STATISTICA, anno LXX, n. 2, 2010

A SIMPLIFIED PROCEDURE OF LINEAR REGRESSION IN A PRELIMINARY ANALYSIS

S. Facchinetti, U. Magagnoli

1. INTRODUCTION

The use of the inferential statistical methodologies needs frequently to define preliminary procedures in order to choose a probabilistic model that has to be used in a second stage of the study. The aim of this study is to solve inferential problems of estimation, especially in case of large data set of observations, such as in economic, physical, biological, enviromental and technological fields.

In fact, from a point of view economic and methodologic, it seems useful to choose in a preliminary stage the models and examine them closely, with more complex methods, in a second stage. For example, in financial field large data sets referred to the different characteristics of the stocks, whose structural links have to be studied, are present. Furthermore, also in biology, large data sets have to be examined, in particular in the observation or in the experimental analyses.

This preliminary stage, that usually is called "identification of the model", consists in two components: structural and stochastic; it is better using simplier methodologies than the one of the likelihood, either from a formal point of view or a computational one. Indeed, in this way many optimal asymptotical properties of the likelihood procedure are lost. Consequently, the use of simplified methods must be evaluated in terms of efficiency loss, comparing, when possible, the information matrix or, simply, the variance or mean square error of the estimated parameters, always considering the unbias and verifying the consistency.

These simplified methods can be applied in order to define the dependence, not fixed in a parametric form, of a curve or a surface; or even when it is necessary substituting complex computations with simplier ones (Cox, 2006).

Moreover, our purpose is to evaluate the median applications of the different subsets of observed values, instead of the location index usually represented by the mean. This kind of procedure is useful in case of anomalous values, given by particular types of observation contaminations. Therefore, this method can be considered as a simplified one in the subject of the "robust" regression (Huber and Ronchetti, 2009).

This paper presents within the second paragraph the detailed operative procedure, the general considered regression model and the assumptions that allow to

verify the properties of the estimator regression coefficients, obtained by the simplified procedure. To obtain the regression function we use either the ordered means or the ordered medians. In particular, a polynomial structure of the first and the second order has been considered, in order to define the functional link between the two variables. Furthermore, it has been assumed that the random variable X and the error component, defined in the regression model, have a normal distribution.

In the third paragraph the results of the estimations obtained varying the parameters of the procedure for the polynomial model either of the first and the second order are presented. A reduced number of data than in the real analysis has been considered in details, because it allows to underline the limits of the procedure and some anomalous situations that can happen. Therefore a comparison between the simulation results and the one obtained by the ordinary least squares procedure is carried out.

Finally, the fourth paragraph gives some remarks about the statistical properties of the proposed procedure, in particular, with relation to the consistency of the estimators of the regression coefficients for the first and the second order models.

2. MODEL, ASSUMPTIONS AND SIMPLIFIED PROCEDURE

Let (X, Y) be the two characteristics of interest whose n observations: $S_n = \{(x_i, y_i); i = 1, 2, ..., n\}$ are known. In most cases the two random variables X and Y may be chosen in a set $(X_1, X_2, ..., X_g)$ of large size g. In particular X is the regressor or explicative variable and Y is the regressed one, whose behaviour will be studied related to X by the model:

$$Y = y(X) + \varepsilon \tag{1}$$

where ε is the error component, independent of X, with $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma_\varepsilon^2$.

The regression function $y(x)$, that generally belongs to the complete polynomial family of order r, is estimated by:

$$y_r(x) = \sum_{s=0}^{r} a_{sr} x^s \tag{2}$$

with $a_{sr} \in \Re^1$ and $a_{rr} \neq 0$ gathered in the vector $\mathbf{a_r} = (a_{0r}, ..., a_{rr})'$ of the parameters of order $(r+1) \times 1$.

The $y(x)$ function and the variance of the error component σ_ε^2 can be see respectively as conditioned mean $E\{Y|x\}$ and conditioned variance $Var\{Y|x\}$ (that is assumed as constant).

Among the considered assumptions the random variable X and the error component ε are assumed with a normal distribution $X \sim N(\mu_X, \sigma_X^2)$, with parameters

$\mu_X = 0$ and $\sigma_X^2 = 1$, without loss of generality, and $\varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$, with parameters $\mu_\varepsilon = 0$ and σ_ε^2 costant.

The proposed regression simplified procedure consists of four steps.

1. Let have n observations of the bivariate random variable (X, Y), $S_n = \{(x_i, y_i);\ i = 1, 2, ..., n\}$, ordered for the increasing values of X ($x_i \le x_{i+1}$), in the interval $I_n = [\min_i (x_i); \max_i (x_i)] \equiv [x_1; x_n]$.

2. The data set S_n is partitioned in $m \ge 2$ subsets $S_{[j]}$ for $j = 1, 2, ..., m$, of size $n_{[j]} = \lfloor n/m \rfloor$ or $n_{[j]} = \lfloor n/m \rfloor + 1$ nearly constant, with $\sum_{j=1}^{m} n_{[j]} = n$, defining $\lfloor u \rfloor$ the whole part of u, with $\bigcup_{j=1}^{m} S_{[j]} \equiv S_n$ and $S_j \cap S_k = \varnothing$ for $j \ne k$.
   Each subset is formed by the couples $(x_i, y_i) \in S_n$ according to the relation:
   $$S_{[j]} = \{(x_i, y_i) : N_{[j-1]} < i \le N_{[j]}\}$$
   where $N_{[j]} = \sum_{k=1}^{j} n_{[k]}$ and $N_{[0]} = 0$.

3. For each subset $S_{[j]}$ the observations are synthesized by the "mean" or by the "median":
   $$\overline{x}_{[j]} = \text{Mean}(x_i : x_i \in S_{[j]}),\ \overline{y}_{[j]} = \text{Mean}(y_i : y_i \in S_{[j]})$$
   $$\tilde{x}_{[j]} = \text{Median}(x_i : x_i \in S_{[j]}),\ \tilde{y}_{[j]} = \text{Median}(y_i : y_i \in S_{[j]}) \qquad (3)$$
   obtaining the points $P_{[j]} = (x_{[j]}, y_{[j]})$, specifically signed $\overline{P}_{[j]} = (\overline{x}_{[j]}, \overline{y}_{[j]})$ or $\tilde{P}_{[j]} = (\tilde{x}_{[j]}, \tilde{y}_{[j]})$.

