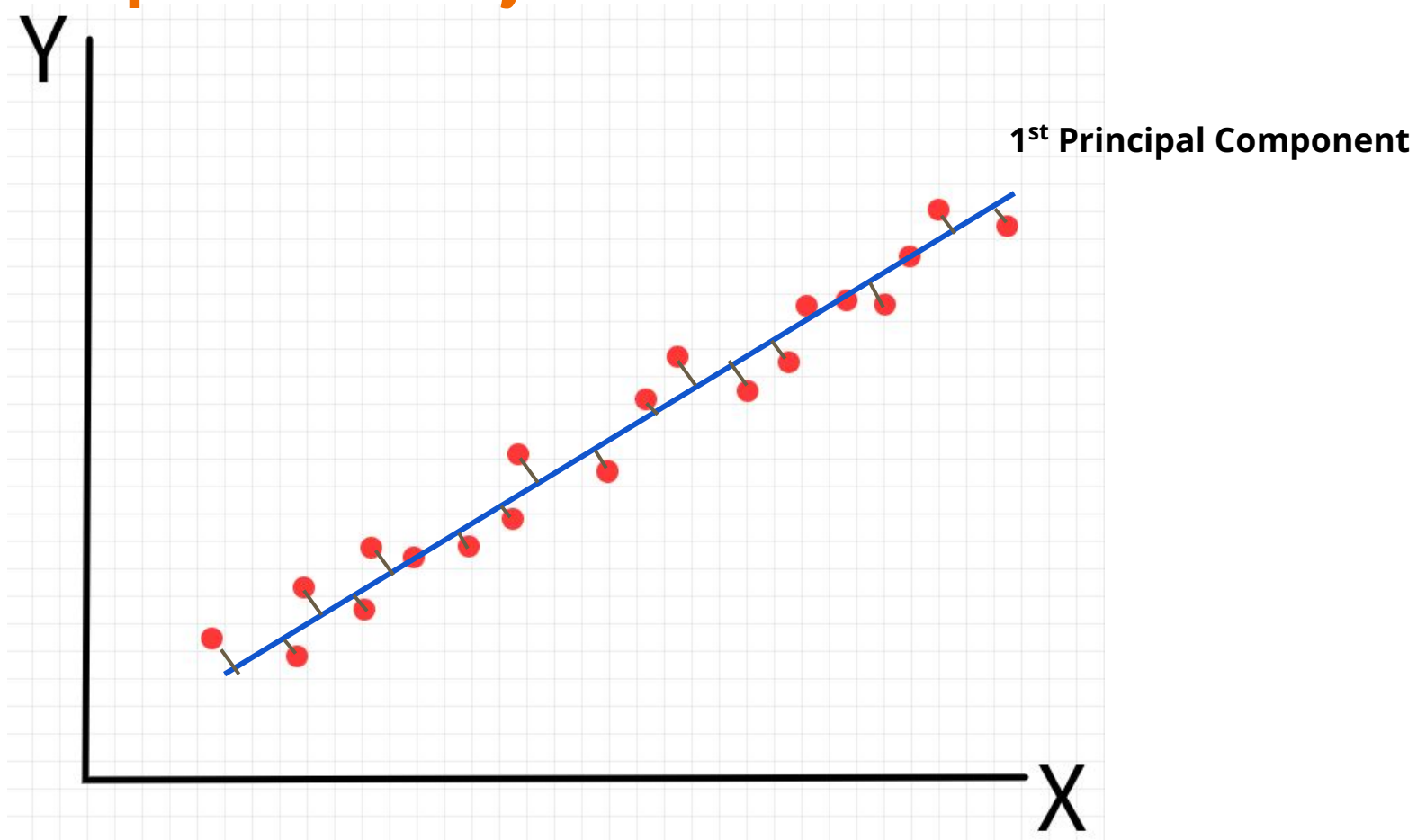
$$\beta_{LS} = (X^T X)^{-1} X^T y$$

# Worksheet b) & c)

# Is Linear Regression just PCA?



# Principal Component Analysis





# Linear Regression



# Assumptions

Our data was generated by a **linear function** plus some **noise**:

$$\vec{y} = h_X(\beta) + \vec{\epsilon}$$

Where **h** is linear in a parameter  **$\beta$** .

Which functions below are linear in  **$\beta$** ?

