# Linear least-square regression

Thibaud Taillefumier

## 1 Problem set-up

Suppose we have a $m$-dimensional vector $\boldsymbol{y} = \{y_1, ..., y_m\}$ whose components represent scalar output measurements to be related to $m$ input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, each of dimension $n$. We want to find weights $\boldsymbol{w} = \{w_1, \ldots, w_m\}$ such that one can predict the measurement outcomes $\boldsymbol{y}$ as a linear combination of the input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$: $y_j \approx \boldsymbol{w}^T \boldsymbol{x}_j = \sum_{i=1}^m x_{ji} w_i$.

In principle, we can repeat measurements at will so that $m$ can be very large, whereas $n$ is set by the complexity of the model and should be assumed comparatively small $n < m$. For instance, think of $n$ as the number of relevant features in the model. Because the vector $\boldsymbol{y}$ lies into a much larger $m$-dimensional space than the at most $n$-dimensional space spanned by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, it is in general impossible to perfectly reconstruct $\boldsymbol{y}$. For this reason, regression methods propose to minimize the prediction error instead of trying to achieve a perfect reconstruction. Specifically, in linear least-square regression, we aim at finding the weights $\boldsymbol{w}$ which minimize the squared prediction error $E(\boldsymbol{w})$. This can be formally stated as follows:

$$\boldsymbol{w}^\star = \arg\min_{\boldsymbol{w}} E(\boldsymbol{w}) \quad \text{with} \quad E(\boldsymbol{w}) = \sum_{j=1}^m \left( y_j - \sum_{i=1}^n x_{ji} w_i \right)^2 .$$

The squared prediction error $E(\boldsymbol{w})$ can be interpreted geometrically as the squared Euclidean length of the residual vector defined by $\boldsymbol{y} - \sum_{j=1}^n w_j \boldsymbol{x}_j$. Thus, $E(\boldsymbol{w})$ can also be written as

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2 ,$$

where $\|u\|$ denotes the length of vector $u$ and where $X$ is the matrix whose rows are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. As expected, the squared prediction error is a non-negative number that is zero only if the output $y$ lies in the span of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. However, this generally does not happen due to the dimensionality mismatch $m > n$ and perfect reconstruction is in general impossible. To address this point, the method of

linear least-square regression resorts to looking for weights $\boldsymbol{w}$ that minimize the Euclidean square length of the residual vector. This approach, which involves a combination of calculus and linear algebra, will yield the best linear prediction of $\boldsymbol{y}$ based on observing $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$.

## 2 Solution via calculus and linear algebra

Throughout this section, bear in mind that the core motivation behind linear least-square regression stems from the dimensionality mismatch $m < n$. This can be stated concretely by saying that the matrix $X$ has many more rows than columns.

The first step of linear least-square regression is to compute the derivative of the squared prediction error $E$ with respect to an arbitrary weight $w_k$, while holding all the other weights fixed:

$$\begin{aligned}\frac{\partial E(\boldsymbol{w})}{\partial w_k} &= \sum_{i=1}^{m} \frac{\partial}{\partial w_k} \left( y_j - \sum_{i=1}^{n} x_{ji} w_i \right)^2 , \\ &= 2 \sum_{i=1}^{m} \left( y_j - \sum_{i=1}^{n} x_{ji} w_i \right) \left[ \frac{\partial}{\partial w_k} \left( y_j - \sum_{i=1}^{n} x_{ji} w_i \right) \right] ,\end{aligned}$$

The term in between square bracket is actually much simpler than it looks as it is the derivative of a linear function of $w_k$ with linear coefficient $x_{ik}$.

$$\frac{\partial}{\partial w_k} \left( y_j - \sum_{i=1}^{n} x_{ji} w_i \right) = x_{jk} \, .$$

The weights $\boldsymbol{w}$ that minimize the squared prediction error $\boldsymbol{E}$ are those weights for which the derivatives of $\boldsymbol{E}$ with respect to any $w_k$ is zero. Based on our computation of the derivative of the squared prediction error, this means that the weights $\boldsymbol{w}$ satisfy the following set $n$ linear equations:

$$\frac{\partial E(\boldsymbol{w})}{\partial w_k} = 2 \sum_{i=1}^{m} \left( y_j - \sum_{i=1}^{n} x_{ji} w_i \right) x_{jk} = 0 \, , \quad \text{with} \quad 1 \leq k \leq n \, .$$

The second step of linear least-square regression is to solve the above set of equations via matrix algebra. To see how, observe that our set of equations can be conveniently expressed in matrix form by using the transpose operation. Indeed, using the fact that $x_{jk} = (X^T)_{kj}$, we can write the system of equation in matrix form as

$$X^T \left( \boldsymbol{y} - X \boldsymbol{w} \right) = 0 \, ,$$

showing that the weights $\boldsymbol{w}$ are solution of the matrix equation

$$X^T X \boldsymbol{w} = X^T \boldsymbol{y} \,.$$

The matrix $X^T X$ is a $n$-by-$n$ square matrix that is invertible if $n \geq m$ (which is true) and if the matrix $X$ has rank $n$, i.e. if the columns of $X$ are linearly independent (which need to be checked). Under this assumption of invertibility, the weight $\boldsymbol{w}$ are obtained via matrix inversion

$$\boldsymbol{w}^\star = (X^T X)^{-1} X^T \boldsymbol{y} \,,$$

thereby answering our problem of linear least-square regression. Moreover, the best approximation to the original vector $\boldsymbol{y}$, denoted by $\boldsymbol{y}^\star$, can be recovered as:

$$\boldsymbol{y}^\star = X \boldsymbol{w}^\star = X (X^T X)^{-1} X^T \boldsymbol{y} \,.$$