4. The so located points $\overline{P}_{[j]}$ (or $\tilde{P}_{[j]}$) in $\Re^2$ give a "piecewise-linear regression" of $(m - 1)$ consecutive segments which can be assumed as a preliminary estimation of the regression function $y(x)$. Assuming a polynomial of order $r = m - 1$, it is possible a fitting through the m points $\overline{P}_{[j]}$ or $\tilde{P}_{[j]}$ that define "the regression procedure ordered by means or by medians", respectively. The parameters (regression coefficients) are obtained as the solutions of the following system of m linear equations in $r + 1$ unknown quantities:
   $$y_{[j]}(x) = \sum_{s=0}^{r} a_{sr} x_{[j]}^{s} \quad \text{for} \quad j = 1, 2, ..., m \qquad (4)$$

or as a matrix:

$$\mathbf{y_{[.]}} = \mathbf{P_{[.]}}\mathbf{a_r} \tag{5}$$

where $\mathbf{y_{[.]}} = (y_{[1]}, y_{[2]}, ..., y_{[m]})'$ is the $m \times 1$ vector of $y_{[j]}$, with $y_{[j]}$ equal to $\overline{y}_{[j]}$ or $\tilde{y}_{[j]}$, $\mathbf{a_r} = (a_{0r}, ..., a_{rr})'$ is the $(r + 1) \times 1$ vector of the regression coefficients and $\mathbf{P_{[.]}}$ is the $m \times (r + 1)$ invertible square matrix of $x_{[j]}$ ($\overline{x}_{[j]}$ or $\tilde{x}_{[j]}$) powers:

$$\mathbf{P_{[.]}} = \{p_{js} = x_{[j]}^{s}; j = 1, 2, ..., m; \;\; s = 0, 1, ..., r\} .$$

As the $\mathbf{P_{[.]}}$ matrix is invertible by construction, the vector of the coefficient estimators $\mathbf{a_r}$ is

$$\mathbf{a_r} = \mathbf{P_{[.]}^{-1}}\mathbf{y_{[.]}} \tag{6}$$

that gives the estimations of $\overline{\mathbf{a}}_\mathbf{r}$ or $\tilde{\mathbf{a}}_\mathbf{r}$, if we consider $\overline{P}_{[j]}$ or $\tilde{P}_{[j]}$, for $j = 1, 2, ..., m$.

If we may suppose that the polynomial function $y(x)$ has given by means of a polynomial in x of order r

$$y_r(x) = \sum_{s=0}^{r} a_{sr} x^s = \mathbf{a_r'}\mathbf{x} \quad \text{with} \quad \mathbf{x} = (1, x, ..., x^r)' \tag{7}$$

the proposed procedure allows to obtain the regression coefficient estimation $\mathbf{a_r}$, referred to the use of the $P_{[j]}$ points for $j = 1, 2, ..., m$ "means" ($\overline{P}_{[j]}$) or "medians" ($\tilde{P}_{[j]}$), respectively.

In this preliminary explorative research we study the polynomial functions of order $r = 1$ and $r = 2$ for their general approximation to the $y(x)$ function (Taylor theorem).

The polynomial function defined in (2) can be expressed by a series of orthogonal polynomial. In particular, among the different groups of orthogonal polynomials, we have considered the Hermite ones, relating to the possible values assumed by X $(-\infty, +\infty)$ and under the assumption of normal distribution for the values of the random variable X, whose standardized density function is consistent with the weight function $w(x) = \exp(-x^2)$.

The general function to define the Hermite polynomials (Abramovitz and Stegum, 1972) is the following:

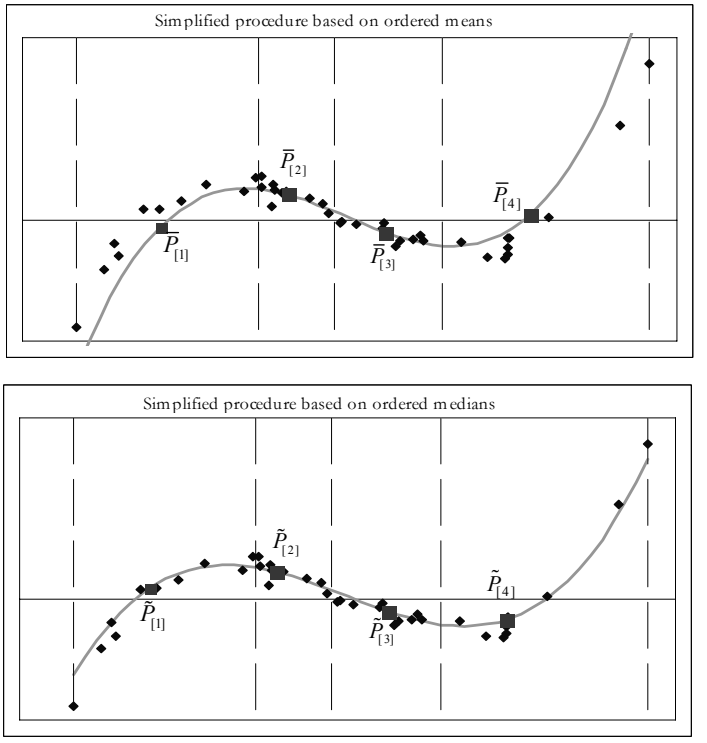

Figure 1 – The proposed simplified procedure: an example for $n = 30$, $m = 4$, $r = 3$.

