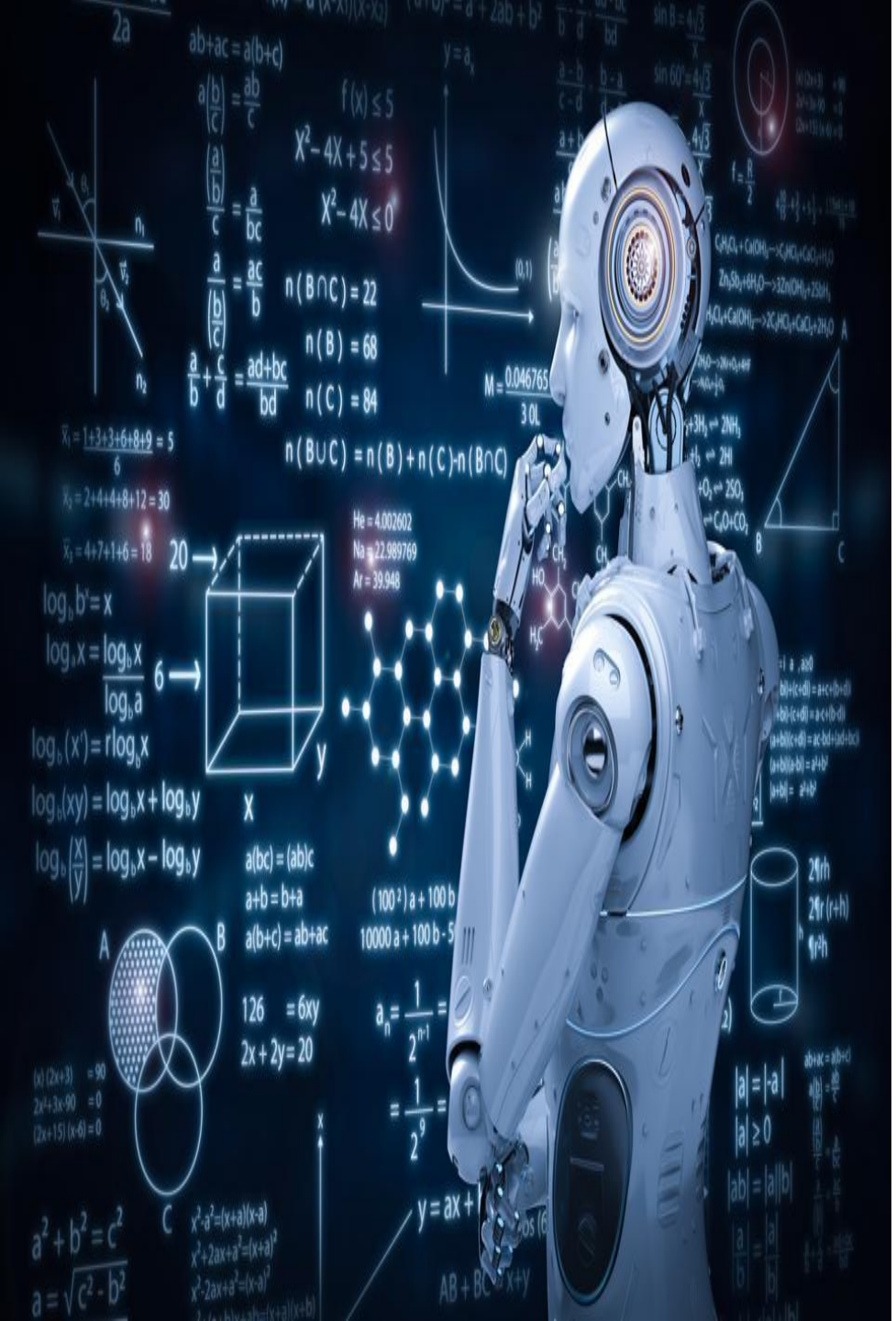


Practical Machine Learning

Day 4: Mar24 DBDA

Kiran Waghmare

Agenda

- Regression
- Types of Regression

Linear model

In regression, the relationship between Y and X is modelled in the following form:

$$Y = a + b * X + E$$

where:

- **Y** is the dependent variable (Income in the example)
- **X** is the independent variable (IQ in the example)
- **a** is an intercept
- **b** is the coefficient
- **E** is an error term for each observation (since there is additional variation not explained by income)

Linear model

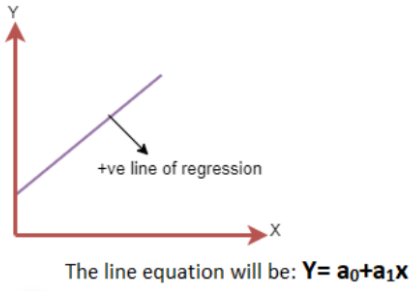
We are not interested in the intercept **a** but only in the coefficient **b**.

The coefficient **b** represents the relationship between X and Y.

