Lecture 2

Simple and Multivariable Regression

Review of Last Lecture

- In the last lecture, we learnt how we could observe and approximate any real-world behaviour e.g. using stochastic models.

- Future prediction using **Markov Chain**.

Today’s Agenda

- Today we will talk about another fundamental AI/ML algorithm:
  - **Linear Regression**
    - Supervised ML algorithm: **A basic building block of deep learning and advanced AI/ML**
    - Linear predictive model

Linear Regression - An example

Problem: Predict the house prices in Boston, MA

Dataset: Boston Housing Dataset

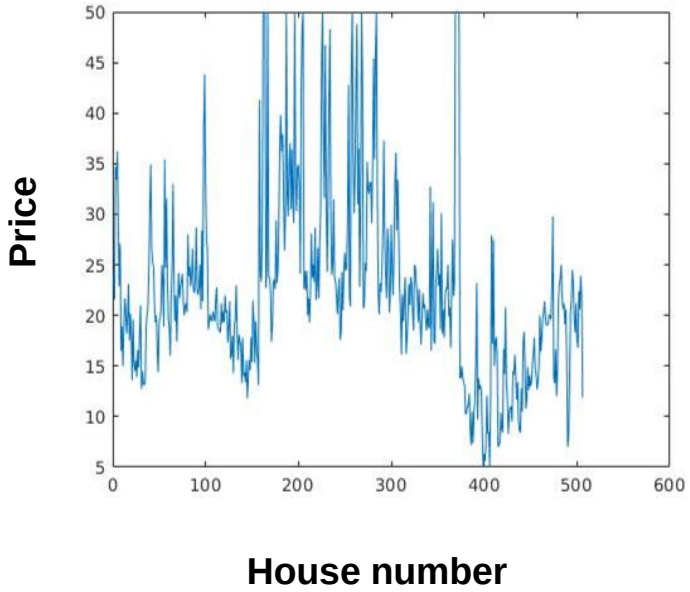
“This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The dataset is small in size with only 506 cases.”

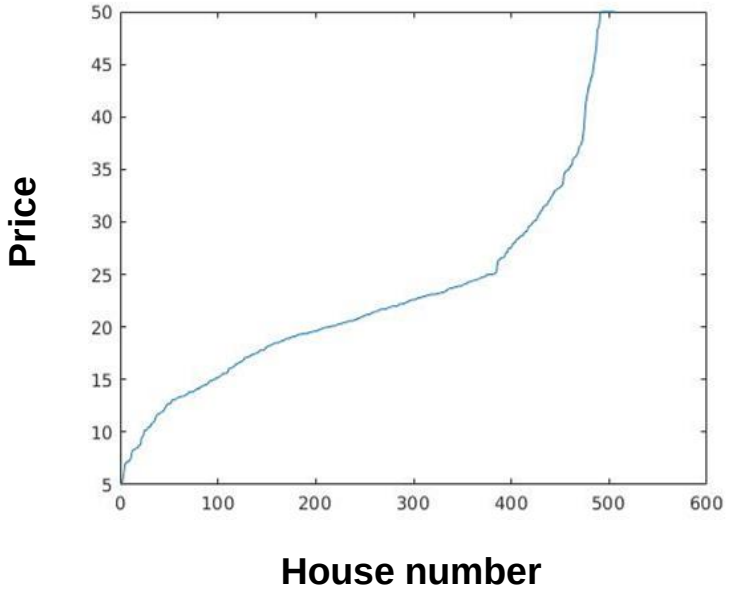
- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per $10,000
- PTRATIO pupil-teacher ratio by town
- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in $1000’s

D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Linear Regression - An example