$$h_r(x) = (-1)^r e^{x^2} \frac{d^r}{dx^r} e^{-x^2} . \tag{8}$$

In particular, for r from 0 to 5:

$$h_0(x) = 1$$
$$h_1(x) = 2x$$
$$h_2(x) = -2 + 4x^2$$
$$h_3(x) = -12x + 8x^3$$
$$h_4(x) = 12 - 48x^2 + 16x^4$$
$$h_5(x) = 120x - 160x^3 + 32x^5$$

and the polynomial function is

$$y_r(x) = \sum_{s=0}^{r} c_s h_s(x) . \tag{9}$$

By the (8) and (9) equations some linear functions, synthesized in the $\mathbf{T}_r$ matrix of order $(r + 1) \times (r + 1)$, $\mathbf{a_r}$ and $\mathbf{c_r} = (c_0, c_1 ..., c_r)'$ coefficients are obtained the ones from the others. In particular, the coefficients $\mathbf{a_r}$ of the polynomial of order r in x are obtained by assigning values to $\mathbf{c_r}$, that defines the function $y(x)$ in terms of Hermite polynomials:

$$\mathbf{a_r} = \mathbf{T}_r \mathbf{c_r} . \tag{10}$$

As $\mathbf{T}_r$ is a triangular matrix $(r + 1) \times (r + 1)$ it is ever invertible, therefore it is possible to obtain the estimations of the $\mathbf{c_r}$ parameters, if the $\mathbf{a_r}$ parameters, have been estimated by the proposed procedure:

$$\mathbf{c_r} = \mathbf{T}_r^{-1} \mathbf{a_r} . \tag{11}$$

In particular, for r from 0 to 5:

$$\mathbf{T}_5 = \begin{pmatrix} 1 & 0 & -2 & 0 & 12 & 0 \\ 0 & 2 & 0 & -12 & 0 & 120 \\ 0 & 0 & 4 & 0 & -48 & 0 \\ 0 & 0 & 0 & 8 & 0 & -160 \\ 0 & 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 32 \end{pmatrix}$$

where the $\mathbf{T}_0$, $\mathbf{T}_1$,..., $\mathbf{T}_4$ matrixes are the sub matrixes of $\mathbf{T}_5$. For example $\mathbf{T}_3 = \{t_{ks} = t_{ks} \in \mathbf{T}_5$, for $k, s = 1, ..., 4\}$.

2.1. *The first-order linear model*

For $r = 1$ the regression function is $y_1(x) = a_{01} + a_{11}x$ and the corresponding random variable is $Y = a_{01} + a_{11} X + \varepsilon$.

Under the assumption of normal distribution of X and ε, the random variable Y is normally distributed, too: $Y \sim N(\mu_Y, \sigma_Y^2)$, with $\mu_Y = a_{01} + a_{11}\mu_X$ and $\sigma_Y^2 = a_{11}^2\sigma_X^2 + \sigma_\varepsilon^2$.

The bidimensional random variable (X, Y), whose observations $\{(x_i, y_i), i = 1, 2, ..., n\}$ give a random simple sample of size n, is binormally distributed $N_2[(\mu_X, \mu_Y), (\sigma_X^2, \sigma_Y^2, \sigma_{XY} = \rho\sigma_X\sigma_Y)]$ with covariance $\sigma_{XY} = a_{11}\sigma_X^2$. Therefore, the parameters of the regression function $y_1(x)$ are given as function of those of the binormal random variable (X, Y):

$$a_{01} = \mu_Y - a_{11}\mu_X = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \qquad a_{11} = \rho\frac{\sigma_Y}{\sigma_X} . \tag{12}$$

In this case the data set S_n, ordered for increasing value of X, is partitioned in two subsets ($m = 2$) of equal size (nearly $n/2$). Then a straight line is drawn through the two points $P_{[j]}$, $j = 1, 2$: considered the ordered "mean" values $\overline{P}_{[j]}$ or the "median" ones $\tilde{P}_{[j]}$.

To verify the statistical properties of the estimators $\hat{a}_{01}$ and $\hat{a}_{11}$, for (X, Y) a standardised binormal distribution with $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$ is assumed, having $\sigma_{XY} = \rho$, $a_{01} = 0$ and $a_{11} = \rho$.

Moreover, replacing the Y with the X and ordering the data set S_n for increasing values of Y, the estimations of the regression coefficients of the straight line $x_1(y) = b_{01} + b_{11}y$ can be obtained.

The estimation of ρ results

$$\hat{\rho} = \begin{cases} 1 \ (\text{or} -1) & \text{if} \quad \operatorname{sign}(\hat{a}_{11})(\hat{a}_{11} \cdot \hat{b}_{11}) > 1 \ \ (\text{or} \ < -1) \\ \operatorname{sign}(\hat{a}_{11})\sqrt{\hat{a}_{11} \cdot \hat{b}_{11}} & \text{if} \quad 0 < \hat{a}_{11} \cdot \hat{b}_{11} < 1 \\ 0 & \text{if} \quad \operatorname{sign}(\hat{a}_{11}) \cdot \operatorname{sign}(\hat{b}_{11}) < 0 \end{cases}$$

where $\operatorname{sign}(a) = a / |a|$ for $a \neq 0$.

2.2. *The second-order linear model*

For $r = 2$ the regression function is $y_2(x) = a_{02} + a_{12}x + a_{22}x^2$ and the corresponding random variable is $Y = a_{02} + a_{12}X + a_{22}X^2 + \varepsilon$. Under the given distribution assumption for the random variable X and the error component ε, without loss of generality, we can assume $\mu_X = 0$ and $\sigma_X^2 = \sigma_\varepsilon^2 = 1$; applying the simplified procedure for ordered "means" or "medians", the data set S_n is partitioned into three subsets ($m = 3$) of equal size (nearly $n/3$) and the three points result $P_{[j]}$ for $j = 1, 2, 3$.

2.3. *Statistical properties of the regression coefficients obtained by the simplified procedure*

In this section, some considerations on the consistency of the estimators of the regression coefficients, obtained by the proposed procedure, are given.

We know that to verify the consistency, the variances of the estimators need to tend to zero, when n diverges, and under the stated distributive assumption, it occurs. Therefore, we only verify the asymptotic unbias of the estimators.

In particular, when $r = 1$ and n diverges, the two ordered data subsets $S_{[1]}$ and $S_{[2]}$ show higher or lower values, respectively, than the threshold value, that for the random variable X corresponds to the 50th percentile. Moreover, since X follows a standardised normal distribution, the threshold value is equal to zero. The

abscissas of the points $\overline{P}_{[1]}$ and $\overline{P}_{[2]}$ ($\overline{x}_{[1]}$ and $\overline{x}_{[2]}$) are, when n diverges, means of a truncated random variable:

$$\overline{x}_{[1]} \to E\{X \mid X \le 0\}; \quad \overline{x}_{[2]} \to E\{X \mid X \ge 0\}.$$

For the properties of the normal distribution of a standardised truncated random variable we have

$$E\{X \mid X \ge 0\} = \frac{\varphi(0)}{\Phi(0)} = \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}} \qquad E\{X \mid X \le 0\} = -\sqrt{\frac{2}{\pi}} \tag{13}$$

where $\varphi(z)$ and $\Phi(z)$ are the probability distribution function and the cumulative distribution function of the standardised normal random variable, respectively.

