9. Multiple Points Regression

George Daniel MATEESCU[1]

Abstract

We analyse the use of linear regression for a set of points $(x,y)\in R^2$ which don't belong to a function graph, i.e. there is more than one value of the endogenous variable y, that corresponds to the independent variable x. Discrete and continuous range of dependent variable are considered.

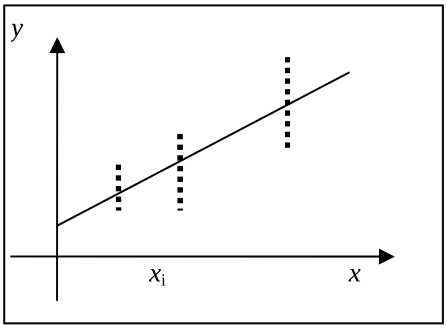
Keywords: regression, multiple values
JEL Classification: C02, C32

I. Discrete Range

As it is known, for a set of observations, $(x_i,y_i)\in R^2, i=1..m$, $x_i \neq x_j$ for $i \neq j$, the linear regression is represented by the function $y=\bar{a}x+\bar{b}$, where:

$$\varphi(\bar{a},\bar{b})=\min\varphi(a,b)$$

$$\varphi(a,b)=\sum_{i=1}^{m}(ax_i+b-y_i)^2$$



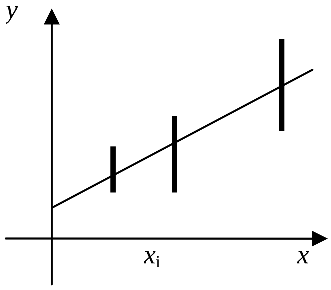
However, there are some time series for which there is many data observed, corresponding to the same "time". As for example, transactions into the exchange market or stock exchange values are variable during one day, etc. Consequently, it is necessary to find a linear function that adjusts multiple data points.

The model will be

$$\min\varphi(a,b)$$
$$\varphi(a,b)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(ax_i+b-y_{i,j})^2$$

where the multiple values $y_{i,j}, j=1..n_i$ correspond to $x_i, i=1..m$.

[1] *Institute for Economic Forecasting, Romanian Academy. E*-mail:*daniel@mateescu.ro*



Thus, it is easy to calculate:

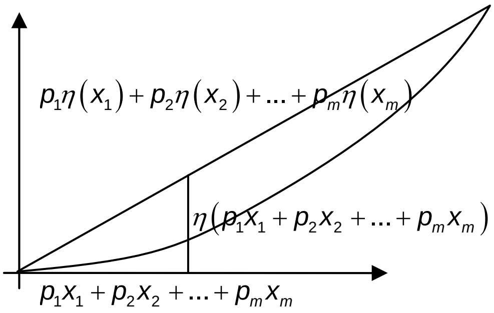
$$\frac{\partial\varphi}{\partial a}(a,b)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}2x_i\left(ax_i+b-y_{i,j}\right)=0\text{, and}$$

$$\frac{\partial\varphi}{\partial b}(a,b)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}2\left(ax_i+b-y_{i,j}\right)=0$$

The following system has an unique solution:

$$a\sum_{i=1}^{m}n_ix_i^2+b\sum_{i=1}^{m}n_ix_i=\sum_{i=1}^{m}\left(x_i\sum_{j=1}^{n_i}y_{i,j}\right)$$

$$a\sum_{i=1}^{m}n_ix_i+bN=\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{i,j}\text{, where }N=n_1+n_2+\ldots n_m\text{.}$$


$p_1\eta(x_1)+p_2\eta(x_2)+\ldots+p_m\eta(x_m)$
$\eta(p_1x_1+p_2x_2+\ldots+p_mx_m)$
$p_1x_1+p_2x_2+\ldots+p_mx_m$

Indeed, the system matrix is

$$A=\begin{pmatrix}\sum_{i=1}^{m}n_ix_i^2 & \sum_{j=1}^{n_i}n_ix_i\\ \sum_{j=1}^{n_i}n_ix_i & N\end{pmatrix}$$

that coincides with the *Hessian* matrix of function φ.

