Tracking changes in the regional social-economic activity by a geometric fitting model: the Romanian case

ANDREEA ILUZIA IACOB*, CIPRIAN COSTIN POPESCU**, GINA CRISTINA DIMIAN*

*Department of Statistics and Econometrics, **Department of Mathematics
The Bucharest Academy of Economic Studies
15-17 Calea Dorobantilor, District 1, Bucharest
ROMANIA
aiacob@ase.ro, ciprian.popescu@csie.ase.ro, ginadimian@gmail.com

Abstract: This work deals with total least squares algorithm in the plane. An analytical method is given as an alternative to single value decomposition approach. A numerical application is also presented. It is referring to the bidimensional modeling of relevant indicators regarding regional social economic evolution. For this purpose, four representative Romanian Counties are studied.

Key-Words: orthogonal regression, hyperplane, fitting line, county, unemployment rate, overdue loan.

1 Introduction

The regression methods are very popular tools in statistical literature. Among them, orthogonal regression (OR) (or total least squares [2], [7]) plays a key role in the field of data analysis, especially when the sense of dependency between measured variables is rather unclear, [3], [4], [5]. Generally speaking, the OR method can be initiated as in the following lines. Let it be a cluster of \( n \) datapoints in \( \mathbb{R}^{p+1} \): \( P_k(x_k^{(i)},...,x_k^{(p)}, y_k) \), \( k = 1,n \). The OR approach is defined as follows:

\[
\min_{a_0,...,a_p} \sum_{k=1}^{n} \text{dist}^2 \left[ \left( x_k^{(i)},...,x_k^{(p)}, y_k \right), g_{a_0,...,a_p} \left( x_k^{(i)},...,x_k^{(p)} \right) \right] \tag{1}
\]

where

\[
g_{a_0,...,a_p} \left( x_k^{(i)},...,x_k^{(p)} \right) = a_0 + a_1x_k^{(1)} + ... + a_px_k^{(p)} \tag{2}
\]

is a function depending on parameters \( a_i \), \( i = 0,p \) (actually a fitting hyperplane [1]) and “dist” is the distance between \( P_k \) and \( g_{a_0,...,a_p} \). The points \( P_k \) are the image in \( \mathbb{R}^{p+1} \) of \( n \) measured values of the studied variables: \( X^{(i)},...,X^{(p)}, Y \). Essentially, the OR regression is divergent from other known regression methods due to the fact that the causality can be neglected. Any of \( X^{(j)} \), \( j = 1,p \) can be permuted with \( Y \). This is an important feature, if we are in the one of various situations when the role of variables (independent or dependent) is hard to assign, or even unaccomplishable. Furthermore, the method works for a fitting function depending on \( t \) parameters, when \( 2 \leq t \leq p + 1 \). For instance, in the plane it is a line but in a tridimensional space may be a line (Fig.2) or a plane [5].

2 Model Formulation

For a better grasp of conceptual disparity between OR and other regression methods, we will bargain...
Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation (Volume I)

for algorithm and will enunciate a short example [4], [5]. Let it be in the plane the following five datapoints: \( P_1(1,4), P_2(3,2), P_3(4,6), P_4(5,8), P_5(7,5) \) which correspond to five observed values of two variables \( X, Y \). For instance, in the classical least squares (LS) we have two possibilities, consequently, two different fitting lines: \( 0.45x - y + 3.2 = 0 \) (if \( Y \) is considered as dependent variable) and \( x - 0.45y - 1.75 = 0 \) (if \( X \) is the dependent variable). Unlike LS, the OR is free of dependence relation between \( X, Y \), and gives a unique line: \( x - y + 1 = 0 \). Notice that all the three lines have one point in common: \( C(4.5) \), called “centroid” (Fig.3). Each of the centroid coordinate is calculated as the arithmetic mean of the same dimension datapoints coordinates.

