Linear regression

Alberto Morcillo Sanz

April 2024

1 Introduction

In statistics, linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

2 Formulation

Given a dataset $\{y_i; x_{i1}, x_{i2}, \cdots, x_{ik}\}$ of k statistical units, a linear regression model assumes that the relationship between the dependent variable y and the vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ε that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form:

$$\hat{y} = X\theta + \varepsilon$$

Where:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

We can say that the vector θ represents the weights and the vector ε represents the biases. Thus each $\hat{y}_i \in \hat{y}$ can be expressed as:

$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_k x_{ik} + \varepsilon_i = \sum_{j=0}^{k} \theta_j x_{ij} + \varepsilon_i$$

3 Training

To find the appropriate weights and biases, we can train the model by minimizing a loss function using Gradient Descent. So the training procedure is something like:

3.1 Training algorithm

- Generate the θ and ε vectors (I like to initially set the value $\frac{1}{k}$ to each element in θ and a random number between 0 and 1 to each element of ε).
- Apply the linear regression equation and compute the loss function L.
- Compute the gradient of the loss function ∇L and apply the Gradient Descent equations:

$$\theta_j^{t+1} = \theta_j^t - \gamma \frac{\partial L}{\partial \theta_j}$$

$$\varepsilon_i^{t+1} = \varepsilon_i^t - \gamma \frac{\partial L}{\partial \varepsilon_i}$$

- Repeat the previous two steps until the loss function is close enough to zero.

3.2 Loss function

In this case, we will consider the MSE function:

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

Thus the gradient of L is defined as:

$$\nabla L = \left(\frac{\partial L}{\partial \theta_j}, \frac{\partial L}{\partial \varepsilon_i} \right) = \frac{1}{n} \left[\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \frac{\partial}{\partial \varepsilon_i} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right]$$

- Partial derivative of L with respect to θ_j:

$$\frac{\partial L}{\partial \theta_j} = \frac{1}{n} \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} (y_i - \hat{y}_i)^2 =$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2 (y_i - \hat{y}_i) \frac{\partial}{\partial \theta_j} (y_i - \hat{y}_i) = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial \theta_j}$$

  - Partial derivative of $\hat{y}_i$ with respect to θ_j:

$$\frac{\partial \hat{y}_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left(\sum_{j=0}^{k} \theta_j x_{ij} + \varepsilon_i \right) = \frac{\partial}{\partial \theta_j} \sum_{j=0}^{k} \theta_j x_{ij} + \frac{\partial \varepsilon_i}{\partial \theta_j} = \sum_{j=0}^{k} x_{ij} \frac{\partial \theta_j}{\partial \theta_j} = \sum_{j=0}^{k} x_{ij}$$

  Then,

$$\boxed{\frac{\partial L}{\partial \theta_j} = -\frac{2}{n} \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) \sum_{j=0}^{k} x_{ij} \right]} \tag{1}$$

- Partial derivative of L with respect to ε_i:

$$\frac{\partial L}{\partial \varepsilon_i} = \frac{1}{n}\frac{\partial}{\partial \varepsilon_i}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \varepsilon_i}\left(y_i - \hat{y}_i\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - \hat{y}_i\right)\frac{\partial}{\partial \varepsilon_i}\left(y_i - \hat{y}_i\right) = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)\frac{\partial \hat{y}_i}{\partial \varepsilon_i}$$

  - Partial derivative of $\hat{y}_i$ with respect to ε_i:

$$\frac{\partial \hat{y}_i}{\partial \varepsilon_i} = \frac{\partial}{\partial \varepsilon_i}\left(\sum_{j=0}^{k}\theta_j x_{ij} + \varepsilon_i\right) = \frac{\partial}{\partial \varepsilon_i}\sum_{j=0}^{k}\theta_j x_{ij} + \frac{\partial \varepsilon_i}{\partial \varepsilon_i} = 1$$

Then,

$$\boxed{\frac{\partial L}{\partial \varepsilon_i} = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)} \tag{2}$$

So, finally ∇L is defined as:

$$\nabla L = \left(\frac{\partial L}{\partial \theta_j}, \frac{\partial L}{\partial \varepsilon_i}\right) = -\frac{2}{n}\left(\sum_{i=1}^{n}\left[\left(y_i - \hat{y}_i\right)\sum_{j=0}^{k} x_{ij}\right], \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)\right)$$

3.3 Gradient Descent

In mathematics gradient descent (also often called steepest descent) is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent. Conversely, stepping in the direction of the gradient will lead to a local maximum of that function; the procedure is then known as gradient ascent.

$$\theta_j^{t+1} = \theta_j^t - \gamma\frac{\partial L}{\partial \theta_j}$$

$$\varepsilon_i^{t+1} = \varepsilon_i^t - \gamma\frac{\partial L}{\partial \varepsilon_i}$$

Where $\gamma \in \mathbf{R}$ is a scalar called learning rate.

3.4 Exponential decay learning rate schedule

A quantity is subject to exponential decay if it decreases at a rate proportional to its current value. Symbolically, this process can be expressed by the following differential equation, where N is the quantity and λ is a positive rate

called the exponential decay constant, disintegration constant, rate constant or transformation constant:

$$\frac{dN}{dt} = -\lambda N \rightarrow N(t) = N_0 e^{-\lambda t}$$

According to the previous solution of the equation, exponential decay learning rate equation is defined as:

$$\gamma = \gamma_0 e^{-\lambda t}$$

Where γ_0 is the initial learning rate, λ is the decay rate and t is the time (epoch).