The ***A*** matrix is positive definite, as a consequence of the *Sylvester* criteria:

$$\Delta_1=\sum_{i=1}^{m}n_ix_i^2>0\text{ and}$$

$$\Delta_2=N\sum_{i=1}^{m}n_ix_i^2-\left[\sum_{i=1}^{m}n_ix_i\right]^2=N^2\left\{\sum_{i=1}^{m}\frac{n_i}{N}x_i^2-\left[\sum_{i=1}^{m}\frac{n_i}{N}x_i\right]^2\right\}$$

We recall the Jensen inequality:

$$\eta(p_1x_1+p_2x_2+\ldots+p_mx_m)\le p_1\eta(x_1)+p_2\eta(x_2)+\ldots+p_m\eta(x_m)$$

where $p_1+p_2+\ldots+p_m=1$ and the equality is true only if the *x*-values are not distinct; for the function $\eta(t)=t^2$ and $p_i=\frac{n_i}{N}$; the result is $\Delta_2>0$.

Finally, we get:

$$a=\frac{\left[\sum_{i=1}^{m}x_i\left(\sum_{j=1}^{n_i}y_{i,j}\right)\right]\cdot N-\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{i,j}\right)\cdot\left(\sum_{i=1}^{m}n_ix_i\right)}{\Delta_2}$$

$$b=\frac{\left(\sum_{i=1}^{m}n_ix_i^2\right)\cdot\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}y_{i,j}\right)-\left(\sum_{i=1}^{m}n_ix_i\right)\cdot\left[\sum_{i=1}^{m}x_i\left(\sum_{j=1}^{n_i}y_{i,j}\right)\right]}{\Delta_2}$$

II. Continuous Range

Furthermore, we consider the continuous range for the dependent variable, i.e. $\left(x_i,\left[u_i,v_i\right]\right), i=1..m$. Consequently, we define a distance type function by:

$$\varphi(a,b)=\sum_{i=1}^{m}\int_{u_i}^{v_i}\left(ax_i+b-t\right)^2 dt$$

and the minimum type problem:

$$\min\varphi(a,b) \qquad (\boldsymbol{P})$$

We calculate:

$$\frac{\partial\varphi}{\partial a}(a,b)=\sum_{i=1}^{m}2x_i\int_{u_i}^{v_i}\left(ax_i+b-t\right)dt=0$$

$$\frac{\partial\varphi}{\partial b}(a,b)=\sum_{i=1}^{m}2\int_{u_i}^{v_i}\left(ax_i+b-t\right)dt=0$$

and consequently:

$$a\sum_{i=1}^{m}(v_i-u_i)x_i^2+b\sum_{i=1}^{m}(v_i-u_i)x_i=\frac{1}{2}\sum_{i=1}^{m}x_i\left(v_i^2-u_i^2\right) \qquad (1)$$

$$a\sum_{i=1}^{m}(v_i-u_i)x_i+b\sum_{i=1}^{m}(v_i-u_i)=\frac{1}{2}\sum_{i=1}^{m}\left(v_i^2-u_i^2\right)$$

For $n_i=v_i-u_i$, the system matrix above has the same properties as in the case of the discrete range, even tough the n_i values are not natural numbers.

The solution, in the continuous case, is:

$$a=\frac{\left(\frac{1}{2}\sum_{i=1}^{m}x_i\left(v_i^2-u_i^2\right)\right)\cdot\left(\sum_{i=1}^{m}(v_i-u_i)\right)-\left(\frac{1}{2}\sum_{i=1}^{m}\left(v_i^2-u_i^2\right)\right)\cdot\left(\sum_{i=1}^{m}(v_i-u_i)x_i\right)}{\Delta_2}$$

$$b=\frac{\left(\sum_{i=1}^{m}(v_i-u_i)x_i^2\right)\cdot\left(\frac{1}{2}\sum_{i=1}^{m}\left(v_i^2-u_i^2\right)\right)-\left(\sum_{j=1}^{n_i}(v_i-u_i)x_i\right)\cdot\left(\frac{1}{2}\sum_{i=1}^{m}x_i\left(v_i^2-u_i^2\right)\right)}{\Delta_2}$$

III. Remarks and Application

Remark 1. If all the intervals $\left[u_i,v_i\right]$ have equal length, $v_i-u_i=c,\forall i=1..m$, then the continuous range case becomes the classical regression, applied to the central observed (dependent) value.

$$a\sum_{i=1}^{m}x_i^2+b\sum_{i=1}^{m}x_i=\sum_{i=1}^{m}x_i\frac{v_i+u_i}{2}$$

$$a\sum_{i=1}^{m} x_i + b\cdot m = \sum_{i=1}^{m}\frac{v_i + u_i}{2}$$

Remark 2. If the intervals (ranges) $[u_i, v_i]$ have the centre lying on the straight line $y = \alpha x + \beta$ then $a = \alpha$ and $b = \beta$.

