Société de Calcul Mathématique SA
Mathematical Modelling Company, Corp.
Decision-making tools, since 1995

The regression line associated with a Simple Random Walk

Bernard Beauzamy

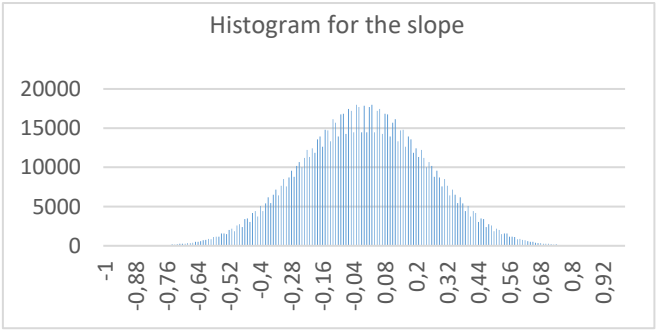
March 2, 2019

Abstract

We investigate the value of the slope of a regression line, when the points are (k, X_k), the X_k being independent random variables with values ± 1 with probability 1/2. We show that this slope is given by the formula :

$$a = \frac{6}{N\left(N^2-1\right)} \sum_{k=1}^{N-1} k\left(N-k\right) X_k$$

Numerical investigation shows a strange behavior; for instance for $N = 20,$ we obtain:



which shows that the slope is irregular. This histogram seems to be made of two different gaussians. Indeed, we observe that, for 1720 situations, we have $a = 0$ and, for 2712 situations, we have $a = \frac{1}{1330}$.

1. General description

Let (x_k, y_k), $k = 1,...,N$, be points in the plane ; the regression line associated to them minimizes the quantity:

$$D = \sum_{k=1}^{N} \left(y_k - (ax_k + b)\right)^2$$

The equations $\frac{\partial D}{\partial a} = 0, \frac{\partial D}{\partial b} = 0,$ lead to the system:

$$\begin{cases} \sum_{k=1}^{N} x_k \left(y_k - (ax_k + b)\right) = 0 \\ \sum_{k=1}^{N} y_k - (ax_k + b) = 0 \end{cases} \quad (1)$$

We set $m_x = \frac{1}{N}\sum_{k=1}^{N} x_k, m_y = \frac{1}{N}\sum_{k=1}^{N} y_k$ and we deduce from (1b):

$$b = m_y - am_x$$

and, from (1a):

$$\sum_{k=1}^{N} x_k \left(y_k - ax_k - m_y + am_x\right) = 0$$

that is:

$$\sum_{k=1}^{N} x_k y_k - ax_k^2 - m_y x_k + am_x x_k = 0$$

So the slope of the line is:

$$a = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

where:

$$\mathrm{cov}(X,Y) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k - m_x m_y$$

$$\mathrm{var}(X) = \frac{1}{N}\sum_{k=1}^{N} x_k^2 - m_x^2$$

2. Case of a time series

Let us consider the case of a time series, with origin of time at T_0. We have:

$$x_k = T_0 + k, \ \ k = 1, ..., N$$

$$m_x = \frac{1}{N}\sum_{i=1}^{N}\left(T_0 + k\right) = T_0 + \frac{N+1}{2}$$

$$\mathrm{var}\left(X\right) = \frac{N^2 - 1}{12}$$

$$\mathrm{cov}\left(X, Y\right) = \frac{1}{N}\sum_{k=1}^{N}\ \left(T_0 + k\right) y_k - \left(T_0 + \frac{N+1}{2}\right) m_y = \frac{1}{N}\sum_{k=1}^{N}\ k y_k - \frac{N+1}{2} m_y$$

and so the slope of the line is:

$$a = \frac{\frac{1}{N}\sum_{k=1}^{N}\ k y_k - \frac{N+1}{2} m_y}{\frac{N^2 - 1}{12}}$$

It is independent of T_0.

