Linear Regression

- Intro to linear regression
- Intro to least square method

DAY

Content

- Problem
- Linear regression
- Least Square Method
- Variance and Covariance

Go-Car

- Go-car adalah taxi online yang sedang popular di Indonesia.
- Banyak orang naik Go-car, karena tarifnya sudah jelas di depan. Namun cara menghitung tarif **belum jelas** karena berbagai factor yang digunakan Go-car dalam menentukan tarifnya.
- Faktor utama adalah **Jarak**, namun ada faktor lain yang tidak kita ketahui yang juga mempengaruhi (mungkin **jumlah permintaan**, **cuaca**, **kemacetan**, dll).*

*Adopted from Machine Learning course IF-TelU

Problem

- "JalanJalan" adalah **perusahaan travel** yang sangat sering menggunakan jasa go-car untuk mengantar tamu-tamu mereka. Karena mereka harus menghitung anggaran bulanan, maka mereka harus **menghitung perkiraan biaya Go-car** jauh hari sebelum digunakan, sehingga pengecekan langsung pada aplikasi go-car dianggap bukan cara yang tepat.
- "JalanJalan" memiliki DATA history biaya go-car selama beberapa bulan terakhir.
- Pertanyaan: Bagaimana cara kita untuk menentukan MODEL dari data history biaya go-car untuk memprediksi tarif bagi perusahaan "JalanJalan" tersebut?