- If **b is positive,** X has a positive effect on Y (as X increases, Y increases);
- If **b is negative**, X has a negative effect on Y (as X increases, Y decreases).

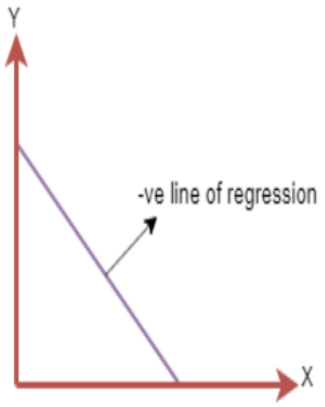
If **b = 0,** there is **no effect** of X on Y.

Linear Regression Line



- A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

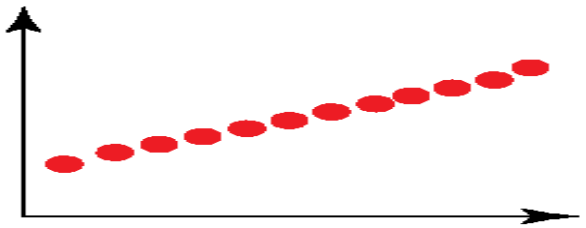


- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

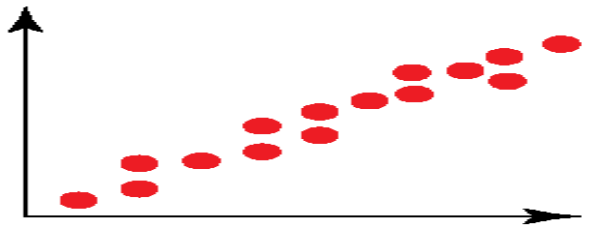
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Correlation coefficient (r)

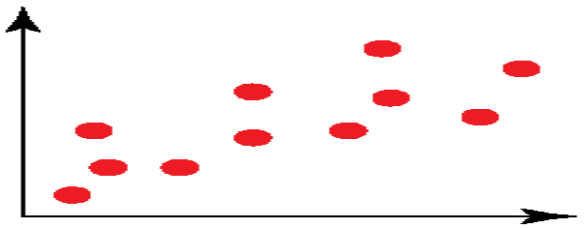
- Correlation coefficient (*r*) describes a linear relationship between `x` and `y` variables. r can range from -1 to 1.
- r > 0 indicates a positive linear relationship between `x` and `y` variables. As one of the variable increases, the other variable also increases. r = 1 is a perfect positive linear relationship
- Similarly, r < 0 indicates a negative linear relationship between `x` and `y` variables. As one of the variable increases, the other variable decreases, and vice versa. r = -1 is perfect negative linear relationship
- r = 0 indicates, there is no linear relationship between the `x` and `y` variables



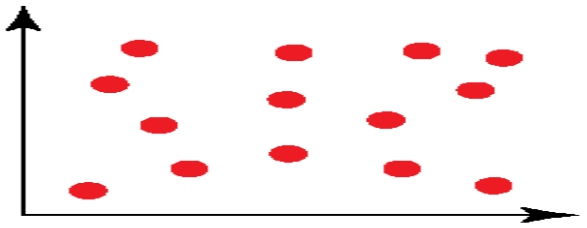
Perfect Positive Correlation



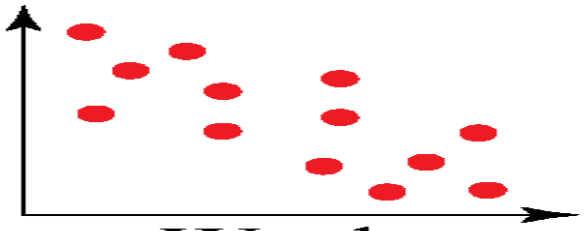
Strong Positive Correlation



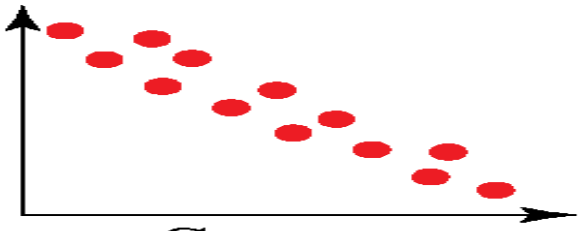
Weak Positive Correlation



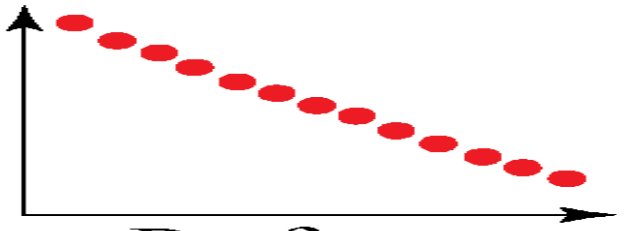
No Correlation



Weak Negative Correlation



Strong Negative Correlation



Perfect Negative Correlation

Coefficient of determination (R-Squared or r-Squared)

- R-Squared (R^2) is a square of correlation coefficient (r) and usually represented as percentages.
- R-Squared explains the variation in the `y` variable that is explained by independent variables in the fitted regression.
- Multiple correlation coefficient (R), which is the square root of the R-Squared, is used to assess the prediction quality of the `y` variable in multiple regression analysis. Its value range from 0 to 1.
- R-Squared can range from 0 to 1 (0 to 100%). R-squared = 1 (100%) indicates that the fitted regression line explains all the variability of Y variable around its mean.

