ESTUDIOS DE ECONOMÍA APLICADA VOL. 21 - 3, 2003. PÁGS. 559-573

# Adaptive regression analysis: theory and applications in econometrics

*GARCÍA PEREZ, J., **CHEBRAKOV, Y.V. and **SHMAGIN, V.V.

*Departamento de Economía Aplicada. Universidad de Almeria. **St-Petersburg Technical University.*

*Campus de la Cañada. 40011 Almería. **Apt.22, Bldg.11/1, Polustrovsky pr., 195221, S.-Petersburg, Russia.

*E-mail: jgarcia@ual.es; **E-mails: gchebra@mail.ru

## ABSTRACT

In this work we (a) discuss some theoretical and computational difficulties of regression analysing dependences, describing the behaviour of the heterogeneous systems, (b) offer a set of new techniques adaptable to regression analysing the heterogeneous dependences and (c) demonstrate the advantages of application of these new techniques in econometrics.

*Key words:* Adaptive regression analysis, heterogeneous econometric systems, Econometric Methods, Single Equuiation Models-General.

## Análisis de regresión adaptada: teoría y aplicaciones en econometría

## RESUMEN

En este trabajo (a) se realiza una discusión sobre las dificultades teóricas y prácticas, en el análsisis de regresión, cuando se trata dedescribir el comportamiento de sistemas heterogéneos (b) se ofrecen un conjunto de nuevas técnicas de regresión adaptada para analizar las dependencias en los sistemas heterogéneos (c) se demuestra la ventaja del uso de estas nuevas técnicas en Econometría.

*Palabras Clave:* Análisis de regresión adaptada, sistemas econométricos heterogéneos, Métodos Econométricos, modelos uniecuacional general.

Clasificación JEL: C20.

Artículo recibido en septiembre de 2003. Aceptado en noviembre de 2003.

## 1. INTRODUCTION

It is well known, many computational problems, arising in the investigation of the multivariate statistical systems (for instance, econometric systems), can hardly be solved without using some kind of fitting techniques. Such problems refer to the construction of mathematical models, description of the investigated systems behaviour, and/or finding the system parameters values [4]. It has been shown in [1–3], that the application of available fitting techniques leads to some theoretical and computational difficulties (paradoxes). They are mainly explained by the fact that the modern statements of data analysis problems are too far from the real situations. In particular, these statements make no provision for facts that:

1. Results of calculations may depend on the way the investigated data are obtained;
2. The given accuracy of estimated values cannot be achieved for the strongly contaminated data;
3. A single solution of the data analysis problems for contaminated data arrays is incomplete one;
4. Every investigated system may contain some heterogeneity, which leads to (i) an adequate, (ii) a removable (local) inadequate or (iii) an irremovable (global) inadequate postulated fitting model, describing the behaviour of the system.

The main goals of this paper are:

a) Discussing some theoretical and computational difficulties of regression analysing dependences, describing the behaviour of the heterogeneous systems;

b) Offering some new methods adaptable to regression analysing the heterogeneous dependences;

c) Illustrating the advantages of application of new regression techniques in econometrics.

## 2. COMPUTATIONAL DIFFICULTIES OF REGRESSION ANALYSING HETEROGENEOUS DEPENDENCES

As generally known [1 – 5], if a dependence $\{y_n, \mathbf{X}_n\}$ ($n = 1, 2, \ldots, N$) and a fitting function $F(\mathbf{A}, \mathbf{X})$ are given, then the main problems of regression analysis theory are finding estimates of $\mathbf{A}'$ and $y'$ and variances of $\delta\mathbf{A}'$ and $\delta(y - y')$, where $\mathbf{A}'$ is an estimate of vector parameter $\mathbf{A}$ of the function $F(\mathbf{A}, \mathbf{X})$ and $\{y_n'\} = \{F(\mathbf{A}', \mathbf{X}_n)\}$. In particular, if $F(\mathbf{A}, \mathbf{X}) = \sum_{l=1}^{L} a_l h_l(\mathbf{X})$ {$F(\mathbf{A}, \mathbf{X})$ is a linear model}, where $h_l(\mathbf{X})$ are some functions on $\mathbf{X}$, then classical least squares (LS) method gives standard regression analysis solution:

$$\mathbf{A}'_{\mathrm{LS}} = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1}\,\mathbf{H}^{\mathrm{T}}\mathbf{Y}\,, \quad (\delta\mathbf{A}')^2 = s/(N-L)\,\mathrm{diag}(\mathrm{H}^{\mathrm{T}}\mathrm{H})^{-1} \qquad [1]$$