*Adopted from Machine Learning course IF-TelU

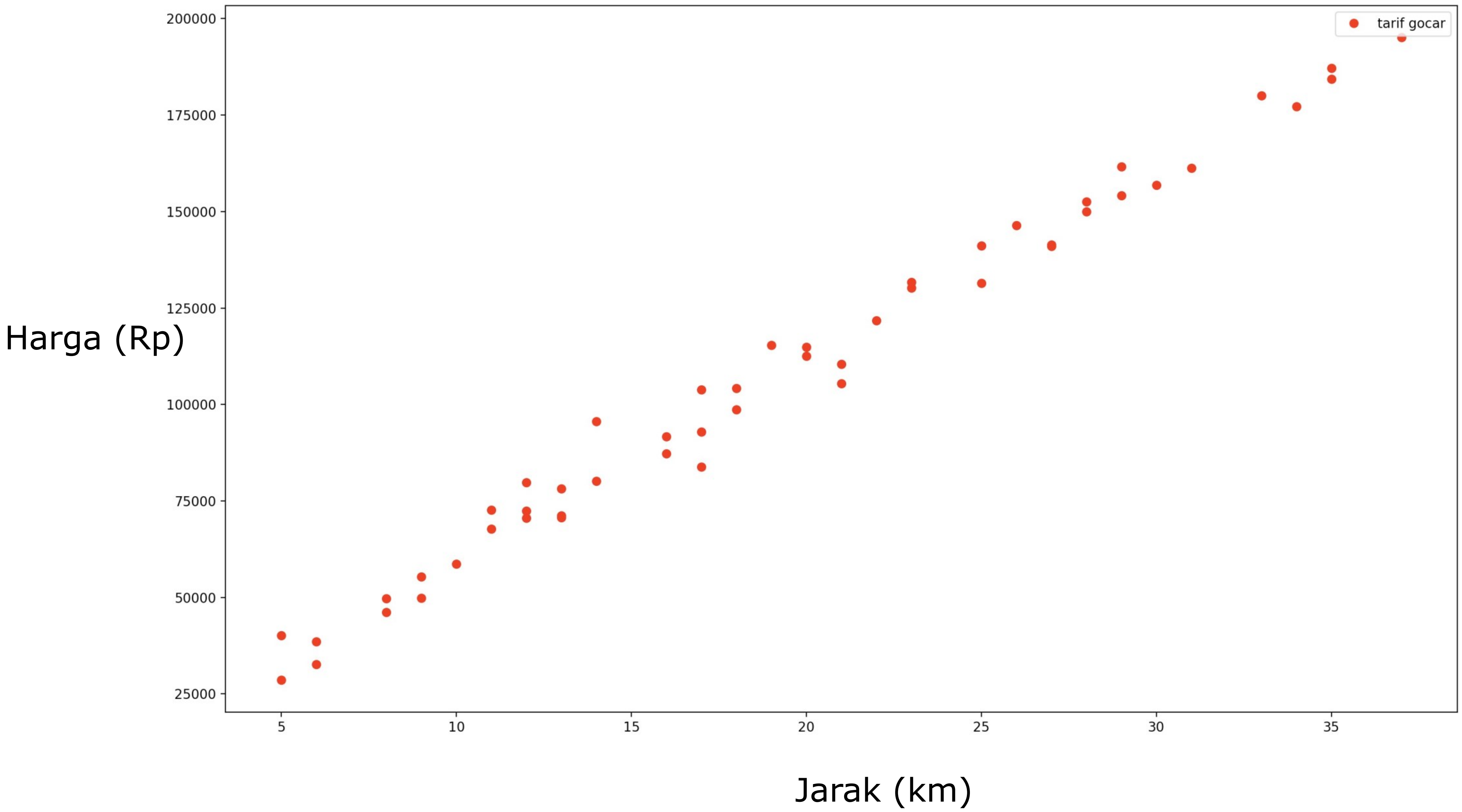
DATA historis Go-car

- Sebagai contoh data go-car dari 15 kali naik Go-car sebagai berikut:

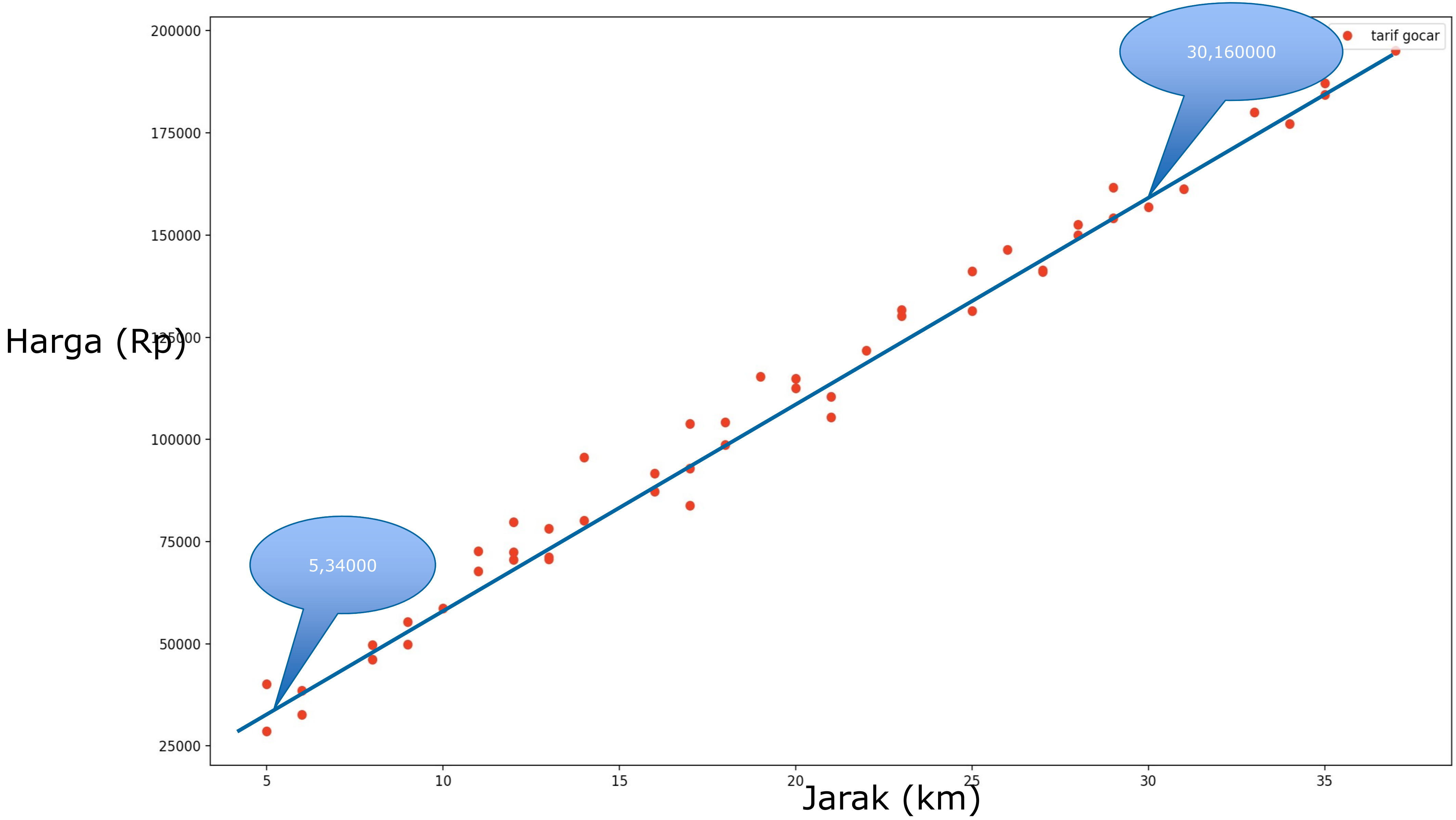
- What can you do with this history data ?