Residuals (regression error)

- Residuals or error in regression represents the distance of the observed data points from the predicted regression line

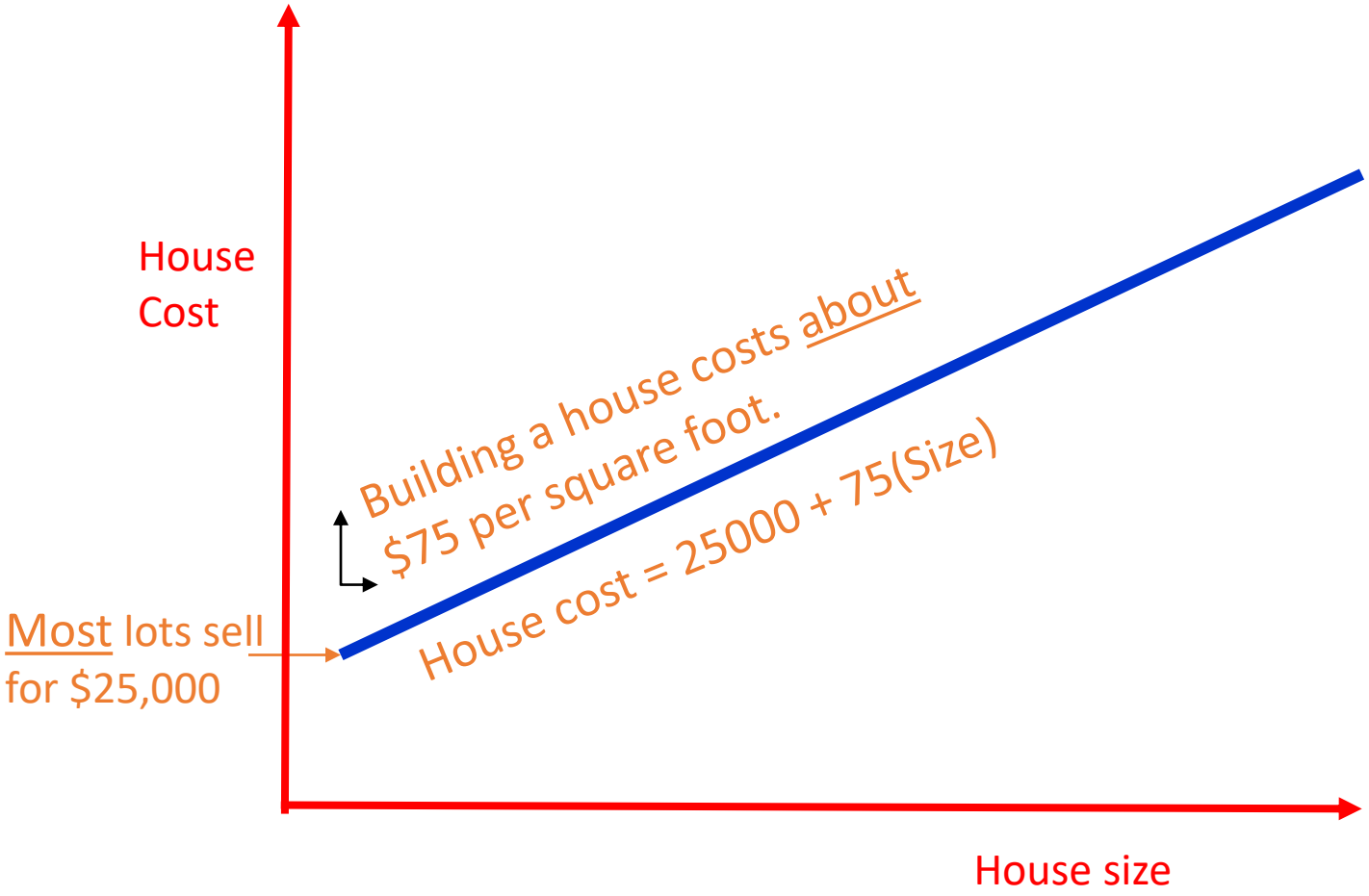
$$residuals = actual\ y(y_i) - predicted\ y\ (\hat{y}_i)$$

Root Mean Square Error (RMSE)

- RMSE represents the standard deviation of the residuals. It gives an estimate of the spread of observed data points across the predicted regression line.