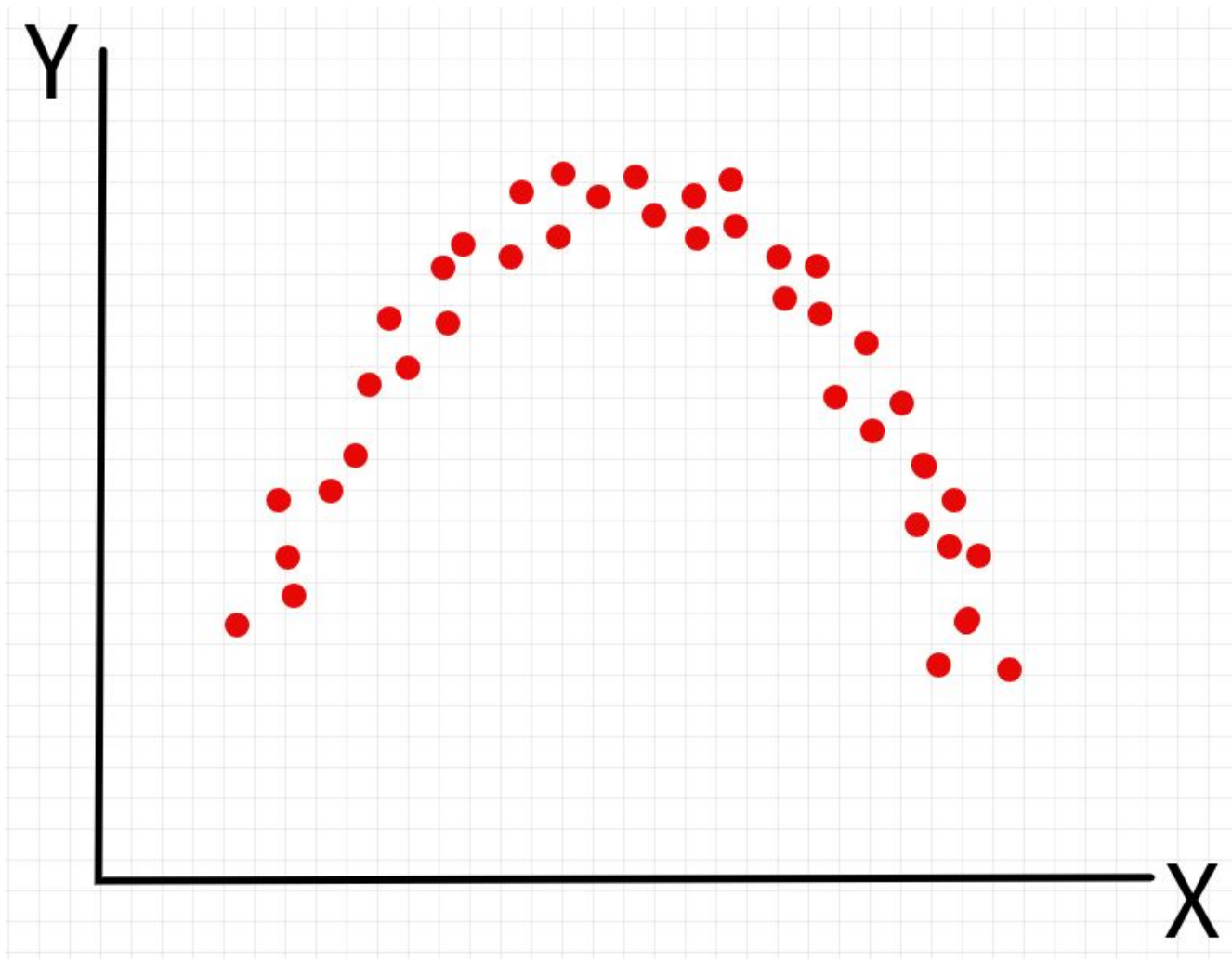
$$h(\beta) = \beta_1 x \quad \checkmark$$