The ordinates of the points $\overline{P}_{[1]}$ and $\overline{P}_{[2]}$ ($\overline{y}_{[1]}$ and $\overline{y}_{[2]}$) are, when n diverges, the conditioned means of the random variable Y:

$$\overline{y}_{[1]} = E\{Y \mid X \le 0\} = E\{a_{01} + a_{11}(X \mid X \le 0) + \varepsilon\} =$$
$$= a_{01} + a_{11}E(X \mid X \le 0) = a_{01} + a_{11}\overline{x}_{[1]}$$

with $E\{\varepsilon\} = 0$, and

$$\overline{y}_{[2]} = E\{Y \mid X \ge 0\} = a_{01} + a_{11}\overline{x}_{[2]}.$$

The simplified procedure allows to obtain the convergence values of the estimators $\hat{a}_{01}$ and $\hat{a}_{11}$ as the solutions of the system of the equations:

$$\begin{cases} \overline{y}_{[1]} = \hat{a}_{01} + \hat{a}_{11}\overline{x}_{[1]} \\ \overline{y}_{[2]} = \hat{a}_{01} + \hat{a}_{11}\overline{x}_{[2]}. \end{cases}$$

They coincide with the parameters a_{01} and a_{11} of the model, so the estimates are consistent.

When $r = 2$ and n diverges, the three ordered data subsets $S_{[1]}$, $S_{[2]}$ and $S_{[3]}$ are characterised by two values of threshold corresponding to the 33.3th percentile and to the 66.7th percentile. By the given assumption of X as standardised normal random variable, the threshold values are, respectively:

$x_{33.3\%} = \Phi(0.333) = -0.43073 = -z_s$ and $x_{66.7\%} = \Phi(0.667) = 0.43073 = z_s$.

The abscissas $\overline{x}_{[1]}$, $\overline{x}_{[2]}$ and $\overline{x}_{[3]}$ of the three points $\overline{P}_{[1]}$, $\overline{P}_{[2]}$ and $\overline{P}_{[3]}$, when n diverges, correspond to the expected values of the truncated normal random variables:

$$\bar{x}_{[1]} \to E\{X \mid X \le -z_{\alpha}\};\ \bar{x}_{[2]} \to E\{X \mid |X| < z_{\alpha}\};\ \bar{x}_{[3]} \to E\{X \mid X \ge z_{\alpha}\}. \quad (14)$$

Furthermore, for the properties of the standardised truncated normal distributions (Johnson, Kotz and Balakrishnan, 1994-1995), we have

$$E\{X \mid X \le -z_{\alpha}\} = -\frac{\varphi(-z_{\alpha})}{1/3} = -1.0908$$
$$E\{X \mid |X| < z_{\alpha}\} = 0$$
$$E\{X \mid X \ge z_{\alpha}\} = 1.0908.$$

Hence:

$$\bar{x}_{[1]} = -1.0908 \qquad \bar{x}_{[2]} = 0 \qquad \bar{x}_{[3]} = 1.0908\,. \quad (15)$$

The second order moments of the three truncated normal random variables are

$$E\{X^2 \mid X \le -z_{\alpha}\} = E\{X^2 \mid X \ge z_{\alpha}\} = \frac{3}{\sqrt{\pi}}\left[\Gamma\left(\frac{3}{2}\right) - \Gamma\left(\frac{3}{2}, \frac{z_{\alpha}^2}{2}\right)\right] = 1.4698$$
$$E\{X^2 \mid |X| < z_{\alpha}\} = \frac{2}{\sqrt{\pi}/3}\left[\Gamma\left(\frac{3}{2}, \frac{z_{\alpha}^2}{2}\right)\right] = 0.0603 \quad (16)$$

where $\Gamma(a)$ and $\Gamma(a, u)$ are the Gamma and incomplete Gamma functions, respectively.

The $\bar{y}_{[1]}$, $\bar{y}_{[2]}$ and $\bar{y}_{[3]}$ ordinates of the three points $\bar{P}_{[1]}$, $\bar{P}_{[2]}$ and $\bar{P}_{[3]}$, when n diverges, correspond to the expected values of $Y = a_{02} + a_{12}X + a_{22}X^2$, where the conditioned means of X and X^2, before defined, have placed instead of X and X^2:

$$\begin{cases} \bar{y}_{[1]} \to E\{Y \mid X \le -z_{\alpha}\} = a_{02} + a_{12}E\{X \mid X \le -z_{\alpha}\} + a_{22}E\{X^2 \mid X \le -z_{\alpha}\} \\ \bar{y}_{[2]} \to E\{Y \mid |X| < z_{\alpha}\} = a_{02} + a_{12}E\{X \mid |X| < z_{\alpha}\} + a_{22}E\{X^2 \mid |X| < z_{\alpha}\} \\ \bar{y}_{[3]} \to E\{Y \mid X \ge z_{\alpha}\} = a_{02} + a_{12}E\{X \mid X \ge z_{\alpha}\} + a_{22}E\{X^2 \mid X \ge z_{\alpha}\}. \end{cases} \quad (17)$$

The $\hat{a}_{02}$, $\hat{a}_{12}$ and $\hat{a}_{22}$ regression estimator coefficients are obtained by the "ordered means" procedure as solutions of the following linear system:

$$\begin{cases} \bar{y}_{[1]} = \hat{a}_{02} + \hat{a}_{12}\bar{x}_{[1]} + \hat{a}_{22}\bar{x}_{[1]}^2 \\ \bar{y}_{[2]} = \hat{a}_{02} + \hat{a}_{12}\bar{x}_{[2]} + \hat{a}_{22}\bar{x}_{[2]}^2 \\ \bar{y}_{[3]} = \hat{a}_{02} + \hat{a}_{12}\bar{x}_{[3]} + \hat{a}_{22}\bar{x}_{[3]}^2. \end{cases} \quad (18)$$

