1. Linear Regression

Linear regression is a statistical method used for modeling the relationship between **a dependent variable** (often denoted as ***y***) and o**ne or more independent variables** (often denoted as ***x***). It assumes a linear relationship between the independent variables and the dependent variable.

The formula for linear regression can be written as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Here:

- y is the dependent variable.
- x is the independent variable.
- β_0 is the y-intercept (value of y when x = 0).
- β_1 is the slope of the regression line (the change in y for a one-unit change in x).
- e is the error term (the difference between the observed value and the predicted value).

The goal of linear regression is to estimate the coefficients *β*0 and *β*1 that minimize the sum of the squared differences between the observed and predicted values of *y*.

L2 loss

The L2 loss, also known as the squared error loss or **mean squared error** (MSE)

$$L_{\mathrm{L2}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

In the context of linear regression, the goal is to minimize the L2 loss function with respect to the model parameters (coefficients) to find the best-fitting line that minimizes the squared differences between the predicted and actual values.

We use the L2 loss in linear regression because it encourages the **residuals** (the differences between observed and predicted values) **to be distributed around zero,** which aligns with the assumptions of linear regression.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) linear regression is a variant of linear regression where instead of using the entire dataset to update the model parameters in each iteration, it uses a single random data point (or a small subset of data points) to update the parameters.

In each iteration:

- Randomly select a data point or a mini-batch of data points from the dataset.
- Compute the gradient of the loss function with respect to the selected data point(s).
- Update the model parameters using the gradient descent algorithm. This involves adjusting the parameters in the opposite direction of the gradient to minimize the loss function.

Mini-batch gradient descent computes gradients and updates parameters based on a small random subset of the data.

Why Linear regression:

- explainable
- cheap to compute