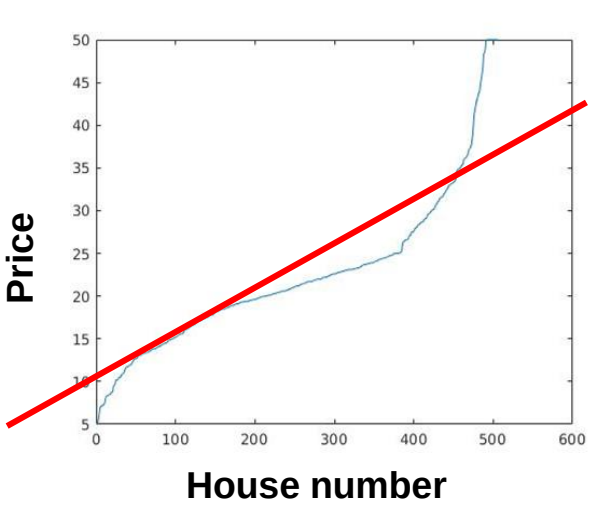
- House prices

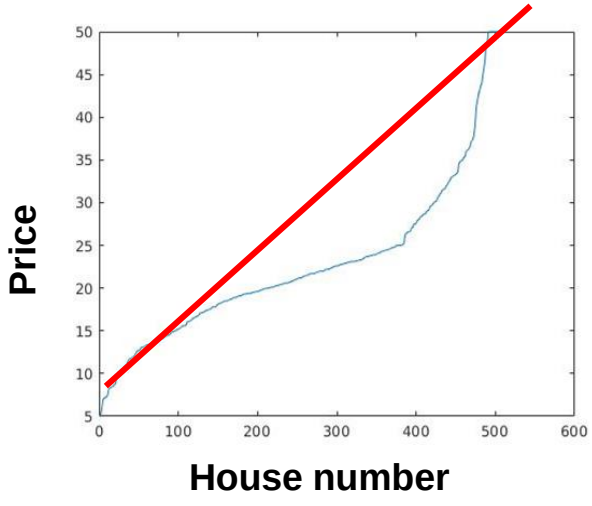


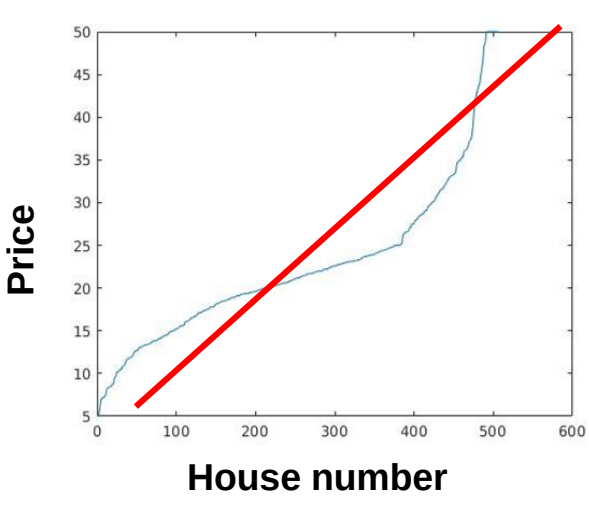


Linear Regression - An example

- How can we solve it?
- Maybe by drawing an optimal line that best fits the dataset?

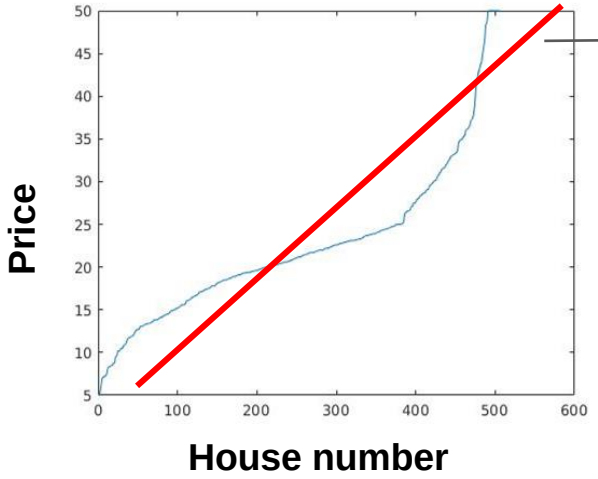






Linear Regression - An example

- What's the objective?