NO	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
KM	5	5	6	6	8	8	9	9	10	11	11	12	12	12	13
TARIF	28600	40100	32600	38600	49700	46100	55300	49800	58700	67700	72600	79800	72400	70600	70700

Data plot


Harga (Rp)
200000
175000
150000
125000
100000
75000
50000
25000
5
10
15
20
25
30
35
tarif gocar
Jarak (km)

Model → GARIS !


tarif gocar
30,160000
5,34000
Harga (Rp)
Jarak (km)
200000
175000
150000
125000
100000
75000
50000
25000
5
10
15
20
25
30
35

Persamaan Garis?

- **$y = a + mx$**
- **Cari m**
- m = slope = $(y2-y1)/(x2-x1)$
- $= (160000-34000)/(30-5)$
- $= 126000/25 = 5040$
- **Cari Persamaan Garis dari 1 titik**
- $m = (y-y1)/(x-x1)$
- $y = y1+m(x-x1)$
- $y = 34000+5040(x-5)$
- $y = 34000 + 5040x - 25200$
- **$y = 8800 + 5040x$**

Linear Regression

- Linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X

- The goal is to make quantitative (real valued) predictions on the basis of a (vector of) features or attributes

- We consider the case $\boldsymbol{x} \in \mathbb{R}^d$ throughout this chapter

Linear Regression

- Modelling the relation that best fit the data using a single line (linear function)

  - β_0 : Population Y-Intercept (intercept, bias, …)
  - β_1 : Population slope (weight vector,…)
  - ε : Random error

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

- Also often written
  - $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} = \Sigma_{j=1}^{d} w_j x_j$
  - $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + a$

(1a)

(1b)

Linear Regression

- Linear regression assumes that…
  - The relationship between X and Y is **linear**
  - Y is distributed normally at each value of X
  - The variance of Y at every value of X is the same (homogeneity of variances)
  - The **observations are independent**
- The learning problem is to determine the parameters $\boldsymbol{w}$ and a based on data

