Linear Regression

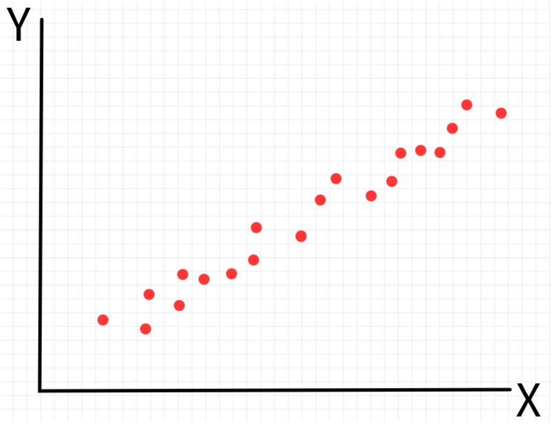
Boston University CS 506 - Lance Galletti

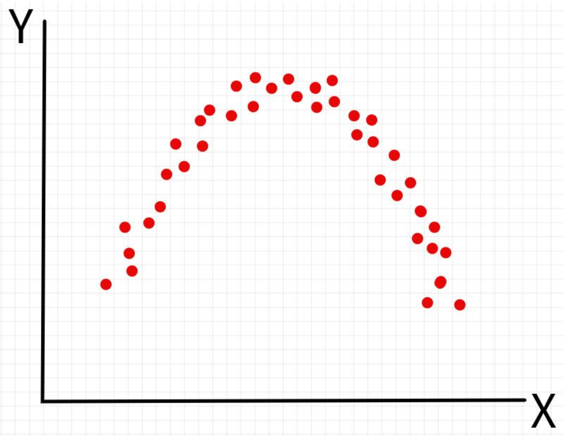
Challenge for those who have LR experience

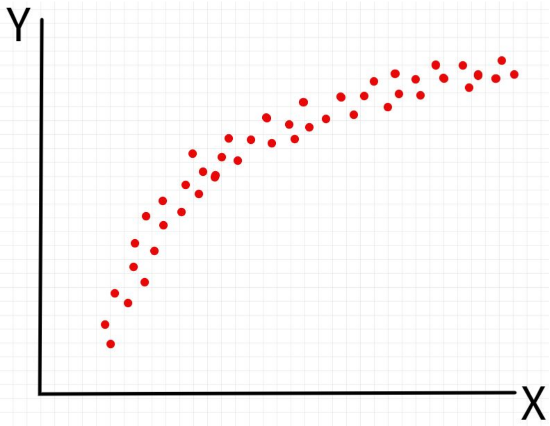
- Find the data.csv file in the regression folder of our course repo
- Challenge:
  - Every day my alarm goes off at seemingly random times…
  - I’ve recorded the times for the past year or so (1 - 355 days)
  - Today is day 356
  - Can you predict when my alarm will ring?

Motivation

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

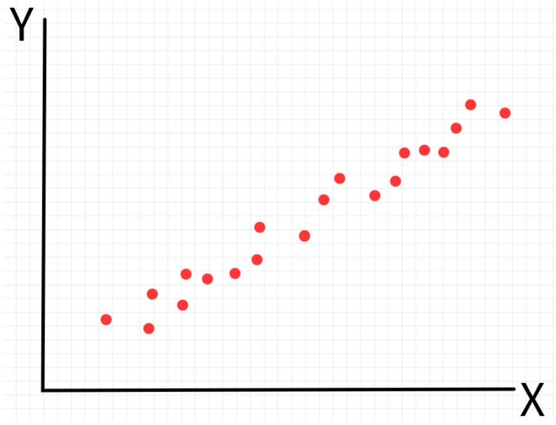


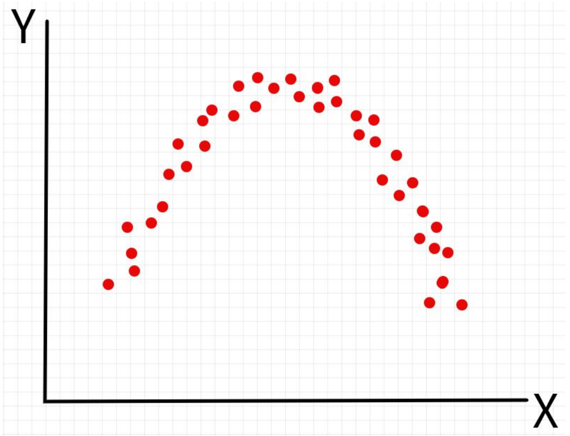


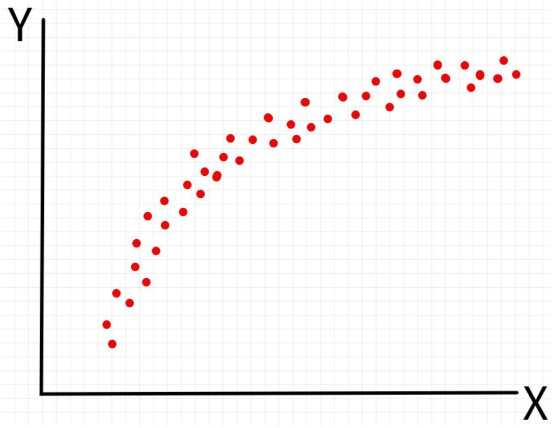


Motivation

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

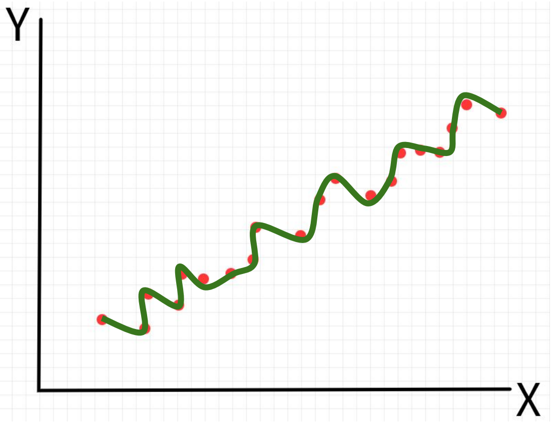


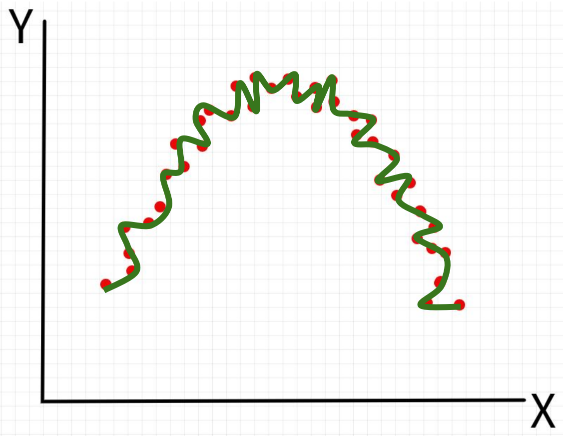


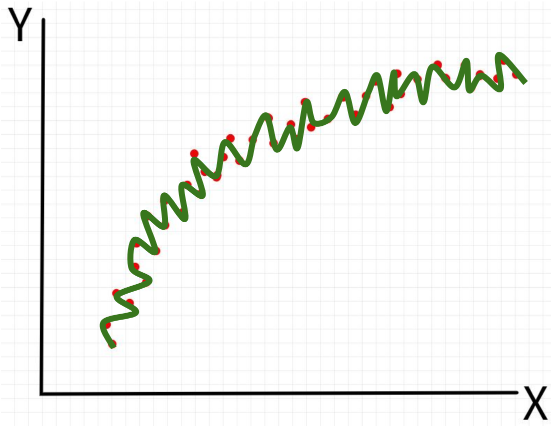


Motivation

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



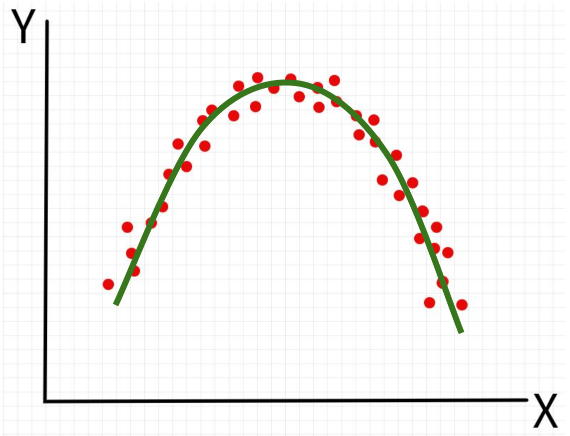


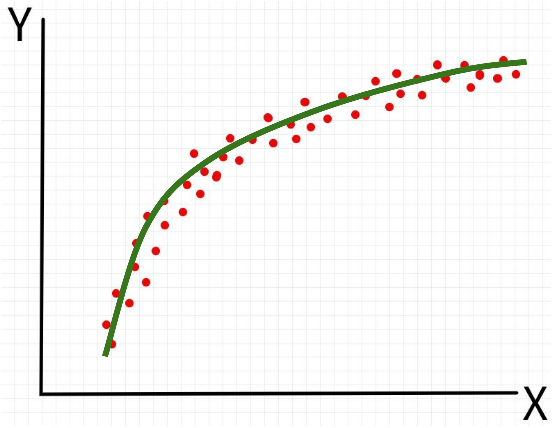


h may be too complex
overfitting - may not perform well on unseen data

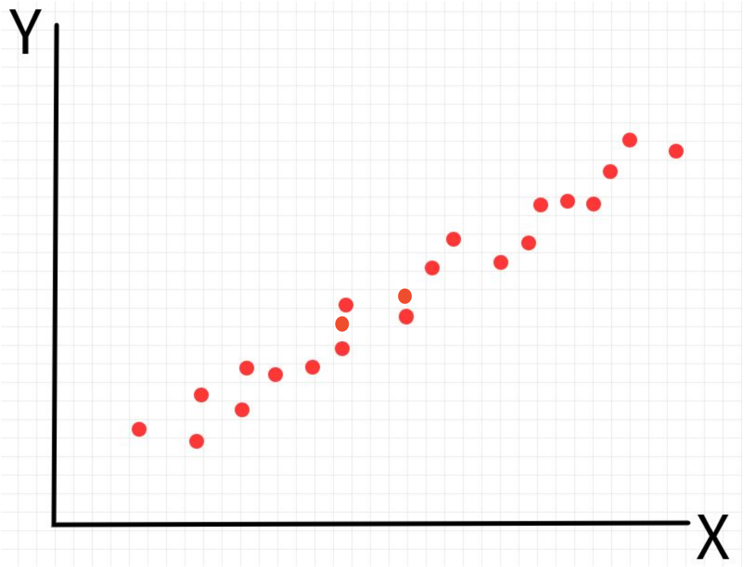
Motivation

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?

