Methods section

Linear Regression

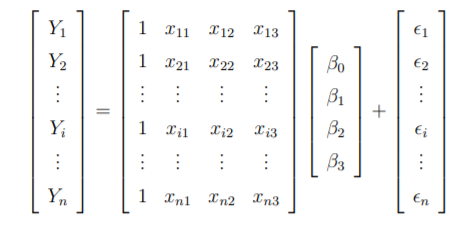
Linear regression is a fitting procedure useful for predicting a quantitative response Y on the basis of predictor variables. If we consider the relationship in the form Y = f(X) + ϵ for some unknown function f, in linear regression we will assume that f can be approximated by a linear function. In this way the linear relationship between Y and X that mathematically can be written as 𝑦=𝑋𝜷+𝝐,

Where: 𝑦=[𝑦0,𝑦1,𝑦2,…,𝑦𝑛−1]𝑇 is the vector of the Y values,

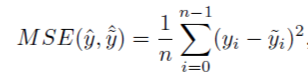
𝜷=[𝛽0,𝛽1,𝛽2,…,𝛽𝑛−1]𝑇 are the regression parameters,

𝝐=[𝜖0,𝜖1,𝜖2,…,𝜖𝑛−1]𝑇 is a mean-zero random error term normally distributed,

X ∈ℝ𝑛×𝑝 is the design matrix containing the information about the predictor variables.

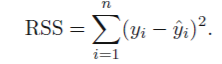
In a matrix form:

In order to quantify the capacity of the model at fitting the data, for a linear regression fit the quality is typically verified using two quantities: the Mean Squared error (MSE) and the R2 statistic.

The MSE considers the average amount that the predicted values will deviate from the true ones, is calculated as: 

The MSE is a measure of the lack of fit of the model to the data. If the predictions obtained using the model are very close to the true outcome values then the MSE will be small, and we can conclude that the model fits the data very well. Otherwise, if ˆyi is very far from yi, then the MSE will be large, indicating that the model doesn’t fit the data well.

The R2 statistic provides an alternative measure of fit in the form of a proportion of variance. So it always takes values between 0 and 1, and is independent of the scale of Y .

It is calculated as:  Where TSS is the Total Sum of Squares  measure of the total variance of Y and RSS is the Residual Sum of Squares defined as:  It represents the amount of variability left unexplained after the regression. Considering that TSS−RSS measures the amount of variability of Y explained by the regression, overall, we can say that R2 measures the proportion of variability explained using X. An R2 statistic close to 1 indicates that there is a large proportion of variability explained by the model, instead a number near 0 indicates that the regression did not explain much of the variability of the response.

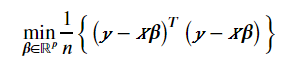
In order to optimize a linear model we studied different approaches for the determination of the regression parameters:

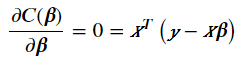
* Ordinary Least Squares (OLS)

In order to optimize the 𝛽𝑖 values we define a function known as cost function, that gives a measure of the distance between the true values 𝑦𝑖 and the predicted values 𝑦tilde:

Immagine che contiene oggetto

Descrizione generata automaticamente   
We want then to minimize the distance by solving the problem:

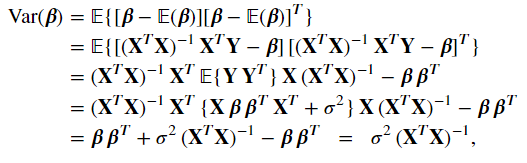


In a matrix form:   
Rewriting it as if 𝑋𝑇𝑋 is invertible, then the solution is: 

We can verify how close the ˆ βs obtained with ols are to the true values β by calculating the quantities:

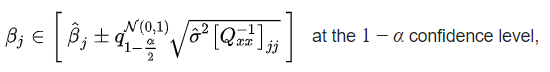


Average of the estimated parameters, confirms that the parameter is unbiased, and:



Variance of the parameters.

The variance can be used to create confidence intervals as a range of values that will contain the true value of the parameter with a probability equal to 1-alpha (alpha small). Mathematically we can express it as



 *q* denotes the [quantile function](https://en.wikipedia.org/wiki/Quantile_function) of a normal distribution, and [·]*jj* is the *j*-th diagonal element of the matrix.