Supervised Learning

Regression Models

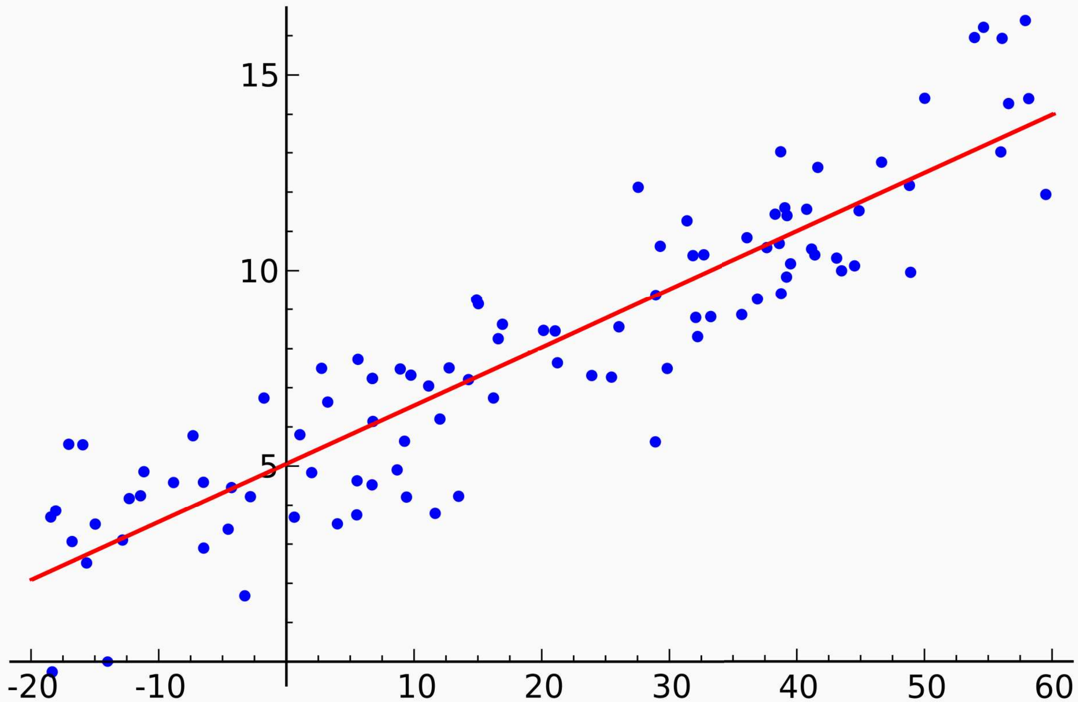
Diego Campos Sobrino

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

Table of contents

In regression problems, we take input variables and try to fit the output onto a continuous expected result function.

$$\boldsymbol{x} \in \mathbb{R}^n \qquad y \in \mathbb{R} \qquad f : \mathbb{R}^n \to \mathbb{R}$$



A linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the *bias* or *intercept*.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

where:

- $\hat{y}$ is the predicted value.
- n is the number of features.
- x_i is the i^{th} feature value.
- θ_j is the j^{th} model parameter (weights).

Linear Regression Vectorized form

In Machine Learning we usually use the vectorized form of the equation

$$\hat{y} = h_\theta(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$$

where:

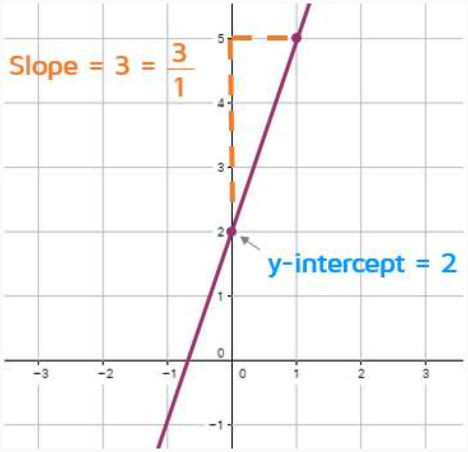
- $\boldsymbol{\theta}$ is the parameter vector containing θ_0 to θ_n
- $\boldsymbol{x}$ is the feature vector containing x_0 to x_n with $x_0 = 1$
- h_θ is the hypothesis function using model parameters $\boldsymbol{\theta}$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{\theta}^T \boldsymbol{x} = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \dots & \theta_n \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\boldsymbol{\theta}^T \boldsymbol{x} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

- Predict a single output value y from a single input value x.
- The input can be seen as the cause and the output the effect.
- We are looking for a function called h_θ that tries to map the input data (the x's) to the output data (the y's).

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$$



- This is like the equation of a straight line with θ_0 as the intercept (bias) and θ_1 as the slope.
- The values of θ_0 and θ_1 define the specific line used to make predictions.
- We give to $h_\theta(x)$ the value of x and receive and estimated output $\hat{y}$.