$$\delta_p(y - y') = \pm t_p\, s \sqrt{1 + \mathbf{H}_i^{\mathrm{T}} (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \mathbf{H}_i}\ ,$$

where $\mathbf{H}$ is a matrix $L \times N$ in size with $n$-th row $(h_1(\mathbf{X}_n), h_2(\mathbf{X}_n), \ldots, h_L(\mathbf{X}_n))$; $\mathbf{H}^{\mathrm{T}}$ is the transposed matrix $\mathbf{H}$; $\mathbf{Y} = \{y_n\}$; $s = \sum_{n=1}^{N} (y_n - y'_n)^2 / (N - L)$ ; $\mathbf{H}_i = (h_1(\mathbf{X}_i), h_2(\mathbf{X}_i), \ldots, h_L(\mathbf{X}_i))$; the value of $t_p$ is determined by $t$-Student distribution table and generally depends on the assigned value of the significance level of $p$ and the value of $N - L$ (a number of freedom degree); when a value of the significance level of $p$ is assigned the notation of $\delta_p(y - y')$ means confidence interval for possible deviations of experimental values of $y$ from computed values $y' = F(\mathbf{A}', \mathbf{X})$.

We found [1 – 3] that three various types of heterogeneity affection may be revealed when dependences $\{y_n, \mathbf{X}_n\}$ are analysed:

a) The heterogeneity has no effect on the results of solving fitting problems.

b) The heterogeneity leads to a removable (local) of fitting function $F(\mathbf{A}, \mathbf{X})$.

c) The heterogeneity leads to a irremovable (global) of fitting function $F(\mathbf{A}, \mathbf{X})$.

In this Section some of such defects of both LS-solution and alternative regression analysis solutions are discussed as arise in the cases (b) and (c).

### 2.1. LS-estimating parameters of inadequate fitting function

α) Let us discuss some defects of the standard solution of the regression problems for a data array $\{y_n, x_n\}$ given in Table 1. This array was obtained by Russian chemist D. I. Mendeleev in 1881, when he investigated the solvability ($y$, relative units) of sodium nitrate ($NaNO_3$) on the water temperature ($x$, °C).

**Table 1. D. I. Mendeleev data array**

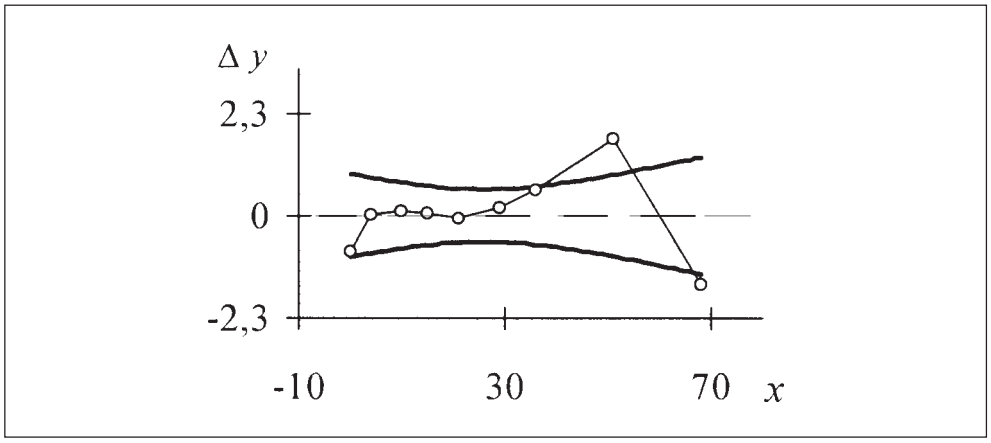
| $n$ | $x_n$ | $y_n$ | $y_n - y_n'$ | $n$ | $x_n$ | $y_n$ | $y_n - y_n'$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 66.7 | –0.80 | 6 | 29 | 92.9 | 0.17 |
| 2 | 4 | 71.0 | 0.02 | 7 | 36 | 99.4 | 0.58 |
| 3 | 10 | 76.3 | 0.10 | 8 | 51 | 113.6 | 1.73 |
| 4 | 15 | 80.6 | 0.05 | 9 | 68 | 125.1 | –1.56 |
| 5 | 21 | 85.7 | –0.07 | - | - | - | - |