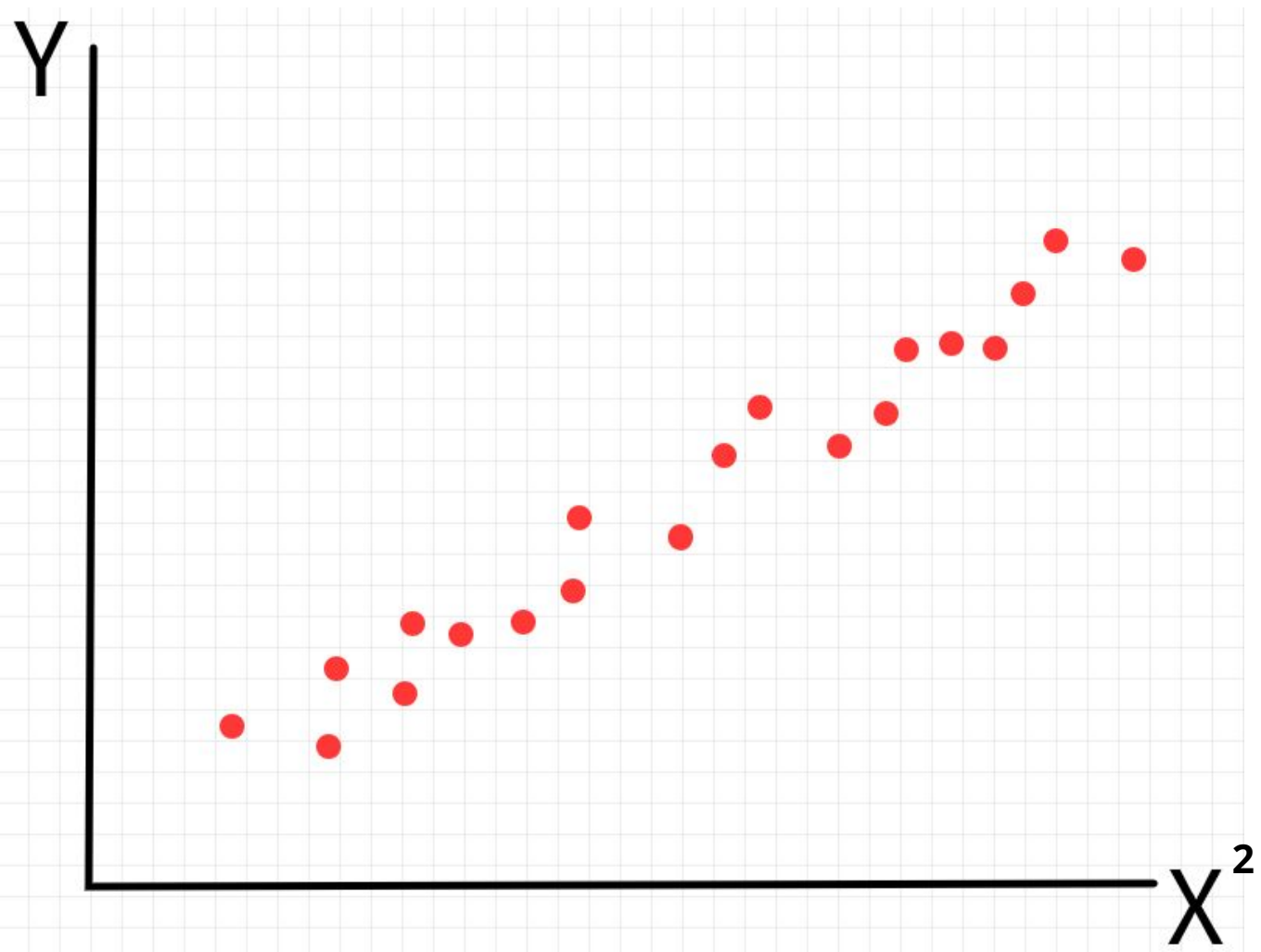
$$h(\beta) = \beta_0 + \beta_1 x \quad \checkmark$$