3. Asymptotic behavior

If the data y_k are bounded, $a \to 0$ when $N \to +\infty$. Indeed, let us assume $\left|y_k\right| \le C$; we have:

$$\left|a\right| \le \frac{\frac{1}{N}\sum_{k=1}^{N}\ kC + \frac{N+1}{2} C}{\frac{N^2 - 1}{12}} = \frac{\left(N+1\right)C}{\frac{N^2 - 1}{12}} \to 0 \text{ when } N \to +\infty.$$

The same conclusion holds if the y_k are slowly increasing, for instance $\left|y_k\right| \le C\sqrt{k}$, with similar computations.

On the other hand, if for all k we have $y_k = k$, we get

$$a = \frac{\frac{1}{N}\sum_{k=1}^{N}\ k^2 - \left(\frac{N+1}{2}\right)^2}{\frac{N^2 - 1}{12}} = 1 \text{ for all } N.$$

Let us see in what cases we may have $a > \varepsilon$, asymptotically when $N \to +\infty$.

The condition $a > \varepsilon$ is equivalent to:

$$\frac{1}{N}\sum_{k=1}^{N} ky_k > \varepsilon\frac{N^2-1}{12} + \frac{N+1}{2}m_y \quad (2)$$

Assume that (2) holds and that $y_k \geq 0$ for all k. Then, for all N:

$$\frac{1}{N}\sum_{k=1}^{N} ky_k > \varepsilon\frac{N^2-1}{12}$$

which gives approximately:

$$\sum_{k=1}^{N} ky_k > \frac{\varepsilon N^3}{12}$$

Let E be the set of indices k such that $y_k \leq \frac{\varepsilon k}{3}$; then:

$$\sum_{\substack{k=1 \\ k\in E}}^{N} ky_k \leq \sum_{\substack{k=1 \\ k\in E}}^{N} \frac{\varepsilon k^2}{5} \leq \frac{\varepsilon}{5}\sum_{k=1}^{N} k^2 \approx \frac{\varepsilon N^3}{15}$$

that is:

$$\sum_{\substack{k=1 \\ k\notin E}}^{N} ky_k \geq \frac{\varepsilon N^3}{12} - \frac{\varepsilon N^3}{15} = \frac{\varepsilon N^3}{60} \quad (3)$$

But if $k \notin E$, this means:

$$y_k > \frac{\varepsilon k}{3} \quad (4)$$

So we see that if (2) holds, a large proportion of the y_k satisfies (4), which means that they increase linearly.

4. Simple Random Walks

Let us now consider the case where the y_k come from a simple random walk. We have:

$$y_k = \sum_{j=1}^{k} X_j$$

where the X_j are independent random variables, taking the values ± 1 with probability $1/2$. Using the formulas of §2, we obtain:

$$a = \frac{\frac{1}{N}\sum_{k=1}^{N} k y_k - \frac{N+1}{2} m_y}{\frac{N^2-1}{12}} \tag{5}$$

But:

$$\begin{aligned} m_y &= \frac{1}{N}\sum_{k=1}^{N} y_k = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{k} X_j = \frac{1}{N}\left(X_1 + X_1 + X_2 + \cdots + X_1 + \cdots + X_N\right) \\ &= \frac{1}{N}\left(N X_1 + (N-1) X_2 + \cdots + X_N\right) = \frac{1}{N}\sum_{k=1}^{N} (N-k+1) X_k \end{aligned}$$