Linear Regression - Univariate

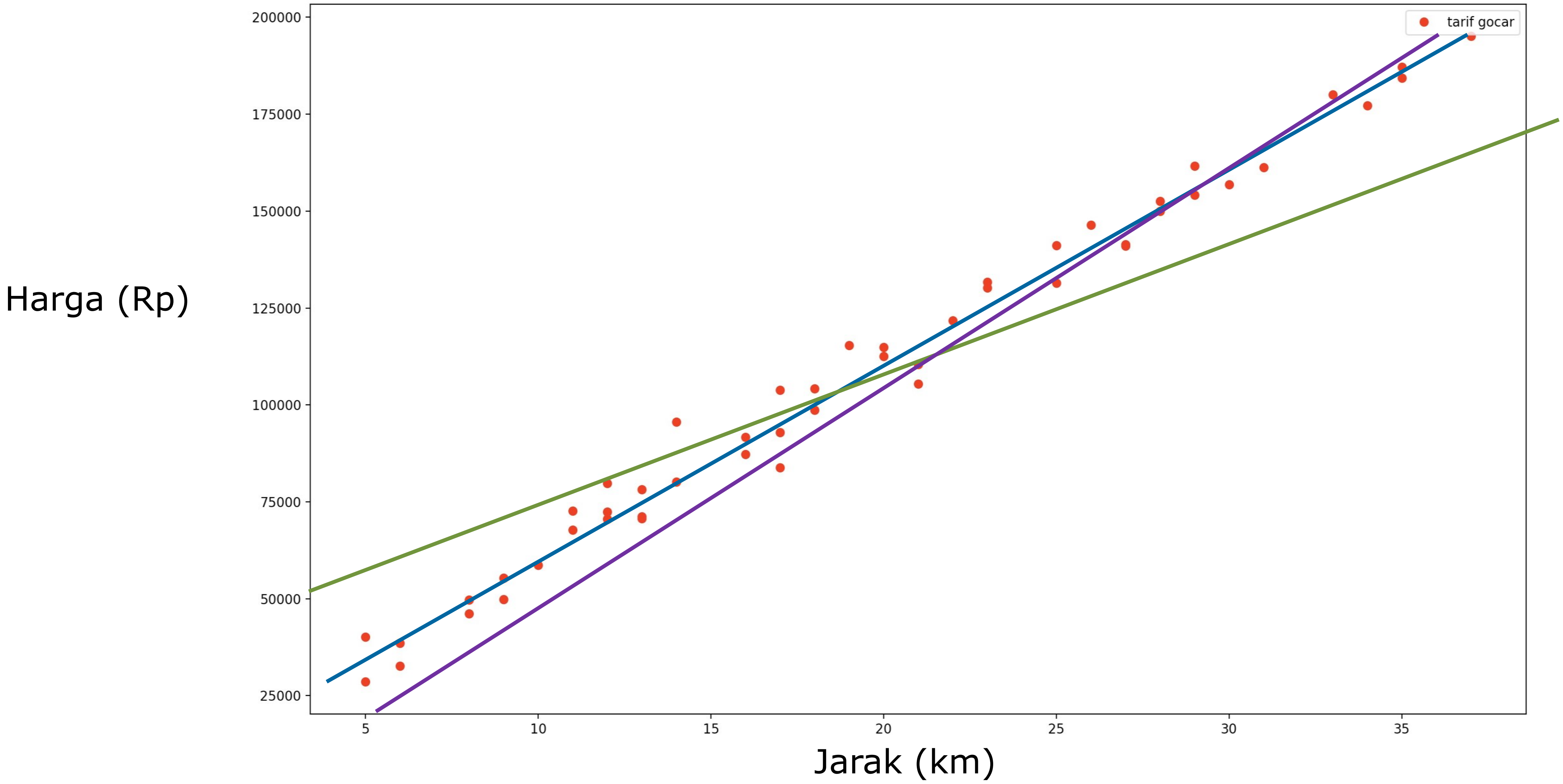
- X is one-dimensional case (d = 1)
  - $X \in \mathbb{R}^1$, $\boldsymbol{y} \in \mathbb{R}^1$,
- If we have n data available,
  - $X = [n \times 1]$, $\boldsymbol{y} = [n \times 1]$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Then the regression for each x_i data point:

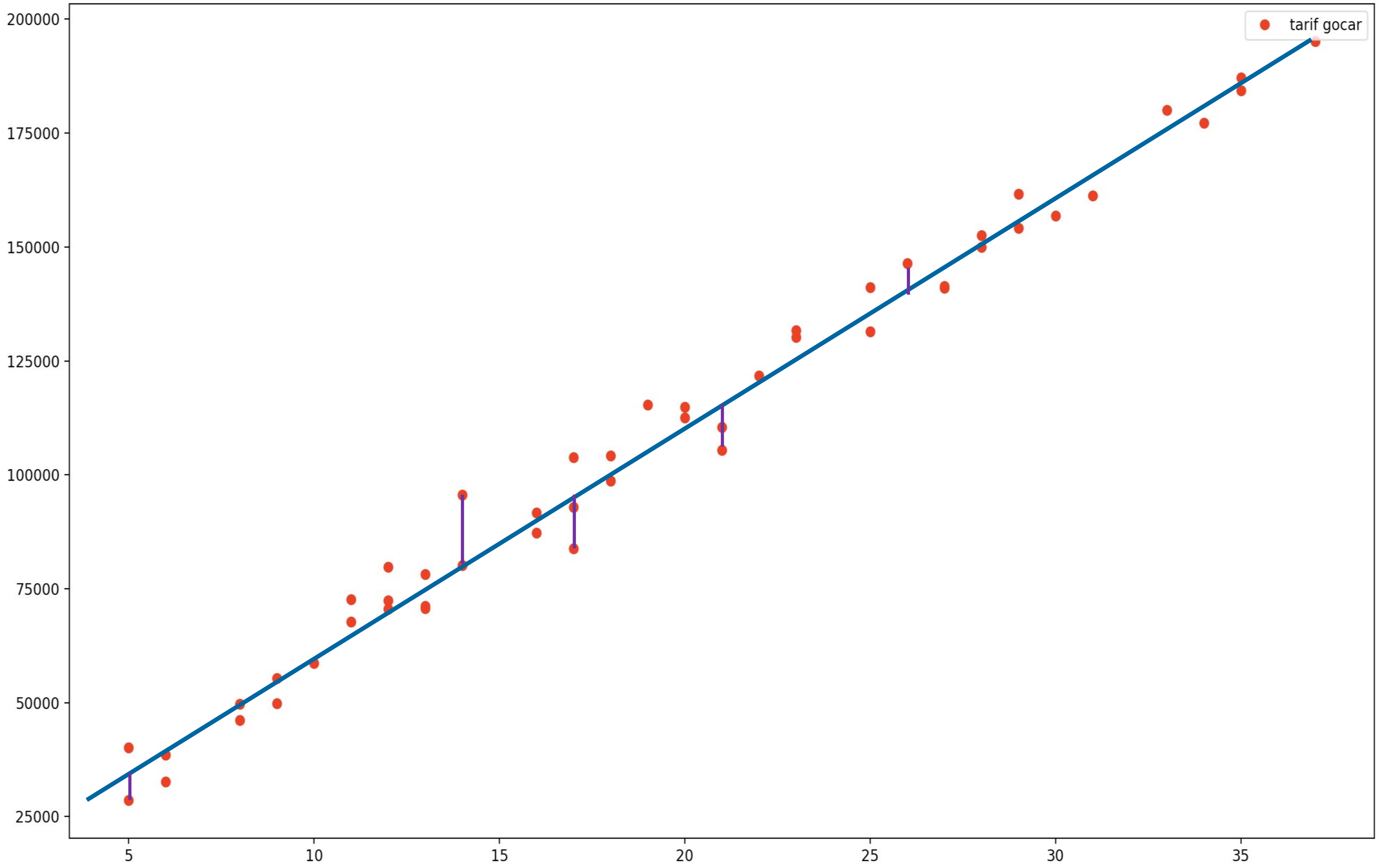
$$f(x) = w\,x + a \qquad\qquad w, a \in \mathbb{R}\;(scalar)$$

Performance Measure : How to choose The Best line ?