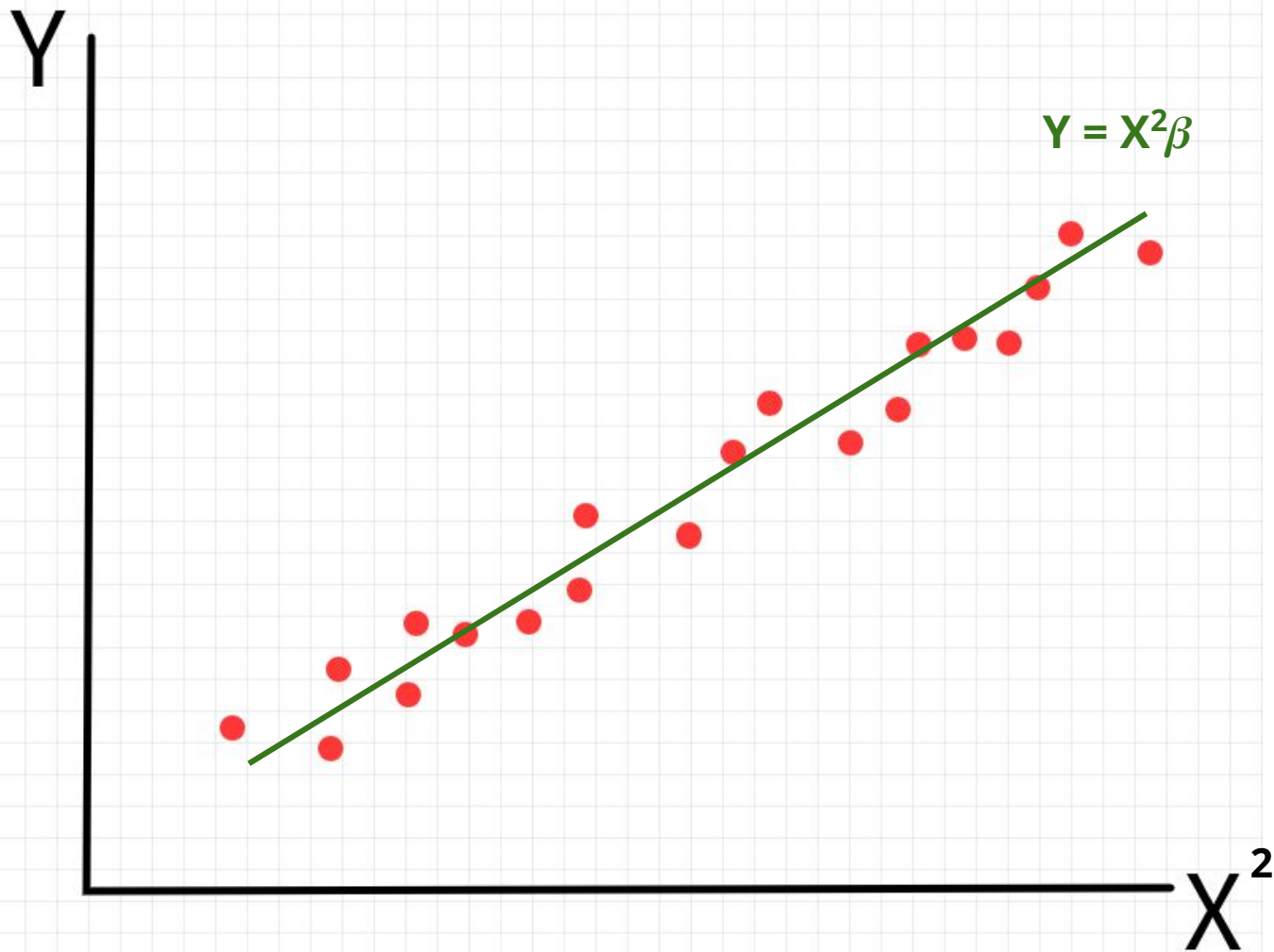
$$h(\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \checkmark$$

