Simple Linear Regression - A Journey Begins

Adam Lee

I. Introduction

Regression is a statistical tool used in many facets of data analysis. Wherever correlated data is concerned, regression methods are often used to effectively and efficiently describe trends between two or more variables. During my brief first insights into machine learning, the initial bridge between statistics and machine learning has been linear regression. In its most simple form, a linear regression model assumes the relationship between a set of dependent variables $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ and a set of independent variables $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ is linear. Here each independent variable $\mathbf{x}_i \in \mathbb{R}^D$ is a D dimension vector of real valued inputs, however for our simple model we consider the case where $x_i \in \mathbb{R}$ such that the **simple linear regression model** has the form

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is our vector of response variables and $\mathbf{x}$ is the vector of single regressors. The intercept of this model β_0 and the slope β_1 are unknown constants and associated error term $\boldsymbol{\varepsilon}$ is treated as a random variable.

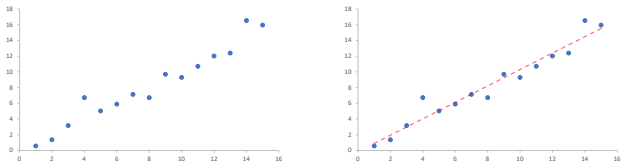

Fig. 1: A scatter-plot of data (left) and an example of a linear regression model (right)

We can consider the error term ε to be a statistical error. Therefore, it accounts for the failure of the model to fit the data exactly. It is common to assume the error term ε is Gaussian with mean $\mu = 0$ and variance $\mathrm{var}(\varepsilon) = \sigma^2$. Assume further that our regressor variable x is fixed, such that the response variable y is now dependant only on ε. Then our mean response variable at any value of x is given by

$$\mathcal{E}[y|x] = \mu_{y|x} = \beta_0 + \beta_1 x \tag{2}$$

with variance

$$\mathrm{var}(y|x) = \sigma^2_{y|x} = \sigma^2. \tag{3}$$

Therefore, we can consider the true regression model $\mu_{y|x}$ to be a function of mean values such that

$$\mu_{y|x}(x_{n+1}) = \beta_0 + \beta_1 x_{n+1} = y_{n+1}, \tag{4}$$

is the expected value y_{n+1} for a new prediction variable x_{n+1} to the data set $\mathbf{x}$. The variance of y_i at a given x_i is thus determined by σ^2. We see this model in fig. 2, with two example training points and our prediction point x_{n+1}.

In general, the response variable y may be related to multiple regressors. Say we have a relation between d regressors then

$$y = \beta_0 + \sum_{i=1}^{d} \beta_i x_i + \varepsilon. \tag{5}$$

This is known as multiple-linear regression and will be discussed in full later.

Clearly, the aim of both simple and multiple linear regression is to minimise the error term ε such that we have a model which best fits our observed data.

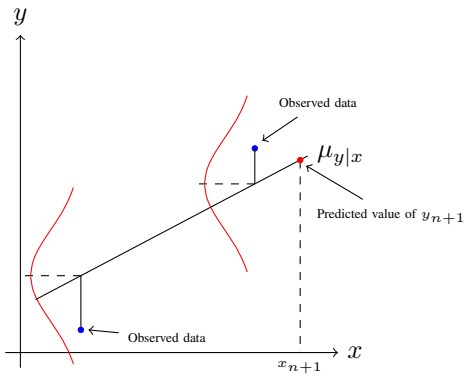

Fig. 2: A visual representation of simple linear regression

II. Simple Linear Regression

Let us first consider the case of simple linear regression, such that we have a single response variable y and a single regressor $x_1 := x$. Our initial model, as we have discussed, is

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{6}$$

where the intercept β_0 and the slope β_1 are unknown constants and the error term ε is a random variable. The parameters $\beta_{0,1}$ must be determined by sampling data and we will consider methods for doing so in later sections. These parameters are known as the regression coefficients and have a very simple practical interpretation. The slope β_1 is the expected change in value of y per unit change in value of x. The intercept β_0 is the expected value of the response given the value of the regressor is zero.

A. Least Squares method for parameter estimation

As we mentioned, the parameters $\beta_{0,1}$ are unknown constants and must be estimated using the sample data we have collected. The *method of least squares* allows us to do so. Let us adopt the notation that our data set $\varkappa$ is a collection of sample pairs

$$\varkappa = \{(x_0, y_0), \ldots, (x_k, y_k)\}. \tag{7}$$

We can input this data into eq. (1) to obtain a system of equations

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 0, \ldots, k. \tag{8}$$

Let us define the Least Squares criterion function

$$\mathbf{E}(\beta_0, \beta_1) := \sum_{i=0}^{k} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{9}$$

Let us further introduce real constants $\hat{\beta}_0, \hat{\beta}_1$ which satisfy the minima conditions

$$\begin{aligned} \frac{\partial \mathbf{E}}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=0}^{k} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial \mathbf{E}}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=0}^{k} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{aligned} \tag{10}$$

We can simplify eq. (10) to gain the least squares normal equations

$$\begin{aligned} (k+1)\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=0}^{k} x_i &= \sum_{i=0}^{k} y_i, \\ \hat{\beta}_0 \sum_{i=0}^{k} x_i + \hat{\beta}_1 \sum_{i=0}^{k} (x_i)^2 &= \sum_{i=0}^{k} y_i x_i. \end{aligned} \tag{11}$$