![Fig.3. OR over against LS](Image)

Some more interesting feature of OR regression are: the orthogonal distance looks more spontaneous then other measurements; the procedure is invariant to the system of coordinates, it’s also work in a case of non-Cartesian coordinates; in the plane, as we have already seen in the previous example, the fitting line is unique, the causality doesn’t matters a bit; the algorithm can be used for any number of dimensions [5]. In the next sections we consider two variables \( X, Y \) whose known (measured) values are: \( (x_k, y_k), k = 1, n \). For implementing the OR algorithm, it’s necessary to consider the datapoints \( P_k(x_k, y_k) \in \mathbb{R}^2 \), for all \( k = 1, n \). The scope is to find the line \( (L') \) such that the sum of squared distances from all \( P_k \) to \( (L') \) is lesser than sum of squared distances from \( P_k \) to any other line \( (L) \). It boils down to the problem:

\[
\min \sum_{k=1}^{n} \text{dist}^2(P_k, (L)) = \sum_{k=1}^{n} \text{dist}^2(P_k, (L'))
\] (3)

### 3 Model Solution

For the points \( P_k(x_k, y_k) \), we consider \( C(x_C, y_C) \) the centroid, whose coordinates are computed as follows:

\[
x_C = \frac{1}{n} \sum_{k=1}^{n} x_k \quad \text{and} \quad y_C = \frac{1}{n} \sum_{k=1}^{n} y_k
\] (4)

It is known that the equation of a non-vertical line \( (L) \) of slope \( m \), passing through a point \( P_0(x_0, y_0) \) is given by

\[
y - y_0 = m(x - x_0)
\] (5)

that is

\[
x = \frac{y - y_0}{m} + x_0
\] (6)

The distance from a point \( P_k(x_k, y_k) \) to \( (L) \) is

\[
\text{dist}(P_k, (L)) = \frac{|mx_k - y_k - m x_0 + y_0|}{\sqrt{m^2 + 1}}
\] (7)

Thus, the sum of squared distances can be put in the form

\[
\sum_{k=1}^{n} \text{dist}^2(P_k, (L)) = \sum_{k=1}^{n} \left( \frac{mx_k - y_k - m x_0 + y_0}{m^2 + 1} \right)^2
\] (8)

On the other hand, we have that

\[
\sum_{k=1}^{n} \left( mx_k - y_k - m x_0 + y_0 \right)^2 = \\
= \sum_{k=1}^{n} \left[ m(x_k - x_C) - (y_k - y_C) \right]^2 \\
- 2 \sum_{k=1}^{n} \left[ m(x_k - x_C) - (y_k - y_C) \right] \left[ m(x_0 - x_C) - (y_0 - y_C) \right] + \\
\sum_{k=1}^{n} \left[ m(x_0 - x_C) - (y_0 - y_C) \right]^2
\] (9)

due to the fact that

\[
\sum_{k=1}^{n} \left[ m(x_k - x_C) - (y_k - y_C) \right] = \\
m \sum_{k=1}^{n} (x_k - x_C) - \sum_{k=1}^{n} (y_k - y_C) = \\
m \left( \sum_{k=1}^{n} x_k - nx_C \right) + \left( \sum_{k=1}^{n} y_k - ny_C \right) = 0
\] (10)

Thus, for an arbitrary but fixed slope \( m \), the following relation holds:

\[
\sum_{k=1}^{n} \text{dist}^2(P_k, (L)) = 
\]
\[ = \frac{1}{m^2 + 1} \sum_{k=1}^{n} [m(x_k - x_C) - (y_k - y_C)]^2 + \]
\[ + \frac{1}{m^2 + 1} \sum_{k=1}^{n} [m(x_0 - x_C) - (y_0 - y_C)]^2 \geq \]
\[ \geq \frac{1}{m^2 + 1} \sum_{k=1}^{n} [m(x_k - x_C) - (y_k - y_C)]^2 \]

with equality if \( x_0 = x_C, y_0 = y_C \). Furthermore, for a vertical line \( (V) \) we obtain:
\[ \sum_{k=1}^{n} \text{dist}^2(P_k, (V)) = \sum_{k=1}^{n} (x_k - x_0)^2 = \]
\[ = \sum_{k=1}^{n} (x_k - x_C)^2 + \sum_{k=1}^{n} (x_0 - x_C)^2 \geq \sum_{k=1}^{n} (x_k - x_C)^2 \]

with equality if \( x_0 = x_C \). In this case, the minimum sum is exactly \( \sum_{k=1}^{n} (x_k - x_C)^2 \). In conclusion, for a given slope \( m \) (\( \infty \) for a vertical line), the centroid \( C(x_C, y_C) \) must be on the fitting line. A proof was also given in [1] where the fact that the centroid belongs to the ordinary LS fitting line was shown, too. From now on, the method based on matrix calculus [1], [2] (single value decomposition algorithm) will be replaced with an analytical approach. To that effect, we consider all the lines which contain the centroid. The next step is to find the slope of one of them, \( (L^*) \), for which sum of squared distances from \( P_k \) \( (k = 1, n) \) is minimum. At first, we consider only the non-vertical lines \( (L) \) which contain the centroid. Thus, a preliminary goal is to find the slope of that non-vertical line \( (L^*) \) for which the sum (regarded as a function of \( m \))
\[ S(m) = \sum_{k=1}^{n} \frac{(mx_k - y_k - mx_C + y_C)^2}{m^2 + 1} \]