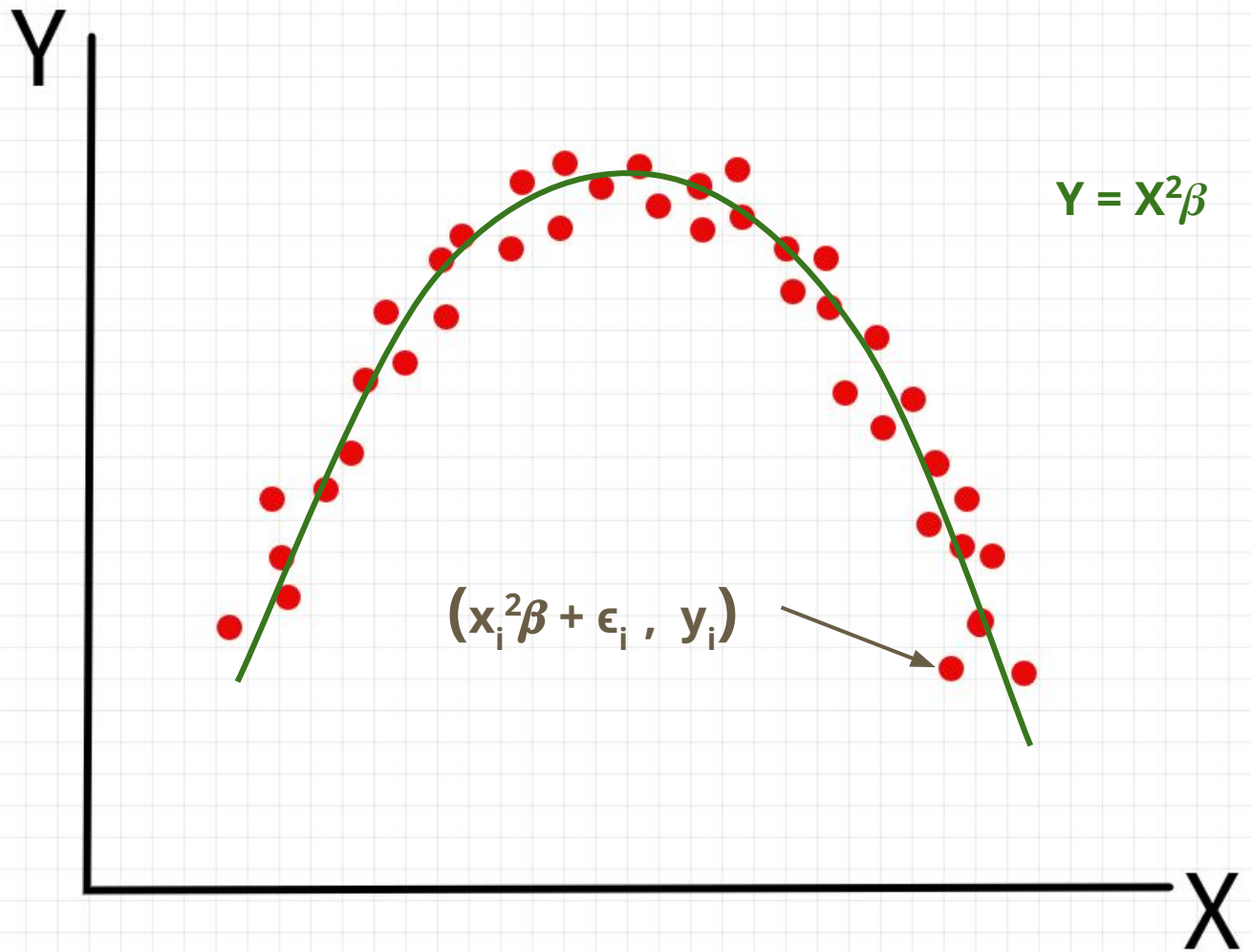
$$h(\beta) = \beta_1 \log(x) + \beta_2 x^2 \quad \checkmark$$