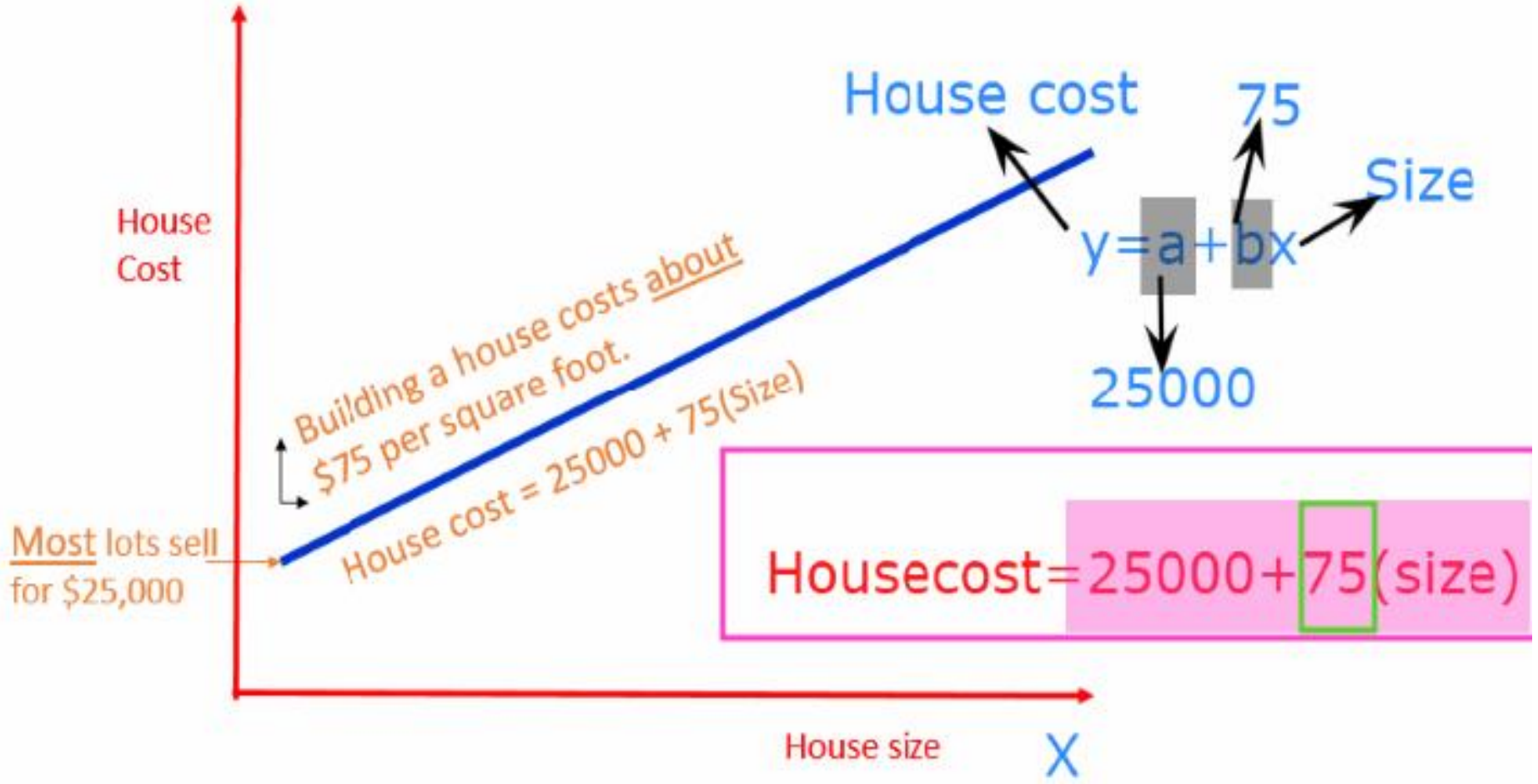
The Model

The model has a deterministic and a probabilistic components

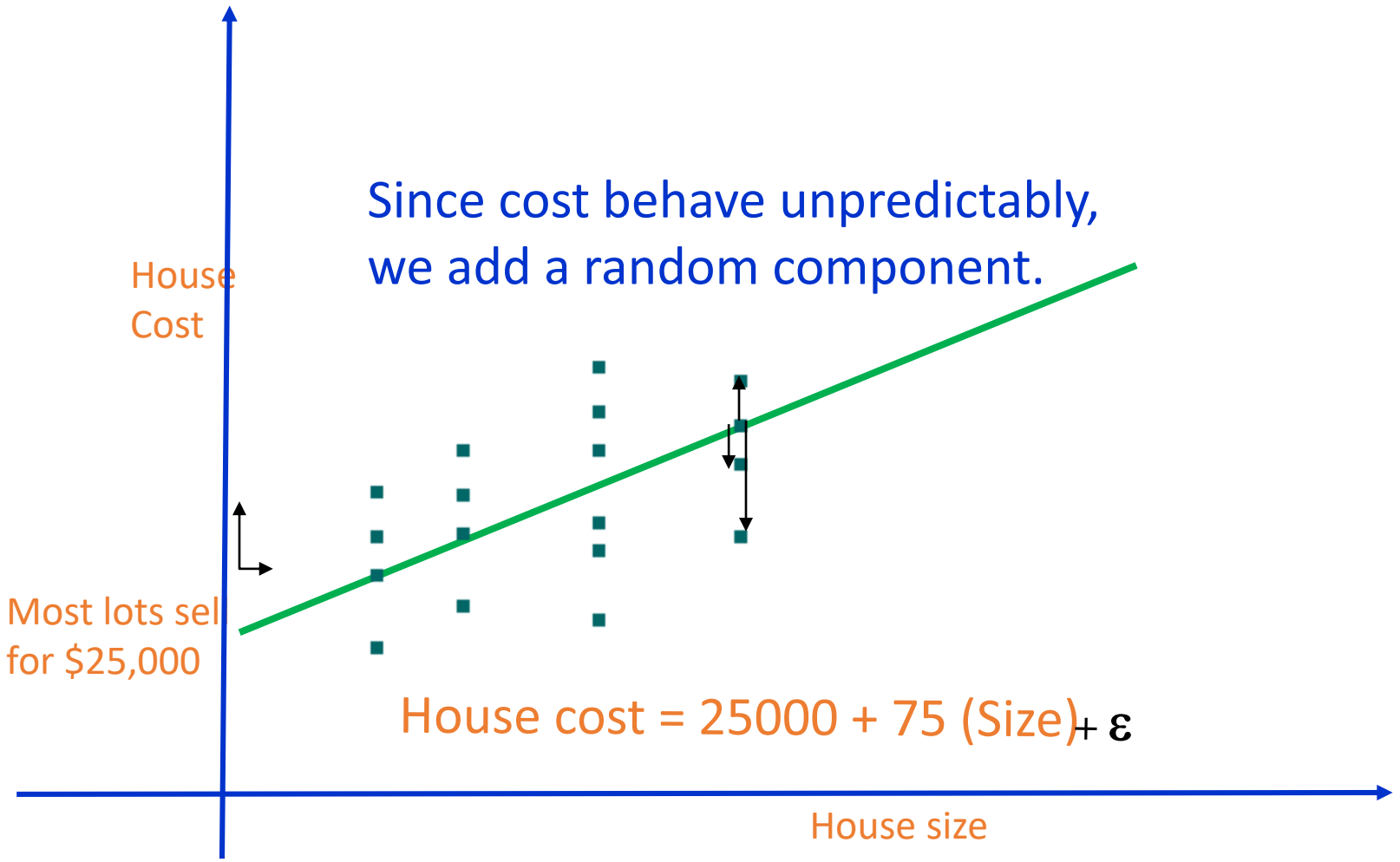


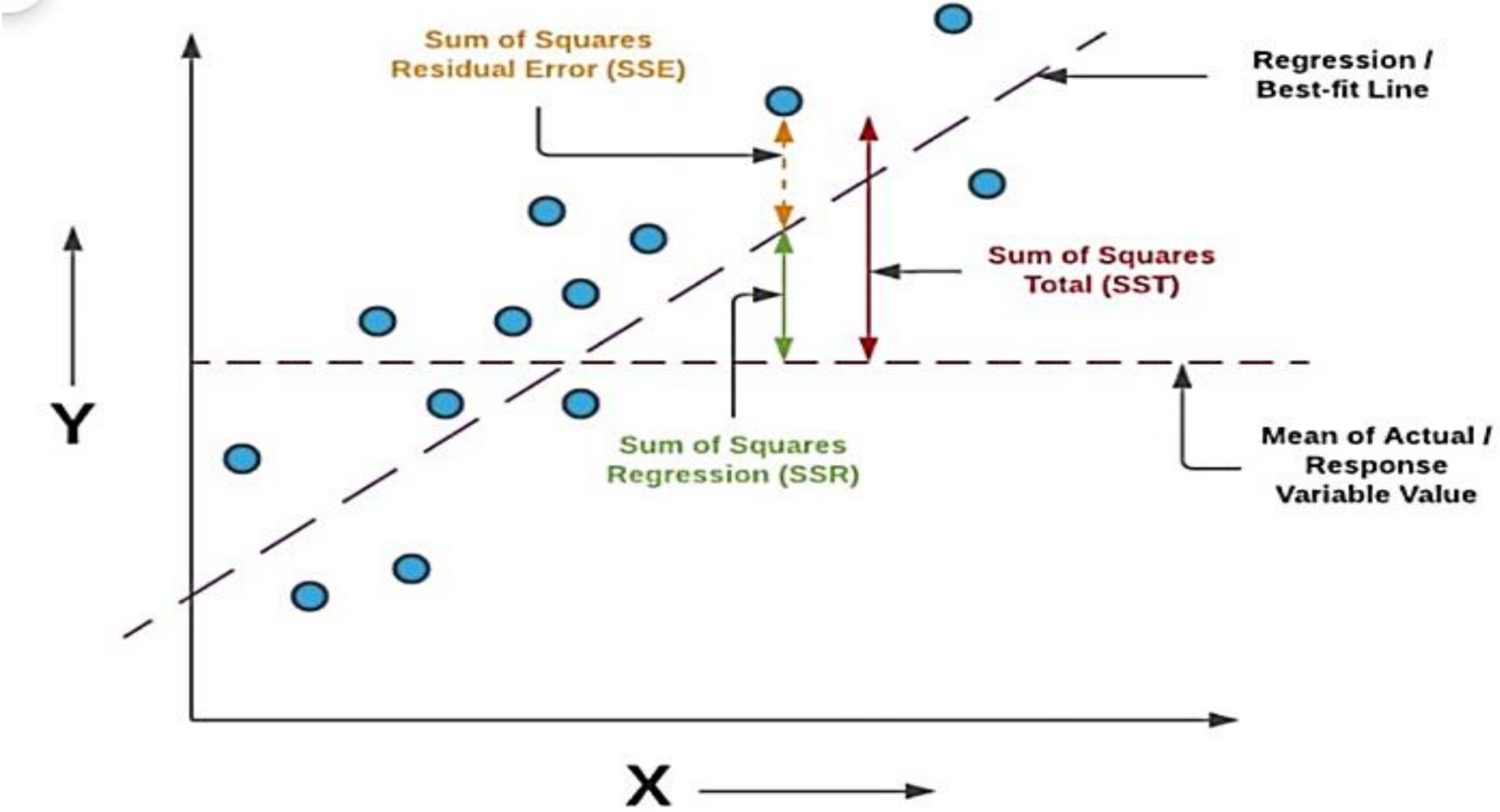
The Model