Error model-1

Harga (Rp)

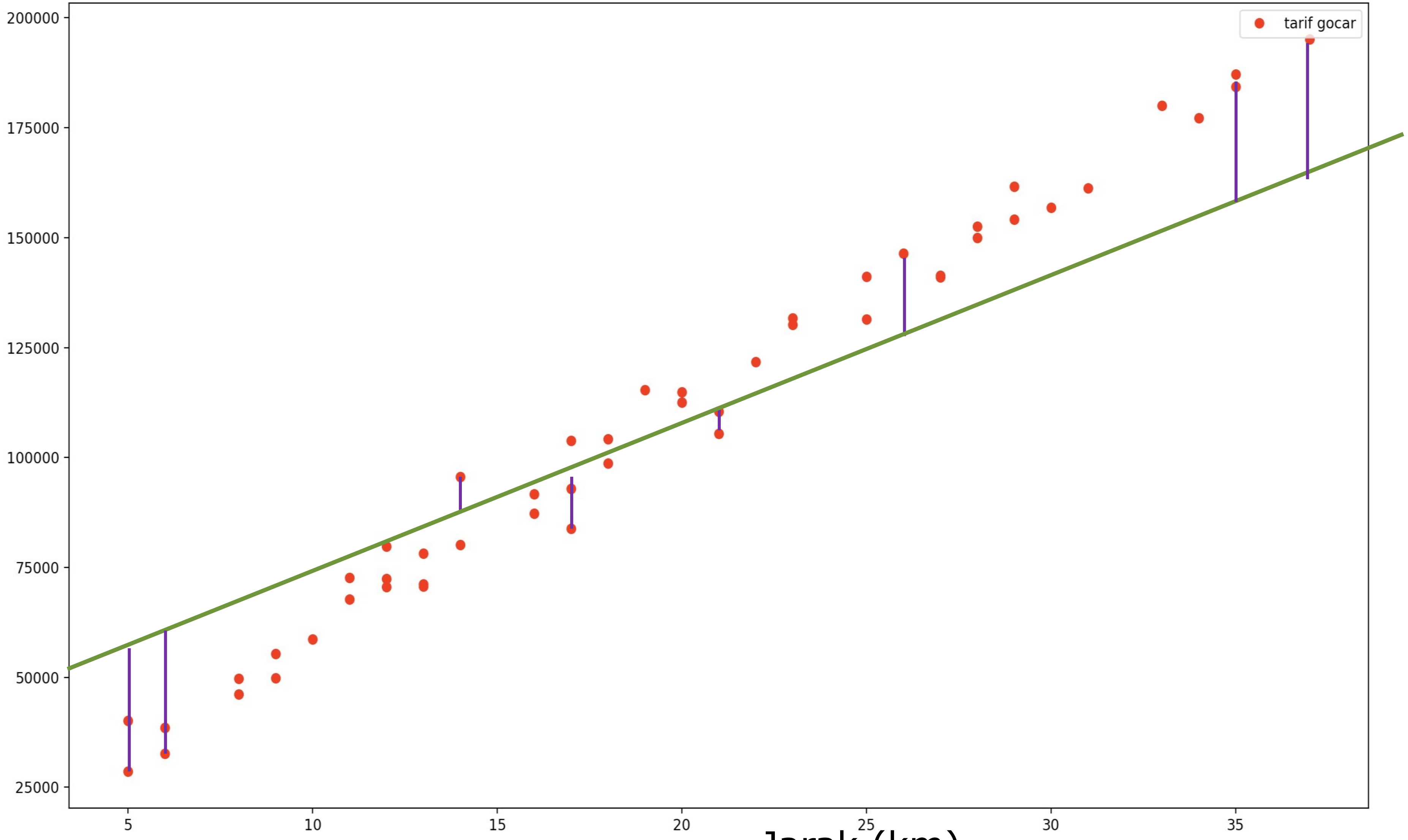


Jarak (km)

Error model-2

Harga (Rp)

Which one
is better??



Jarak (km)

Concept

Good Model ??

- Smallest error
- Which error?

Error/Residual

- **$y = 8800 + 5040x$,** y adalah tarif **prediksi**
- Karena kita punya data Jarak dan Tarif (**sebenarnya**), yang juga biasa dinotasikan dengan x dan y, maka ada 2 y.
- Bedakan
- $\hat{y}$: Nilai Prediksi (*predicted*)
- y : Nilai Sebenarnya (*observed*)
- Maka Error untuk data pertama adalah :
- $\boldsymbol{\varepsilon_1 = y_1 - \hat{y}_1}$

Error untuk semua ?