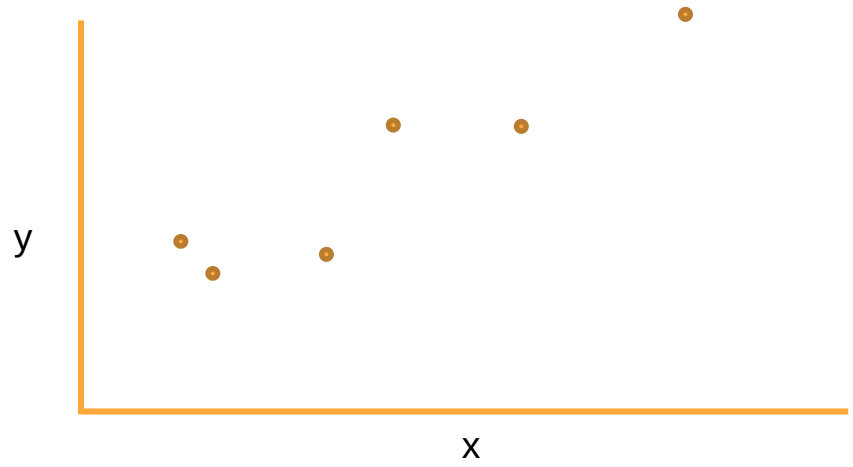
$\hat{Y} = mx + c$

- $\hat{Y}$: Approximated value
- m: Slope
- x: Input
- c: y-intercept

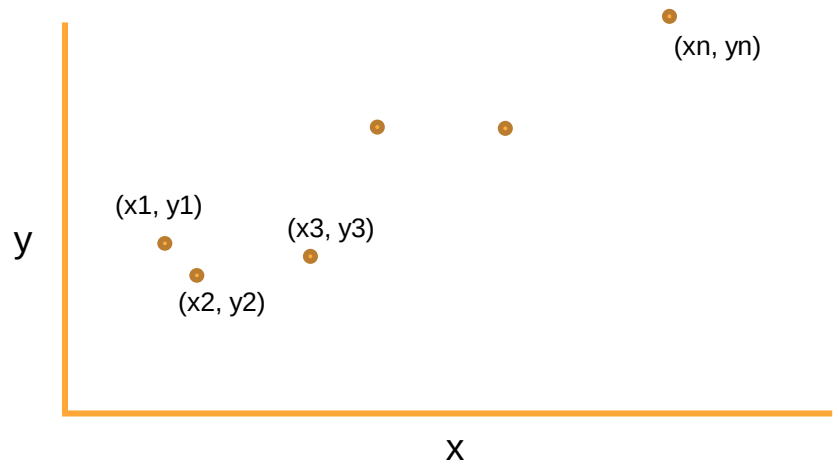
- Approximation with a **least squared error (LSE)!**

Simple Linear Regression: Closed-Form Equation (Analytical Solution)

- Let’s first try to find out a closed-form solution to the simple regression problem given below:

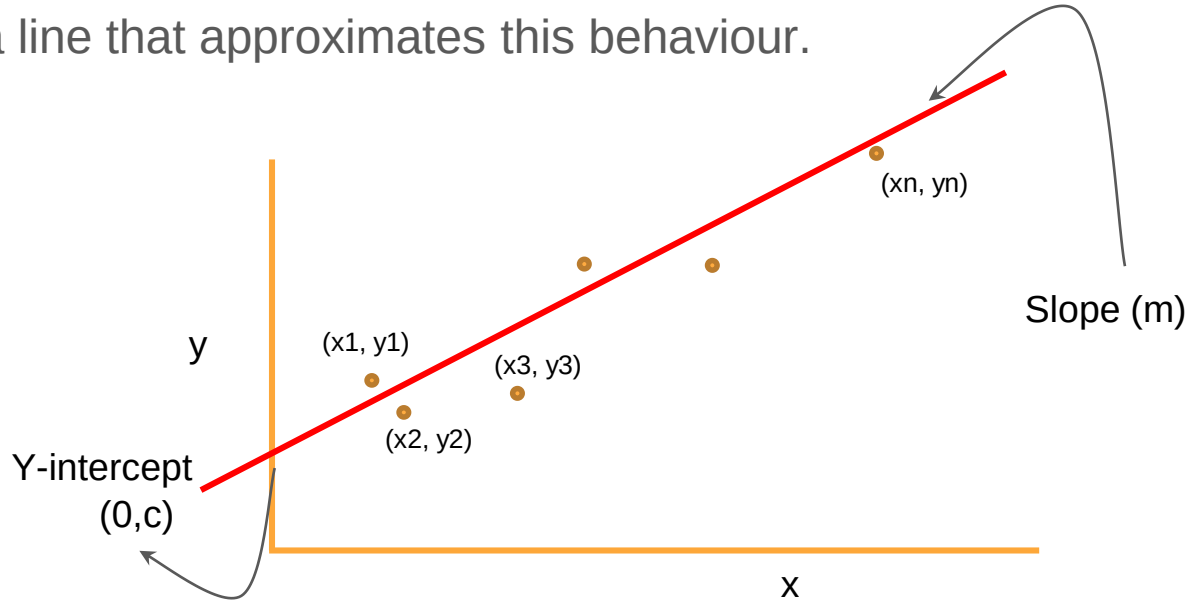


Simple Linear Regression: Closed-Form Equation (Analytical Solution)



Simple Linear Regression: Closed-Form Equation (Analytical Solution)

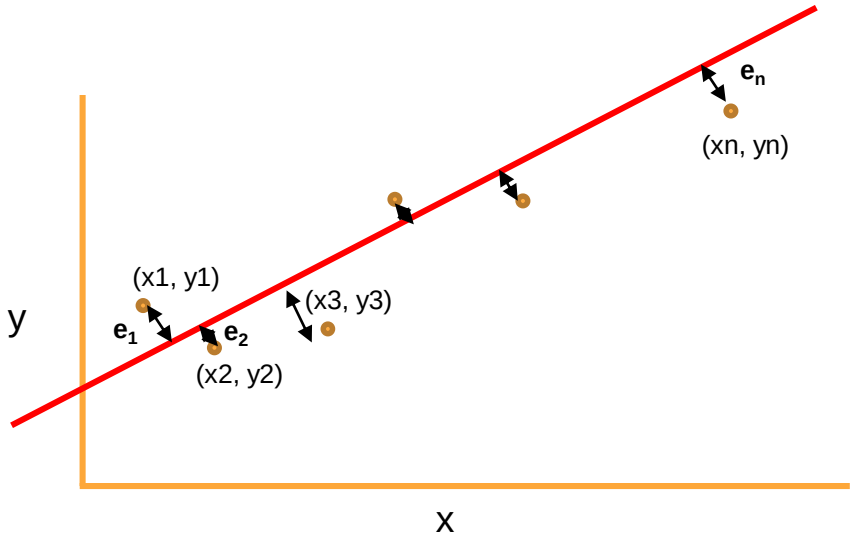
- Draw a line that approximates this behaviour.



Simple Linear Regression: Closed-Form Equation (Analytical Solution)