The model has a deterministic and a probabilistic components



However, house cost vary even among same size houses!




Sum of Squares
Residual Error (SSE)
Regression /
Best-fit Line
Sum of Squares
Total (SST)
Sum of Squares
Regression (SSR)
Mean of Actual /
Response
Variable Value
Y
X

Estimating the Coefficients

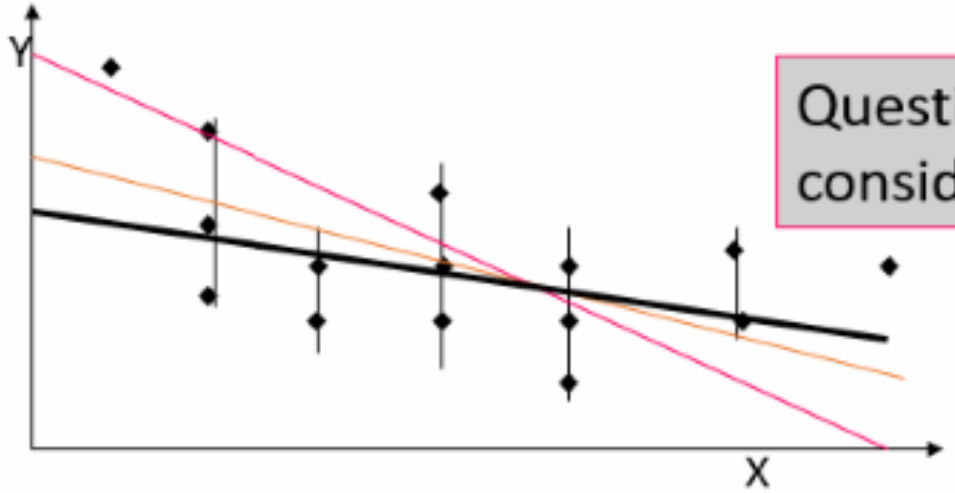
Cost function

minimum

Best fit line for Regression

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.

