# Regression on intervals

*George Daniel Mateescu*[1]

**Abstract**. *In some previous papers* ([3],[4]) *we introduced and used a regression suitable for data series where the depended variable is not a value but a set of values. These values may be a discrete set or a continuous data. The economic correspondent of this mathematical approach is the exchange rate, where values are spread into an interval, during one day. Also, the stock exchange market is an example where indicators values are continuously variable during a day, etc.*

**Keywords**: regression

**JEL Classification**: C02, C22, C58

## The Model

In this paper we propose a model for study the dependence between two data series consisting in intervals:

$$[t_i, \tau_i],\ [u_i, v_i],\ i = 1..n$$

And we will search a linear dependency

$$[t_i, \tau_i] \to [at_i + b, a\tau_i + b]$$

where *a* and *b* are real (numbers) parameters.

In order to compare observed values, $[u_i, v_i]$ with the calculated $[at_i + b, a\tau_i + b]$ we

[1] Institute for Economic Forecasting, Romanian Academy, e-mail: dan@mateescu.ro

used the distance:

$$\left(u_i - at_i - b\right)^2 + \left(v_i - a\tau_i - b\right)^2$$

By consequence, the model will be:

$$\left[t_i, \tau_i\right],\ \left[u_i, v_i\right],\ i = 1..n$$

$$\min_{a,b} \sum_{i=1..n} \left[\left(u_i - at_i - b\right)^2 + \left(v_i - a\tau_i - b\right)^2\right]$$

Let us consider the linear (affine) function:

$$\varphi\left(a,b\right) = \sum_{i=1..n} \left[\left(u_i - at_i - b\right)^2 + \left(v_i - a\tau_i - b\right)^2\right]$$

and

$$\frac{\partial\varphi}{\partial a}\left(a,b\right) = \sum_{i=1..n} \left[-2t_i\left(u_i - at_i - b\right) - 2\tau_i\left(v_i - a\tau_i - b\right)\right] =$$
$$= 2a\sum_{i=1..n}\left(t_i^2 + \tau_i^2\right) + 2b\sum_{i=1..n}\left(t_i + \tau_i\right) - 2\sum_{i=1..n}\left(t_i u_i + \tau_i v_i\right)$$

$$\frac{\partial\varphi}{\partial b}\left(a,b\right) = \sum_{i=1..n} \left[-2\left(u_i - at_i - b\right) - 2\left(v_i - a\tau_i - b\right)\right] =$$
$$= 2a\sum_{i=1..n}\left(t_i + \tau_i\right) + 2b\cdot 2n - 2\sum_{i=1..n}\left(u_i + v_i\right)$$

We have to solve the system:

$$\frac{\partial\varphi}{\partial a} = 0$$
$$\frac{\partial\varphi}{\partial b} = 0$$

i.e:

$$a\sum_{i=1..n}\left(t_i^2 + \tau_i^2\right) + b\sum_{i=1..n}\left(t_i + \tau_i\right) = \sum_{i=1..n}\left(t_i u_i + \tau_i v_i\right)$$
$$a\sum_{i=1..n}\left(t_i + \tau_i\right) + 2bn = \sum_{i=1..n}\left(u_i + v_i\right)$$

and take into consideration the Hessian matrix:

$$\begin{pmatrix} \sum_{i=1..n}\left(t_i^2+\tau_i^2\right) & \sum_{i=1..n}\left(t_i+\tau_i\right) \\ \sum_{i=1..n}\left(t_i+\tau_i\right) & 2n \end{pmatrix}$$

If we split the Hessian into a sum:

$$\begin{pmatrix} \sum_{i=1..n}\left(t_i^2+\tau_i^2\right) & \sum_{i=1..n}\left(t_i+\tau_i\right) \\ \sum_{i=1..n}\left(t_i+\tau_i\right) & 2n \end{pmatrix} = \begin{pmatrix} \sum_{i=1..n} t_i^2 & \sum_{i=1..n} t_i \\ \sum_{i=1..n} t_i & n \end{pmatrix} + \begin{pmatrix} \sum_{i=1..n} \tau_i^2 & \sum_{i=1..n} \tau_i \\ \sum_{i=1..n} \tau_i & n \end{pmatrix}$$

then, taking into account the Schwartz inequality

$$\left(\sum_{i=1..n} t_i \cdot 1\right)^2 \le \left(\sum_{i=1..n} t_i^2\right)\cdot\left(\sum_{i=1..n} 1^2\right)$$

$$\left(\sum_{i=1..n} t_i\right)^2 \le n\left(\sum_{i=1..n} t_i^2\right)$$

it results that the Hessian matrix is positive definite. As consequence, the solution of the system equations:

$$\frac{\partial\varphi}{\partial a}=0$$

$$\frac{\partial\varphi}{\partial b}=0$$

is the minimum of the function $\varphi(a,b)=\sum_{i=1..n}\left[\left(u_i-at_i-b\right)^2+\left(v_i-a\tau_i-b\right)^2\right]$.