is minimized. The first order derivative of \( S(m) \) with respect to \( m \) is given by
\[ S'(m) = \frac{2}{(m^2 + 1)^2} \sum_{k=1}^{n} [m(x_k - x_C) - (y_k - y_C)] \cdot [x_k - x_C + m(y_k - y_C)] \]

If, for all \( k = 1, n \) we denote \( a_k = x_k - x_C \) and \( b_k = y_k - y_C \), it results that
\[ m^2 \sum_{k=1}^{n} a_k b_k + m \sum_{k=1}^{n} (a_k^2 - b_k^2) - \sum_{k=1}^{n} a_k b_k = 0 \]
\[ (16) \]

or
\[ cm^2 + dm - c = 0 \]
\[ (17) \]

where
\[ c = \sum_{k=1}^{n} a_k b_k, d = \sum_{k=1}^{n} (a_k^2 - b_k^2) \]

The discriminant of the quadratic polynomial from (17) can be computed as \( \Delta = d^2 + 4c^2 \geq 0 \). Note that \( \Delta = 0 \) if and only if \( c = d = 0 \). If \( c, d \) are simultaneously null, from (13) we get
\[ S(m) = \frac{1}{m^2 + 1} \sum_{k=1}^{n} (ma_k - b_k)^2 = \]
\[ = \frac{1}{m^2 + 1} \left( m^2 \sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} b_k^2 - 2m \sum_{k=1}^{n} a_k b_k \right) = \]
\[ = \sum_{k=1}^{n} (x_k - x_C)^2 \sum_{k=1}^{n} (y_k - y_C)^2 \]

Thus \( (L^*) \) can be chosen as any line passing through \( C(x_C, y_C) \). In this case the solution is not unique.

After equalizing to zero the first order derivative we obtain
\[ m^2 \sum_{k=1}^{n} a_k b_k + m \sum_{k=1}^{n} (a_k^2 - b_k^2) - \sum_{k=1}^{n} a_k b_k = 0 \]
\[ (16) \]

or
\[ cm^2 + dm - c = 0 \]
\[ (17) \]

For example, we consider \( P_k(1,0), P_k(0,1), P_k(-1,0), P_k(0,-1) \). Thus the centroid is \( C(0,0) = O(0,0) \). The sum of squared distances from \( P_k, k = 1,4 \) to any line through the \( O(0,0) \) is equal to 2 (Fig.4). If at least one of \( c, d \) is non-zero then \( \Delta > 0 \). For \( c \neq 0 \), the equation (17) has two distinct real roots \( m_1, m_2, m_1 \neq m_2 \). The second order derivative of \( S(m) \) is:
When \( c < 0 \), the minimum is attained in the smallest root, namely in \( m_2 \). When \( c > 0 \), the minimum is attained in the biggest root, which is equal to \( m_2 \). In conclusion, the minimum of \( S(m) \) is attained in \( m_2 \).

Thus \( (L^*) \) has the equation

\[
y - y_C = m_2(x - x_C)
\]

(33)

and the minimum sum of squared distances is:

\[
S(m_2) = \frac{1}{m_2^2 + 1}\sum_{k=1}^{n}(m_2a_k-b_k)^2
\]

(34)

Remember that the previous result is valid only for non-vertical lines. We choose the final solution, \( m^* \), by including the vertical line which also contains the centroid. If

\[
\frac{1}{m_2^2 + 1}\sum_{k=1}^{n}(m_2a_k-b_k)^2 \geq \sum_{k=1}^{n}(x_k-x_C)^2
\]

(35)

the OR line \( (L') \) is the vertical line passing through centroid. Conversely, \( (L') \) coincide with \( (L^*) \).