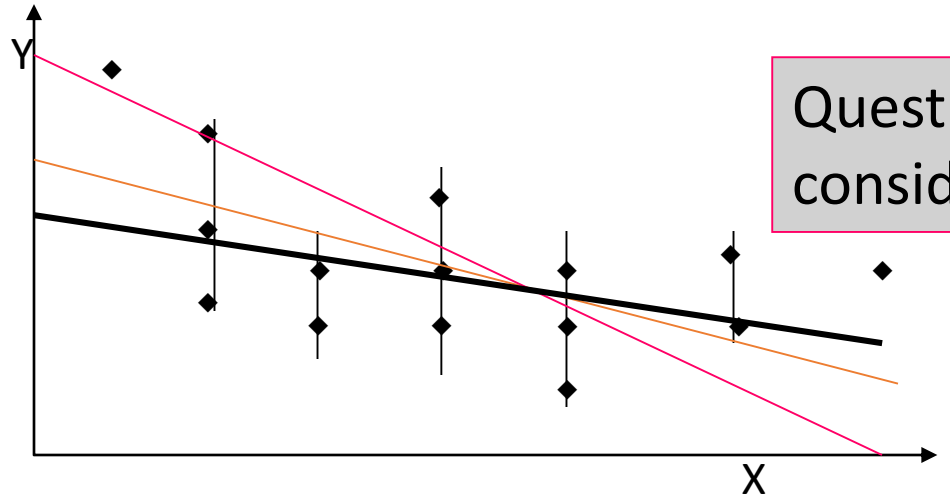
MSE: Mean Squared Error



Question: What should be considered a good line?

Estimating the Coefficients

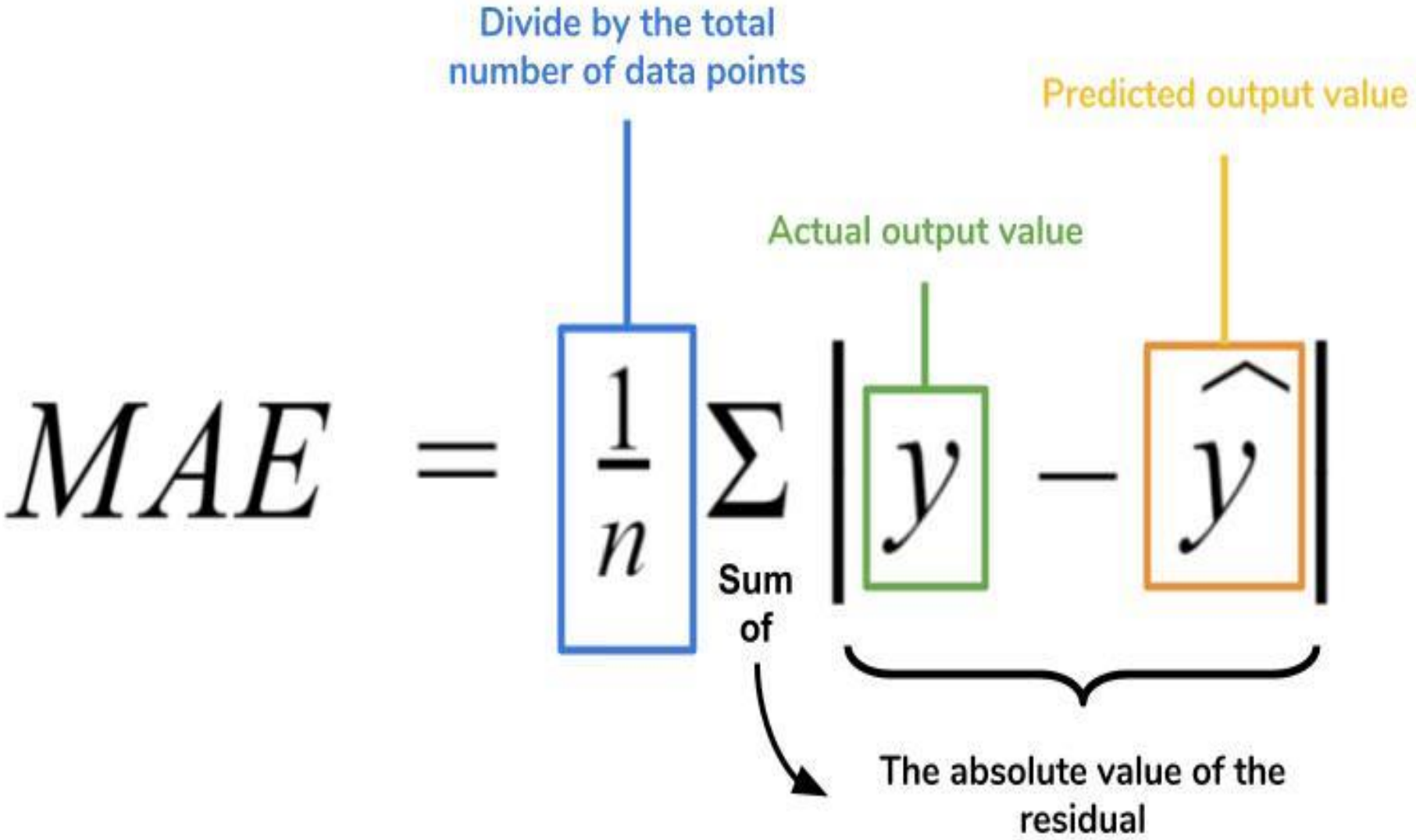
- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



Question: What should be considered a good line?