Subtracting the respective equations of the (17) from the (18) and remembering the results of the (15):

$$\begin{cases} (\hat{a}_{02} - a_{02}) + (\hat{a}_{12} - a_{12})\bar{x}_{[1]} + \hat{a}_{22}\bar{x}_{[1]}^2 - a_{22}E\{X^2 \mid X \le -z_s\} = 0 \\ (\hat{a}_{02} - a_{02}) - a_{22}E\{X^2 \mid |X| < z_s\} = 0 \\ (\hat{a}_{02} - a_{02}) - (\hat{a}_{12} - a_{12})\bar{x}_{[1]} + \hat{a}_{22}\bar{x}_{[1]}^2 - a_{22}E\{X^2 \mid X \le -z_s\} = 0. \end{cases}$$

Furthermore, when n diverges, from (16) it is possible to obtain the estimators $\hat{a}_{02}$, $\hat{a}_{12}$ and $\hat{a}_{22}$, as function of a_{02}, a_{12} and a_{22}:

$$\begin{cases} \hat{a}_{02} = a_{02} + a_{22}E\{X^2 \mid |X| < z_s\} = a_{02} + 0.0603a_{22} \\ \hat{a}_{12} = a_{12} \\ \hat{a}_{22} = a_{22}[E\{X^2 \mid X \le -z_s\} - E\{X^2 \mid |X| < z_s\} / \bar{x}_{[1]}^2] = 1.1846a_{22}. \end{cases} \tag{19}$$

We note that the estimators $\hat{a}_{02}$ and $\hat{a}_{22}$ do not result consistent, while $\hat{a}_{12}$ converges to the corresponding parameter, when n diverges.

So we have to evaluate the bias of the estimators.

In particular, if the Hermite coefficients are $c_0 = c_1 = 0$ and $c_2 = 1$, corresponding to the parameters $a_{02} = -2$, $a_{12} = 0$, $a_{22} = 4$ and the mean values of the estimators $\hat{a}_{02}$, $\hat{a}_{12}$ and $\hat{a}_{22}$ are equal to:

$$\hat{a}_{02} = -1.7587$$
$$\hat{a}_{12} = 0$$
$$\hat{a}_{22} = 4.7385.$$

3. NUMERICAL SIMULATION: SOME RESULTS

In order to give not only methodological indications, but also operative ones, a Monte Carlo numerical simulation has been carried out.

The aim is to obtain the distribution of the estimator parameters of the polynomial model, to evaluate their properties and to compare them with the corresponding parameters obtained by the least square method.

To underline the properties of the simplified procedure $N = 1000$ replications have been made. Moreover, to emphasise the bias and the dispersion of the estimated parameters, we have considered low sample size $n = 10, 20, 30$.

This simulation has been carried out considering a sample size reduced with respect to a large data set, because from one side we have no problems relating to the number of the terms in calculating means or medians and on the other side the reduced number emphazises the possible biases in the estimations of the parameters of the regression models.

The models and the basic assumptions are those specified in the second paragraph, with particular reference to the polynomial models of order $r = 1$ and $r = 2$.

3.1. *The first-order linear model*

For the polynomial of the first order, we have considered ρ equal to 0, 0.25, 0.5 and 0.75[1]. The simplified procedure is the one given in paragraph 2.1..

In particular, for $n = 30$ and $\rho = 0.5$ the following Figure 2 shows the estimated models, considering the results obtained by the "means" procedure and by the "medians" procedure, with a comparison with the results obtained by the ordinary least square method.

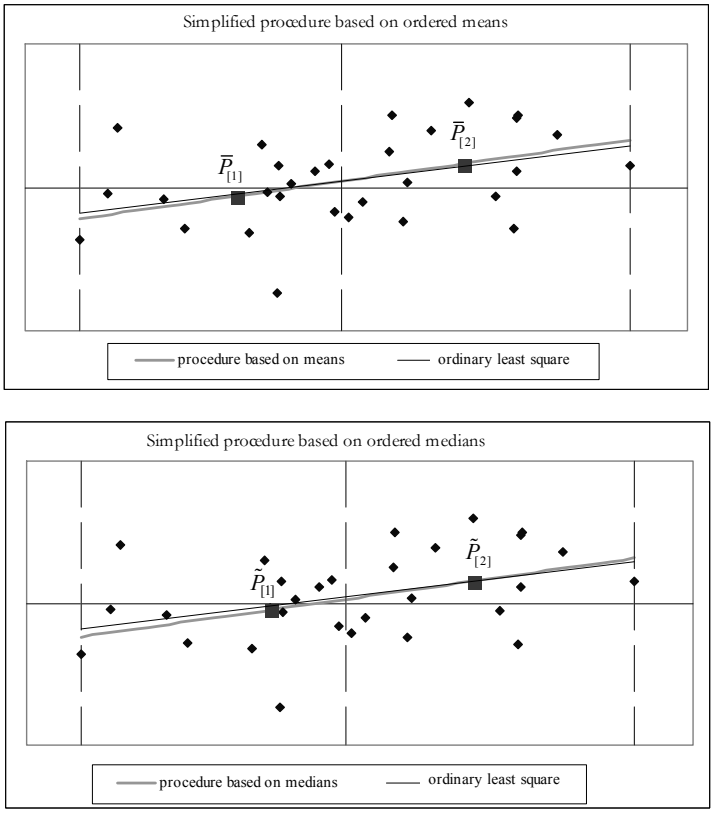

Figure 2 – The proposed simplified procedure: an example for $n = 30$, $m = 2$, $r = 1$.

In the graphs we can observe the partition in two subsets ($m = 2$) of size $n/2$ and the mean and the median points of each subset that define the models.

[1] The results obtained by simulations for $\rho > 0$ are equivalent to the corresponding ones for $\rho < 0$ and therefore it is not necessary to analyse them.

To verify the fitness of the proposed procedure, the tables 1, 2 and 3 give, for some values of n and ρ, the bias and the standard deviations (SD) of the estimated parameters for the ordered means procedure, the ordered medians procedure and the least square method, respectively.

The tables give only the values of the estimates of the regression coefficients for the function $y_1(x) = a_{01} + a_{11}x$, because those referred to the function $x_1(y) = b_{01} + b_{11}y$ are quite similar.