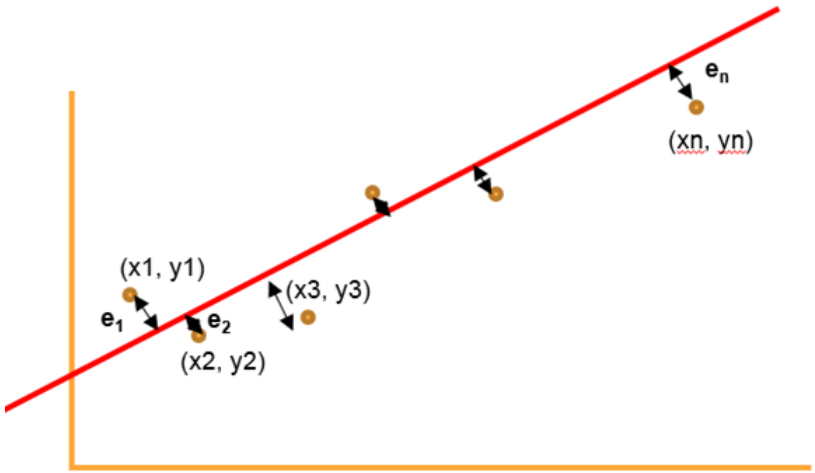
- What's the error between the line and each data point:

e_1, e_2, …., e_n

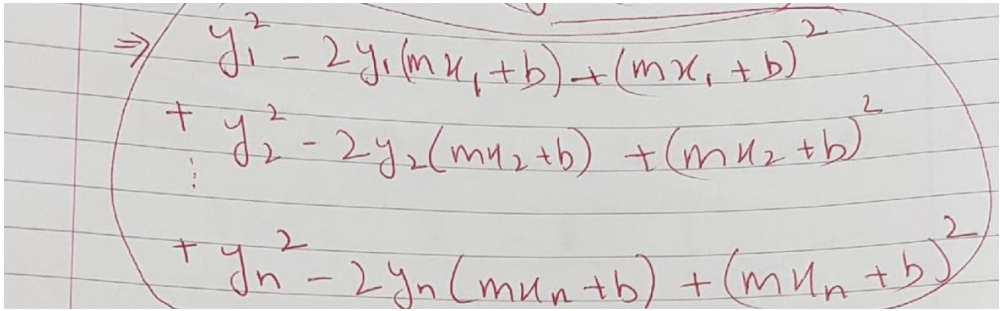


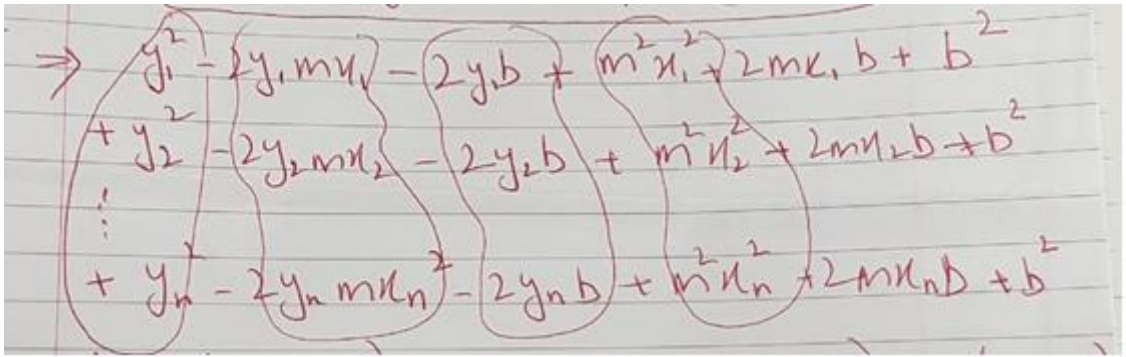
Simple Linear Regression: Closed-Form Equation (Analytical Solution)

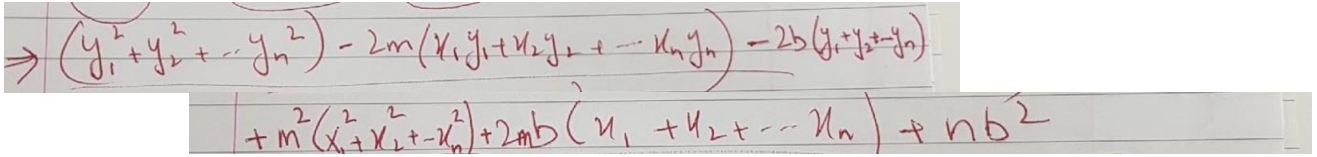
- $SE = e_1^2 + e_2^2 + e_3^2 +, \ldots\ldots, + e_n^2$
- $SE = (y_1 - (mx_1+b))^2 + (y_2 - (mx_2+b))^2 + \ldots\ldots, + (y_n - (mx_n+b))^2$