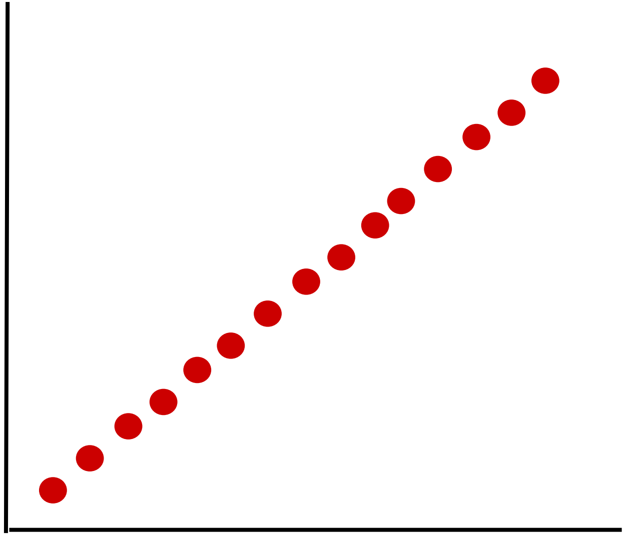




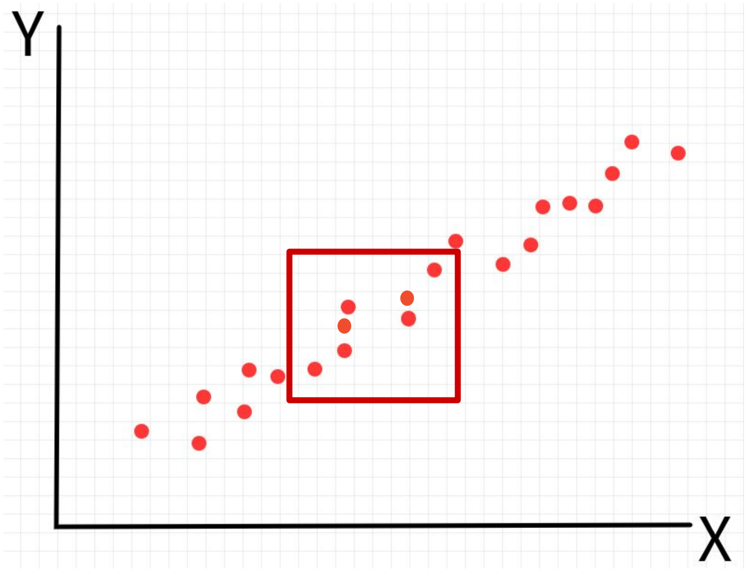
Assumptions



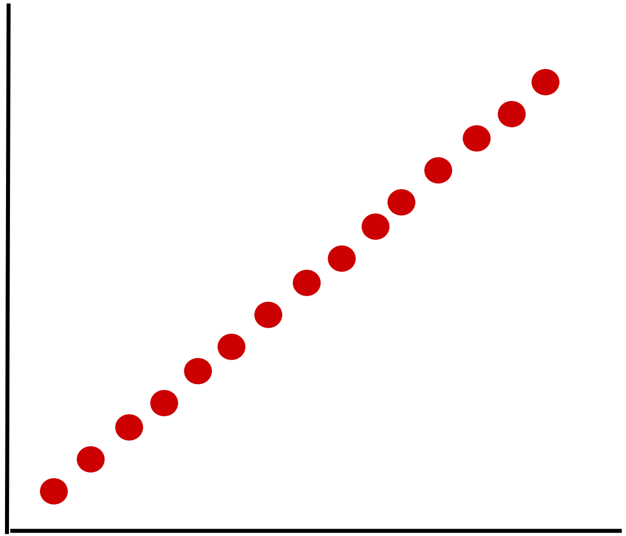
VS



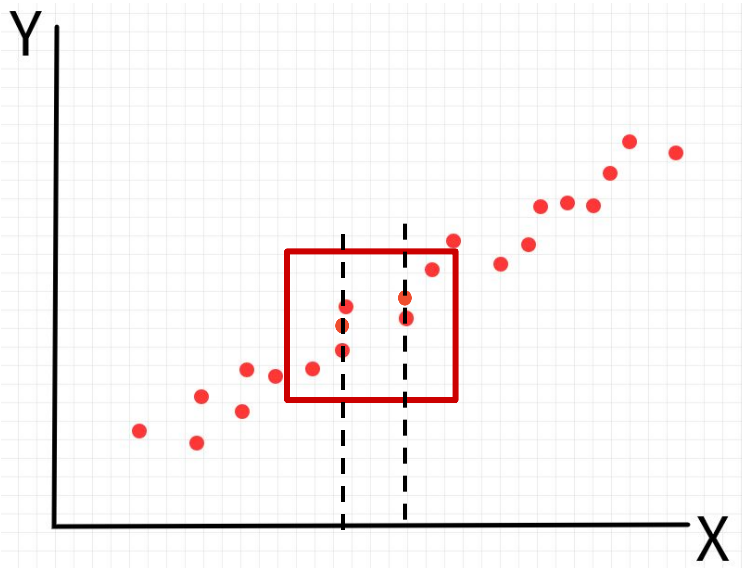
Assumptions



VS

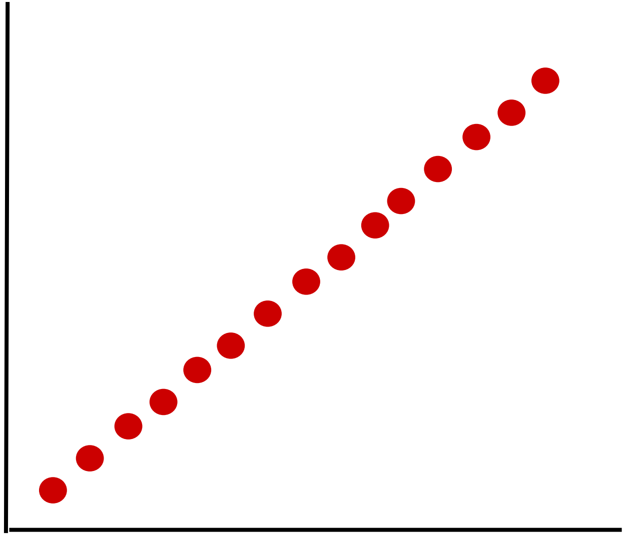


Assumptions

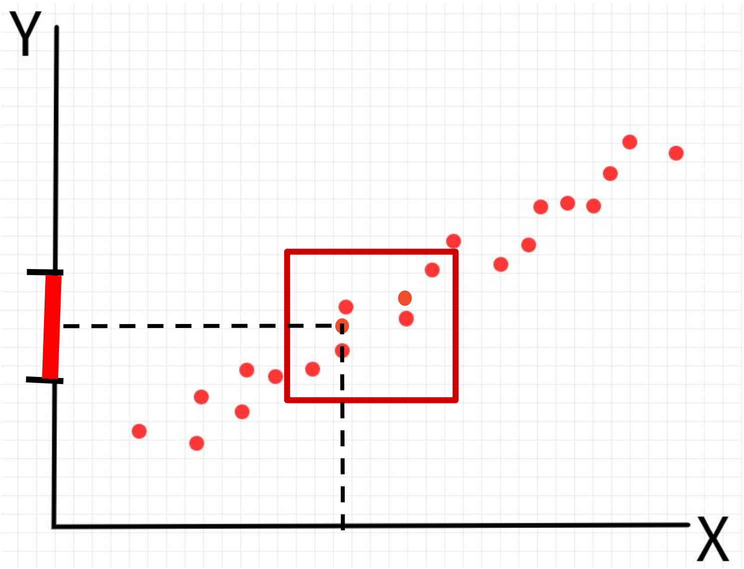


for a given x, can't get an exact y

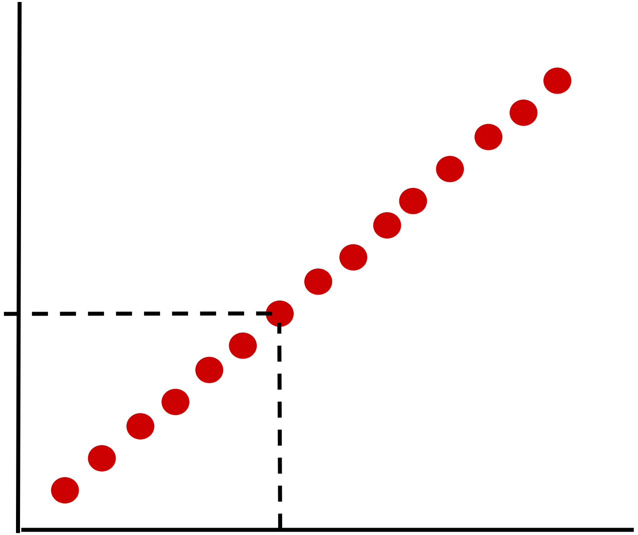
VS



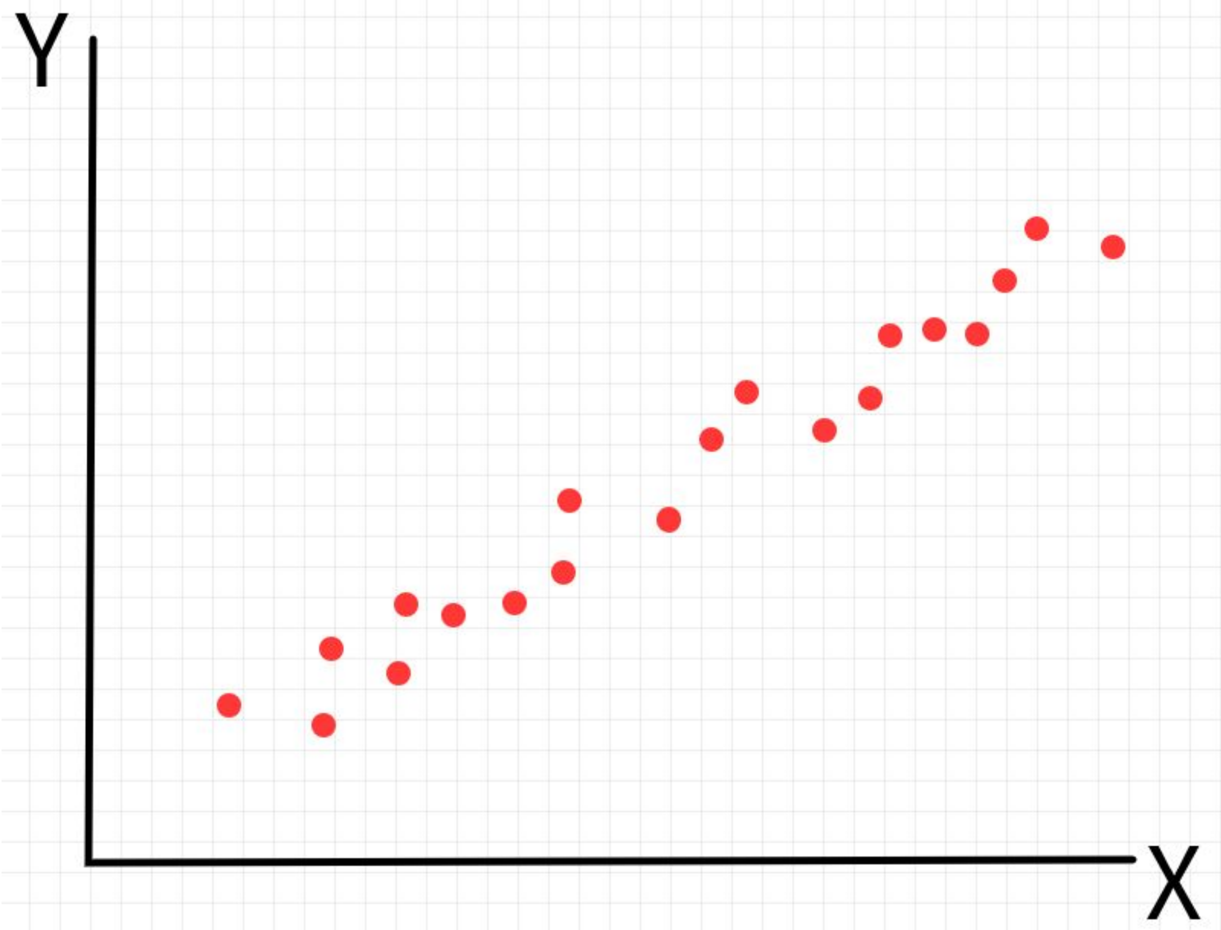
Assumptions



VS

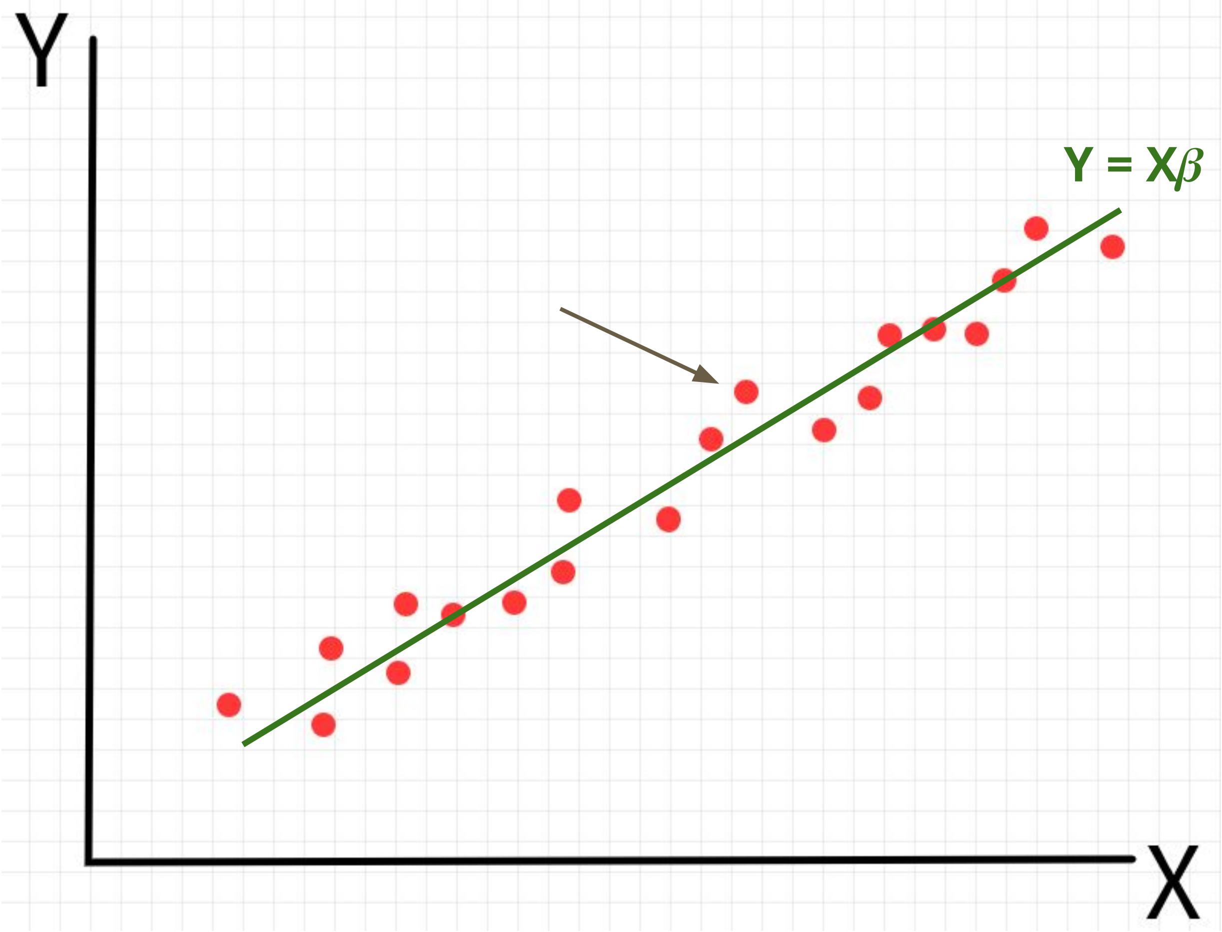


Assumptions

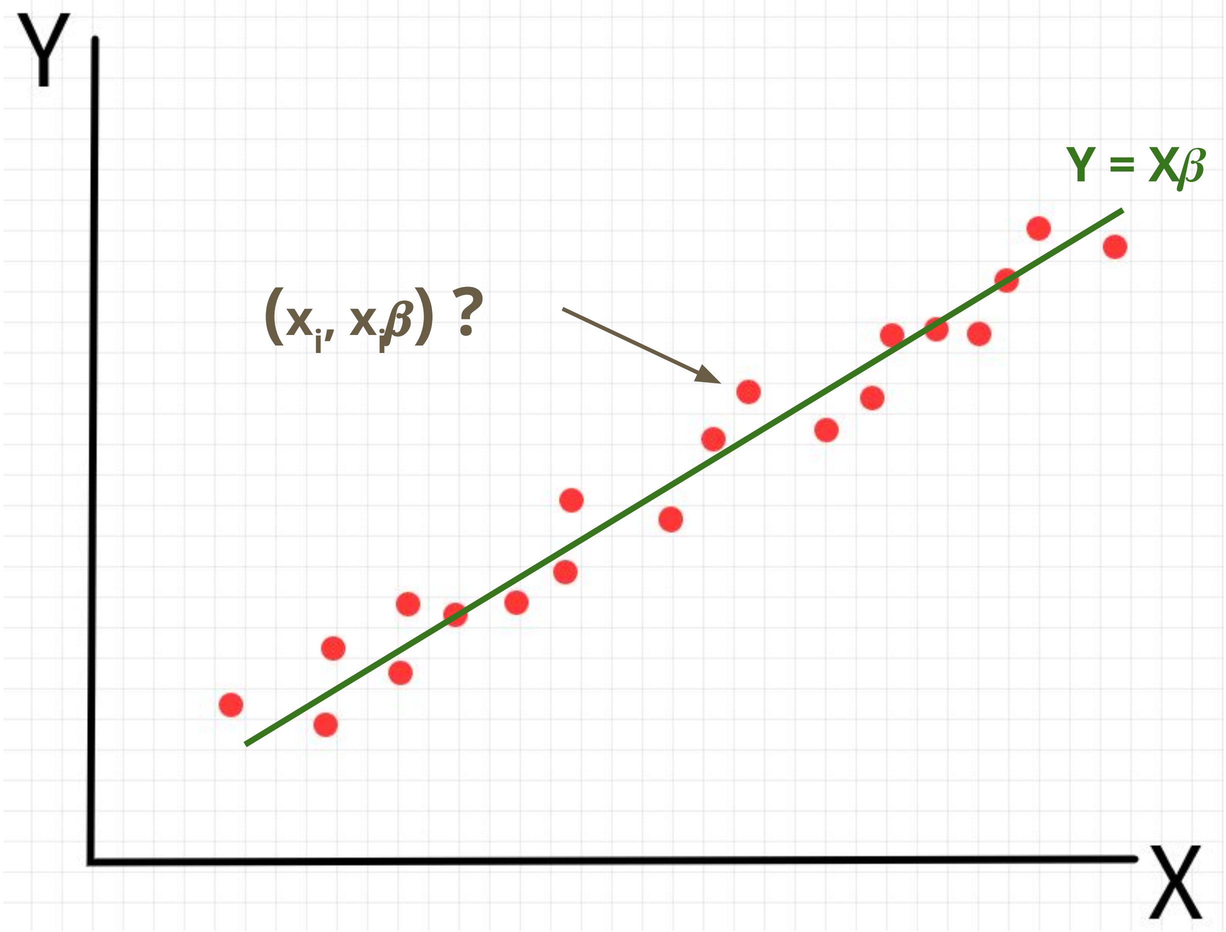


Assumptions

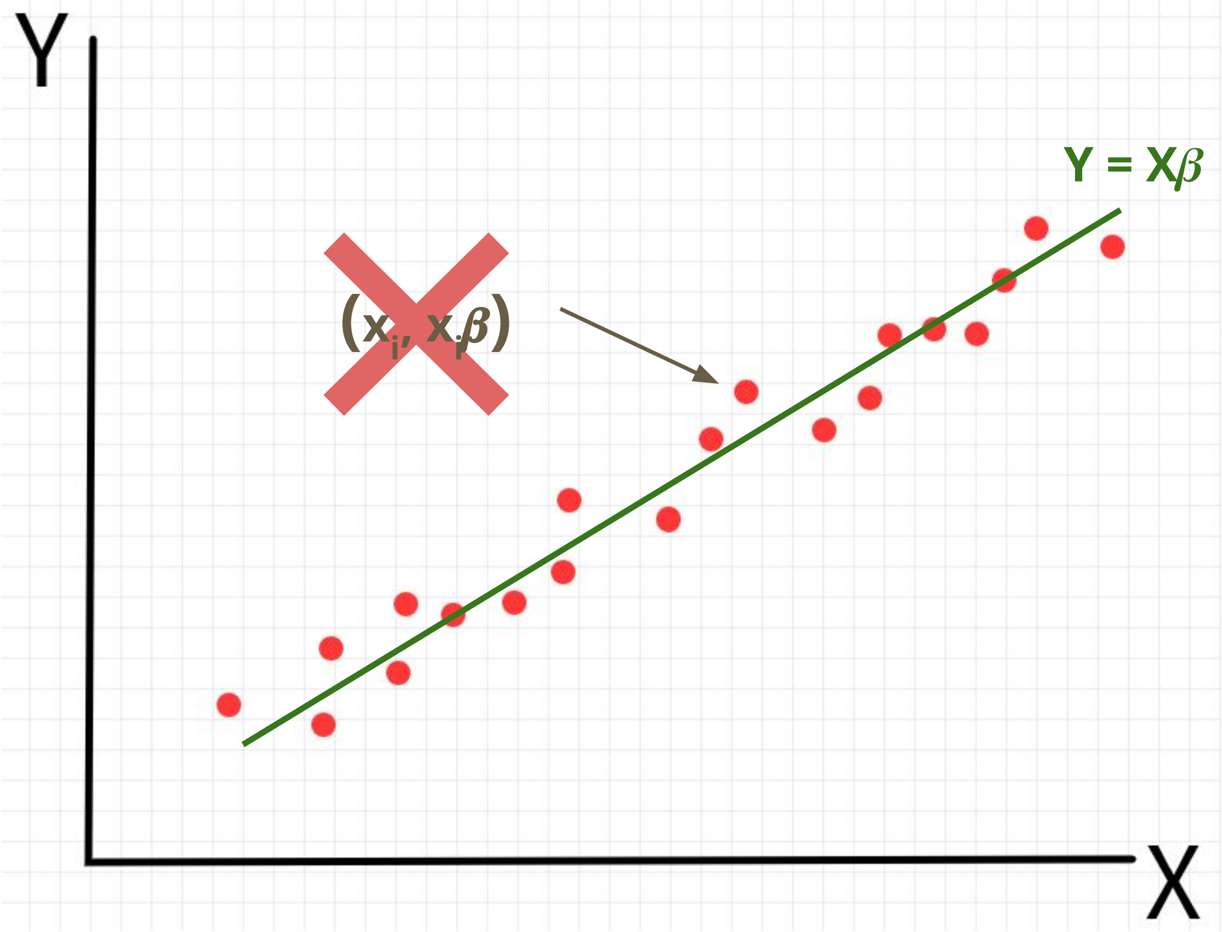
Assumptions



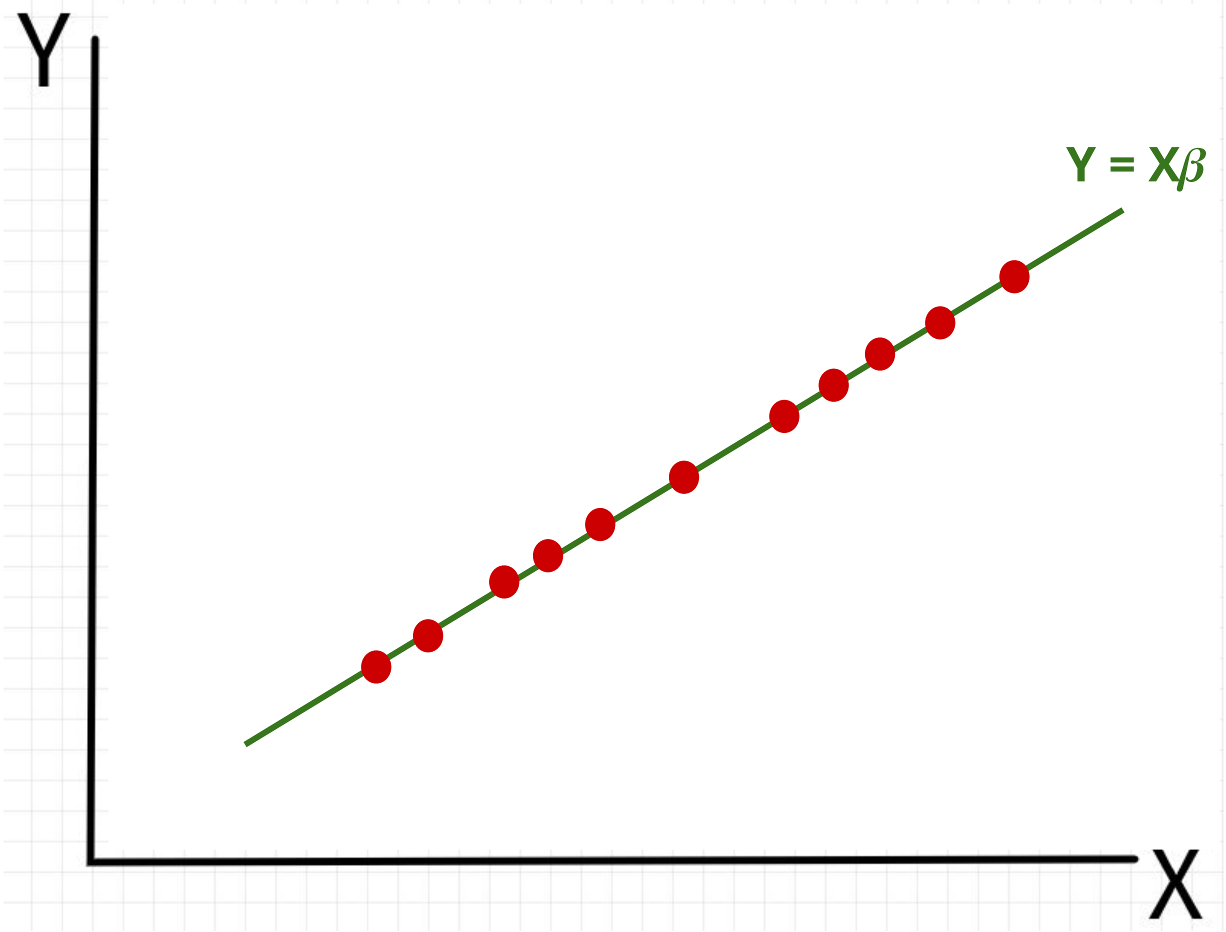
Assumptions



