HOW TO CHANGE THE STATEMENT OF A REGRESSION TASK A LITTLE AND GET A DEMING REGRESSION, OR HOW TO ESTIMATE A CLOUD OF POINTS

by

Panchenko Yehor

Abstract. — In this article, I approximate the points by a straight line under the assumption that there are no errors. Then I compare this result with the solution of the Deming Regression task.

Contents

1. Introduction

The statement of regression task is to estimate $f(x) = E(y/\xi = x)$. In this definition, we consider that x - is a non-random value. Now we change the statement to the task of Deming Regression. Let us estimate another function $g(x) = E(y/\xi \sim N(x, \sigma^2))$.

I wish to thank I.Y.Kaniovska for the précis [Kan21], thanks to which I have learned Statistics.

You can ask: "What it means?". The motivation may be as follows. Let's consider a 2d case. Suppose we have a cloud of points. We want to draw a line that describes the behaviour of these points in the best way (Fig. 1).

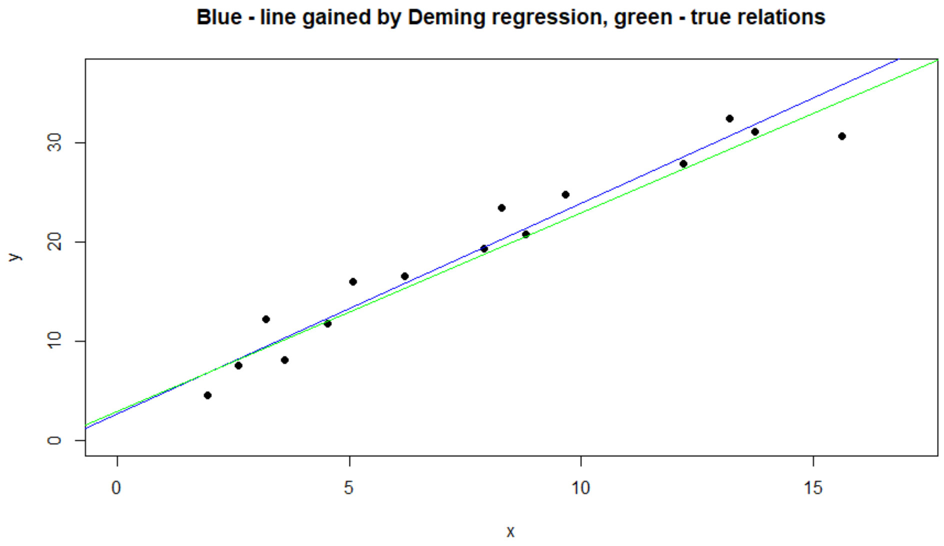

Figure 1. Two lines that describe points

When we consider the usual task of regression, we suppose that only y contains a noise. However, there is another point of view - when both x and y contain a noise (Fig. 2).

It would be useful in the numerical calculation because every measuring contains a mistake. Especially, it relates to computer calculation, where we cannot operate with all real numbers. However, I am sure that there is more application and experts in their domain could find something useful.

However, I want to solve another task and then compare results with Deming Regression. Suppose that points **don't** have the errors, or, equivalently, that $\sigma^2 = 0$. The main goal would be to find a line that describes a cloud of points and compare results for the posed task and Deming Regression.

2. Statement of the Deming Regression task

Let us consider n points - (x_i, y_i) for $i = \overline{1, n}$. Suppose that each point has an error - i.e. $\xi_i \sim N(X_i, \sigma^2)$ and $\eta_i \sim N(Y_i, \sigma^2)$ and (x_i, y_i) is a realization of ξ_i and η_i respectively.