Suppose we have the following set of training data:

x	y
0	4
1	7
2	7
3	8

We can make a random guess about our h_θ function with $\theta_0 = 2$ and $\theta_1 = 2$, then the hypothesis function becomes $h_\theta(x) = 2 + 2x$.

- What would be the predictions $\hat{y}$'s for our x's with this model?
- How far off are our predictions $\hat{y}$'s from the actual y's?
- Can we do better?

We can try different combinations of θ_0 and θ_1 to find the straight line that "fit" better our data points.

- Measures the accuracy of our hypothesis function.
- Average-ish of all the results of the hypothesis $\hat{y}$'s compared to actual y's.

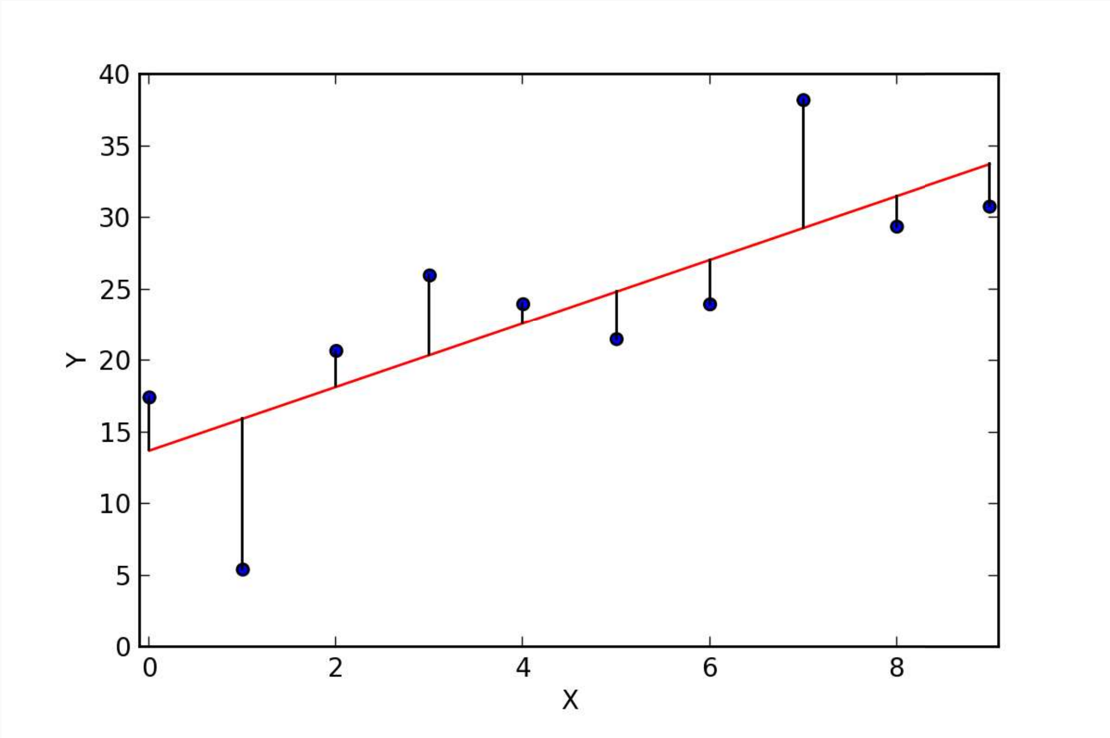
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

  - θ_0 and θ_1 are the parameters of the model.
  - m is the number of examples (observations) in the data.
  - $\hat{y}_i$ is the predicted value of the i-th example.
  - y_i is the actual value of the i-th example.

- Mean Squared Error (MSE) halved as convenience for computation (more on that later).
- Since $\hat{y}_i = h_\theta(x_i)$ sometimes is written as:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

Ordinary Least Squares (OLS) method to estimate the unknown parameters.



$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

- The best possible line will be such so that the average squared vertical distances of the scattered points from the line will be the least.
- The differences between predicted and actual values are called residuals $e_i = \hat{y}_i - y_i$.
- If the line pass through all points of the training set then $J(\theta_0, \theta_1) = 0$.

To find the value of θ_0 and θ_1 that minimizes J there is a closed-form solution.

For univariate linear regression:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

$$S_{xx} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})(x_i - \bar{x}) \qquad S_{xy} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})$$

$$\theta_1 = \frac{S_{xy}}{S_{xx}} \qquad \theta_0 = \bar{y} - \theta_1\bar{x}$$

$$\hat{y} = \theta_0 + \theta_1 x$$

Later we'll see how to solve $\boldsymbol{\theta}$ for multivariate data (normal equation).