By the substitution $t = \tau + \frac{v_i + u_i}{2} = \tau + \alpha x_i + \beta$, in each intergral from the definition of the function φ, we have:

$$\varphi(a,b) = \sum_{i=1}^{m}\int_{u_i}^{v_i}(ax_i + b - t)^2\,dt = \sum_{i=1}^{m}\int_{\frac{u_i - v_i}{2}}^{\frac{v_i - u_i}{2}}\left[(ax_i + b) - \tau - (\alpha x_i + \beta)\right]^2 d\tau$$

$$\varphi(a,b) = \sum_{i=1}^{m}\int_{\frac{u_i - v_i}{2}}^{\frac{v_i - u_i}{2}}\left\{\tau^2 + 2\tau\left[(\alpha x_i + \beta) - (ax_i + b)\right] + \left[(\alpha x_i + \beta) - (ax_i + b)\right]^2\right\}d\tau$$

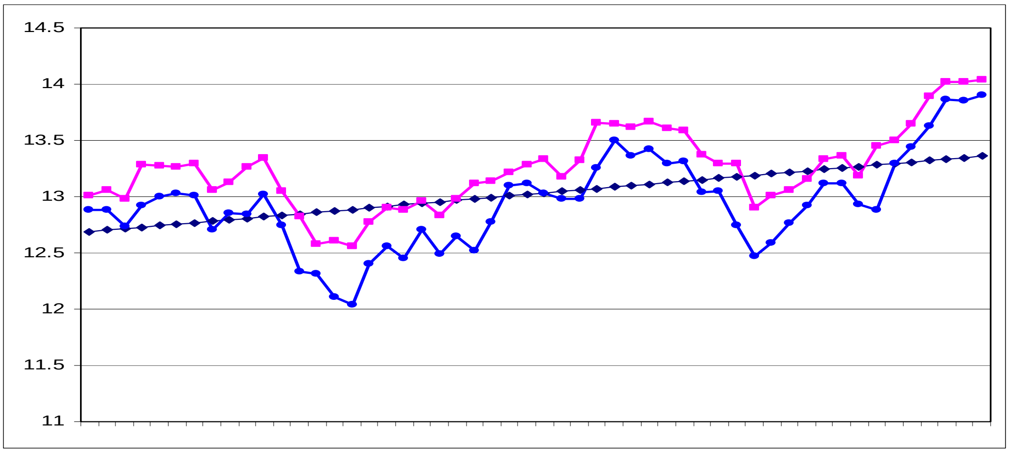
$$\varphi(a,b) = \sum_{i=1}^{m}\left\{\frac{2}{3}\left(\frac{v_i - u_i}{2}\right)^3 + \left[(\alpha x_i + \beta) - (ax_i + b)\right]^2 (v_i - u_i)\right\}$$

We can observe that all the values above are positive, so the minimum will be obtained if $(\alpha x_i + \beta) - (ax_i + b) = 0, \forall i = 1..m$, or:

$$(\alpha - a)x_1 = b - \beta,\ (\alpha - a)x_2 = b - \beta,\ \ldots\ (\alpha - a)x_m = b - \beta$$

Taking into account the condition $x_i \neq x_j$, it results $a = \alpha$ and $b = \beta$.

Application. We used daily maximum and minimum index values for *Dow Jones*(*), between 2/20/2012 and 2/11/2013. Data was divided by 1000 and we found *a*=0.131562 and *b*=12.67668. Maximum, minimum and regression line, calculated in the assumption of continuous range, are represented in the following graph:



(*) Data from http://investing.money.msn.com/investments/market-summary/