The Linear Regression Of Weighted Segments

George Daniel Mateescu[1]

Abstract. We proposed a regression model where the dependent variable is made not up of points but in segments. This situation corresponds to the markets throughout the day are observed values. Currency market or the stock exchange are examples in which values are different and variable throughout the day.

Keywords: regression, weighted segments, center of mass

JEL Classification: C02, C22, C58

The Model

In a previous paper ([3]) we introduced the notion of multiple points regression. Such regression refers to a set of data of the form:

$$\left(t_i, \left[x_i, y_i\right]\right), i = 1..n$$

The t_i values can be variables of time, and $\left[x_i, y_i\right]$ is a range of corresponding values. Such a situation is found for example in daily values of trading on the currency market, the stock market, etc.

For such data sets we used a calculation scheme for determining a regression straight shape

$$f\left(t\right) = at + b$$

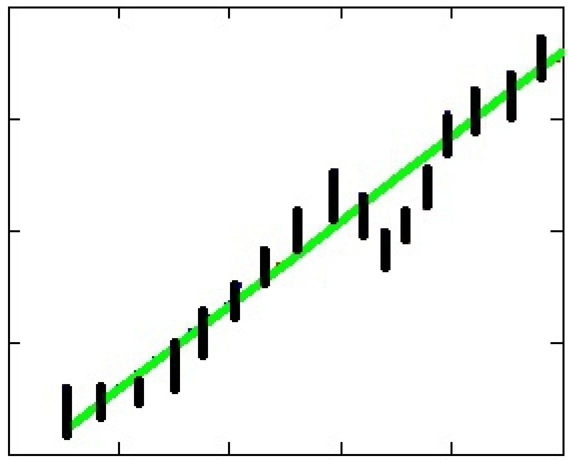
Where the coefficients *a* and *b* are obtained from the condition:

$$\min_{a,b} \sum_{i=1..n} \int_{x_i}^{y_i} \left(at_i + b - \tau\right)^2 d\tau$$

[1] e-mail: dan@mateescu.ro

This model found some properties of classical regression, for example if equidistant intervals are taken, but also in other situations, including the case of higher order regressions.

In the following we analyze a form for calculating weighted regression, to eflecte better reality. Thus, if foreign exchange transactions may be of interest not only exchange rate level or range during the day but the volume of transactions carried out at a certain value. In fact, the volume level can be reported even the most important indicator.



In this article, we put the question of reporting the "center of gravity" or "weighted average" of transactions.

Mathematically, each interval $\left[x_i, y_i\right]$ associate a density function $f_i:\left[x_i, y_i\right]\rightarrow\mathbb{R}_+$ for which we calculate "y" coordinate of the center of mass:

$$y_{G_i}=\frac{1}{M}\int_{[x_i,y_i]}yf_i ds \ ,$$

where M is the total mass of the segment $\left[x_i, y_i\right]$, $M_i=\int_{[x_i,y_i]}f_i ds$, and the notation $\int_{[x_i,y_i]}Fds$ is the line integral of the function F, along the $\left[x_i, y_i\right]$ segment.

We used parameterization $\begin{cases} x=t_i \\ y=\tau,\tau\in\left[x_i, y_i\right]\end{cases}$ and got:

$$M_i=\int_{x_i}^{y_i} f_i\left(\tau\right)d\tau \text{ si } y_{G_i}=\frac{1}{M}\int_{x_i}^{y_i}\tau f_i\left(\tau\right)d\tau$$

In applications, we will use the model:

$$\min_{a,b}\sum_{i=1..n}\left(at_i+b-y_{G_i}\right)^2$$

This normalized form corresponds to the limit $x_i\rightarrow y_i$, with classical regression when $x_i=y_i$