4 Romanian case study

Orthogonal regression can be successfully employed in the process of defining the economic state of a region in a specified time period. In such a context, a new point of view regarding space description of national economies was given in [5]. In the present work we will also make effective use of the OR algorithm, but in a bidimensional context. We choose four representative Romanian Counties (Bucharest, Cluj, Timis and Iasi, formally denoted by C1, C2, C3, respectively C4) and two indicators, fairly relevant [8] for the economic crisis times: the unemployment rate [6] and the overdue loans, both calculated from September 2009 to May 2010. For a better visualization and comparison of regional economic status, the processed data can be displayed in the same framework, plane in our case (Fig.11). The use the OR approach brings along some advantages: instead of a fractionated information about each aspect (Fig.5-6) the OR provides information about selected indicators as a whole (Fig.7-10); as is suggested in [5], new economic indicators can be initiated after cases; in our circumstances such a indicator may be the normal line to the fitting line, the slopes, angles between regression lines etc. (Fig.11). Table 1 shows the unemployment rate (source: Romanian Ministry of Labor, Family and Social Protection / National Agency for Employment). Tables 2 and 3 show the overdue loans, expressed in RON (source: National Bank of Romania, Loans and Deposits by County / Financial Behaviour of Households and Companies by County).
Table 1. Unemployment rate for C1, C2, C3, C4

<table>
<thead>
<tr>
<th>Month</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 2009</td>
<td>2.00</td>
<td>5.60</td>
<td>4.00</td>
<td>7.10</td>
</tr>
<tr>
<td>Oct 2009</td>
<td>2.10</td>
<td>5.90</td>
<td>4.20</td>
<td>7.20</td>
</tr>
<tr>
<td>Nov 2009</td>
<td>2.20</td>
<td>6.10</td>
<td>4.40</td>
<td>7.40</td>
</tr>
<tr>
<td>Dec 2009</td>
<td>2.30</td>
<td>6.30</td>
<td>4.40</td>
<td>7.30</td>
</tr>
<tr>
<td>Jan 2010</td>
<td>2.40</td>
<td>6.50</td>
<td>4.40</td>
<td>7.40</td>
</tr>
<tr>
<td>Feb 2010</td>
<td>2.40</td>
<td>6.70</td>
<td>4.60</td>
<td>7.90</td>
</tr>
<tr>
<td>Mar 2010</td>
<td>2.53</td>
<td>6.87</td>
<td>4.50</td>
<td>8.18</td>
</tr>
<tr>
<td>Apr 2010</td>
<td>2.50</td>
<td>6.50</td>
<td>4.60</td>
<td>8.10</td>
</tr>
<tr>
<td>May 2010</td>
<td>2.50</td>
<td>6.20</td>
<td>4.40</td>
<td>7.80</td>
</tr>
</tbody>
</table>

The variation (rate) of overdue loans at a time \( t \) (in this case: month) is the ratio (see Table 4)

\[
r_t = \frac{V_t - V_{t-1}}{V_{t-1}}
\]

where \( V_t \) is the overdue loans volume at \( t \).

Table 4. Variation of overdue loans for C1, C2, C3, C4

<table>
<thead>
<tr>
<th>Month</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oct 2009</td>
<td>0.120</td>
<td>-0.188</td>
<td>-0.054</td>
<td>0.084</td>
</tr>
<tr>
<td>Nov 2009</td>
<td>0.027</td>
<td>0.228</td>
<td>0.102</td>
<td>0.162</td>
</tr>
<tr>
<td>Dec 2009</td>
<td>0.048</td>
<td>-0.214</td>
<td>-0.101</td>
<td>0.084</td>
</tr>
<tr>
<td>Jan 2010</td>
<td>0.044</td>
<td>0.136</td>
<td>0.247</td>
<td>0.042</td>
</tr>
<tr>
<td>Feb 2010</td>
<td>0.071</td>
<td>0.084</td>
<td>0.118</td>
<td>0.041</td>
</tr>
<tr>
<td>Mar 2010</td>
<td>0.032</td>
<td>0.042</td>
<td>-0.037</td>
<td>0.307</td>
</tr>
<tr>
<td>Apr 2010</td>
<td>0.113</td>
<td>0.102</td>
<td>0.010</td>
<td>0.020</td>
</tr>
<tr>
<td>May 2010</td>
<td>0.081</td>
<td>0.019</td>
<td>0.044</td>
<td>0.170</td>
</tr>
</tbody>
</table>

For collected data, the implementation of the orthogonal regression method ensues as in the following lines. First, to make a choice, consider \( X \) variable being the unemployment rate and \( Y \) as the overdue loans rate. In the case of Bucharest (C1), the centroid is the point \((2.3663, 0.067)\). After some calculation we obtain equation (17) as

\[
-0.00455m + 0.15894m + 0.00455 = 0
\]