As stated in [5], when $p = 0.9$ the standard solution (1) for the data array $\{y_n, x_n\}$ has form:

$$y' = 67.5 + 0.871\, x\ ,\ \delta\mathbf{A}' = (0.5;\ 0.2), \qquad \textbf{[2]}$$

$$\delta_{0.9}\,(y - y') = \pm 0.593\ \sqrt{1 + (x - 26)^2 / 451.1}\ .$$

***Figure 1. The plots of confidence interval of the deviation of y from y ' (heavy lines) and residuals y – y ' (circles) for D. I. Mendeleev data array.***




We show the plots of $\delta_{0.9}(y - y')$ on $x$ by heavy lines in Figure 1 and $\{y_n - y_n', x_n\}$ by the circles. Since the plot of $\{y_n - y_n', x_n\}$ steps over the heavy lines in Figure 1, the first theoretical and computational difficulty is revealed:

*The standard solution, suggested in [5] for determining the confidence interval of the deviations of y from y ', is out of character with the data array $\{y_n, x_n\}$ given in Table 1.*

It follows from results presented in Table 1 and/or Figure 1, if one assumes that $\delta(y - y') \geq \max|y_n - y_n'| = 1.73$ then the broken connections of the confidence interval $\delta(y - y')$ with D. I. Mendeleev data array will be pieced up [1 – 3]. But in this case

*The standard values of* $\delta\mathbf{A}'$*, calculated by (1), are out of character with the data array also.*

β) Let the data array $\{y_n, x_n\}$ be following:

$$\{y_n\} = (5.66;\ 10.39;\ 16.00;\ 22.36;\ 29.39;\ 37.04;\ 45.25;\ 54.00;\ 63.25),\ \{x_n\} = (2;\ 3;\ \&;\ 10). \qquad \textbf{[3]}$$

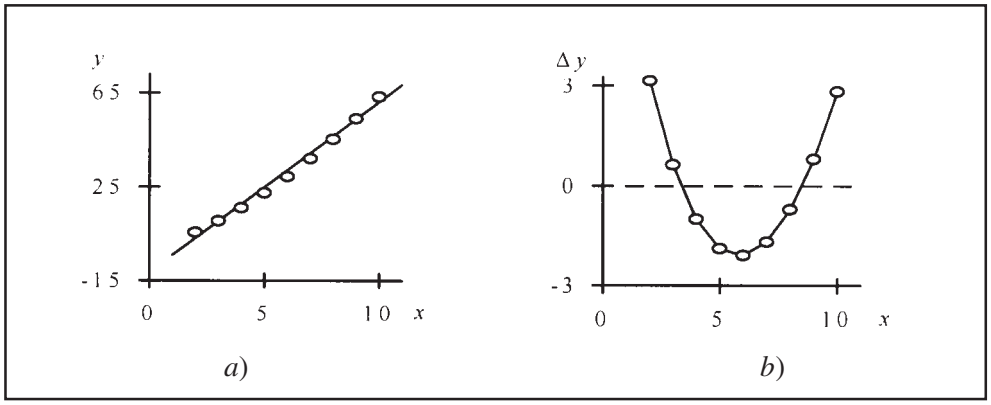
As it may be calculated by [1], if the fitting function $F(\mathbf{A}, \mathbf{X}) = a_0 + a_1 x$, then $\mathbf{A}'_{LS} = (-11.95;\ 7.24)$ for the data array [3]. Plots of the dependence [3] (circles) and the function $F(\mathbf{A}, \mathbf{X}) = -11.95 + 7.24\,x$ (continuous line) are shown in Figure 2(*a*); the plot of residues $\{y_n + 11.95 - 7.24x_n;\ x_n\}$ is shown in Figure 2(*b*).

There are three criteria that the fitting function $a_0 + a_1 x$ is inadequate for the data array (3):

a) The plot of residues, shown in Figure 1(*b*), has the form of a parabola, and, consequently, a second degree multinomial must have the better fitting properties for the dependence (3) then the linear model;

b) The data array (3) was generated by the function $f(x) = 2x^{1.5}$:

*Figure 2.* ***(a) – Plots of the dependence (3) (circles) and the function –11.95 +7.24 x (continuous line); (b) – The plot of residues*** $\{y_n + 11.95 - 7.24x_n;\ x_n\}$



