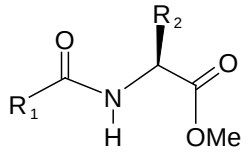
Multiple Linear Least-Squares Regression.

Estimation of regression parameters

There are many situations where more than one independent variable determines the observed signal. An important field of science where such situations frequently arise is drug design based on quantitative structure-activity relationships (QSAR). The ultimate goal of QSAR-based drug design is to find out which structural properties confer the drug highest potency or lowest toxicity. The drug's potency is here a dependent variable, and the structural properties, also called molecular descriptors, are the independent variables. The experimental signal that measures the potency could be, for example, the binding affinity of a drug candidate to its target protein. Some molecular descriptors for substituents are the lipophilicity (π) of a group, polarizability (α) of a group, and electron-withdrawing ability (σ) of a group. To predict which structures have the highest potency, parameters in the regression model relating multiple descriptors to the potency must be determined.

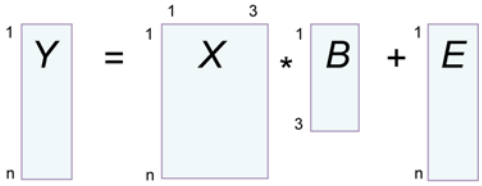
We will illustrate such analysis with a real, albeit simple example. Consider the binding affinity in a series of bi-substituted N-Acetyl L-amino acid methyl esters for the digestive protein chymotrypsin. The active site of chymotrypsin has a hydrophobic binding pocket that accommodates the side chain R_2. Thus one would expect that more hydrophobic substituents at position R_2 increase binding. The R_1 group makes several weaker, non-specific van der Waals interactions with the protein. Because the strength of van der Waals interactions depends on polarizability, one would predict that more polarizable substituents at R_1 increase binding affinity. The simplest approach is to assume that the binding free energy, which is proportional to an experimentally determined quantity log $1/K_d$, is linearly dependent on the hydrophobicity and polarizability. To determine the coefficients, we need to measure binding affinities for a set of compounds that have different substituents at R_1 and R_2, and then find the best linear equation that relates known hydrophobicity and polarizability descriptors to the observed binding affinity. For each compound in a series we can write:



$$\text{Log } 1/K_d\,(1) = 1 \cdot b_0 + \pi_1 \cdot b_1 + \alpha_1 \cdot b_2$$
$$\text{Log } 1/K_d\,(2) = 1 \cdot b_0 + \pi_2 \cdot b_1 + \alpha_2 \cdot b_2$$
$$\text{Log } 1/K_d\,(3) = 1 \cdot b_0 + \pi_3 \cdot b_1 + \alpha_3 \cdot b_2$$
$$\text{Log } 1/K_d\,(4) = 1 \cdot b_0 + \pi_4 \cdot b_1 + \alpha_4 \cdot b_2$$
$$\text{Log } 1/K_d\,(n) = 1 \cdot b_0 + \pi_n \cdot b_1 + \alpha_n \cdot b_2$$

The parameters b_0, b_1, and b_2 can be determined via multiple lest-squares linear regression. The algebraic equations for multiple linear least squares are rather complex and we are not presenting them here. However, the problem can be cast into matrix notation, where the solution is both elegantly simple and computationally efficient. The observed binding data can be represented as a column vector $\boldsymbol{Y}$ with n rows, values of two structural descriptors as a matrix $\boldsymbol{X}$ with three columns and n rows, and the unknown parameters as a column vector $\boldsymbol{B}$ with three rows. The errors in each measurement of binding affinity can

be grouped into a column vector ***E*** with n rows. The matrix algebra tells us that the binding affinity vector ***Y*** can be calculated by multiplying descriptor matrix ***X*** with the parameter vector ***B*** and adding a vector of residual errors ***E***



In the condensed matrix notation, the above reads $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{B} + \boldsymbol{E}$

The best parameters are determined by finding a vector ***B*** that minimizes the squared residuals in matrix ***E***. The matrix formula to calculate the least-squares estimate of vector ***B*** is:

$$B = \left(X^T X\right)^{-1} X^T Y \qquad \text{Eq. 14}$$

This means that the parameter vector ***B*** is obtained by multiplying the transpose of matrix ***X*** with the matrix ***X***, inverting the product matrix, then multiplying the inverted product matrix with the transposed matrix ***X***, and finally multiplying the result of previous operations with the matrix ***Y***. Other statistical characteristics, such as the standard deviation for each parameter, can be also calculated.

Multiple linear regression problems are well suited for computers. One program that can easily solve multi-variable linear least-squares problems is *Mathematica* by Wolfram Research (http://www.wolfram.com/). This program is installed on the SGI computers in Laboratory of Computational Chemistry and Biochemistry. *Mathematica* offers couple of ways to perform multiple linear regression. First, Mathematica can directly deal with matrixes, and one can write the matrix equation above as `fitB = Inverse[Transpose[matX] . matX] . Transpose[matX]] . vecY` The function `PseudoInverse` can be used instead of `Inverse` with matrixes that are difficult to invert.

Second, one can use the built-in function `Regress`. It takes three sets of arguments. The first set is the data to be analyzed, the second set is the description of the regression model, and the third set is the list of independent variables. The data set is a matrix with the dependent variable in the last column. The description of the linear regression model requires the list of descriptors in the model, including the intercept term. The list on independent variables just gives the descriptors.

```
<<"Statistics`LinearRegression`"
Hydrophobicity = {1.5,2.2,3.1,3.8,4.1,4.5,4.9,5.3}
Polarizability = {11, 32, 17, 38, 22, 10, 18, 6}
Binding = {69.7, 85.8, 79.7, 96.3, 89.3, 84.6, 93.1, 88.1}
DataLR1 = Transpose[{Hydrophobicity, Binding}];
DataLR2 = Transpose[{Polarizability, Binding}];
DataMLR = Transpose[{Hydrophobicity, Polarizability, Binding}];
LinReg1 = Regress[DataLR1, {1, x}, x]
LinReg2 = Regress[DataLR2, {1, x}, x]
MLRReg = Regress[DataMLR, {1, x1, x2}, {x1, x2}]
```