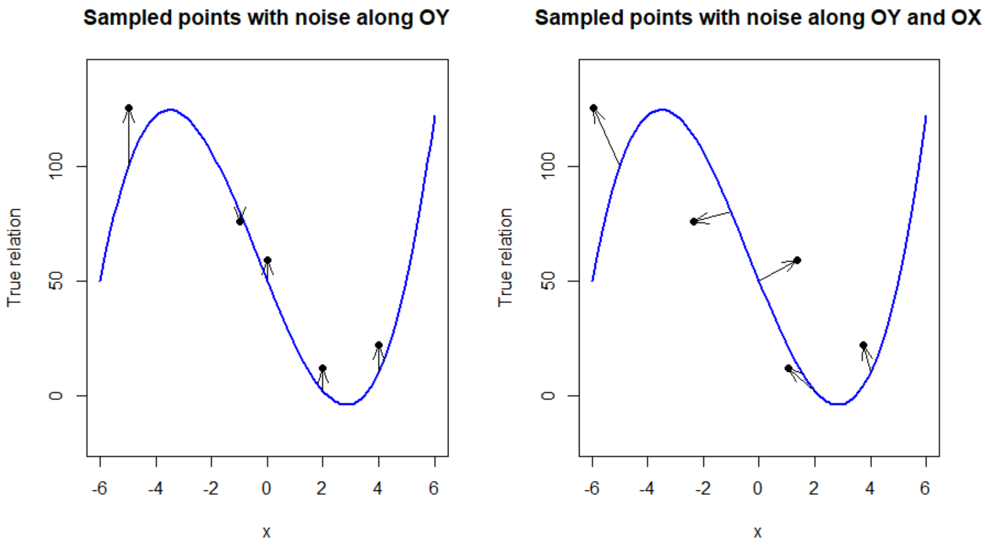

Figure 2. Errors along one or both axis

Consider a function $g(x) = E(y/\xi \sim N(x, \sigma^2))$. Let's estimate function g linearly - $g(x) = kx + b$. Our goal is to find values k and b what minimizes error, knowing points (x_i, y_i) and deviation σ.

3. Statement of the posed task

Let us consider n points - (x_i, y_i) for $i = \overline{1, n}$. Suppose that each point **hasn't** an error - i.e. $\xi_i \sim N(X_i, 0)$ and $\eta_i \sim N(Y_i, 0)$ and (x_i, y_i) is a realization of ξ_i and η_i respectively.

Let's find a function $g(x) = kx + b$ that estimates a cloud of points. Our goal is to find values k and b what minimizes error, knowing points (x_i, y_i).

4. The functional that should be minimized

Let's consider $\varepsilon_{y,i}$ and $\varepsilon_{x,i}$ - distances between i-th point and line (See Fig.3 for more details).

To be precise:

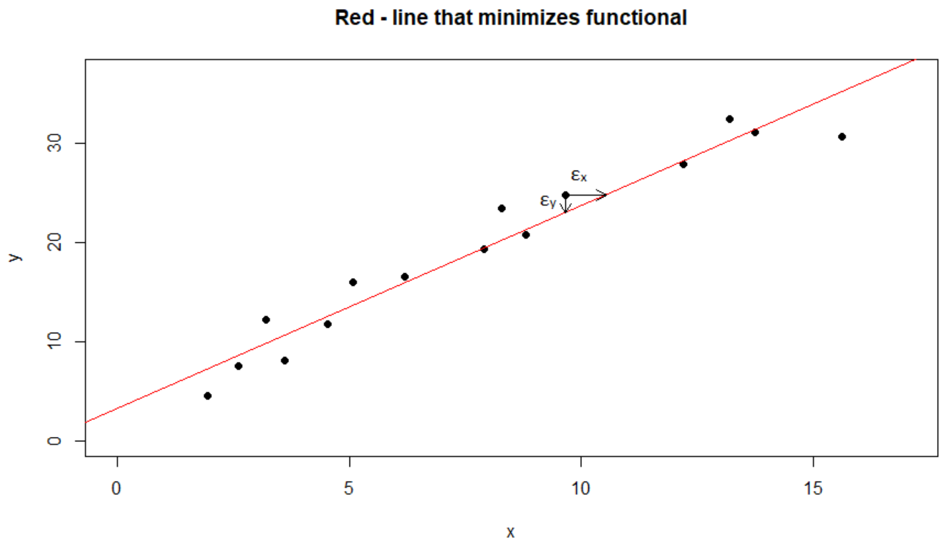

Figure 3. Distance to the line

$$\varepsilon_{y,i} = y_i - kx_i - b;$$
$$\varepsilon_{x,i} = \frac{y_i - b}{k} - x_i.$$

The functional that we want to minimize is $F(k,b) = \sum(\varepsilon_y^2 + \varepsilon_x^2)$ (when index i is omitted - it means sum over all $i = \overline{1,n}$).

Let's expand ε_x^2 and ε_y^2.

$$\varepsilon_{y,i}^2 = (y_i - kx_i - b)^2; \qquad \varepsilon_{x,i}^2 = (\frac{y_i - b}{k} - x_i)^2 = \frac{\varepsilon_{y,i}}{k}^2;$$

Therefore, F(k, b) can be represented like $F(k,b) = (1 + \frac{1}{k^2}) \sum(y_i - kx_i - b)^2$.