Finally, the OR fitting line, \( (L_{C1}) \), has the equation

\[
0.028604x + y - 0.13469 = 0
\]

Fig.5. Monthly unemployment rate in the four selected Counties

Table 2. Overdue loans for C1, C2

<table>
<thead>
<tr>
<th>Month</th>
<th>Bucharest County</th>
<th>Cluj County</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 2009</td>
<td>843 821 191.29</td>
<td>165 946 251.08</td>
</tr>
<tr>
<td>Oct 2009</td>
<td>945 299 430.83</td>
<td>134 662 041.64</td>
</tr>
<tr>
<td>Nov 2009</td>
<td>970 951 031.32</td>
<td>165 365 550.30</td>
</tr>
<tr>
<td>Dec 2009</td>
<td>1 018 196 131.12</td>
<td>1 29 831 805.57</td>
</tr>
<tr>
<td>Jan 2010</td>
<td>1 063.3×10^6</td>
<td>147.6×10^6</td>
</tr>
<tr>
<td>Feb 2010</td>
<td>1 138.8×10^6</td>
<td>160.1×10^6</td>
</tr>
<tr>
<td>Mar 2010</td>
<td>1 175.9×10^6</td>
<td>166.9×10^6</td>
</tr>
<tr>
<td>Apr 2010</td>
<td>1 309.3×10^6</td>
<td>184.0×10^6</td>
</tr>
<tr>
<td>May 2010</td>
<td>1 415.4×10^6</td>
<td>187.6×10^6</td>
</tr>
</tbody>
</table>

Table 3. Overdue loans for C3, C4

<table>
<thead>
<tr>
<th>Month</th>
<th>Timis County</th>
<th>Iasi County</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 2009</td>
<td>164 084 965.42</td>
<td>88 634 781.74</td>
</tr>
<tr>
<td>Oct 2009</td>
<td>155 194 079.14</td>
<td>96 158 106.72</td>
</tr>
<tr>
<td>Nov 2009</td>
<td>171 137 377.35</td>
<td>111 741 516.61</td>
</tr>
<tr>
<td>Dec 2009</td>
<td>153 696 191.36</td>
<td>121 173 706.75</td>
</tr>
<tr>
<td>Jan 2010</td>
<td>191.7×10^6</td>
<td>126.3×10^6</td>
</tr>
<tr>
<td>Feb 2010</td>
<td>214.4×10^6</td>
<td>131.6×10^6</td>
</tr>
<tr>
<td>Mar 2010</td>
<td>206.4×10^6</td>
<td>172.1×10^6</td>
</tr>
<tr>
<td>Apr 2010</td>
<td>208.5×10^6</td>
<td>175.6×10^6</td>
</tr>
<tr>
<td>May 2010</td>
<td>217.8×10^6</td>
<td>205.6×10^6</td>
</tr>
</tbody>
</table>

Fig.6. Monthly changes in overdue loans in analyzed Counties

Fig.7. OR graphical solution for Bucharest County
For C2, C3, C4, the centroids are the points (6.3838, 0.026125), (4.4375, 0.041125), respectively (7.66, 0.11375), and the corresponding OR lines ($L^{*}_{Cj}$), $j = 2, 4$, have the following equations:

\[ 0.2002x - y - 1.2519 = 0 \]  \hspace{1cm} (39)
\[ 0.50124x - y - 2.1831 = 0 \]  \hspace{1cm} (40)
\[ 0.083461x - y - 0.52556 = 0 \]  \hspace{1cm} (41)

Both lines which are obtained from each of real roots $m_1, m_2$ are shown in Fig.7-10 (the fitting line as a thick line). The slope analysis can gives information about the monotonicity direction of $X$ relative to $Y$ and about the joint fluctuation speed.

By analyzing the slopes of the fitting lines, one can overall conclude that in Cluj, Timis and Iasi the unemployment rate has the same type of monotonicity as the rate of overdue loans. Bucharest is somehow atypical: the slope is light negative and the unemployment rate is quasi-stationary by comparison with the second variable, yet both have contrary monotonicity directions.

**Acknowledgements**

This work was supported by CNCSIS – UEFISCSU, project number PNII – IDEI code 1793/2008, financing contract no. 862/2009.

**References:**