Moreover:

$$\begin{aligned} \sum_{k=1}^{N} k y_k &= \sum_{k=1}^{N} k \sum_{j=1}^{k} X_j = X_1 + 2\left(X_1 + X_2\right) + \cdots + k\left(X_1 + \cdots + X_k\right) + \cdots + N\left(X_1 + \cdots + X_N\right) \\ &= \left(1 + 2 + \cdots + N\right) X_1 + \left(2 + \cdots + N\right) X_2 + \cdots + \left(k + \cdots + N\right) X_k + \cdots + N X_N \\ &= \sum_{k=1}^{N} \left(\frac{N(N+1)}{2} - \frac{k(k-1)}{2} \right) X_k \end{aligned}$$

and we get:

$$\frac{1}{N}\sum_{k=1}^{N} k y_k = \frac{1}{N}\sum_{k=1}^{N} \left(\frac{N(N+1)}{2} - \frac{k(k-1)}{2} \right) X_k = \frac{N+1}{2}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} \frac{k(k-1)}{2} X_k$$

which gives:

$$\begin{aligned} a &= \frac{12}{N^2-1}\left(\frac{N+1}{2}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} \frac{k(k-1)}{2} X_k - \frac{N+1}{2}\frac{1}{N}\sum_{k=1}^{N} (N-k+1) X_k \right) \\ &= \frac{6}{N^2-1}\left(\left(-\frac{N+1}{N} \right)\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} k(k-1) X_k + \frac{N+1}{N}\sum_{k=1}^{N} k X_k \right) \end{aligned}$$

and finally:

$$a = \frac{6}{N\left(N^2-1\right)}\sum_{k=1}^{N} (k-1)(N+1-k) X_k$$

We will write the slope under the form:

$$a = \frac{6}{N\left(N^2-1\right)} \sum_{k=1}^{N-1} k\left(N-k\right) X_k \tag{6}$$

The expectation of a is zero, by symmetry. Since $\mathrm{var}\left(X_k\right)=1$ for all k, we have:

$$\mathrm{var}\left(a\right) = \left(\frac{6}{N\left(N^2-1\right)} \right)^2 \sum_{k=1}^{N} k^2\left(N-k\right)^2 = \frac{1}{5} \frac{N^2+1}{N\left(N^2-1\right)} \to 0 \text{ lorsque } N \to +\infty.$$

Let us consider a few simple cases:

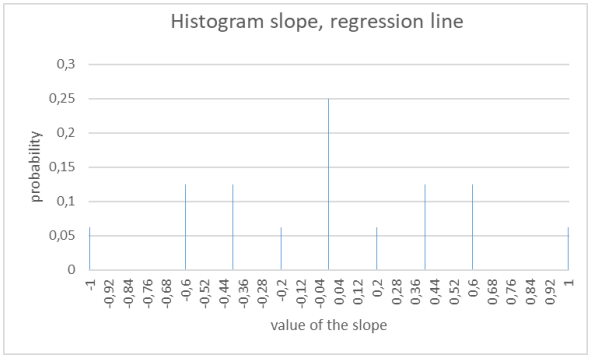
– Case $N=2$: $a = X_1$, values ± 1 with proba 1/2

– Case $N=3$: $a = \frac{1}{2}\left(X_1 + X_2\right)$, values -1 (proba 1/4), 0 (proba 1/2), 1 (proba 1/4)

– Case $N=4$: $a = \frac{1}{10}\left(3X_1 + 4X_2 + 3X_3\right)$; the values are 1 (proba 2/16), 0.4 (proba 4/16), 0.2 (proba 2/16), -0.2 (proba 2/16), -0.4 (proba 4/16), -1 (proba 2/16)

– Case $N=5$: $a = \frac{1}{20}\left(4X_1 + 6X_2 + 6X_3 + 4X_4\right)$; the values are 1 (proba 2/32), 0.6 (proba 4/32), 0.4 (proba 4/32), 0.2 (proba 2/32), 0 (proba 8/32), and the same for negative values.