We can solve the first of the pair of equations for $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{12}$$

where $\bar{a} := \frac{1}{k+1} \sum_{i=0}^{k} a_i$ is the average value of all data points a_i. We now solve the second equation by substituting in eq. (12) as follows

$$(\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=0}^{k} x_i + \hat{\beta}_1 \sum_{i=0}^{k} x_i^2 = \sum_{i=0}^{k} x_i y_i, \tag{13}$$

and then solving with respect to $\hat{\beta}_1$ we obtain

$$\hat{\beta}_1 = \frac{\sum_{i=0}^{k} x_i y_i - \frac{\sum_{i=0}^{k} x_i y_i}{k+1}}{\sum_{i=0}^{k} x_i^2 - \frac{\sum_{i=0}^{k} x_i^2}{k+1}}. \tag{14}$$

The numerator of eq. (13) is nothing other than the corrected sum of dot products between x_i and y_i, while the denominator is the corrected sum of squares for x_i and thus we can introduce a concise notation for $\hat{\beta}_1$, namely

$$\hat{\beta}_1 = \frac{\sum_{i=0}^{k} y_i (x_i - \bar{x})}{\sum_{i=0}^{k} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}. \tag{15}$$

Then, our fitted simple linear regression model is nothing else than

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{16}$$

III. Example 1

Constructing an example for the method of least squares is quite simple. Consider the data-set fig. 3, where each y_i was sampled from a normal distribution centred on x_i with uniform variance $\sigma^2 = 2$

Observation, i	x_i	y_i
1	1	0.5678
2	2	1.3519
3	3	3.1623
4	4	6.7229
5	5	5.0327
6	6	5.9015
7	7	7.1302
8	8	6.7212
9	9	9.6973
10	10	9.3017
11	11	10.7025
12	12	12.0149
13	13	12.3884
14	14	16.5306
15	15	15.9602

Fig. 3: Example data set

The general trend for this data is clear, and from the distribution we used to generate the samples y_i we would expect the parameter values to be $\hat{\beta}_0 \approx 0$ and $\hat{\beta}_1 \approx 1$. Quick calculation concludes that we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{292.5196}{280.00} \approx 1.0447. \tag{17}$$

Then we have

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \\ &= 8.2118 - 1.0447 \cdot 8 \approx -0.1459. \end{aligned} \tag{18}$$

We have our two parameters $\hat{\beta}_0, \hat{\beta}_1$ and can fit our regression line $\hat{y}$ accordingly. We refer back to our expectations for these values, and conclude this would be an appropriate regression line for this simple case. This example is in fact the same as fig. 1, we present the figure again in fig. 4, extending the regression line to observe the y-intercept clearly.

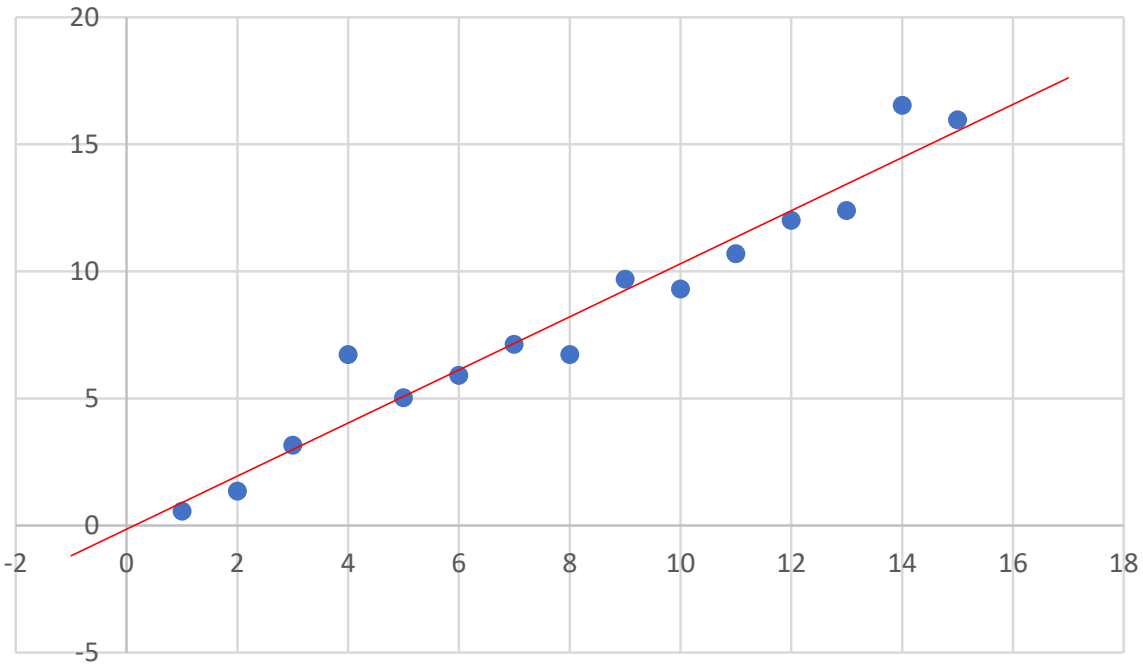

Fig. 4: The scatterplot for fig. 3 with regression line $\hat{y}$ in red.