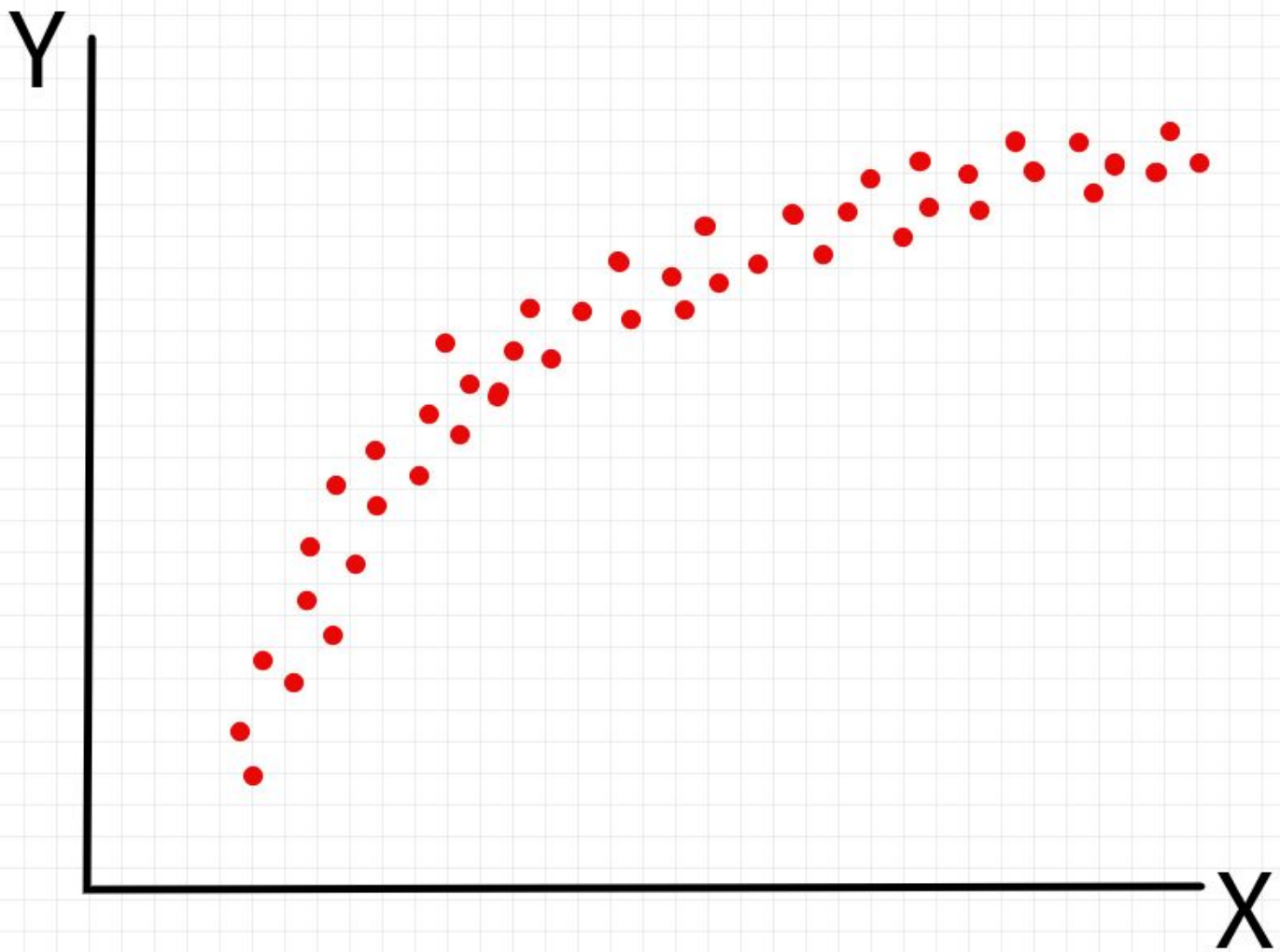
$$h(\beta) = \beta_0 + \beta_1 x + \beta_1^2 x \quad \times$$