TABLE 1

Synthesized results of the simulation referred to the simplified procedure based on ordered means: N = 1000, r = 1

		$\hat{a}_{01}$		$\hat{a}_{11}$		$\hat{\rho}$	
n	ρ	BIAS	SD	BIAS	SD	BIAS	SD
10	0	0.009	0.352	0.000	0.483	0.003	0.353
10	0.25	0.008	0.341	0.000	0.467	-0.030	0.344
10	0.5	0.009	0.294	0.000	0.328	-0.019	0.257
10	0.75	0.006	0.233	0.000	0.319	-0.031	0.232
20	0	0.005	0.236	0.006	0.299	0.004	0.241
20	0.25	0.005	0.228	0.006	0.289	-0.021	0.236
20	0.5	0.004	0.204	0.006	0.259	-0.022	0.221
20	0.75	0.003	0.156	0.004	0.197	-0.013	0.149
30	0	0.004	0.188	0.009	0.237	0.008	0.192
30	0.25	0.004	0.182	0.009	0.229	-0.013	0.199
30	0.5	0.004	0.163	0.008	0.205	-0.008	0.172
30	0.75	0.003	0.124	0.006	0.157	-0.006	0.115

TABLE 2

Synthesized results of the simulation referred to the simplified procedure based on ordered medians: N =1000, r = 1

		$\hat{a}_{01}$		$\hat{a}_{11}$		$\hat{\rho}$	
n	ρ	BIAS	SD	BIAS	SD	BIAS	SD
10	0	0.018	0.452	0.019	0.745	-0.006	0.485
10	0.25	0.018	0.446	0.064	0.737	-0.009	0.470
10	0.5	0.012	0.421	0.105	0.683	0.006	0.416
10	0.75	0.010	0.345	0.119	0.556	0.020	0.296
20	0	0.011	0.280	0.011	0.424	0.007	0.355
20	0.25	0.011	0.274	0.055	0.413	0.021	0.327
20	0.5	0.010	0.255	0.087	0.385	0.037	0.284
20	0.75	0.005	0.215	0.103	0.327	0.054	0.183
30	0	0.004	0.237	0.015	0.354	0.010	0.300
30	0.25	0.001	0.236	0.065	0.347	0.032	0.286
30	0.5	-0.002	0.221	0.100	0.327	0.058	0.242
30	0.75	0.002	0.185	0.120	0.286	0.074	0.150

TABLE 3

Synthesized results of the simulation referred to the ordinary least square method: N = 1000, r = 1

		$\hat{a}_{01}$		$\hat{a}_{11}$		$\hat{\rho}$	
n	ρ	BIAS	SD	BIAS	SD	BIAS	SD
10	0	0.010	0.340	0.000	0.378	-0.002	0.329
10	0.25	0.010	0.329	0.000	0.366	-0.013	0.310
10	0.5	0.009	0.294	0.000	0.328	-0.019	0.257
10	0.75	0.007	0.225	0.000	0.250	-0.018	0.169
20	0	0.006	0.232	0.000	0.232	-0.002	0.219
20	0.25	0.005	0.225	0.000	0.224	-0.005	0.205
20	0.5	0.005	0.201	0.000	0.201	-0.008	0.169
20	0.75	0.004	0.153	0.000	0.153	-0.008	0.106
30	0	0.005	0.186	-0.010	0.180	0.000	0.177
30	0.25	0.005	0.180	-0.001	0.175	-0.161	0.087
30	0.5	0.004	0.161	-0.001	0.156	-0.006	0.135
30	0.75	0.003	0.123	-0.001	0.119	-0.005	0.083

We can observe that the estimator parameters defined by the means procedure present a smaller bias than those based on the medians. Moreover in the three tables there are two particular situations in which the value assigned to the ρ estimates has conventionally defined:

1. when the product of the estimated regression coefficients exceed the unity: in this case $\hat{\rho}$ is assumed equal to ± 1 according to the common sign of the two coefficients;
2. when the product of the estimated regression coefficients is negative: in this case $\hat{\rho}$ is assumed equal to 0.

The distribution of the ρ estimators, obtained by simulation, has characteristics referring to a mixture variable with two components: one continuous and one discontinuous. Furthermore, it allows to evaluate the fraction of simulations number in which the situations 1. and 2., corresponding to the discontinuous variable, happen.

In Table 4 the quantities in the $N = 1000$ simulations with $\hat{\rho}$ equal to – 1, 0, 1, referred to the two simplified procedures by ordered means and medians, are given.

TABLE 4

Simulations number with $\hat{\rho} = -1, 0, 1$ for $N = 1000$

		Procedures by ordered means			Procedures by ordered medians		
n	ρ	$\hat{\rho} = -1$	$\hat{\rho} = 0$	$\hat{\rho} = 1$	$\hat{\rho} = -1$	$\hat{\rho} = 0$	$\hat{\rho} = 1$
10	0	4	289	13	4	0	16
10	0.25	0	251	20	0	0	30
10	0.5	0	119	41	0	0	80
10	0.75	0	19	150	0	0	266
20	0	0	289	0	0	178	0
20	0.25	0	184	1	0	120	1
20	0.5	0	51	3	0	43	7
20	0.75	0	0	15	0	5	76
30	0	0	267	0	0	0	0
30	0.25	0	144	0	0	0	0
30	0.5	0	13	0	0	0	2
30	0.75	0	0	0	0	0	38

We observe that, considering models with theoretical values $\rho > 0$, for $\hat{\rho} = -1$, the number of simulations is nearly zero for both the proposed procedures, since only for $n = 10$ and $\rho = 0$ there is a number of simulations equal to 4 on 1000. Instead, for $\hat{\rho} = 1$ we observe that the number of simulations notably decreases at the increasing of n and at the decreasing of the correlation coefficient ρ. In particular, for $\rho = 0.75$ and $n = 10$ the number of simulations is 150, for the ordered means procedure, and 260 for the ordered medians procedure. For $n = 30$ the number of simulations is reduced to 0 and 38, respectively.

For $\hat{\rho} = 0$ the simulation values are different in the procedure based on ordered medians at $n = 10$ and 30 in respect of $n = 20$: in the first two cases the number of simulations is null, while for $n = 20$ it is notable. The different behaviour is due to the occurrence that in the first two cases the number of subsamples

is odd ($n/2$ = 5 and 15, respectively), hence the medians of the subgroup (in terms either of x and y) are directly located by a point in the plane (x, y). On the contrary, for a number of subsamples as n = 20, the median point is located by the half-sum of the two central values, observed for the x and for the y, in the subsamples themselves. Such different behaviour can be observed in Table 5, in which the number of simulations is considered for $\hat{\rho}$ = 0, $n/2$ = 5, 6, ..., 15 and ρ = 0 and 0.75, that confirms the results in Table 4.