The following task is $(1 + \frac{1}{k^2}) \sum(y_i - kx_i - b)^2 \underset{k,b}{\rightarrow} min.$

5. Partial derivative by b

Let's consider $\frac{\partial F}{\partial b}$ and solve $\frac{\partial F}{\partial b} = 0$.

$$\frac{\partial F}{\partial b} = 2(1 + \frac{1}{k^2}) \sum (b + kx_i - y_i);$$
$$\frac{\partial F}{\partial b} = 0 \iff b = \frac{\sum (y_i - kx_i)}{n} \text{ or } k = \frac{\sum (y_i - b)}{\sum x_i}.$$

The essence meaning of parameter b is clear - for fixed k, parameter b is chosen to minimize the quadratic function. These things are very similar to MSE in Linear Regression - in fact, we solve two MSE along each axis, and those minimums are reached at the same point.

Also, we can notice that k and b are linearly dependent. It means that by substituting k with b or vice versa, we don't change the degree of a polynomial.

6. Partial derivative by k

Let's consider $\frac{\partial F}{\partial k}$ and solve $\frac{\partial F}{\partial k} = 0$.

$$\frac{\partial F}{\partial k} = \frac{-2}{k^3} \sum (y_i - kx_i - b)^2 + 2(1 + \frac{1}{k^2}) \sum (b + kx_i - y_i)x_i;$$
$$\frac{\partial F}{\partial k} = 0 \iff (k^3 + k) \sum (b + kx_i - y_i)x_i - \sum (b + kx_i - y_i)^2 = 0.$$

An important notice is that here we should find the root(s) of the polynomial of 4-th degree. It could be done using the Ferrari method, but there would be complicated calculations.

Also, I have seen a lot of calculations in [Jen07]. I indeed advise you to read this article, if you want to see how it can be done by honest and straightforward calculations. But in this article, I want to describe ideas that I used when I had tried to derive the formula for k.

6.1. Analysis of behaviour of $\frac{\partial F}{\partial k}$. — I calculated all derivatives of $\frac{\partial F}{\partial k}$, but here I didn't reach any results because the calculation is a huge problem, and I hadn't found out any pattern. I provide these derivatives:

$$\frac{\partial F}{\partial k^2} = 3k^2 \sum (b + kx_i - y_i)x_i + (k^3 + k) \sum x_i^2;$$
$$\frac{\partial F}{\partial k^3} = 6k \sum (b + kx_i - y_i)x_i + (6k^2 + 1) \sum x_i^2;$$
$$\frac{\partial F}{\partial k^4} = (24k + 1) \sum x_i^2 + \sum 6x_i(b - y_i);$$
$$\frac{\partial F}{\partial k^4} = 0 \iff k = \frac{\sum 6x_i(y_i - b) - \sum x_i^2}{24 \sum x_i^2}.$$

6.2. Let x_n and y_n free. — The equality $\frac{\partial F}{\partial k} = 0$ is equivalent to $\sum(b + kx_i - y_i) \cdot (k^3x_i + y_i - b) = 0$.

Let's find out that this sum could be rewritten via a dot product.

$$\begin{aligned}
&\sum(b + kx_i - y_i) \cdot (k^3x_i + y_i - b) = 0 \iff \\
&(b + kx_i - y_i, k^3x_i + y_i - b) = 0 \iff \\
&(\varepsilon_y, k^3x_i + y_i - b) = 0.
\end{aligned}$$

It means that the vector of errors along axis OY is orthogonal to the strange vector.

Suppose that all variables are fixed except the last point (x_n, y_n). Therefore, we can find a relation between coordinates of the last point and finally compare this relation with factual coordinates of this point. We can obtain an equation that, maybe, can be solved easier.

6.3. Analysis of coefficients at k^j. — The equality can be rewritten in the form

$$k^4\sum x_i^2 + k^3\sum x_i(b - y_i) - k\sum x_i(b - y_i) - \sum(y_i - b)^2 = 0.$$

Consider two arrays: $\alpha_i = x_i$ and $\beta_i = b - y_i$. Then the equality can be to shorten to

$$k^4\sum \alpha_i^2 + (k^3 - k)\sum \alpha_i\beta_i - \sum \beta_i^2 = 0.$$

The equation in this form indeed motivates to solve his.

6.4. Ferrari method. — Just substitute b with k in the equation from the previous subsection. You have the equation of 4-th degree and the two coefficients sum to zero. It, maybe, make your calculations easier.