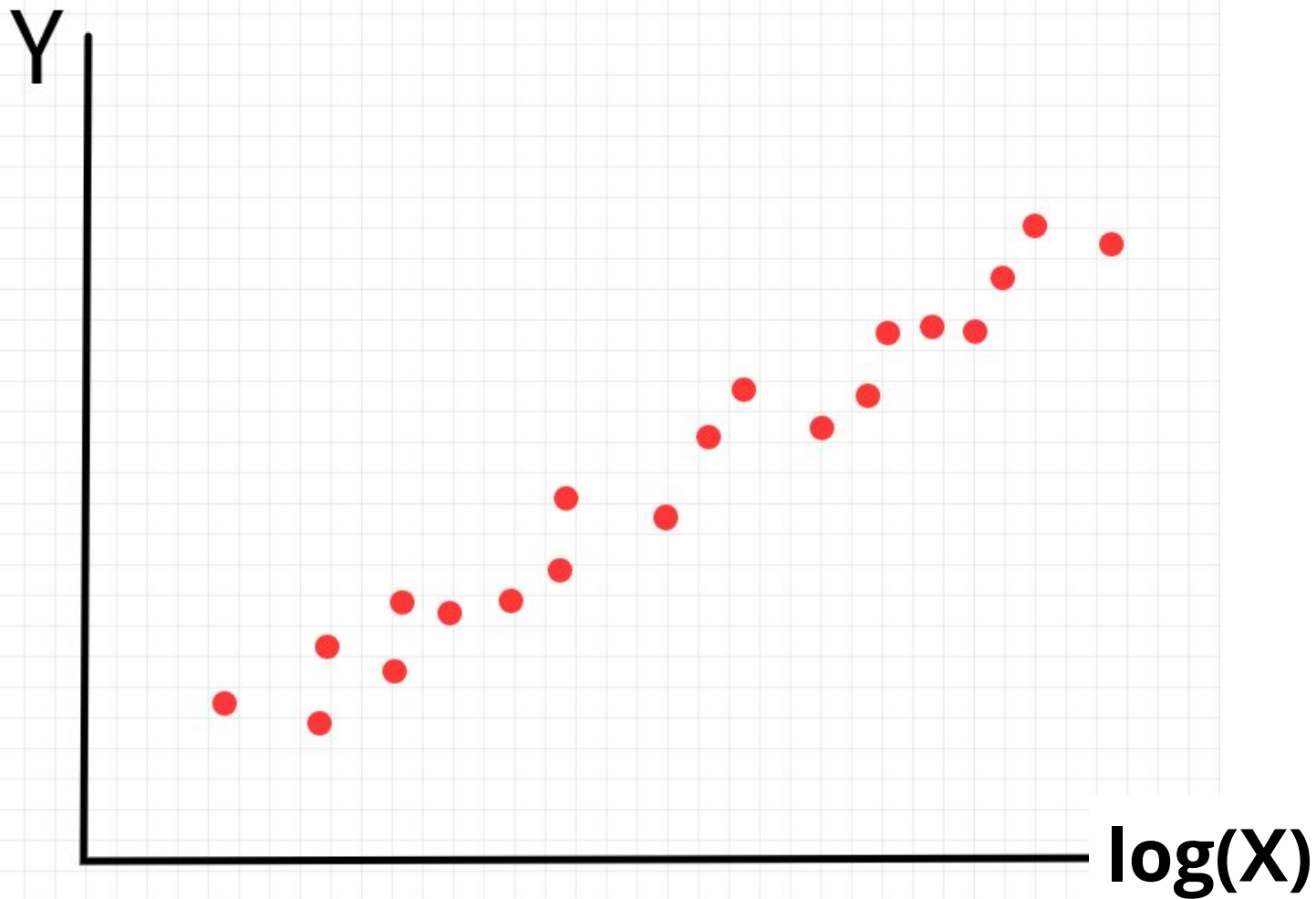






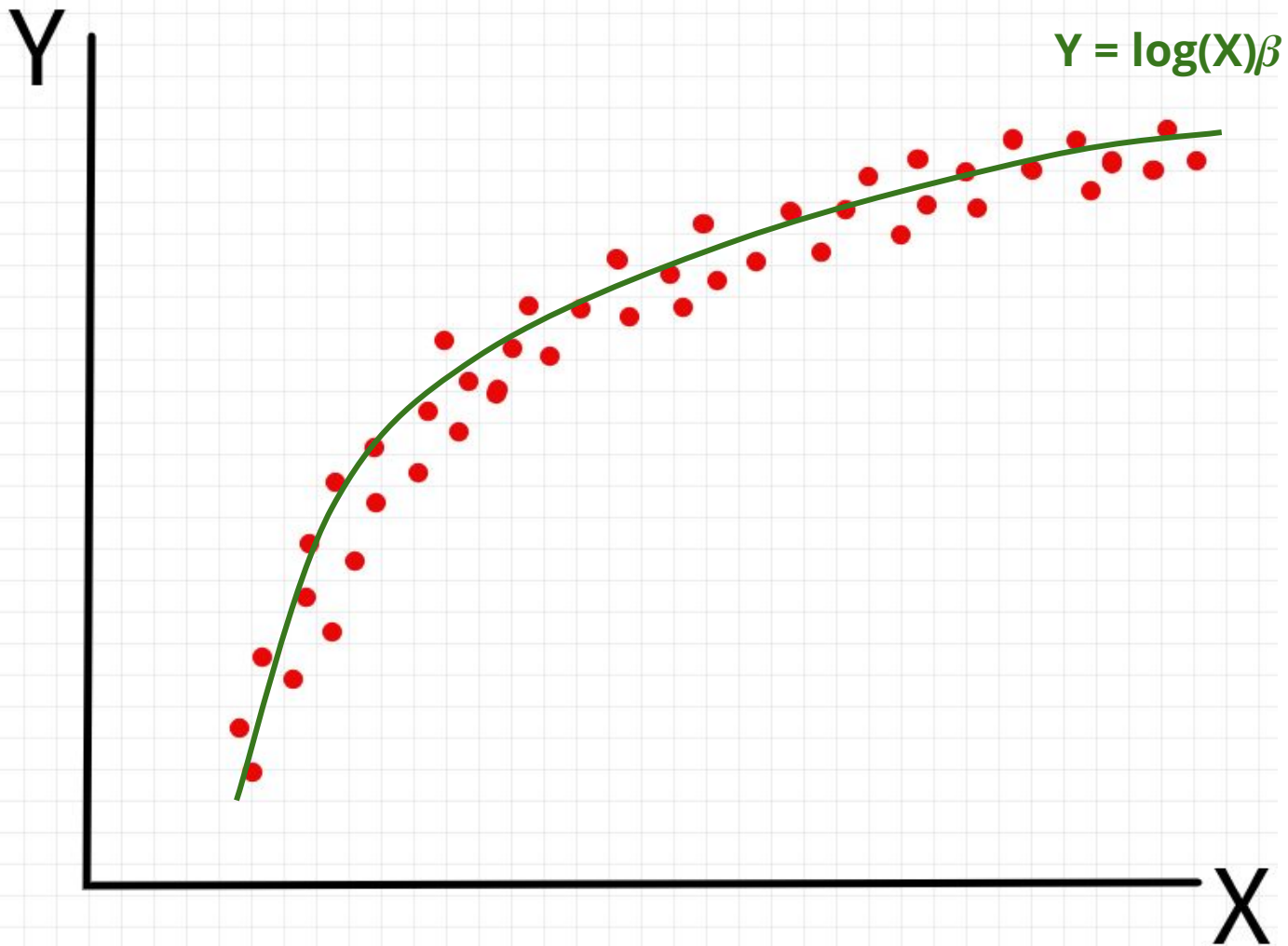












**Worksheet d) -> f)**

# Maximum Likelihood

Another way to define this problem is in terms of probability.

Define  $\mathbf{P}(\mathbf{Y} \mid \mathbf{h})$  as the probability of observing  $\mathbf{Y}$  given that it was sampled from  $\mathbf{h}$ .

Goal: Find  $\mathbf{h}$  that maximizes the probability of having observed our data.

# Maximum Likelihood

Maximize  $L(\mathbf{h}) = P(\mathbf{Y} \mid \mathbf{h})$

Since  $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2)$  and  $\mathbf{Y} = \mathbf{X}\beta + \epsilon$  then  $\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2)$ .

# Maximum Likelihood

Since  $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2)$  and  $\mathbf{Y} = \mathbf{X}\beta + \epsilon$  then  $\mathbf{Y} \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2)$ .

$$\begin{aligned}\beta_{MLE} &= \arg \max_{\beta} \frac{1}{\sqrt{(2\pi)^n \sigma^n}} \exp\left(-\frac{\|y - X\beta\|_2^2}{2\sigma^2}\right) \\ &= \arg \max_{\beta} \exp(-\|y - X\beta\|_2^2) \\ &= \arg \max_{\beta} -\|y - X\beta\|_2^2 \\ &= \arg \min_{\beta} \|y - X\beta\|_2^2 \\ &= \beta_{LS} = (X^T X)^{-1} X^T y\end{aligned}$$