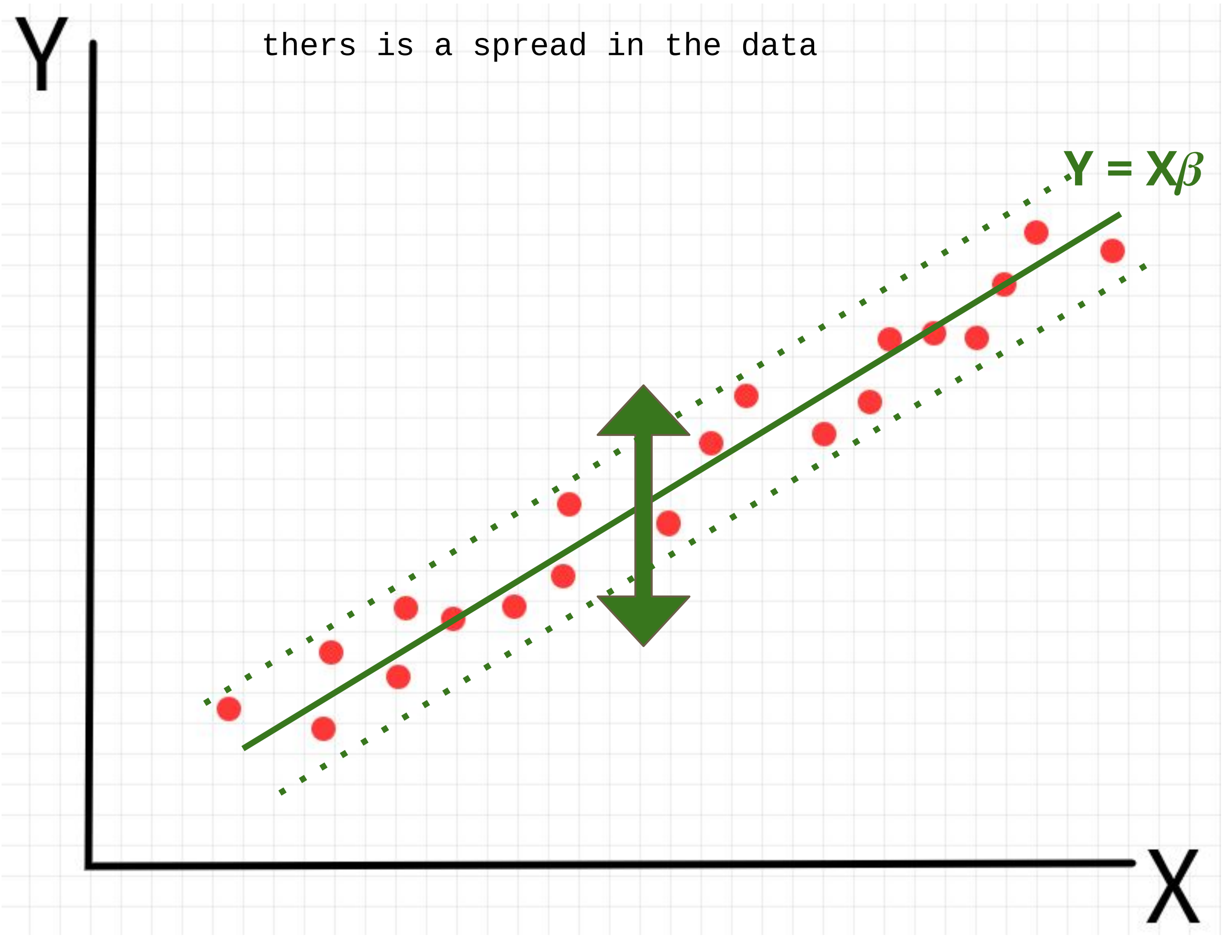
Assumptions



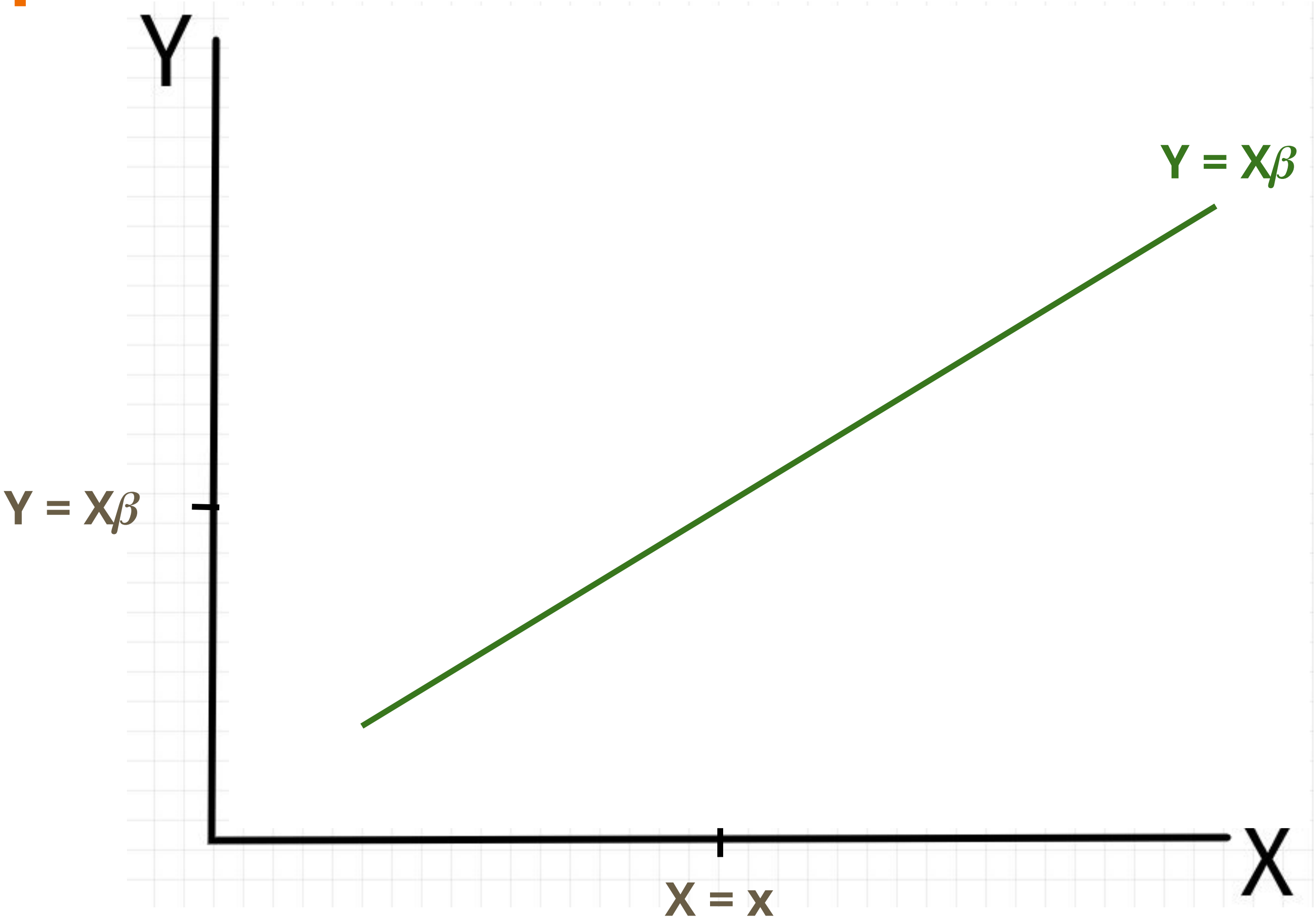
Assumptions



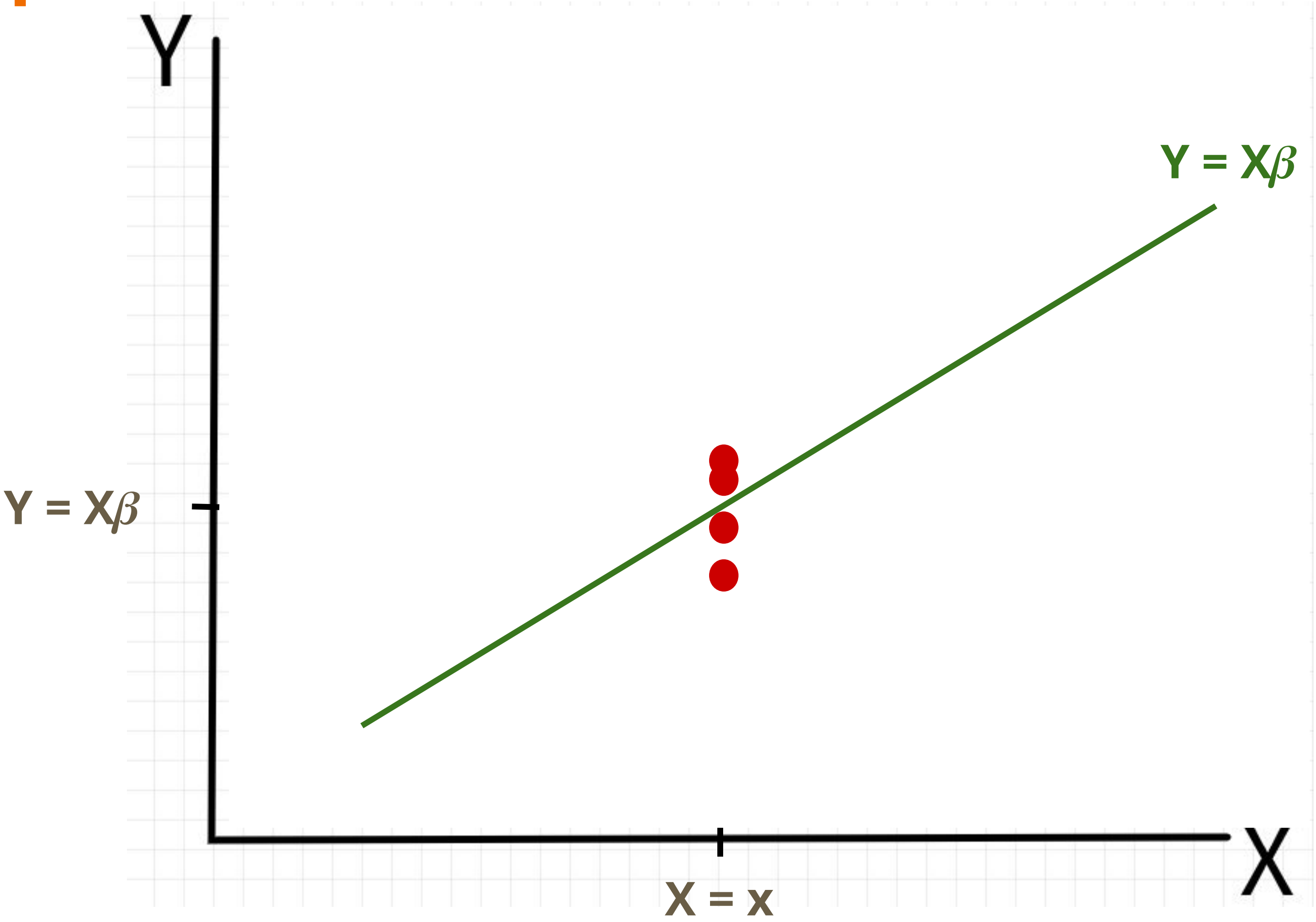
Assumptions



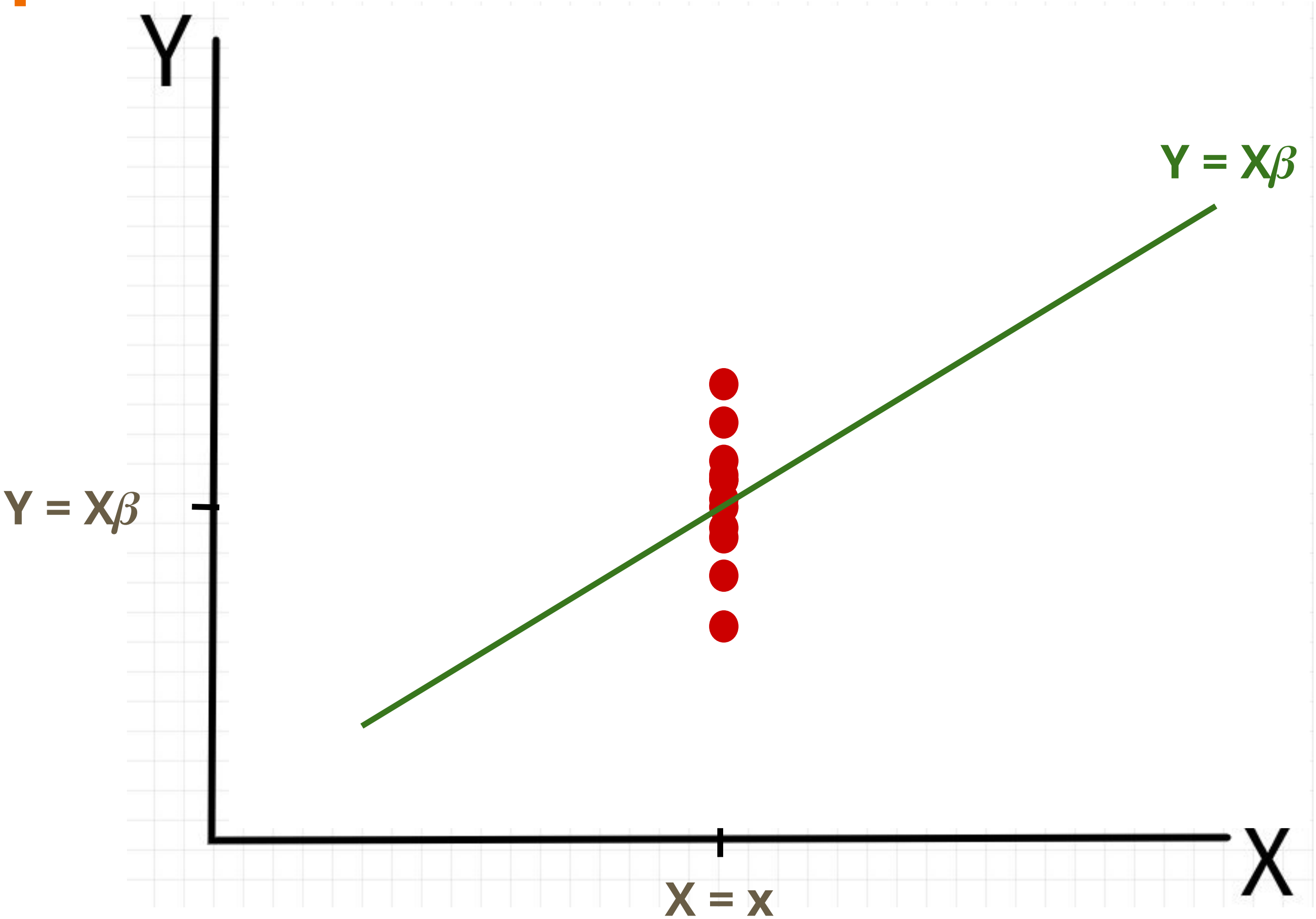
Assumptions



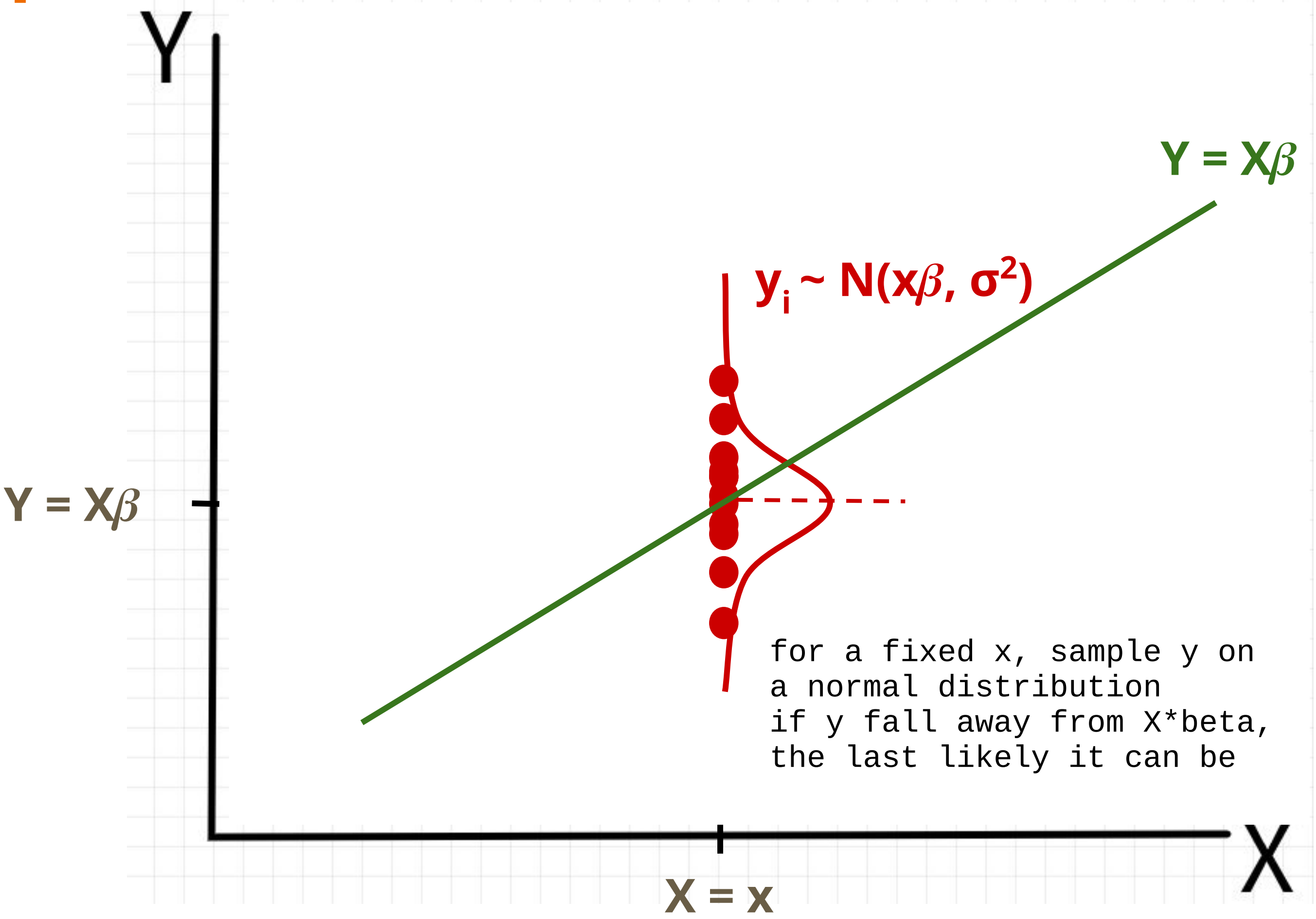
Assumptions



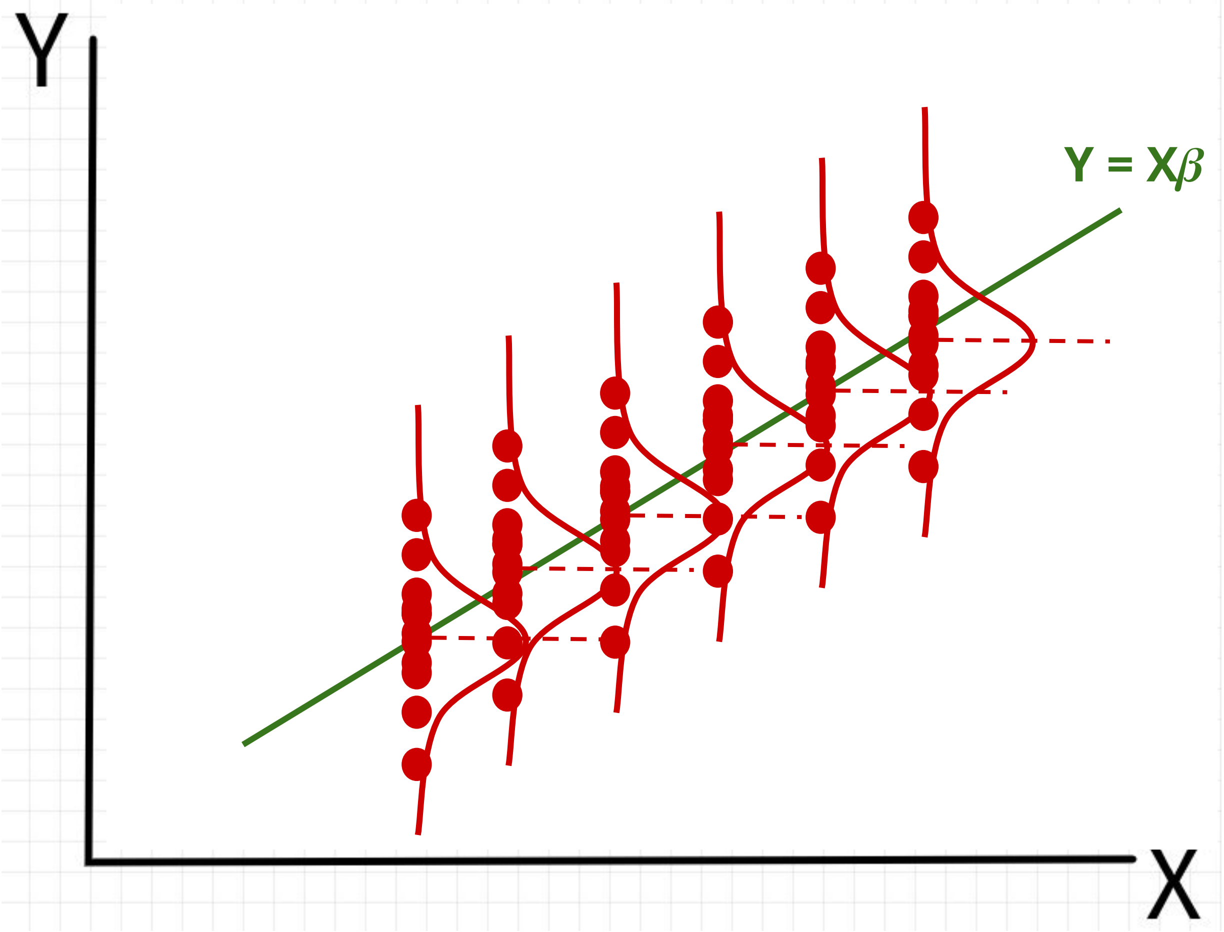
Assumptions



Assumptions

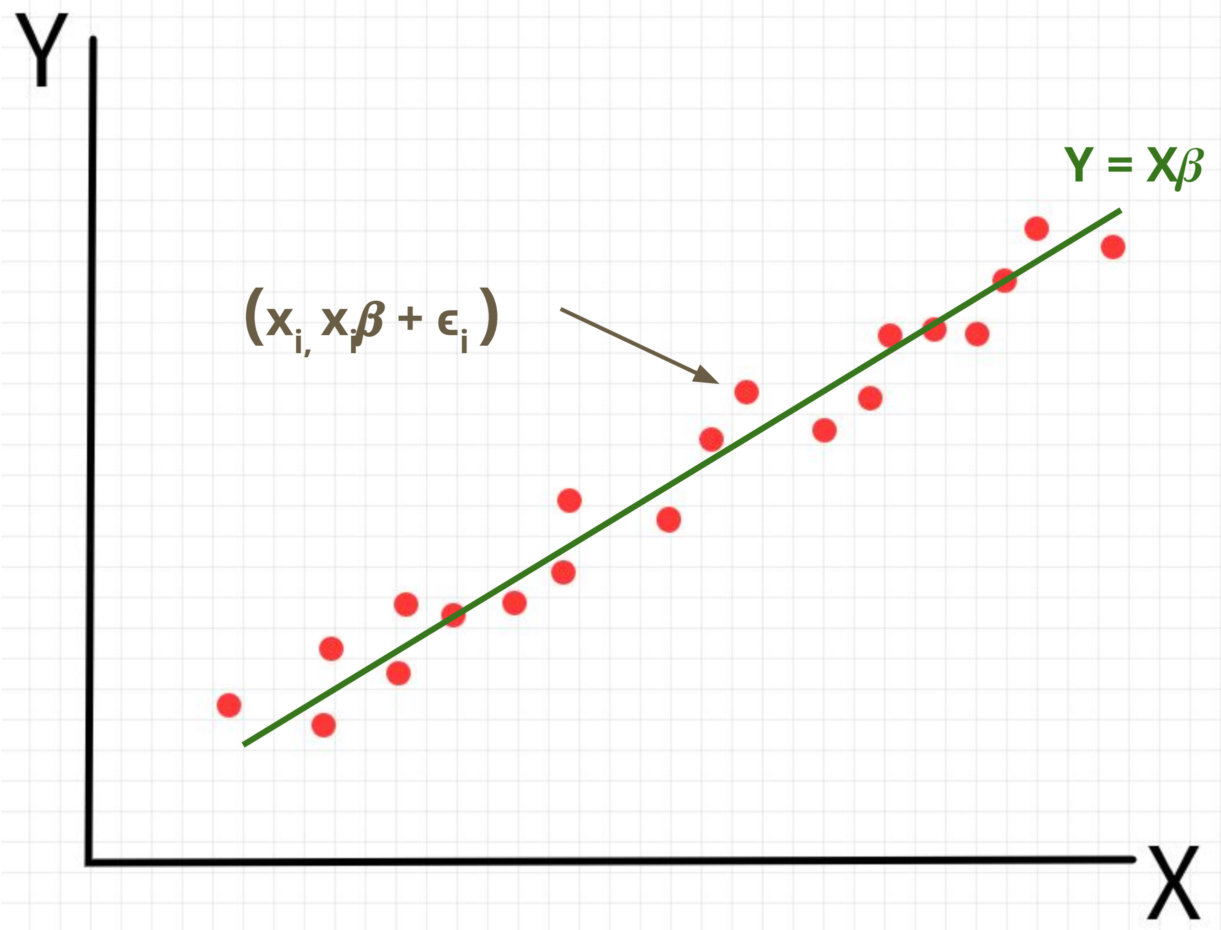


Assumptions

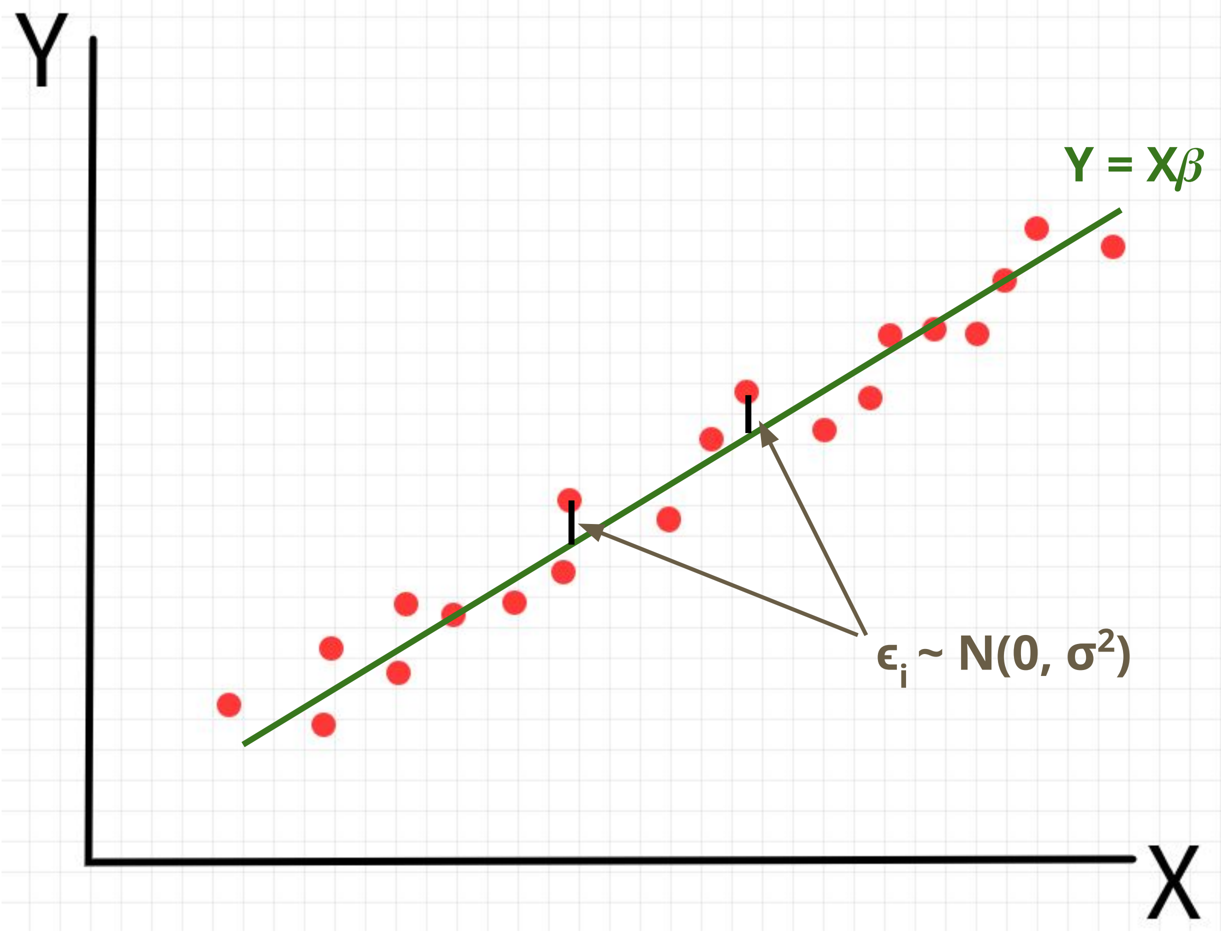


Assumptions

Assumptions



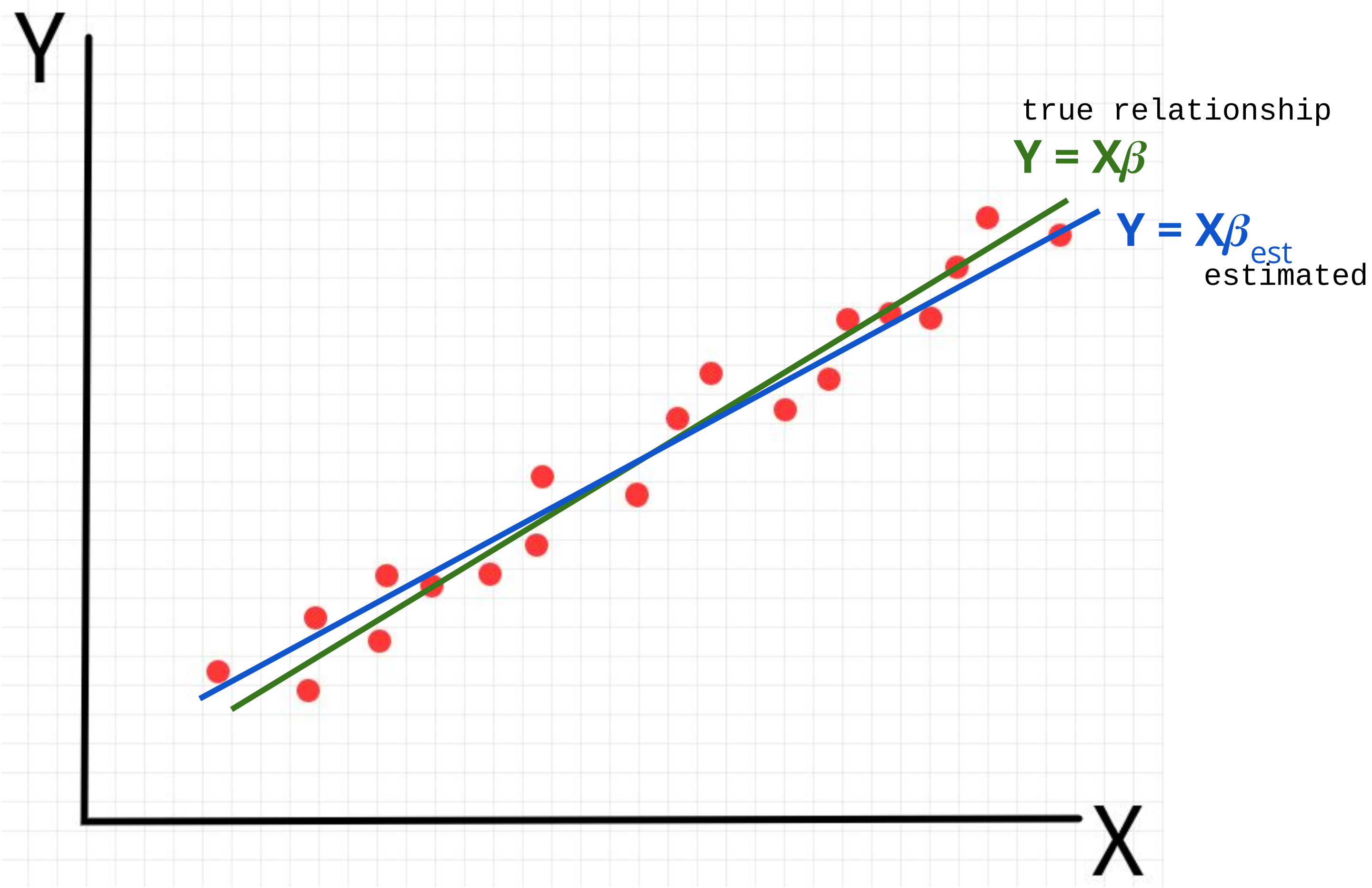
Assumptions



Assumptions

Where does this randomness come from?

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

Cost Function

Given our data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

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

Compare $h(x_i)$ to y_i for all i.

Cost Function

Given our data: **$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$**

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

Compare **$h(x_i)$** to **y_i** for all **i**.

Goal: For a given distance function **d**, find **h** where **L** is smallest.

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

Worksheet a)

Assumptions

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

Least Squares

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

(x_i, y_i) are from our dataset

Least Squares

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

$$= \arg\min_{\boldsymbol{\beta}} \|\vec{y} - h_\beta(\mathrm{X})\|_2^2$$

$$= \arg\min_{\boldsymbol{\beta}} \|y - X\beta\|_2^2$$

Least Squares

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

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

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

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

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

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

Worksheet b) & c)

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 ***β***.
Which functions below are linear in ***β***?

