Title: Linear predictor function

URL: https://en.wikipedia.org/wiki/Linear_predictor_function

PageID: 35272263

Categories: Category:Machine learning, Category:Regression analysis

---

In statistics and in machine learning , a linear predictor function is a linear function (linear combination) of a set of coefficients and explanatory variables (independent variables), whose value is used to predict the outcome of a dependent variable . [1] This sort of function usually comes in linear regression , where the coefficients are called regression coefficients . However, they also occur in various types of linear classifiers (e.g. logistic regression , [2] perceptrons , [3] support vector machines , [4] and linear discriminant analysis [5]), as well as in various other models, such as principal component analysis [6] and factor analysis . In many of these models, the coefficients are referred to as "weights".

Definition

The basic form of a linear predictor function f (i) {\displaystyle f(i)} for data point i (consisting of p explanatory variables), for i = 1, ..., n , is

where x i k {\displaystyle x_{ik}} , for k = 1, ..., p , is the value of the k -th explanatory variable for data point i , and β 0 , … , β p {\displaystyle \beta _{0},\ldots ,\beta _{p}} are the coefficients (regression coefficients, weights, etc.) indicating the relative effect of a particular explanatory variable on the outcome .

Notations

It is common to write the predictor function in a more compact form as follows:

The coefficients β 0 , β 1 , ..., β p are grouped into a single vector β of size p + 1.

For each data point i , an additional explanatory pseudo-variable x i 0 is added, with a fixed value of 1, corresponding to the intercept coefficient β 0 .

The resulting explanatory variables x i0 (= 1), x i 1 , ..., x ip are then grouped into a single vector x i of size p + 1.

Vector Notation

This makes it possible to write the linear predictor function as follows:

using the notation for a dot product between two vectors.

Matrix Notation

An equivalent form using matrix notation is as follows:

where β {\displaystyle {\boldsymbol {\beta }}} and x i {\displaystyle \mathbf {x} _{i}} are assumed to be a (p+1) -by-1 column vectors , β T {\displaystyle {\boldsymbol {\beta }}^{\mathrm {T} }} is the matrix transpose of β {\displaystyle {\boldsymbol {\beta }}} (so β T {\displaystyle {\boldsymbol {\beta }}^{\mathrm {T} }} is a 1-by- (p+1) row vector), and β T x i {\displaystyle {\boldsymbol {\beta }}^{\mathrm {T} }\mathbf {x} _{i}} indicates matrix multiplication between the 1-by- (p+1) row vector and the (p+1) -by-1 column vector, producing a 1-by-1 matrix that is taken to be a scalar .

Linear regression

An example of the usage of a linear predictor function is in linear regression , where each data point is associated with a continuous outcome y i , and the relationship written

where ε i {\displaystyle \varepsilon _{i}} is a disturbance term or error variable — an unobserved random variable that adds noise to the linear relationship between the dependent variable and

predictor function.

Stacking

In some models (standard linear regression, in particular), the equations for each of the data points i = 1, ..., n are stacked together and written in vector form as

where

The matrix X is known as the design matrix and encodes all known information about the independent variables . The variables ε i {\displaystyle \varepsilon _{i}} are random variables , which in standard linear regression are distributed according to a standard normal distribution ; they express the influence of any unknown factors on the outcome.

This makes it possible to find optimal coefficients through the method of least squares using simple matrix operations. In particular, the optimal coefficients β ^ {\displaystyle {\boldsymbol {\hat {\beta }}}} as estimated by least squares can be written as follows:

The matrix (X T X) − 1 X T {\displaystyle (X^{\mathrm {T} }X)^{-1}X^{\mathrm {T} }} is known as the Moore–Penrose pseudoinverse of X . The use of the matrix inverse in this formula requires that X is of full rank , i.e. there is not perfect multicollinearity among different explanatory variables (i.e. no explanatory variable can be perfectly predicted from the others). In such cases, the singular value decomposition can be used to compute the pseudoinverse.