# An Unbiased Estimator

$\beta_{LS}$  is an unbiased estimator of the true  $\beta$ . That is  $E[\beta_{LS}] = \beta$ .

$$E[\beta_{LS}] = E[(X^T X)^{-1} X^T y]$$

# An Unbiased Estimator

$\beta_{LS}$  is an unbiased estimator of the true  $\beta$ . That is  $E[\beta_{LS}] = \beta$ .

$$\begin{aligned} E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\ &= (X^T X)^{-1} X^T E[y] \end{aligned}$$



# An Unbiased Estimator

$\beta_{LS}$  is an unbiased estimator of the true  $\beta$ . That is  $E[\beta_{LS}] = \beta$ .

$$\begin{aligned} E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\ &= (X^T X)^{-1} X^T E[y] \\ &= (X^T X)^{-1} X^T E[X\beta + \epsilon] \end{aligned}$$

# An Unbiased Estimator

$\beta_{LS}$  is an unbiased estimator of the true  $\beta$ . That is  $E[\beta_{LS}] = \beta$ .

$$\begin{aligned} E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\ &= (X^T X)^{-1} X^T E[y] \\ &= (X^T X)^{-1} X^T E[X\beta + \epsilon] \\ &= (X^T X)^{-1} X^T X\beta + E[\epsilon] \\ &= \beta \end{aligned}$$

**Worksheet g) ->**