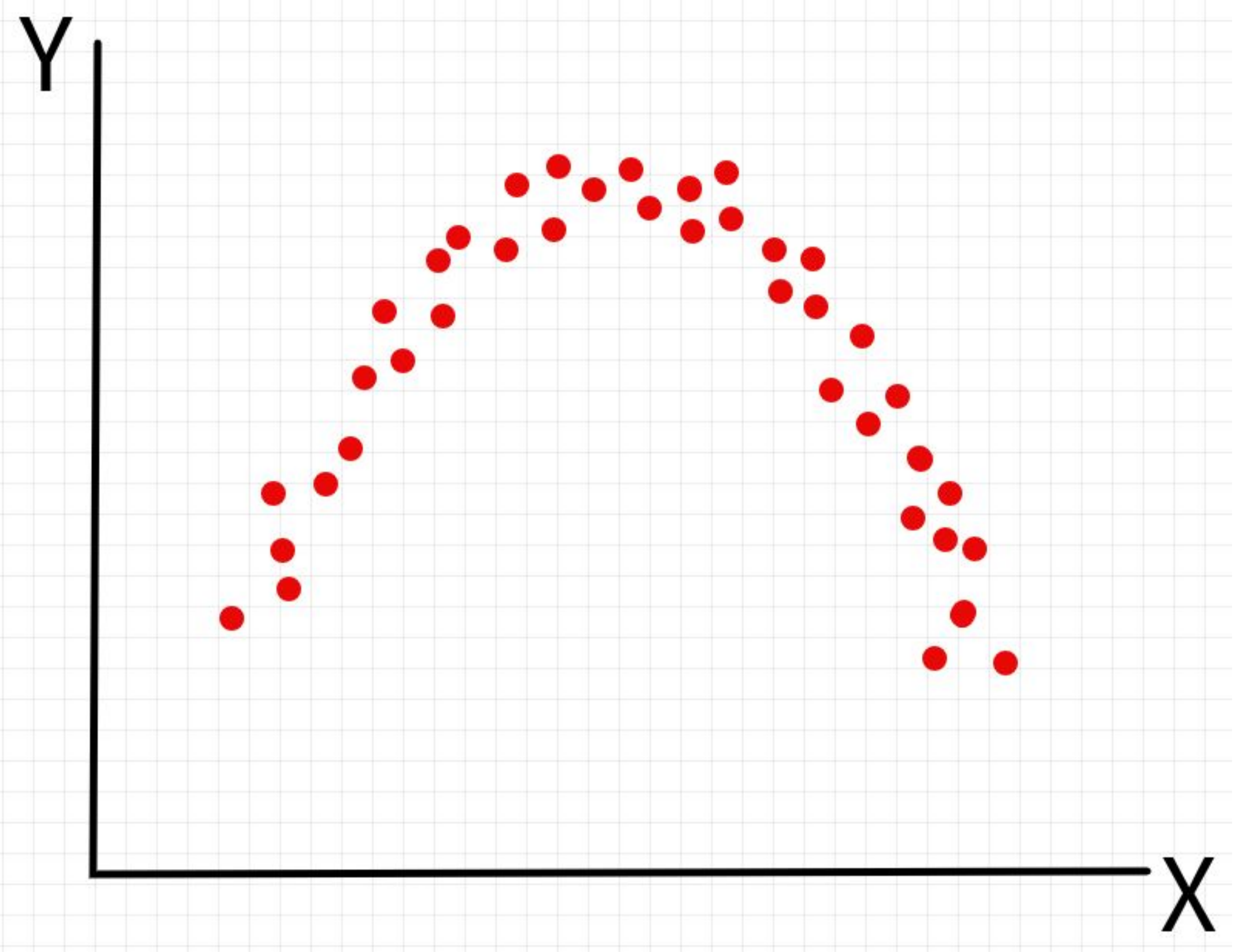
$h(\beta) = \beta_1 x$ ✔

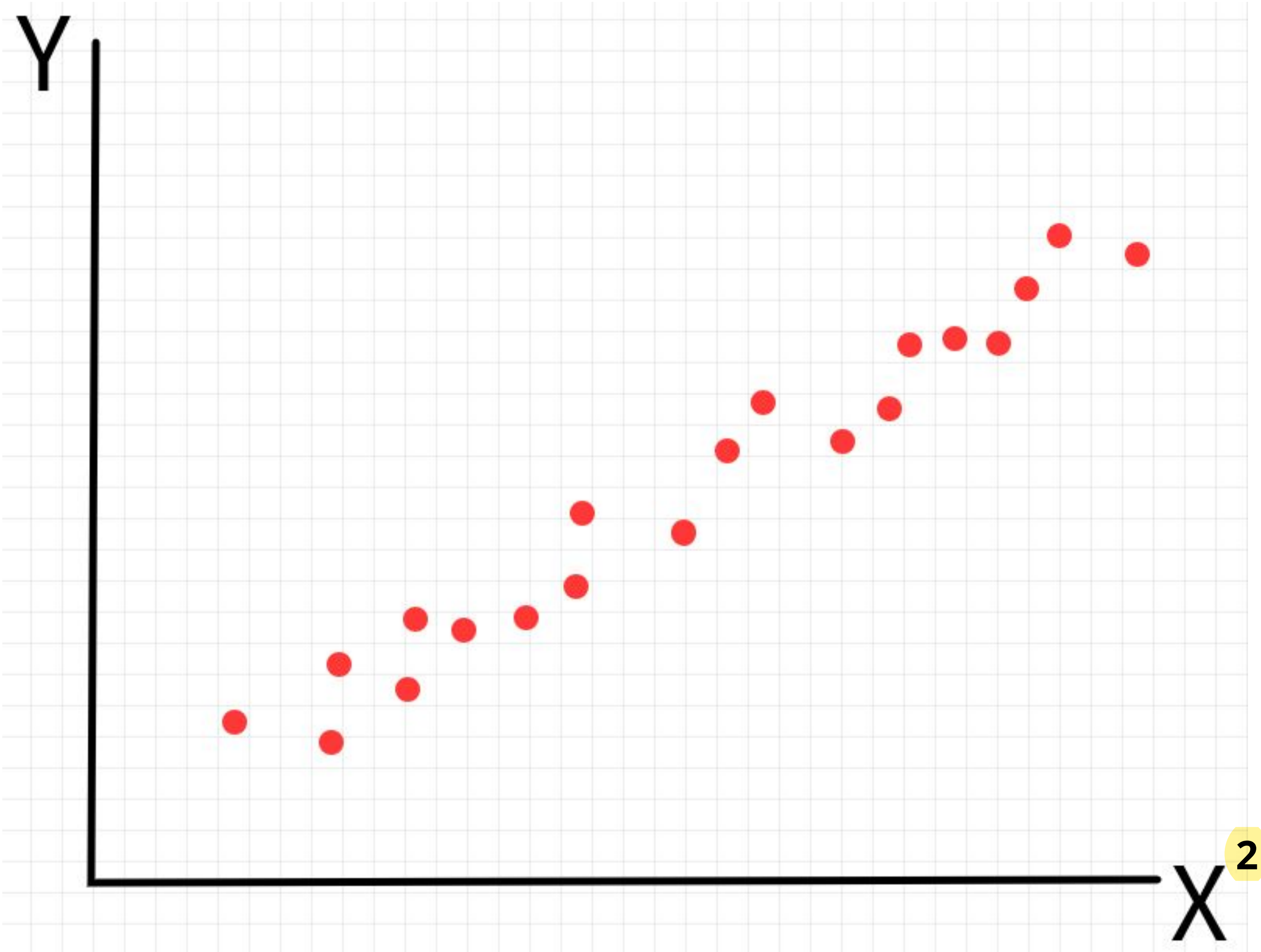
$h(\beta) = \beta_0 + \beta_1 x$ ✔

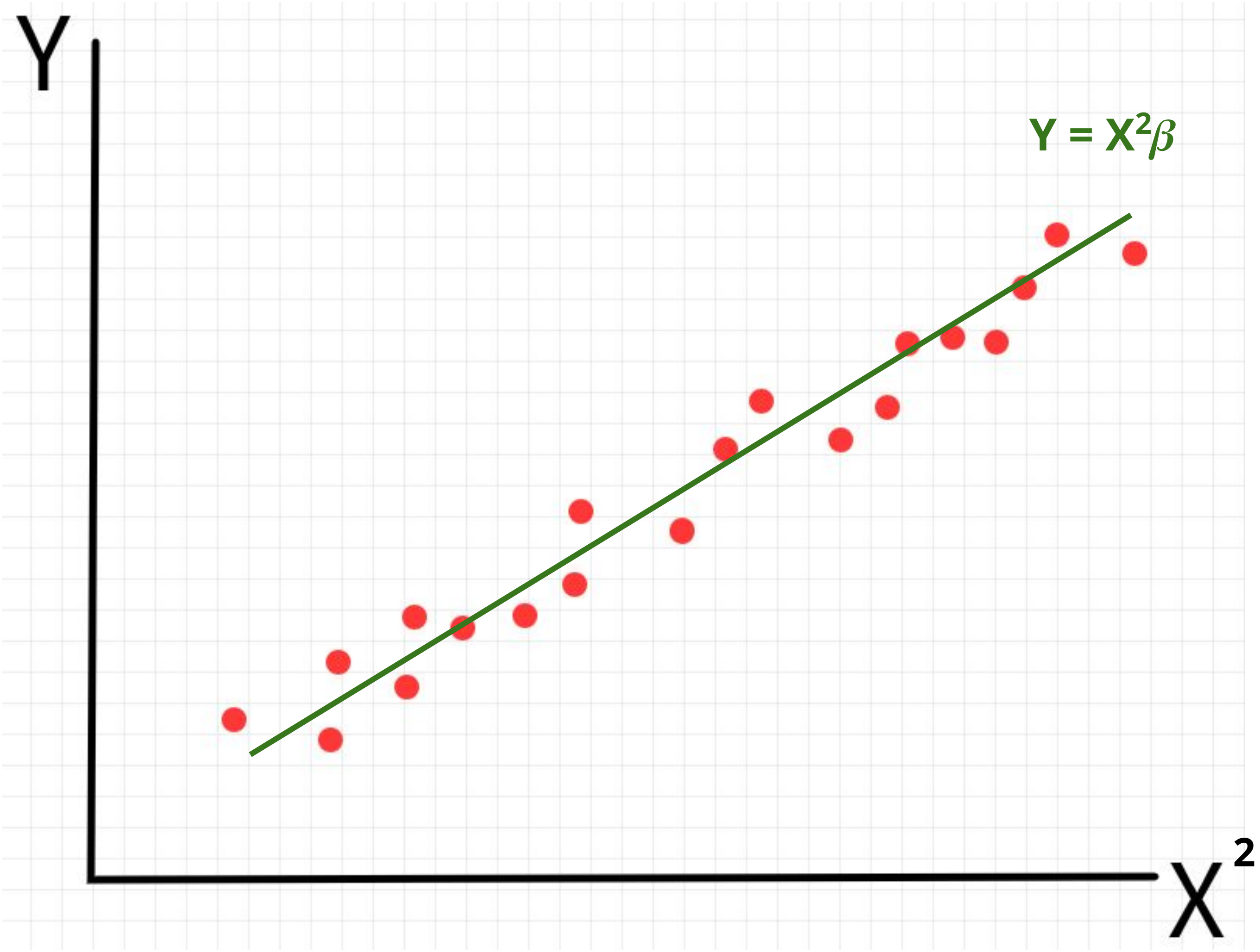
$h(\beta) = \beta_0 + \beta_1 x + \beta_2 x^2$ ✔

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

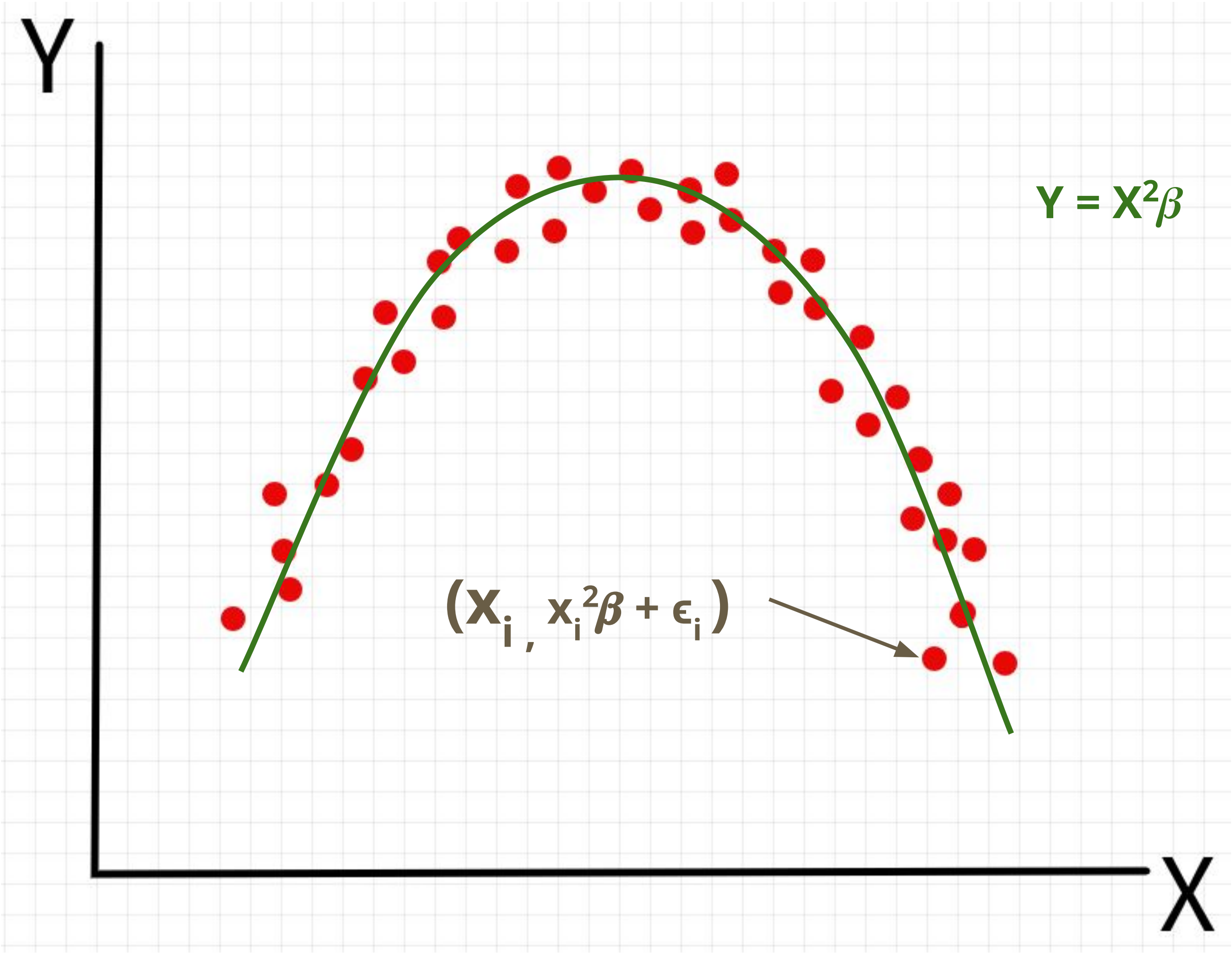
$h(\beta) = \beta_0 + \beta_1 x + \beta_1^2 x$ ✘

