LINEAR REGRESSION

Fall semester 2025

https://www.phys.uconn.edu/~rozman/Courses/P2200_25F/

Last modified: October 15, 2025

Suppose we conduct an experiment where we observe n data pairs and call them (x_i, y_i), $i = 1, \ldots, n$. We want to describe the underlying relationship between y_i and x_i involving the error of the measurements, ε_i, by the following relation:

$$y_i = \alpha + \beta x_i + \varepsilon_i. \tag{1}$$

This relationship between the true (but unobserved) parameters α and β and the data points is called a linear regression model.

Out goal is to find estimated values, $\widehat{\alpha}$ and $\widehat{\beta}$, for the parameters α and β which would provide the "best" fit in some sense for the data points (x_i, y_i),. We chose the best fit in the least-squares sense: the best-fit line minimizes the sum of squared residuals, $\widehat{\varepsilon}_i$, which are the differences between measured and predicted values of the dependent variable y:

$$\widehat{\varepsilon}_i = y_i - \widehat{\alpha} - \widehat{\beta} x_i. \tag{2}$$

That is, we are looking for the values $\widehat{\alpha}$ and $\widehat{\beta}$ that are the solutions of the following minimization problem: find

$$\min_{\widehat{\alpha}, \widehat{\beta}} Q(\widehat{\alpha}, \widehat{\beta}), \tag{3}$$

where

$$Q(\widehat{\alpha}, \widehat{\beta}) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2. \tag{4}$$

To find a minimum, we take partial derivatives of Q with respect to $\widehat{\alpha}$ and $\widehat{\beta}$ and equate

them to zeros. First, let's take the derivative with respect to $\widehat{\alpha}$:

$$\frac{\partial}{\partial \widehat{\alpha}} Q(\widehat{\alpha}, \widehat{\beta}) = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right) = 0, \tag{5}$$

or

$$\sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right) = 0. \tag{6}$$

Rearranging the terms, we get

$$\sum_{i=1}^{n} \widehat{\alpha} = \sum_{i=1}^{n} y_i - \widehat{\beta} \sum_{i=1}^{n} x_i. \tag{7}$$

Since $\sum_{i=1}^{n} \widehat{\alpha} = n\widehat{\alpha}$,

$$n\widehat{\alpha} = \sum_{i=1}^{n} y_i - \widehat{\beta} \sum_{i=1}^{n} x_i, \tag{8}$$

or

$$\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i - \widehat{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right). \tag{9}$$

Introducing $\bar{x}$ and $\bar{y}$, the average values of the x_i and y_i, respectively:

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i, \tag{10}$$

we obtain:

$$\widehat{\alpha} = \bar{y} - \widehat{\beta} \bar{x}, \tag{11}$$

or

$$\bar{y} = \widehat{\alpha} + \widehat{\beta} \bar{x}. \tag{12}$$

The relation Eq. (12) can be interpreted as follows: the best fit line passes through the "center of mass" of the data points.

Using the result Eq. (11), let's rewrite $Q(\widehat{\alpha}, \widehat{\beta})$ as follows:

$$Q(\widehat{\alpha}, \widehat{\beta}) = \sum_{i=1}^{n} \left(y_i - \left(\bar{y} - \widehat{\beta} \bar{x} \right) - \widehat{\beta} x_i \right)^2 = \sum_{i=1}^{n} \left((y_i - \bar{y}) - \widehat{\beta} (x_i - \bar{x}) \right)^2. \tag{13}$$

Now, take the derivative with respect to $\widehat{\beta}$:

$$\frac{\partial}{\partial \widehat{\beta}} Q(\widehat{\alpha}, \widehat{\beta}) = -2 \sum_{i=1}^{n} \Big((y_i - \bar{y}) - \widehat{\beta}\,(x_i - \bar{x}) \Big)(x_i - \bar{x}) = 0. \tag{14}$$

Rearranging the terms,

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \widehat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0, \tag{15}$$

or

$$\widehat{\beta} = \frac{\displaystyle\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{16}$$

We can now substitute $\widehat{\beta}$ to Eq. (11) to determine $\widehat{\alpha}$.

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}. \tag{17}$$

The definitions Eq. (10), and the formulas Eqs. (16) and (17) solve the problem of finding least squares fit to the data.

Finally, let's state without derivation that the standard error of $\widehat{\beta}$ is

$$\sigma_{\widehat{\beta}} = \sqrt{\frac{\displaystyle\sum_{i=1}^{n} \widehat{\varepsilon}_i^{\,2}}{(n-2)\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{18}$$

With the probability larger that 95% the true value of β is contained in the interval:

$$\beta \in \left[\widehat{\beta} - 2\sigma_{\widehat{\beta}},\, \widehat{\beta} + 2\sigma_{\widehat{\beta}} \right]. \tag{19}$$

With the probability larger that 99% the true value of β is contained in the interval:

$$\beta \in \left[\widehat{\beta} - 3\sigma_{\widehat{\beta}},\, \widehat{\beta} + 3\sigma_{\widehat{\beta}} \right]. \tag{20}$$