We notice that, mathematically we obtain the linear regression model for the data serie of centers of gravity $\left(t_i, y_{G_i}\right), i=1..n$

$$\min_{a,b} \sum_{i=1..n} \left(at_i + b - y_{G_i}\right)^2$$

For applications, we will use the solutions by applying the known steps:

$$\varphi(a,b) = \sum_{i=1..n} \left(at_i + b - y_{G_i}\right)^2$$

$$\frac{\partial \varphi}{\partial a} = \sum_{i=1..n} t_i \left[\left(at_i + b\right) - y_{G_i}\right] = 0$$

$$\frac{\partial \varphi}{\partial b} = \sum_{i=1..n} \left(at_i + b - y_{G_i}\right) = 0$$

or

$$a\sum_{i=1..n} t_i^2 + b\sum_{i=1..n} t_i = \sum_{i=1..n} y_{G_i} \cdot t_i$$

$$a\sum_{i=1..n} t_i + b \cdot n = \sum_{i=1..n} y_{G_i}$$

and by applying Cramer's rule, it results:

$$a = \frac{\begin{vmatrix} \sum_{i=1..n} y_{G_i} \cdot t_i & \sum_{i=1..n} t_i \\ \sum_{i=1..n} y_{G_i} & n \end{vmatrix}}{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} t_i \\ \sum_{i=1..n} t_i & n \end{vmatrix}} \; ; \; b = \frac{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} y_{G_i} \cdot t_i \\ \sum_{i=1..n} t_i & \sum_{i=1..n} y_{G_i} \end{vmatrix}}{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} t_i \\ \sum_{i=1..n} t_i & n \end{vmatrix}}$$

Applications

First, we consider the case of uniformly distributed values. This corresponds cu equal volumes for each value into the interval $\left[x_i, y_i\right]$. In our model of weighted segments this represents the case of constant density, i.e. $y_{G_i} = \dfrac{x_i + y_i}{2}$.

We calculate the regression line by using Mathcad functions, as is shown in that follow

$i := 1..10$ $\qquad\qquad$ $n := 10$

$$t := \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{pmatrix} \qquad x := \begin{pmatrix} 2 \\ 2.1 \\ 2.2 \\ 2.3 \\ 2.6 \\ 2.9 \\ 3.1 \\ 2.8 \\ 3.4 \\ 3.5 \\ 3.8 \end{pmatrix} \qquad y := \begin{pmatrix} 2.2 \\ 2.3 \\ 2.3 \\ 2.5 \\ 2.7 \\ 3.1 \\ 3.3 \\ 2.9 \\ 3.6 \\ 3.7 \\ 4 \end{pmatrix}$$

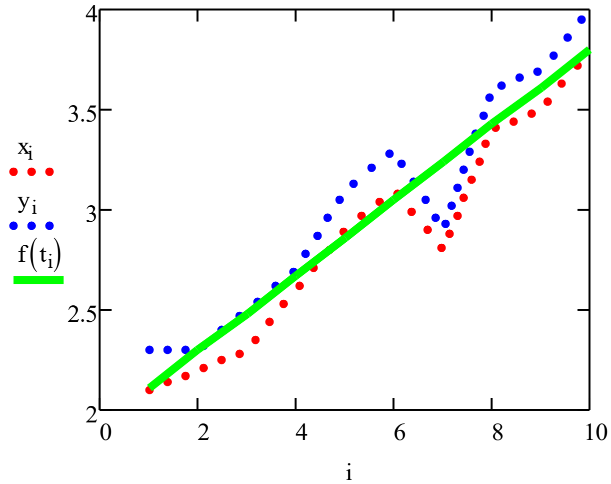
$$a = \frac{\begin{vmatrix} \sum_{i=1..n} \frac{y_i + x_i}{2} \cdot t_i & \sum_{i=1..n} t_i \\ \sum_{i=1..n} \frac{y_i + x_i}{2} & n \end{vmatrix}}{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} t_i \\ \sum_{i=1..n} t_i & n \end{vmatrix}}$$

$$a := \frac{\left|\begin{bmatrix} \sum_i \left(\frac{y_i + x_i}{2} \cdot t_i\right) & \sum_i t_i \\ \sum_i \frac{y_i + x_i}{2} & n \end{bmatrix}\right|}{\left|\begin{bmatrix} \sum_i (t_i)^2 & \sum_i t_i \\ \sum_i t_i & n \end{bmatrix}\right|}$$

$$b = \frac{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} \frac{y_i + x_i}{2} \cdot t_i \\ \sum_{i=1..n} t_i & \sum_{i=1..n} \frac{y_i + x_i}{2} \end{vmatrix}}{\begin{vmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} t_i \\ \sum_{i=1..n} t_i & n \end{vmatrix}}$$