TABLE 5

Simulations number with $\hat{\rho} = 0$ for N = 1000 and n/2 = 5, 6, ..., 15

$n/2 = 2k+1$	ρ	$\hat{\rho}$ = 0	$n/2 = 2k$	ρ	$\hat{\rho}$ = 0
5	0	0	6	0	218
5	0.75	0	6	0.75	26
7	0	0	8	0	185
7	0.75	0	8	0.75	4
9	0	0	10	0	178
9	0.75	0	10	0.75	5
11	0	0	12	0	157
11	0.75	0	12	0.75	2
13	0	0	14	0	147
13	0.75	0	14	0.75	0
15	0	0			
15	0.75	0			

Moreover, referring to the most important parameter a_{11}, the relative efficiencies σ^2_{OLS}/σ^2 are calculated, where σ^2 is the variance of the estimator $\hat{a}_{11}$, obtained by the proposed procedure, and σ^2_{OLS} is the variance of the same estimator obtained by the ordinary least square method.

The results are given in Table 6.

TABLE 6

Relative efficiencies of $\hat{a}_{11}$ obtained by the proposed procedure (means or medians) in comparison with the ordinary least square method

n	ρ	Relative efficiency obtained by the mean procedure	Relative efficiency obtained by the median procedure
10	0	0.612	0.257
10	0.25	0.614	0.247
10	0.5	0.616	0.231
10	0.75	0.614	0.202
20	0	0.602	0.299
20	0.25	0.601	0.294
20	0.5	0.602	0.273
20	0.75	0.603	0.219
30	0	0.577	0.259
30	0.25	0.584	0.254
30	0.5	0.579	0.228
30	0.75	0.575	0.173

We observe that the proposed procedure presents less efficiency in respect of the least square method. In particular the procedure based on means is more efficient than the one based on medians.

The next figure, that gives the distributions of the estimations of the parameter a_{11} obtained by the three methods when $n = 30$ and $\rho = 0.5$ and 0.75, confirms those data.

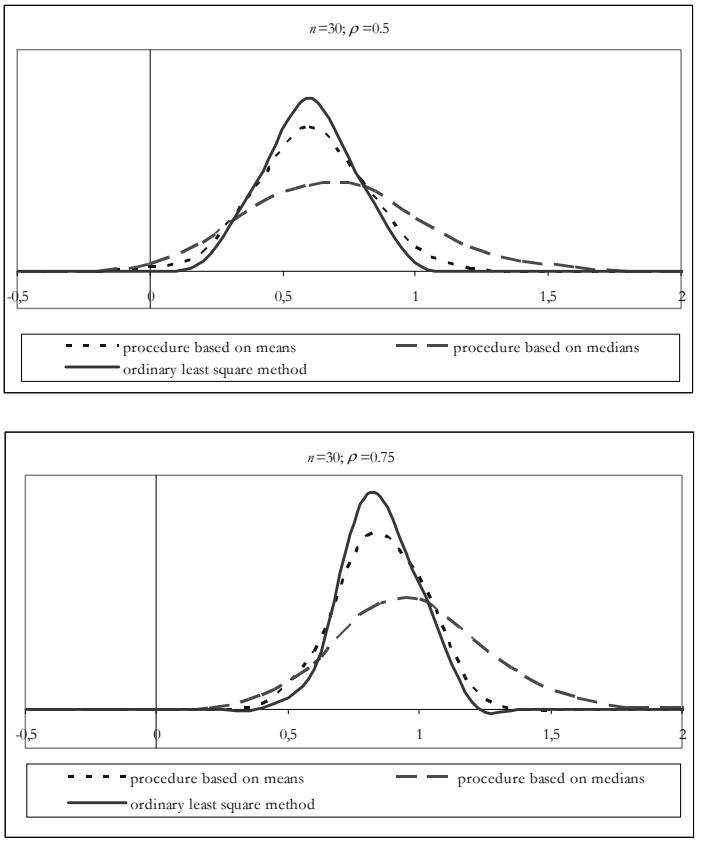

Figure 3 – Empirical distribution of $\hat{a}_{11}$ for $N = 1000$, $n = 20$ and $\rho = 0.5$ and 0.75.

3.2. *The second-order linear model*

We consider only some situations because we have observed asymptotic distorsions, as shown in paragraph 2.3.. In particular, the results of the simulation for $N = 1000$ and $n = 30$, assuming as model the Hermite polinomial of the second order with $c_0 = c_1 = 0$ and $c_2 = 1$, corresponding to $a_{02} = -2$, $a_{12} = 0$, $a_{22} = 4$ are given.

The Figure 4 shows an example of a model estimated by the proposed procedure, considering the ordered means and the ordered medians in comparison with the values obtained by the least square method.

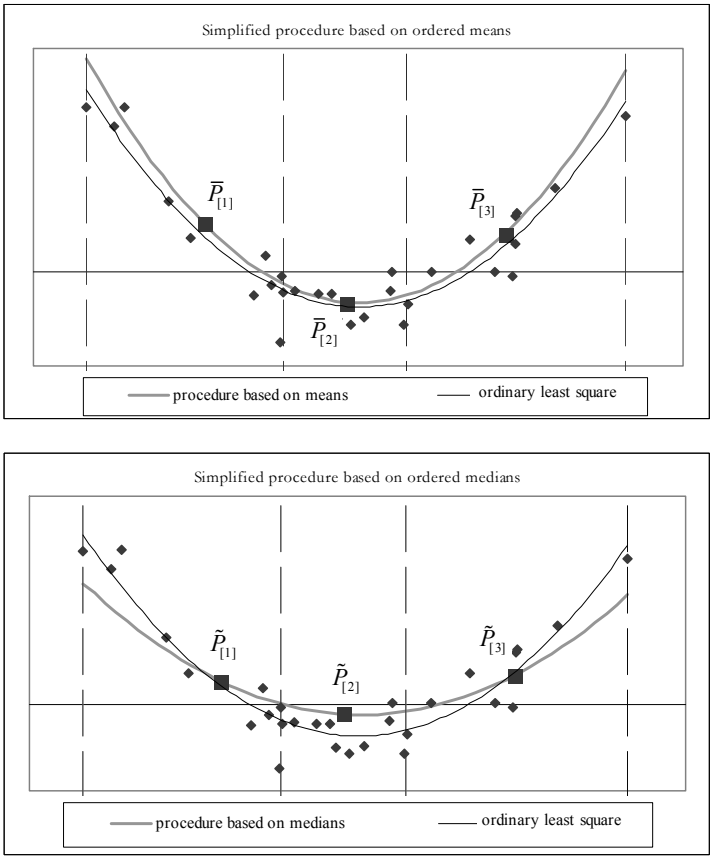

Figure 4 – The proposed simplified procedure: an example for $n = 30$, $m = 3$ and $r = 2$.

