Regression and Uncertainties

(from a work made for the CEA/INSTN, April 2008)

by the

Société de Calcul Mathématique SA

Summary

We have a measurement device, the output of which is assumed to be linear. How should we take into account the uncertainties upon the values to be measured x and upon the results of the measurements y? In practice, we have a "regression line", made from all couples (x, y) ; we show how to define a probability law on all these lines.

However, we insist upon the fact that the response of a measurement device is never linear, and the precision is usually worse on the extremities of the measure scale. So, the use of a straight line is not a good idea, since it masks these two realities : non-linearity, differences in precision over the whole scale. The correct mathematical approach is that of "calibration tables".

I. Usual construction

We have N points in the plane, with coordinates $(x_n, y_n)_{n=1,...,N}$. The regression line is the line with equation $y = a\,x + b$ which minimizes the quantity :

$$U(a,b) = \sum_{n=1}^{N} \left(y_n - ax_n - b \right)^2$$

Partial derivatives may be written :

$$\frac{\partial U}{\partial a} = -2 \sum_{n=1}^{N} x_n \left(y_n - ax_n - b \right)$$
$$\frac{\partial U}{\partial b} = -2 \sum_{n=1}^{N} \left(y_n - ax_n - b \right)$$

So we get the system :

$$\sum_{n=1}^{N} x_n \left(y_n - ax_n - b \right) = 0$$
$$\sum_{n=1}^{N} \left(y_n - ax_n - b \right) = 0$$

or :

$$a \sum_{n=1}^{N} x_n^2 + b \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} x_n y_n$$
$$a \sum_{n=1}^{N} x_n + Nb = \sum_{n=1}^{N} y_n$$

We set :

$$\alpha = \sum_{n=1}^{N} x_n \ , \ \beta = \sum_{n=1}^{N} x_n^2 \ , \ \gamma = \sum_{n=1}^{N} x_n y_n \ , \ \delta = \sum_{n=1}^{N} y_n \ ;$$

The above equations can be written :

$$a\beta + b\alpha = \gamma$$
$$a\alpha + Nb = \delta$$

From the second one, we deduce :

$$b = \frac{\delta}{N} - \frac{a\alpha}{N}$$

and, putting back into the first :

$$a\beta + \left(\frac{\delta}{N} - \frac{a\alpha}{N} \right) \alpha = \gamma$$

or :

$$a\left(\beta - \frac{\alpha^2}{N} \right) = \gamma - \frac{\delta\alpha}{N}$$

and therefore :

$$a = \frac{\gamma N - \delta\alpha}{\beta N - \alpha^2} \tag{1}$$

and :

$$b = \frac{\delta\beta - \alpha\gamma}{\beta N - \alpha^2} \tag{2}$$

If we have explicit values for x_n and y_n, we get an explicit value for a and b.

We now see how to introduce uncertainties both on x and y.

II. Introduction of uncertainties

Now, we have at our disposition some information of uncertainty on each x_n and each y_n, under the form of a probability law.

We could :

- Construct a unique regression line, from the average points : this is not satisfactory, because this gives no information at all about the dispersion.

- Construct a unique regression line, minimizing the average distance.

The probability laws on each x_n and each y_n give a probability law on the distance :

$$U(a,b)=\sum_{n=1}^{N}\left(y_n-a\,x_n-b\right)^2$$

and one can find the coefficients a and b which minimize the average distance :

$$U_{moy}(a,b)=\frac{1}{m}\sum_{j_1=1}^{m}\ \frac{1}{2\varepsilon_1}\int_{\xi_1-\varepsilon_1}^{\xi_1+\varepsilon_1}\left(y_{1,j_1}-ax_1-b\right)^2 dx_1+\frac{1}{m}\sum_{j_2=1}^{m}\ \frac{1}{2\varepsilon_2}\int_{\xi_2-\varepsilon_2}^{\xi_2+\varepsilon_2}\left(y_{2,j_2}-ax_2-b\right)^2 dx_2+\cdots$$

In this formula, we assume for example that each y_i takes m positions with same probability and each x_n follows a uniform law on the interval $\left[\xi_i-\varepsilon_i,\xi_i+\varepsilon_i\right]$.

This second possibility is different from the first one : the line constructed that way is not the line constructed from the average points. It has the same drawbacks : no probability law, no measure of dispersion.

So, these two procedures cannot be recommended. We now turn to the satisfactory solution.

Probabilistic regression line

We let :

$$a=\varphi(x_1,y_1,x_2,y_2,...)$$

$$b=\psi(x_1,y_1,x_2,y_2,...)$$

be the explicit expressions of a and b, as functions of x_i,y_i, given by (1) et (2).
Then, a probability law on each x_i,y_i gives a probability law on a and b : this is the probabilistic regression line. Then, we take the expectation of a and b for the best choice for the line.

With the previous laws, we have :

$$Ea=\frac{1}{2\varepsilon_1}\int_{\xi_1-\varepsilon_1}^{\xi_1+\varepsilon_1}\ \frac{1}{m}\sum_{j_1=1}^{m}\ \frac{1}{2\varepsilon_2}\int_{\xi_2-\varepsilon_2}^{\xi_2+\varepsilon_2}\ \frac{1}{m}\sum_{j_2=1}^{m}\ \frac{1}{2\varepsilon_n}\int_{\xi_n-\varepsilon_n}^{\xi_n+\varepsilon_n}\ \frac{1}{m}\sum_{j_n=1}^{m}\ \varphi(x_1,y_{1,j_1},x_2,y_{2,j_2},...,x_m,y_{n,j_n})dx_1dx_2\cdots dx_n$$

and the same for b.

Doing things this way, we keep all lines, and we have a law upon the coefficient a and a law on the coefficient b ; we can write confidence intervals, and so on.

If the points x_n, y_n take only discrete value, what we do is as follows : the enumerate all possible configurations for x_n, y_n, et construct the line for each configuration.

In practice, the number of configurations is quite high, so one has to use a random method in order to reconstruct the law upon a, b.