Finding the best fit line:

- **main goal** is **to find the best fit line** that means
  - the **error between predicted values and actual values** should be minimized.
  - The best fit line will have the least error.
- To find the **best fit line,** so to calculate this we use **cost function.**
- **Cost function-**
- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the **cost function is used to estimate the values of the coefficient for the best fit line.**
  - Cost **function optimizes the regression coefficients** or weights.
  - It measures how a linear regression model is **performing.**
- Use to find the accuracy of the **mapping function**,
  - which maps the input variable to the output variable.
  - This mapping function is also known as **Hypothesis function**.
- For Linear Regression, we use the **Mean Squared Error (MSE)** cost function,
  - which is the average of squared error occurred between the predicted values and actual values.


Divide by the total number of data points
Predicted output value
Actual output value
$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$
Sum of
The absolute value of the residual

Gradient Descent:

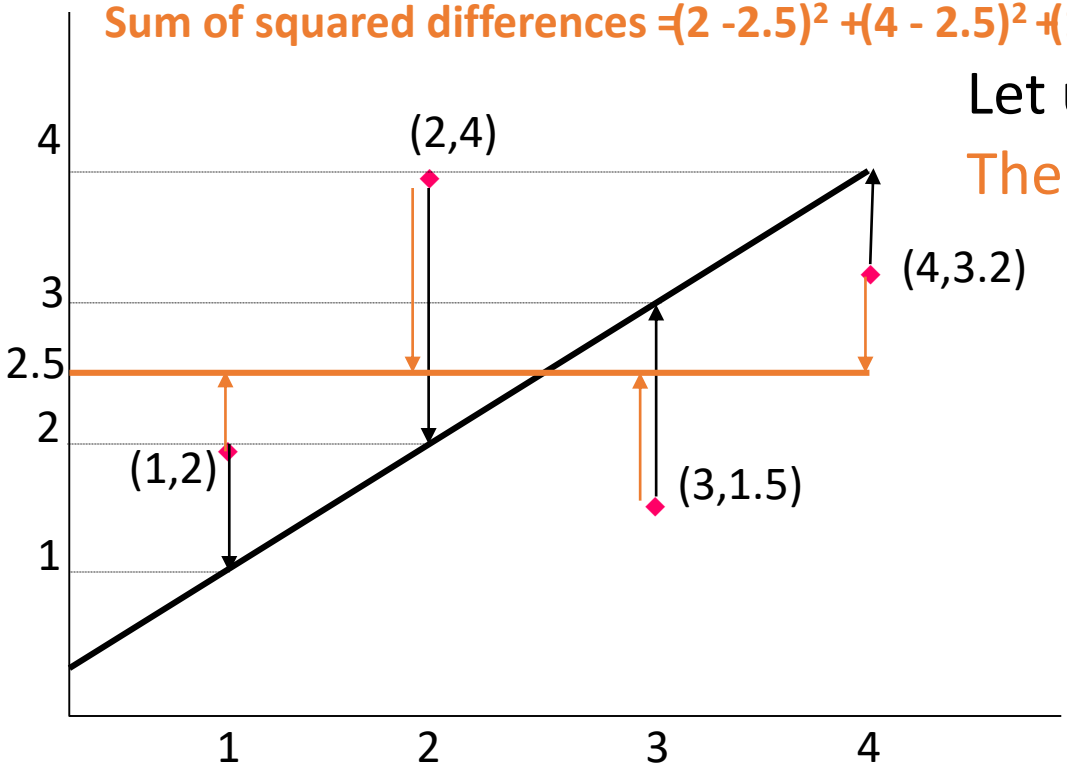
- Gradient descent is used **to minimize the MSE by calculating the gradient of the cost function.**
- A regression model uses gradient descent **to update the coefficients of the line** by reducing the cost function.
- It is **done by a random selection of values of coefficient** and then iteratively update the values to reach the minimum cost function.
- **Model Performance:**
- The Goodness of fit determines how the line of regression fits the set of observations.
- The process of **finding the best model out of various models** is called **optimization**.

The Least Squares (Regression) Line

A good line is one that **minimizes the sum of squared** differences between the points and the line.

Sum of squared differences $= (2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences $= (2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$



Let us compare two lines

The second line is horizontal

The smaller the sum of squared differences the better the fit of the line to the data.

Types of Linear Regression (LR)?

- Univariate LR: Linear relationships between `y` and `x` variables can be explained by a single `x` variable

$$y = a + bX + \epsilon$$

Where, a = y-intercept, b = slope of the regression line (unbiased estimate) and ϵ = error term (residuals)

- Multiple LR: Linear relationships between `y` and `x` variables can be explained by multiple `x` variables

$$y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n + \epsilon$$

Where, a = y-intercept, b = slope of the regression line (unbiased estimate) and ϵ = error term (residuals)

- The y-intercept (a) is a constant and slope (b) of the regression line is a regression coefficient.

Types of Linear Regression

- Linear regression can be further divided into two types of the algorithm:
- **Simple Linear Regression:**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.