Linear Methods for Regression

A linear regression model assumes that the regression function is linear in the inputs . We have an input vector and want to predict a real-valued output . The linear regression model has the form

(1)

The linear model either assumes that the regression function is linear, or that the linear model is a reasonable approximation. The ’s are unknown parameters and the variables can come from different sources:

* quantitative inputs
* transformations of quantitative inputs, such as log, square root or square
* basis expansions such as , , leading to a polynomial representation
* numeric or “dummy” coding of the levels of qualitative inputs. For example, if is a five-level factor input, we might create , such that . Together this group of represents the effect of by a set of level-dependent constants, since in , one of the s is one, and the others are zero.
* interactions between variables , for example .

No matter the source of the , the model is linear in the parameters.

Typically, we have a set of training data from which to estimate the parameters . Each is a vector of feature measurements for the th case. The most popular estimation method is *least squares*, in which we pick the coefficients to minimize the residual sum of squares

(2)

From a statistical point of view, this criterion is reasonable if the training observations represent independent random draws from their population. Even if the ’s were not drawn randomly, the criterion is still valid if the ’s are conditionally independent given the inputs . Note that (2) makes no assumptions about the validity of model (1); it simply finds the best linear fit to the given dataset.

Let us denote by the matrix with each row an input vector (with s in the first column), and similarly let the y be the -vector of outputs in the training set. Then we can write the residual sum of squares as

(3)

is a quadratic function with parameters stored in . Differentiating with respect to we obtain

, (4)

Let us assume that has full column rank[[1]](#footnote-1), and hence is positive definite, we set the first derivative to zero:

(5)

to obtain the unique solution

# References

[1] Chapter 3 of [The Elements of Statistical Learning; Data Mining, Inference, Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Second Edition, 2017](https://github.com/dimitarpg13/statistical_learning_and_kernel_methods/blob/main/literature/books/EelementsOfStatisticalLearning_print12.pdf)

1. Later we will revisit this assumption and will study the case when X is rank-deficient i.e. [↑](#footnote-ref-1)