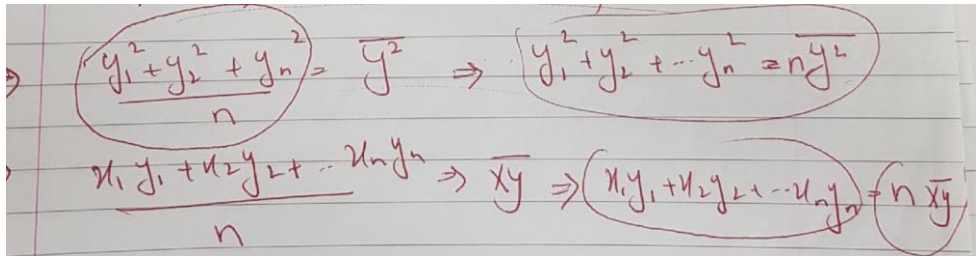


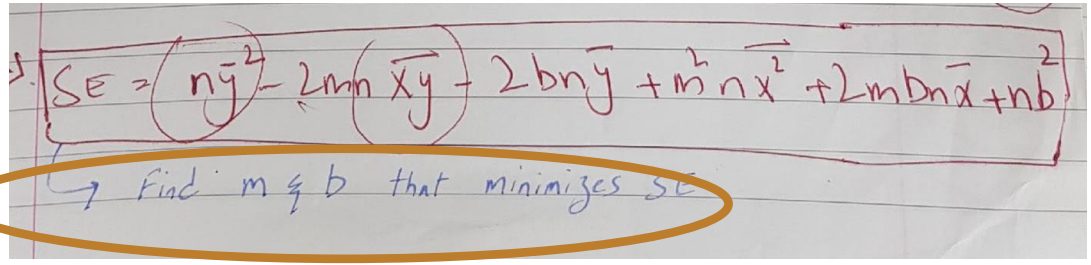
- $SE = e_1^2 + e_2^2 + e_3^2 +, \ldots\ldots, + e_n^2$
- $SE = (y_1 - (mx_1+b))^2 + (y_2 - (mx_2+b))^2 + \ldots\ldots, + (y_n - (mx_n+b))^2$
- **Find m and b that minimizes the SE!**





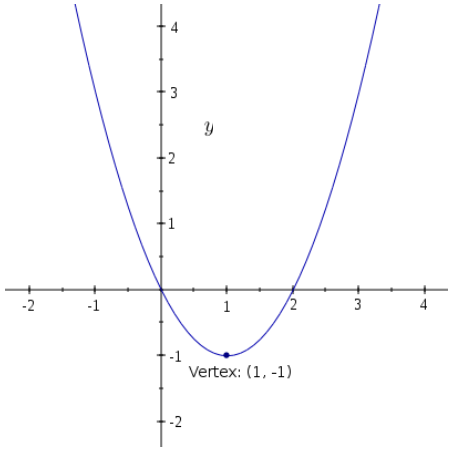

$\Rightarrow (y_1^2 + y_2^2 + \cdots y_n^2) - 2m(x_1y_1 + x_2y_2 + \cdots x_ny_n) - 2b(y_1 + y_2 + \cdots y_n)$
$+ m^2(x_1^2 + x_2^2 + \cdots x_n^2) + 2mb(x_1 + x_2 + \cdots x_n) + nb^2$


$\frac{y_1^2 + y_2^2 + y_n^2}{n} = \overline{y^2} \Rightarrow y_1^2 + y_2^2 + \cdots y_n^2 = n\overline{y^2}$
$\frac{x_1y_1 + x_2y_2 + \cdots x_ny_n}{n} \Rightarrow \overline{xy} \Rightarrow x_1y_1 + x_2y_2 + \cdots x_ny_n = n\overline{xy}$


$SE = n\overline{y^2} - 2mn\overline{xy} + 2bn\overline{y} + m^2n\overline{x^2} + 2mbn\overline{x} + nb^2$
↳ Find m & b that minimizes SE

How?