- Jarak/Gap antara *observed* dengan *predicted*
- $\boldsymbol{\varepsilon_1 = y_1 - \hat{y}_1}$, jika $\mathbf{y_1} > \mathbf{\hat{y}_1}$ positive, jika $\mathbf{\hat{y}_1} > \mathbf{y_1}$ negative
- Error negative ? + Error Positive ? = 0 Error ? **Something is wrong**
- $\boldsymbol{\varepsilon_1 = (y_1 - \hat{y}_1)^2}$
- **Total error/Sum Square Error (SSE)=** $\boldsymbol{\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n}$
- **Performance Measure = SSE, semakin kecil semakin baik**

HOW??

This week: Least Square Method!

Objective

-

- **(SSE)=** $\boldsymbol{\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n}$

  – Find w and a that minimize this function

$$\boxed{E(w,a) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

Intermezzo:

ARE you familiar with:

- Minimization problem??
- Derivative??
- Find critical point of a function??

Least Square Error

- Minimization ➔ derivative

$$\boxed{E(w,a) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- Solve the minimization problem by setting the partial derivatives to zero (2)
  - Denote the solution by $(\hat{w}, \hat{a})$

$$\boxed{E(w,a) = \sum_{i=1}^{n} (y_i - wx_i - a)^2} \qquad (2b)$$

Least Square Error

- From the first derivative of (2b) w.r.t a, We have

$$\boxed{\frac{\partial E(w,a)}{\partial a} = -2\sum_{i=1}^{n}(y_i - wx_i - a)} \quad (3)$$

- and setting (3) to zero gives
  - And $\bar{y} = \frac{1}{\mathrm{n}}\Sigma_{\mathrm{i}}\mathrm{y}_{\mathrm{i}}$ and $\bar{x} = \frac{1}{n}\Sigma_i x_i$

$$\boxed{\hat{a} = \bar{y} - w\bar{x}} \quad (4)$$

Detail:

$0 = \ -2\sum_{i=1}^{n}(y_i - wx_i - a)$

$0 = \sum_{i=1}^{n}(y_i) - w\sum_{i=1}^{n}(x_i) - an$

$0 = \frac{1}{\mathrm{n}}\Sigma_{\mathrm{i}}\mathrm{y}_{\mathrm{i}} - w\frac{1}{n}\Sigma_i x_i - a$

Least Square Error

- From the first derivative of (2b) w.r.t w, We have

- Plugging in $a = \hat{a}$ in (5) and setting the derivative to zero gives us

$$\boxed{\frac{\partial E(w,a)}{\partial w} = -2\sum_{i=1}^{n} x_i(y_i - wx_i - a)} \quad (5)$$

$$\boxed{0 = \sum_{i=1}^{n} x_i(y_i - wx_i - \bar{y} + w\bar{x})} \quad (5b)$$

Least Square Error

- For which we can solve (5b) to

$$\boxed{\hat{w} = \frac{\sum_{i=1}^{N} x_i(y_i - \bar{y})}{\sum_{i=1}^{N} x_i(x_i - \bar{x})}} \qquad (6)$$

Detail:

$$0 = \sum_{i=1}^{n} x_i(y_i - wx_i - \bar{y} + w\bar{x})$$

$$0 = \sum_{i=1}^{n} x_i(y_i - \bar{y} + w\bar{x} - wx_i)$$

$$0 = \sum_{i=1}^{n} x_i\big(y_i - \bar{y} + w(x_i - \bar{x})\big)$$

$$0 = \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \sum_{i=1}^{n} wx_i(x_i - \bar{x})$$

Least Square Error

- Challenge : ubah (6) jadi (7)

$$\boxed{\hat{w} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \qquad (7)$$

Aside: Variance and Covariance

- Since

$$\boxed{cov(X,Y) = \sum_{i}^{N} \frac{(x_i - \bar{X})(y_i - \bar{Y})}{N-1}}$$

- Notice that from (8), we have

(0b)

$$\boxed{\hat{w} = \frac{cov(X,Y)}{cov(X,X)}} \qquad (9)$$