Specifically, we note the data partitioned in three subsets ($m = 3$) of size $n/3$, the mean and the median points of each subset that define the model.

Moreover, Table 7 gives some statistics referred to the estimations of the model parameters obtained by the ordered means procedure and the ordinary least square method.

TABLE 7

Synthesised results of the simulations referred to the simplified procedure based on the ordered means and to the ordinary least square method for N = 1000 and r = 2

	Simplified procedure based on ordered means			Ordinary least square method		
	$\hat{a}_{02}$	$\hat{a}_{12}$	$\hat{a}_{22}$	$\hat{a}_{02}$	$\hat{a}_{12}$	$\hat{a}_{22}$
Mean	-1.6939	0.0085	4.6625	-1.9957	0.0011	3.9889
Median	-1.7009	0.0063	4.6535	-1.9958	0.0046	3.9907
Variance	0.1357	0.2994	0.3604	0.0567	0.0452	0.0294
SD	0.3684	0.5472	0.6003	0.2382	0.2126	0.1714
BIAS	0.3061	0.0085	0.6625	0.0043	0.0011	-0.0111
MSE	0.2294	0.2995	0.7993	0.0568	0.0452	0.0295

We observe that the bias of the estimates of the regression coefficients obtained by the proposed procedure are confirmed. In this case the estimates of a_{02} and a_{22} have values particularly high: 0.3061 and 0.6625 with respect to those obtained by the least square method: 0.0043 and – 0.0111. The last values are only due to the casualness of the simulation procedure and to the dimension of the replications $N = 1000$. In fact, if we compare the mean values of the regression coefficients obtained by simulation with those obtained in asymptotic conditions (see paragraph 2.3.), we have 0.0648 for a_{02} and – 0.0760 for a_{22}. These values are really reduced, and therefore due only to the casualness of the simulation.

4. FINAL REMARKS AND CONCLUSIONS

The proposed simplified regression procedure can be useful in a preliminary stage to choose the polynomial regression model, because the partition of the data (x, y), referred to the explicative variable X, is defined relating to the number of the parameters of the chosen model. Hence in the general case the complete polynomial model is formed by $m = r + 1$ subsets, of size n/m. Therefore it is possible to start from the same data set to obtain the estimates of the regression models of order 1, 2, ..., r and to stop the procedure when the estimate of the regression coefficient a_{rr} has to be considered small ($\hat{a}_{rr} \sim 0$).

To state the order r of the regression polynomial, the "step-wise" procedure can be made like in the ordinaty least square method. In particular it is possible to apply the Student's t test or of the Snedecor's F test to verify the null hypothesis $a_{rr} = 0$. Note that to use this procedure, the variance of the estimeted parameter a_{rr} has to be obtained by numerical simulations.

The preliminary study to evaluate the inferential properties of the proposed procedure, by ordered means or medians and restricted to the polynomials of order 1 and 2 , has already shown some characteristics of the estimators, but it seems necessary to verify the extension for polynomial models of higher order.

For the model of order $r = 1$, we do not observe a bias of the estimators and comparatively, there is a higher relative efficiency of the procedure based on the means in respect of that based on the medians. So, in this first study we have verified for the polynomial model of order $r = 2$ only the ordered means procedure. In such a situation, we have numerically found that the coefficient estimators are generally biased and we have given a theoretical demonstration.

Some observations, relating to the possibility to estimate the correlation between the explicative variables X and Y, not only in the linear function $y = a_{01} + a_{11}x$ but also in the analogous model $x = b_{01} + b_{11}y$, by means of the simplified procedure, are given in the paper.

Finally, we think that this study could be complete with:

- the analysis of polynomial models of order higher than two;
- a study about alternative assumptions in respect of the normal distribution for the explicative variable and, eventually, for the error component, that in this

research has considered normally distributed, independent of the explicative variable and homoscedastic;
- the extension of the comparison to the use of a loss function of order 1 and comparing the inferential properties of the model parameter with the one obtained by the quantile regression;
- the extension of the proposed methodology to the situation with more than one explicative variable.

Dipartimento di Scienze statistiche SILVIA FACCHINETTI
Università Cattolica del S. Cuore – Milano UMBERTO MAGAGNOLI

REFERENCES

M. ABRAMOVITZ, I.E. STEGUM, (1972), *Handbook of Mathematical Functions*, 10th printing, National Bureau of Standards, – United States Department of Commerce, Washington D.C., USA.

D.R. COX, (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge, UK.

P.J. HUBER, E.M. RONCHETTI, (2009), *Robust Statistics*, 2nd Edition, Wiley, New York, USA.

N.L. JOHNSON, S. KOTZ, N. BALAKRISHNAN, (1994), *Continuous Univariate Distributions*, Vol. 1, Wiley, New York, USA.

N.L. JOHNSON, S. KOTZ, N. BALAKRISHNAN, (1995), *Continuous Univariate Distributions*, Vol. 2, Wiley, New York, USA.

SUMMARY

A simplified procedure of linear regression in a preliminary analysis

The analysis of a statistical large data-set can be led by the study of a particularly interesting variable Y – regressed – and an explicative variable X, chosen among the remained variables, conjointly observed. The study gives a simplified procedure to obtain the functional link of the variables $y = y(x)$ by a partition of the data-set into m subsets, in which the observations are synthesized by location indices (mean or median) of X and Y. Polynomial models for $y(x)$ of order r are considered to verify the characteristics of the given procedure, in particular we assume $r = 1$ and 2. The distributions of the parameter estimators are obtained by simulation, when the fitting is done for $m = r + 1$. Comparisons of the results, in terms of distribution and efficiency, are made with the results obtained by the ordinary least square methods. The study also gives some considerations on the consistency of the estimated parameters obtained by the given procedure.