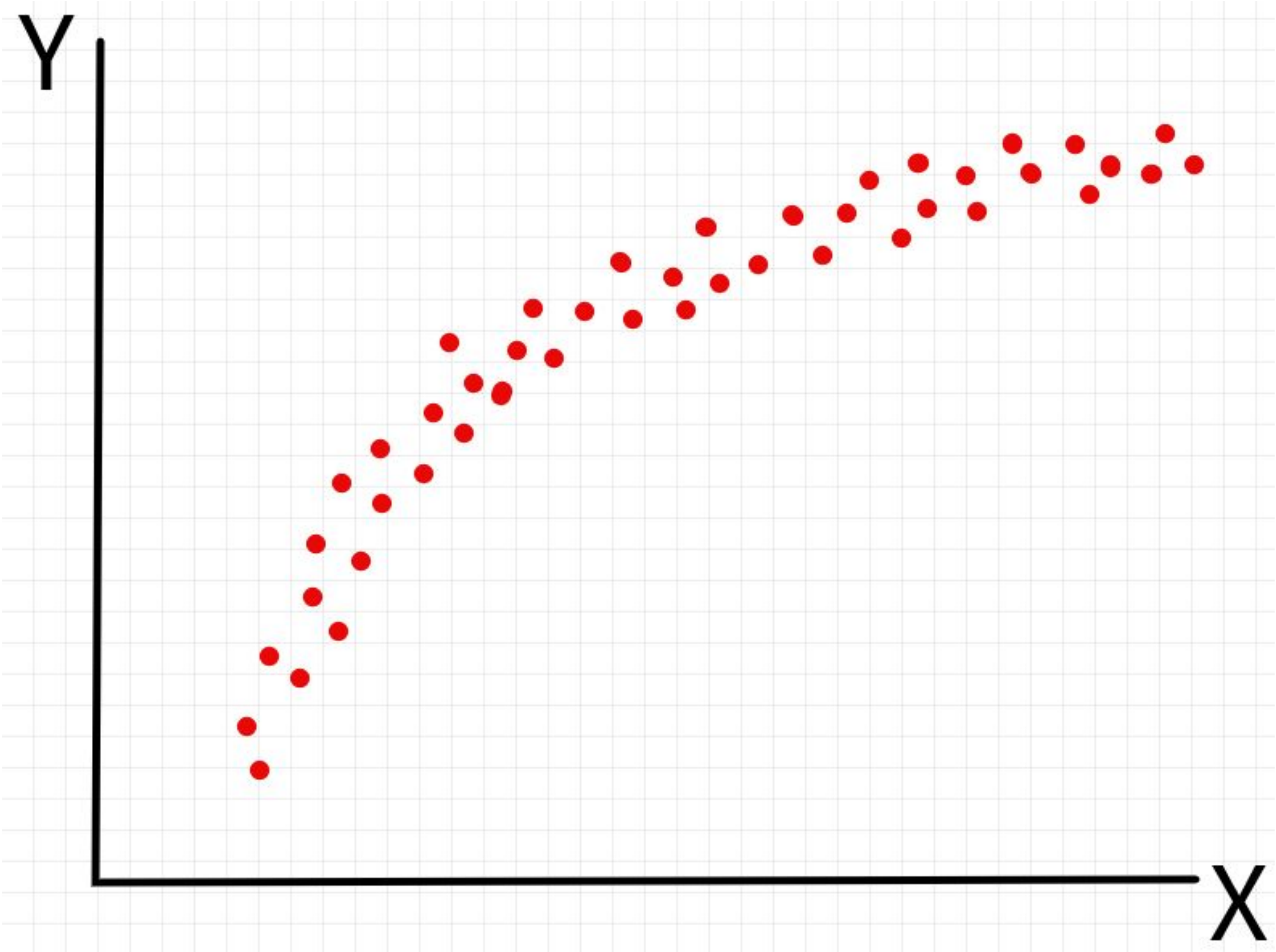

Y
X


Y
X



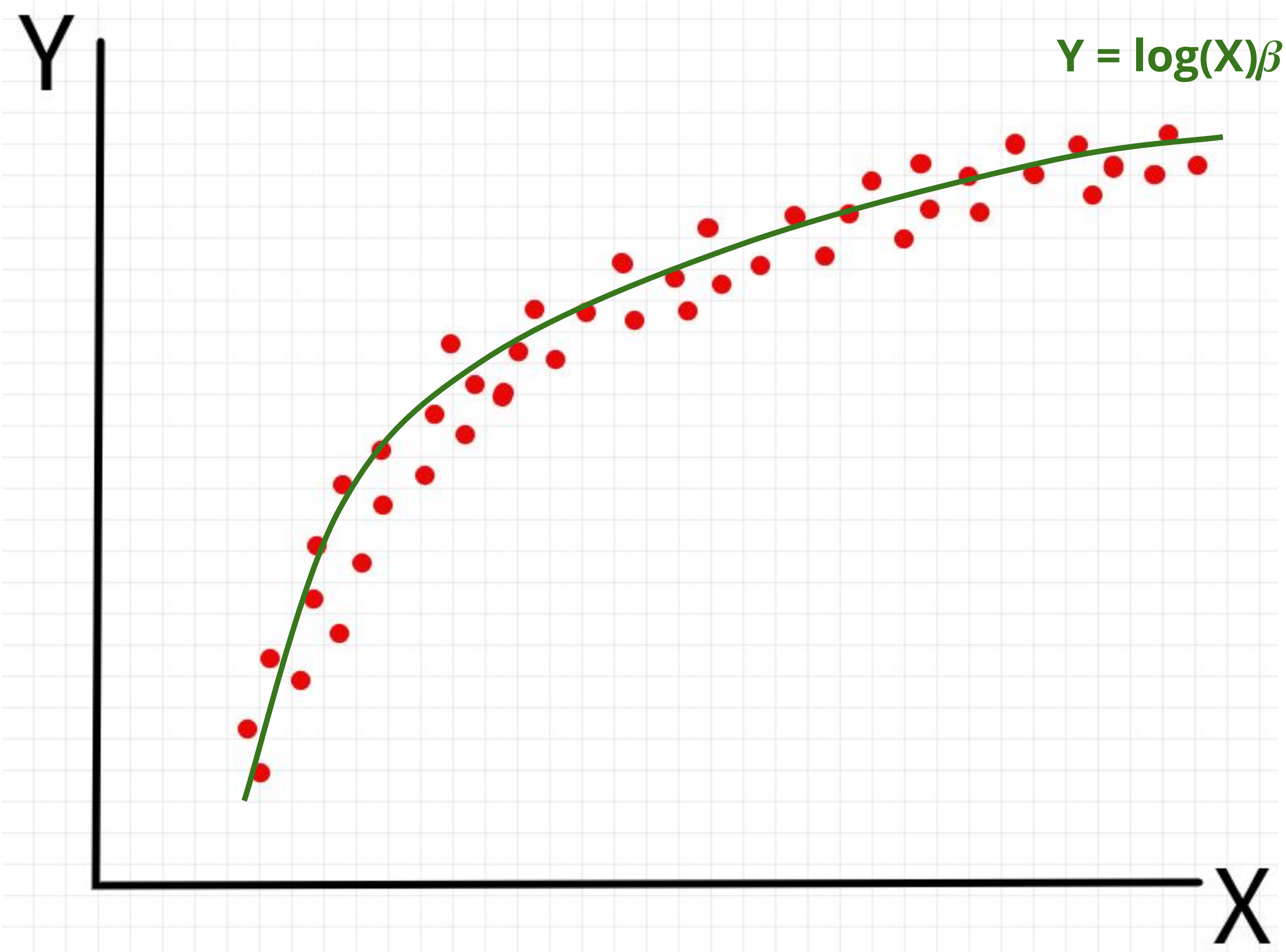
Y
Y = X²β
X²


Y
X
Y = X²β
(xᵢ , xᵢ²β + ϵᵢ)


Y
X


Y
log(X)


Y
Y = log(X)β
log(X)



Y
X
Y = log(X)β

Worksheet d) -> f)

Maximum Likelihood

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

Define **P(Y | h)** as the probability of observing **Y** given that it was sampled from **h**.