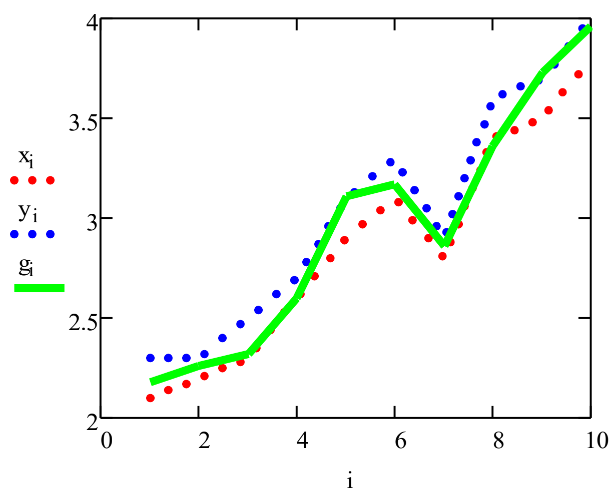
$$b := \frac{\left|\begin{bmatrix} \sum_i (t_i)^2 & \sum_i \left(\frac{y_i + x_i}{2} \cdot t_i\right) \\ \sum_i t_i & \sum_i \frac{y_i + x_i}{2} \end{bmatrix}\right|}{\left|\begin{bmatrix} \sum_i (t_i)^2 & \sum_i t_i \\ \sum_i t_i & n \end{bmatrix}\right|}$$

$a = 0.188$ $\qquad$ $b = 1.92$ $\qquad$ $f(u) := a \cdot u + b$



But, one significant case is the situation of variable volumes, i.e. variable density. In such a case, y_{G_i} may not be equal to $\frac{x_i + y_i}{2}$. As for example, we generated a normal distributed center of mass

$g_1 := \text{rnorm}\left(1, \frac{x_1 + y_1}{2}, 0.1\right)_0$	$g_2 := \text{rnorm}\left(1, \frac{x_2 + y_2}{2}, 0.1\right)_0$	$g_3 := \text{rnorm}\left(1, \frac{x_3 + y_3}{2}, 0.1\right)_0$
$g_4 := \text{rnorm}\left(1, \frac{x_4 + y_4}{2}, 0.1\right)_0$	$g_5 := \text{rnorm}\left(1, \frac{x_5 + y_5}{2}, 0.1\right)_0$	$g_6 := \text{rnorm}\left(1, \frac{x_6 + y_6}{2}, 0.1\right)_0$
$g_7 := \text{rnorm}\left(1, \frac{x_7 + y_7}{2}, 0.1\right)_0$	$g_8 := \text{rnorm}\left(1, \frac{x_8 + y_8}{2}, 0.1\right)_0$	$g_9 := \text{rnorm}\left(1, \frac{x_9 + y_9}{2}, 0.1\right)_0$
$g_{10} := \text{rnorm}\left(1, \frac{x_{10} + y_{10}}{2}, 0.1\right)_0$		



and the corresponding regression line.

Due to lack of real data (stocks, forex market) we cannot calculate some real values for the centre of mass, but we believe such data may be obtained and used.

References.

1. Dolado, J.J. and Lütkepohl, H., 1996. Making Wald Test Work for Cointegrated VAR Systems. Econometric Reviews, 15, pp.369-386.
2. Lütkepohl, H. and Kratzig, M., 2004. Applied Time Series Econometrics (Themes in Modern Econometrics). Cambridge University Press.
3. Mateescu, G.D., 2013. Multiple Points Regression, Romanian Journal of Ecomomic Forecasting, 2/2013, p.141-144