Preprocessing of explanatory variables

When a fixed set of nonlinear functions are used to transform the value(s) of a data point, these functions are known as basis functions . An example is polynomial regression , which uses a linear predictor function to fit an arbitrary degree polynomial relationship (up to a given order) between two sets of data points (i.e. a single real-valued explanatory variable and a related real-valued dependent variable), by adding multiple explanatory variables corresponding to various powers of the existing explanatory variable. Mathematically, the form looks like this:

In this case, for each data point i , a set of explanatory variables is created as follows:

and then standard linear regression is run. The basis functions in this example would be

This example shows that a linear predictor function can actually be much more powerful than it first appears: It only really needs to be linear in the coefficients . All sorts of non-linear functions of the explanatory variables can be fit by the model.

There is no particular need for the inputs to basis functions to be univariate or single-dimensional (or their outputs, for that matter, although in such a case, a K -dimensional output value is likely to be treated as K separate scalar-output basis functions). An example of this is radial basis functions (RBF's), which compute some transformed version of the distance to some fixed point:

An example is the Gaussian RBF, which has the same functional form as the normal distribution :

which drops off rapidly as the distance from c increases.

A possible usage of RBF's is to create one for every observed data point. This means that the result of an RBF applied to a new data point will be close to 0 unless the new point is near to the point around which the RBF was applied. That is, the application of the radial basis functions will pick out the nearest point, and its regression coefficient will dominate. The result will be a form of nearest neighbor interpolation , where predictions are made by simply using the prediction of the nearest observed data point, possibly interpolating between multiple nearby data points when they are all similar distances away. This type of nearest neighbor method for prediction is often considered diametrically opposed to the type of prediction used in standard linear regression: But in fact, the transformations that can be applied to the explanatory variables in a linear predictor function are so powerful that even the nearest neighbor method can be implemented as a type of linear regression.

It is even possible to fit some functions that appear non-linear in the coefficients by transforming the coefficients into new coefficients that do appear linear. For example, a function of the form a + b 2 x i 1 + c x i 2 {\displaystyle a+b^{2}x_{i1}+{\sqrt {c}}x_{i2}} for coefficients a , b , c {\displaystyle a,b,c} could be transformed into the appropriate linear function by applying the substitutions b ′ = b 2 , c ′

= c , {\displaystyle b'=b^{2},c'={\sqrt {c}},} leading to a + b ′ x i 1 + c ′ x i 2 , {\displaystyle a+b'x_{i1}+c'x_{i2},} which is linear. Linear regression and similar techniques could be applied and will often still find the optimal coefficients, but their error estimates and such will be wrong.

The explanatory variables may be of any type : real-valued , binary , categorical , etc. The main distinction is between continuous variables (e.g. income, age, blood pressure , etc.) and discrete variables (e.g. sex, race, political party, etc.). Discrete variables referring to more than two possible choices are typically coded using dummy variables (or indicator variables), i.e. separate explanatory variables taking the value 0 or 1 are created for each possible value of the discrete variable, with a 1 meaning "variable does have the given value" and a 0 meaning "variable does not have the given value". For example, a four-way discrete variable of blood type with the possible values "A, B, AB, O" would be converted to separate two-way dummy variables, "is-A, is-B, is-AB, is-O", where only one of them has the value 1 and all the rest have the value 0. This allows for separate regression coefficients to be matched for each possible value of the discrete variable.

Note that, for K categories, not all K dummy variables are independent of each other. For example, in the above blood type example, only three of the four dummy variables are independent, in the sense that once the values of three of the variables are known, the fourth is automatically determined. Thus, it's really only necessary to encode three of the four possibilities as dummy variables, and in fact if all four possibilities are encoded, the overall model becomes non-identifiable . This causes problems for a number of methods, such as the simple closed-form solution used in linear regression. The solution is either to avoid such cases by eliminating one of the dummy variables, and/or introduce a regularization constraint (which necessitates a more powerful, typically iterative, method for finding the optimal coefficients). [7]

See also

Linear model

Linear regression

References