Goal: Find **h** that maximizes the probability of having observed our data.

Maximum Likelihood

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

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

Maximum Likelihood

Since **ϵ ~ N(0, σ²)** and **Y = X*β* + ϵ** then **Y ~ N(X*β*, σ²)**.

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

same as least square

An Unbiased Estimator

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

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

An Unbiased Estimator

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

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

y is the only thing contains
randomness

An Unbiased Estimator

$\boldsymbol{\beta}_{\mathbf{LS}}$ is an unbiased estimator of the true $\boldsymbol{\beta}$. That is **E[$\boldsymbol{\beta}_{\mathbf{LS}}$]=$\boldsymbol{\beta}$**.

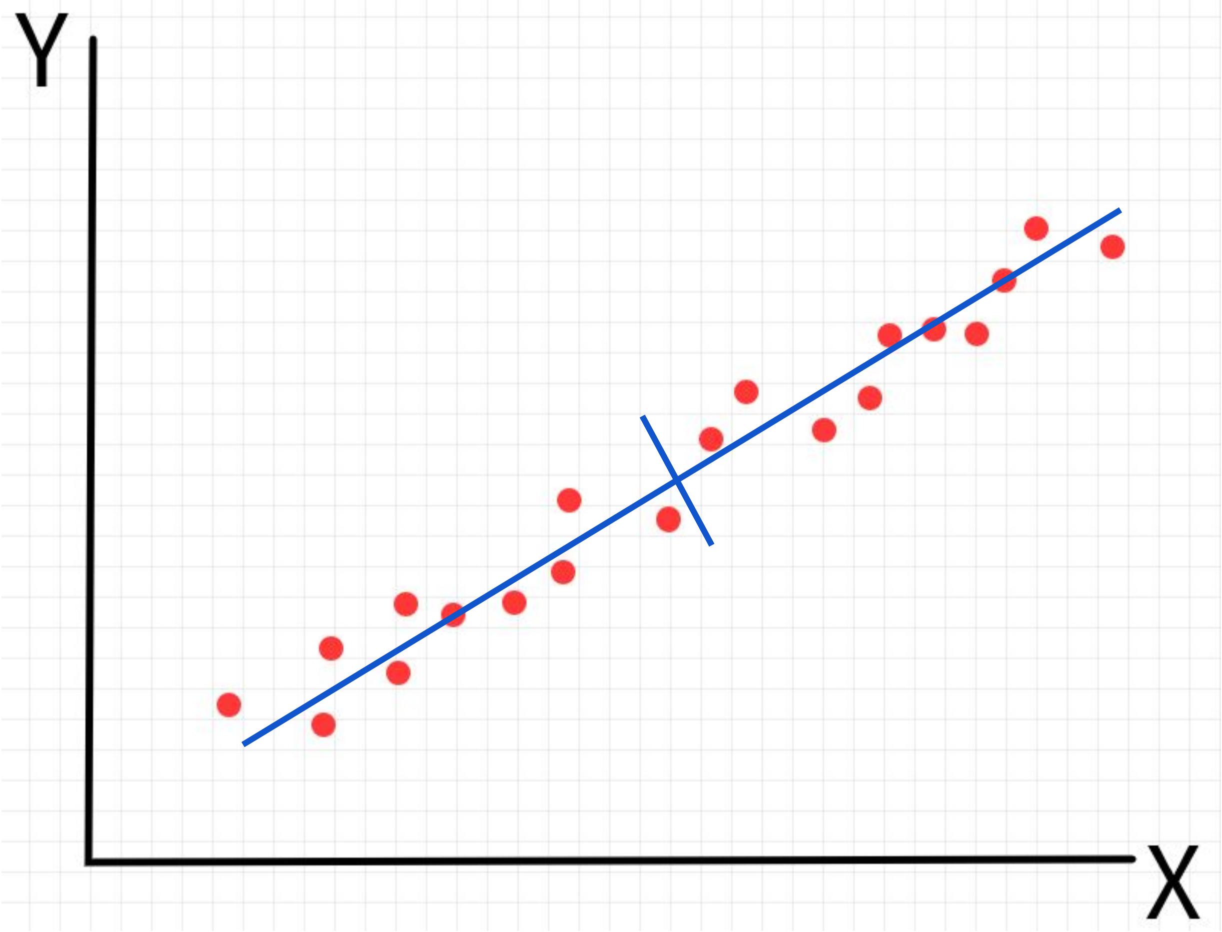
$$
\begin{aligned}
E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\
&= (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T E[X\beta + \epsilon]
\end{aligned}
$$

An Unbiased Estimator

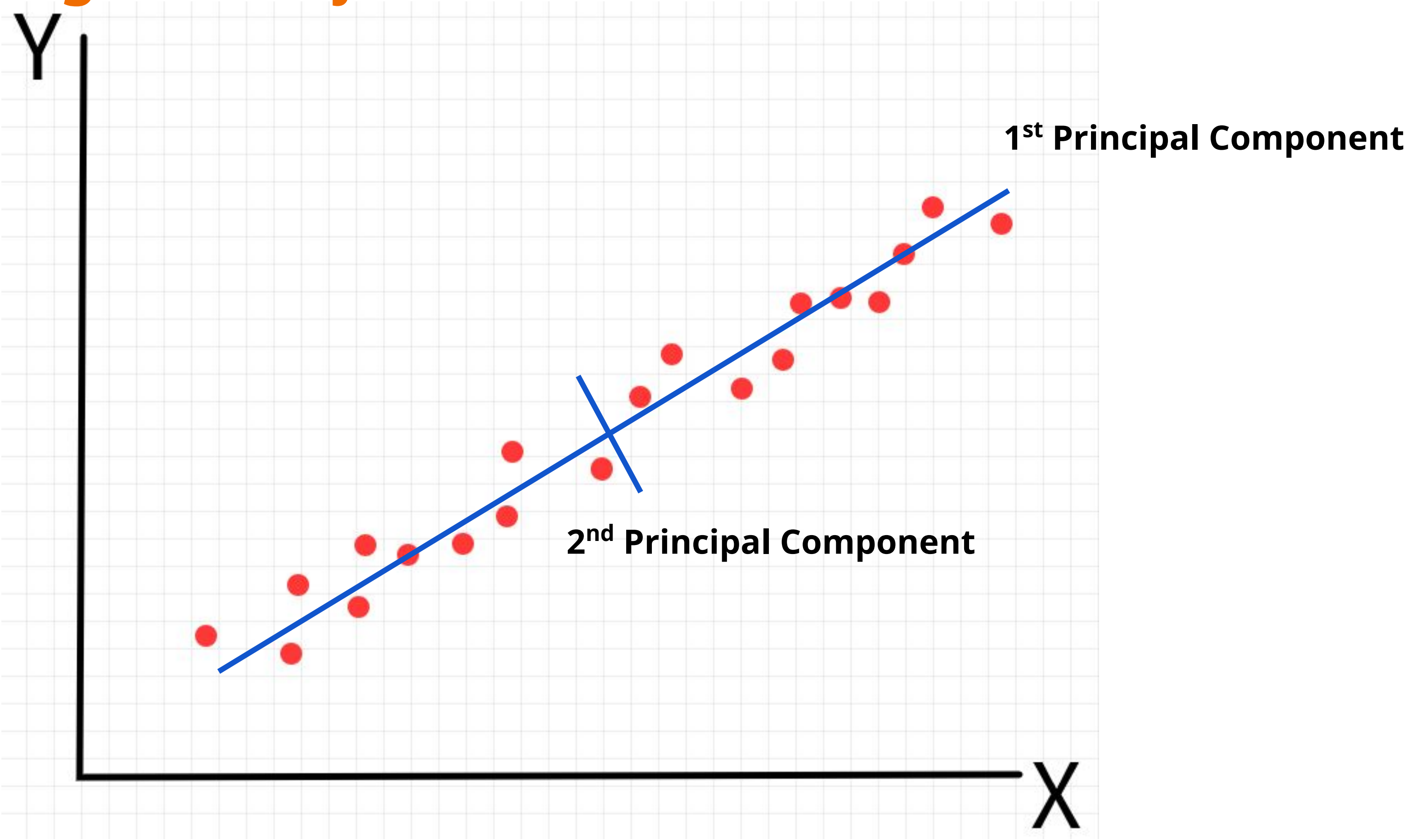
$\boldsymbol{\beta}_{\mathbf{LS}}$ is an unbiased estimator of the true $\boldsymbol{\beta}$. That is $\mathbf{E}[\boldsymbol{\beta}_{\mathbf{LS}}]=\boldsymbol{\beta}$.

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

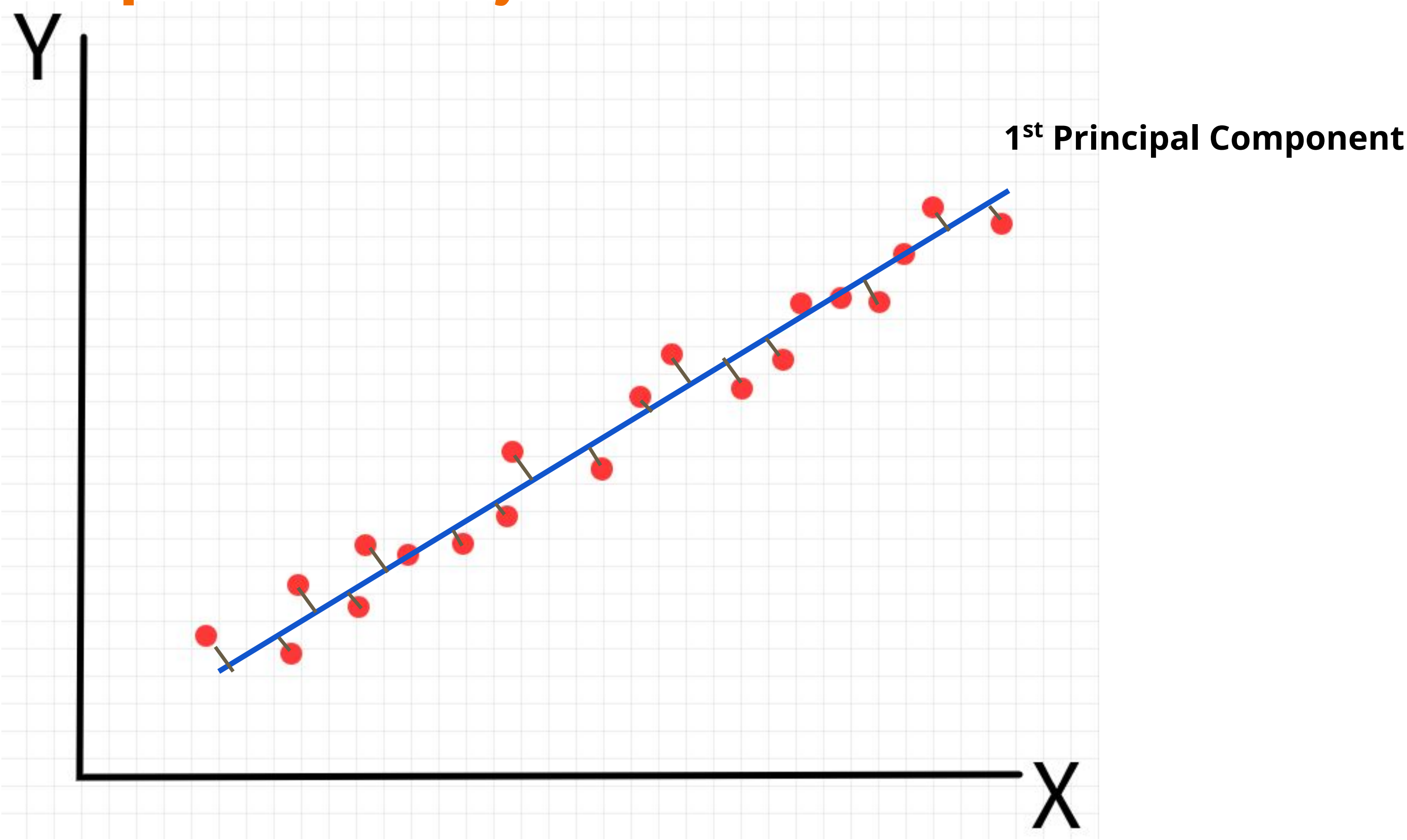
Worksheet g)


Y
X

Is Linear Regression just PCA?



Principal Component Analysis



Linear Regression