6.5. Continue to analyse coefficients at k^j. — Since we have two dependent variables - k and b, let's eliminate b using their relationship. After that, we obtain a polynomial of 4-th degree. His coefficients are (letter with overline is arithmetic means):

$$\begin{aligned}
&k^4 : \sum(x_i - \overline{x})^2; \\
&k^3 : -\sum(x_i - \overline{x})(y_i - \overline{y}); \\
&k^2 : 0; \\
&k : \sum(x_i - \overline{x})(y_i - \overline{y}); \\
&k^0 : -\sum(y_i - \overline{y})^2.
\end{aligned}$$

We obtain the equation of 4-th degree without the coefficient at k^2. Notice that all coefficients look like variance or covariance. Also, this equation has a positive real root: at 0 the polynomial is less than zero and, the coefficient at the highest degree is positive.

This quartic equation is:

$$\sum(x_i-\overline{x})^2k^4-\sum(x_i-\overline{x})(y_i-\overline{y})k^3+\sum(x_i-\overline{x})(y_i-\overline{y})k-\sum(y_i-\overline{y})^2=0.$$

7. Comparing results with Deming Regression

Let's summarize a little what we have done. We found a line that minimizes functional $F(k, b)$. In other words, we obtained a line that goes through a cloud of points. We should make an experiment to reveal differences between two lines - gained by posed task and gained by the solution of Deming Regression. I will use formulas from [Jen07].

Let's compare coefficients k and b with coefficient α and β from [Jen07]. The first observation is that free coefficients in both articles have the same relationship with a slope coefficient. Also, the question could be revealed: "We didn't find a root analytically, but, maybe, the formula for β is a root of our quadric equation?". The answer is: "No." It looks naive to expect that result coincides. We don't use any probability, in contrast to the Deming Regression task statement. In Fig.5, you could see two straight lines that show the results of two tasks (I chose λ to be 1 because I haven't used any kind of asymmetry between x and y in the problem statement).

8. What's further?

The next step is to provide some experiments to reveal differences between slope coefficients.

9. Conclusion

What have we done??

To begin with, we defined a task to solve. The task was to find a line that describes a cloud of points in the assumption, that there are no errors. Then we concocted out the relations between b and k and derived an equation for k. The roots of this equation are critical points, and at some points, the minimum is achieved. Also, how I mentioned, that this equation always has a positive root.

At the second, we took a second to compare results with the solution of the Deming Regression task and found out that relations between k and b are the same. However, a simple experiment revealed that two straight lines, gained by two tasks, are different.

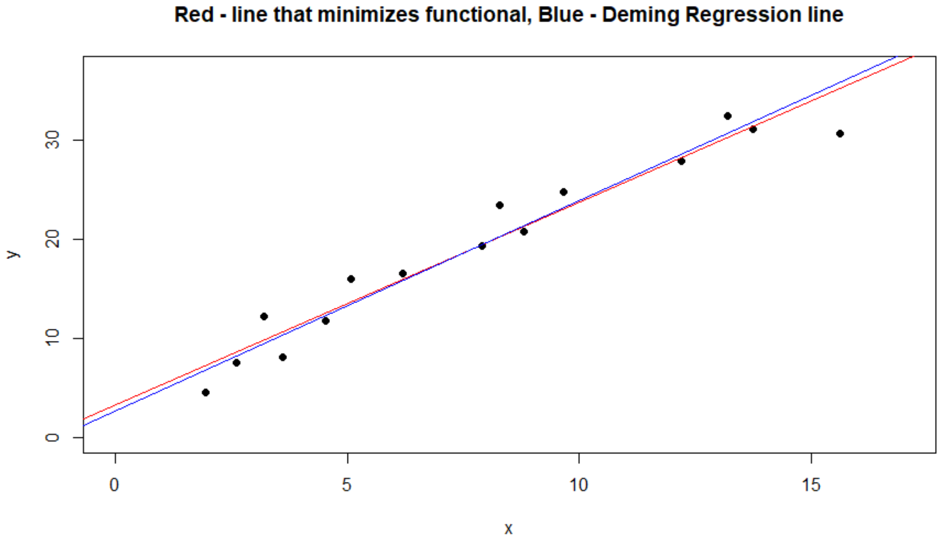

Figure 4. Lines that describe points

At the third, the work hasn't ended yet. Further researches should be with more rigorous details that I missed. Also, there should be more statistically significant conclusions about the differences between the two solutions.

References

[Jen07] A. C. Jensen – *Deming regression*, 2007.

[Kan21] I. Y. Kaniovska – *Synopsis of the probability theory and mathematical statistics lectures*, 2021.

July 20, 2021

• *E-mail :* `panchenko.yehor@lll.kpi.ua`