Here is the histogram, in this case:



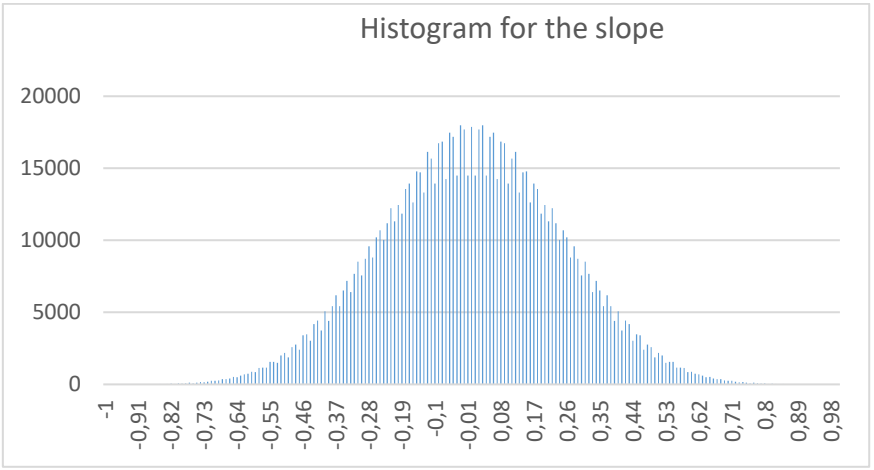
We observe that, if N is odd, the quantity $S = \sum_{j=0}^{N} j\left(N-j\right)\varepsilon_j$ (with $\varepsilon_j = \pm 1$) takes only even values. Indeed, in the above sum,

– either j is even, and then the corresponding term is even ;
– or j is odd, but then $N-j$ is even.

Case $N = 20$

We have $2^{20} = 1\,048\,576$ possible values for the X_k, $k = 1,...,20$; each will give a different slope for the regression line. All these slopes are between -1 and 1. We have $\frac{6}{N\left(N^2-1\right)} = \frac{1}{1330}$.

We obtain the following histogram:



This histogram is "discontinuous" and seems to be made of two different gaussians. Indeed, we observe that, for 1720 situations, we have $a = 0$ and, for 2712 situations, we have $a = \frac{1}{1330}$.

The explanation comes from Number Theroy. The sum $S = \sum_{j=0}^{N} j\left(N-j\right)X_j$ is a combination, with coefficients ± 1, of the numbers 19, 36, 51,64, 75, 84, 91, 96, 99, 100, 99, 96, 91, 84, 75, 64, 51, 36, 19.

If the sum is equal to 0, regrouping the pairs of equal numbers, we get:

$$S' = \sum_{j=0}^{N/2-1} \varepsilon_j j\left(N-j\right) + \varepsilon_{\frac{N}{2}} \frac{1}{2}\left(\frac{N}{2}\right)^2 = 0$$

where $\varepsilon_j = -1,0,1.$ We use only now the numbers 19, 36, 51,64, 75, 84, 91, 96, 99, 100/2, that is 19, 36, 50, 51, 64, 75, 84, 91, 96, 99.

Here is an example :

$$19 - 50 + 51 + 84 + 91 - 96 - 99 = 0$$

If the sum is equal to 2, regrouping the pairs of equal numbers, we get also:

$$S'' = \sum_{j=0}^{N/2-1} \varepsilon_j j(N-j) + \varepsilon_{\frac{N}{2}} \frac{1}{2}\left(\frac{N}{2}\right)^2 = 1$$

where $\varepsilon_j = -1, 0, 1.$ We use only the numbers 19, 36, 50, 51, 64, 75, 84, 91, 96, 99.

Let us look at the sets of numbers for which the first one (that is 39) has a positive coefficient, we find 35 for which the sum is 0 and 61 for which the sum is 1. The reason for this difference comes from the fact that, starting with two sets of sum 1, we can generate a set of sum 0, as soon as $\varepsilon_j \varepsilon'_j \geq 0$: the difference $\varepsilon_j - \varepsilon'_j$ will be 1, 0, -1. There are 195 couples of this type, for instance :

19, 36,50,0,0,0,0,91,-96,-99
and
19,0,50,51,64,0,-84,0,0,-99,
and the difference is:
0,36,0,-51,-64,0,84,91,-96,0.