Linear regression is a model used in statistics and supervised machine learning, and with the use of kernels or other basis function expansions, it can also be used to model non-linear relationships, as well as for classification when its Gaussian output is replace with a Bernoulli or multinoulli distribution.

The linear regression model is of the form $p(y|x, \theta) = N(y|w^T x, \sigma^2)$, and in order to use it to model non-linear relationships, we have to replace x with some non-linear function of the inputs $\varphi(x)$, which means that we use $p(y|x, \theta) = N(y|w^T \varphi(x), \sigma^2)$, which is known as basis function expansion. This is still considered linear regression because the model is still linear in the parameters w.

Computing the maximum likelihood estimation is a common way to estimate the parameters of a statistical model, which is defined as $\hat{\theta} \triangleq \arg\max_\theta \log p(D|\theta)$. It is commonly assumed that the training examples involved are independent and identically distributed, a trait that is usually abbreviated to iid. Instead of maximizing the log-likelihood in this scenario, it is equivalent to minimize the negative log likelihood, or the NLL, which is sometimes more convenient when using optimization software packages designed to find the minima of functions rather than the maxima. RSS is an acronym which stands for the residual sum of squares, and which is defined by $RSS(w) \triangleq \sum_{i=1}^{N} (y_i - w^T x_i)^2$. The RSS is also referred to as the sum of squared errors, or SSE, and the SSE divided by N is called the mean square error, also known as the MSE. Additionally, it can also be written as the square of the l_2 norm of the vector of residual errors, which tells us that the MLE for w is the one that minimizes the RSS, so the overall method is known as least squares.

When it comes to robust linear regression, it is common to model the noise in regression models using a Gaussian distribution with zero mean and constant variance. Here, maximizing likelihood is equivalent to minimizing the sum of the squared residuals. However, if there are outliers in the data, this can cause a poor fit. A method of achieving robustness to outliers is by replacing the Gaussian distribution for the response variable with a distribution that has heavy tails, because this type of distribution will assign higher likelihood to outliers without the need to disturb the straight line. One such possibility is the Laplace distribution. However, this is a non-linear objective function, which can be hard to optimize, but this can be remedied by converting the NLL to a linear objective, subject to linear constraints, using the split variable trick.