Assume this to be a function of SE: f(m) or f(b)

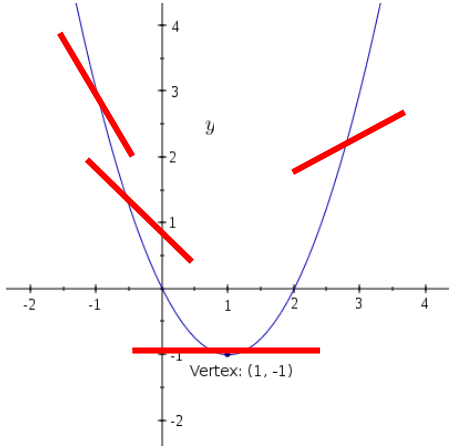


Where is the minima?

Minima?

When derivative is ‘0’

Or ‘0’ slope



- $\frac{d}{dm}(SE) = 0$ _____ (1)

- $\frac{d}{db}(SE) = 0$ _____ (2)

Linear Regression

Solving (1) and (2) gives:

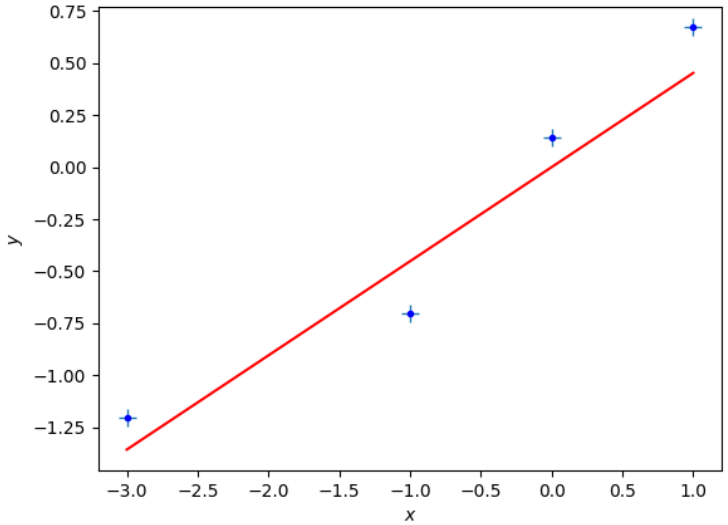
$$\mathbf{b} = \bar{Y} - m\bar{X}$$

$$m = \frac{\overline{Y}\,\overline{X} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}}$$

$$\mathbf{m = Cov(X, Y) / Cov\ (X, X)}$$

Linear Regression – WP task 1

1. Find and draw the equation for the regression line that best fits the following dataset: **{X, Y} = {(-3, -1.2),(-1, -0.7),(0, 0.14),(1, 0.67)}**



Linear Regression – WP task 2

2. Generalize the concept of (1) for more than 1 input variable e.g., 14 variables in case of Boston housing dataset.

- **2.1.** Write a multivariable regression equation
- **2.2.** Explain (2.1) in terms of
    i. Dependent and independent variables (discussed in lecture 1)
    ii. Multivariable and their weights (significance)
    iii. Slopes
    iv. Intercept

Linear Regression – WP task 3

3. Coding task: Prediction model for the house prices in Boston, MA

Hint:

- $m = \frac{\bar{Y}\bar{X} - \overline{XY}}{(\bar{X})^2 - \overline{X^2}} = \frac{X^T Y}{X^T X} = (X^T Y)^{-1} . X^T X$
- m is a vector
- Use np.linalg for calculating the inverse (if Python is the language of programming)

Also, explain which variable has the most weight (significance) and why?