In order to find the minimum, we will solve the system of equations:

$$a\sum_{i=1..n}\left(t_i^2+\tau_i^2\right)+b\sum_{i=1..n}\left(t_i+\tau_i\right)=\sum_{i=1..n}\left(t_iu_i+\tau_iv_i\right)$$

$$a\sum_{i=1..n}\left(t_i+\tau_i\right)+2bn=\sum_{i=1..n}\left(u_i+v_i\right)$$

which has the solution:

$$a = \frac{2n\sum\limits_{i=1..n}\left(t_i u_i + \tau_i v_i\right) - \sum\limits_{i=1..n}\left(u_i + v_i\right)\sum\limits_{i=1..n}\left(t_i + \tau_i\right)}{2n\sum\limits_{i=1..n}\left(t_i^2 + \tau_i^2\right) - \left(\sum\limits_{i=1..n}\left(t_i + \tau_i\right)\right)^2}$$

$$b = \frac{\sum\limits_{i=1..n}\left(t_i^2 + \tau_i^2\right)\sum\limits_{i=1..n}\left(u_i + v_i\right) - \sum\limits_{i=1..n}\left(t_i + \tau_i\right)\sum\limits_{i=1..n}\left(t_i u_i + \tau_i v_i\right)}{2n\sum\limits_{i=1..n}\left(t_i^2 + \tau_i^2\right) - \left(\sum\limits_{i=1..n}\left(t_i + \tau_i\right)\right)^2}$$

**Numerical Approach**

We used unadjusted stock prices for Sundrug Co., Ltd. and Suzuken Co., Ltd. from the Tokyo Stock Exchange. There was a random selection of these companies, only for numerical purposes. Our assumption is there is no direct connection between stock prices for the selected companies. In this case, the model of intervals linear regression may act as an expression of interchangeability of values. In other words the values for the second data series may (or may be not) be obtained from de first series by an affine operation.

| | | Sundrug Co., Ltd. | | | Suzuken Co., Ltd. | |
|---|---|---|---|---|---|---|
| date | | max | min | | max | min |
| 9/1/2017 | | 4565 | 4515 | | 4050 | 4000 |
| 9/4/2017 | | 4535 | 4485 | | 4040 | 3985 |
| 9/5/2017 | | 4505 | 4400 | | 3980 | 3885 |
| 9/6/2017 | | 4510 | 4400 | | 3935 | 3875 |
| 9/7/2017 | | 4575 | 4500 | | 3960 | 3920 |
| 9/8/2017 | | 4595 | 4520 | | 3960 | 3895 |
| 9/11/2017 | | 4655 | 4580 | | 3975 | 3910 |
| 9/12/2017 | | 4665 | 4600 | | 3995 | 3935 |
| 9/13/2017 | | 4665 | 4580 | | 3975 | 3935 |
| 9/14/2017 | | 4765 | 4665 | | 3965 | 3930 |
| 9/15/2017 | | 4750 | 4640 | | 3975 | 3940 |

Data may be found on:

https://www.quandl.com/data/TSE/9989-Sundrug-Co-Ltd-9989 and

https://www.quandl.com/data/TSE/9987-Suzuken-Co-Ltd-9987

In order to calculate a linear regression for intervals we used *Mathcad* software and we calculate the coefficients $a = 0.156$ and $b = 3242$.

**In conclusion**, we proposed a new tool for investigating data series consisting in intervals, instead of values. Such a tool may be used in data series analysis for money exchange rates, stock exchange, etc.

For the moment, we don't have a tool in order to analyze any kind of statistical correlation between data series consisting of intervals, but we will investigate in the future.

**References**.


1. Dolado, J.J. and Lütkepohl, H., *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
4. Mateescu, G.D., *The Linear Regression Of Weighted Segments*, Working Papers of Institute for Economic Forecasting, http://www.ipe.ro/RePEc/WorkingPapers/wpiecf160720.pdf