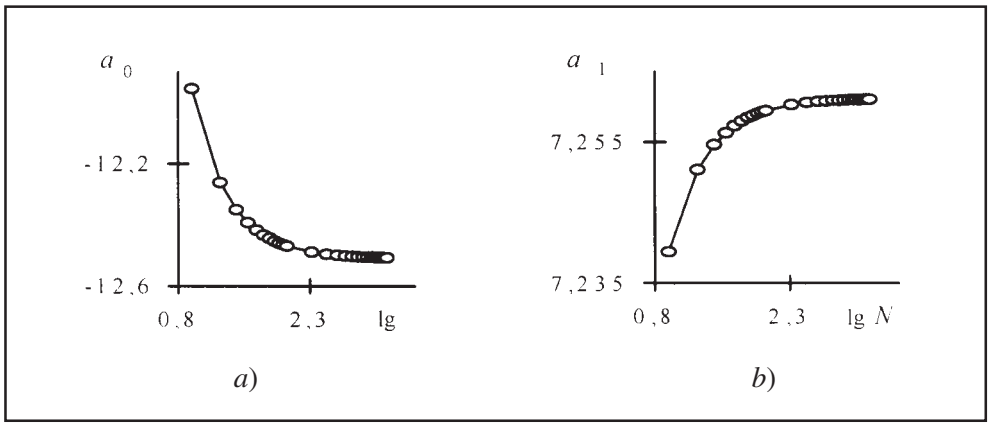

$$\{y_n\} = \{[k\, f(x_n)] / k\}, \qquad [4]$$

where square brackets mean the integer part, a factor $k = 100$ and its presence in (4) is necessary for calculating all values of $y_n$ within error e = 0.01;

c) The maximal value of $\Delta y$ in Figure 2(*b*) exceeds appreciably the value of the computational error ε for the values of $y$ ($\varepsilon = 0.01$).

Dependences of LS-estimations of linear model parameters for the values of the function $2x^{1.5}$ computed at $N$ points uniformly in $x$ on the interval [2; 10], are shown in Figure 3. Limiting values of parameters $a_1$ and $a_0$ (when $N \to \infty$) are marked by the dotted lines.

***Figure 3. Dependences of LS-estimations of linear model parameters for the values of the function $2x^{1.5}$, computed for N points uniformly in x on the interval [2; 10].***



We may conclude from analysing of dependences shown in Figure 3

a) *Inadequacy of the fitting function $a_0+a_1x$ leads to the distortion of LS-estimations of its parameters*;

b) *The distortion of LS-estimations may be decreased by means of increasing numbers of records*.

### 2.2. Alternative estimating parameters of inadequate fitting function

P. Huber [6] noted that, as the rule, 5 – 10% of all observations are anomalous. There are at least two strategies in order to decrease influence of anomalous observations (outliers) on estimations of fitting function parameters:

1) Remove all outliers from the data array;

2) Compute the values of **A'** by means of M-robust estimators.

We note

a) As stated in [7] one of two such equivalent combined statistical procedures should be adopted for revealing outliers in the data array $\{y_n, \mathbf{X}_n\}$, in which parameter estimates, minimising the median (MED) of the array $\{(y_n - y_n')^2\}$ or the sum (SUM) of $K$ first elements of the same array, are considered as the best ones;

b) If $F(\mathbf{A}, \mathbf{X})$ is a linear function, then a robust M-estimate of **A'** is found as decision of one from two minimisation problems [6].

$$S_\varphi(\mathrm{A}) = \sum_{n=1}^{N} \varphi(y_n - y_n') \Rightarrow \min \quad \text{or} \quad \partial S_\varphi / \partial a_l = 0, \qquad \textbf{[5]}$$

where function $\varphi(r)$ is symmetric concerning Y-axis, continuously differentiable with a minimum at zero and $\varphi(0) = 0$.

It follows from the values $\{y_n - y'_n\}$ adduced in Table 1, that the correspondence between experimental and computed values of $y$ is slightly getting worse at the beginning and end of D. I. Mendeleev data array. Namely, we may suppose that records 1, 7, 8 and 9 are outliers.

Let us test effectiveness of both methods, mentioned in points (a) and (b), on D. I. Mendeleev data array.

$\alpha$) If two equivalent combined statistical procedures, described in point (a), are used then following results are obtained:

*i*) Both procedures MED and SUM reveal only three outliers: records 7, 8 and 9;

*ii*) If records 7, 8 and 9 are eliminated from D. I. Mendeleev data array, then none of records of the truncated data array is outlier for procedure MED, but procedure SUM reveals another two outliers, which are records 2 and 6 into the initial data array.

Thus, we may conclude

*A set of revealed outliers depends sometimes on a type of the used statistical procedures*.

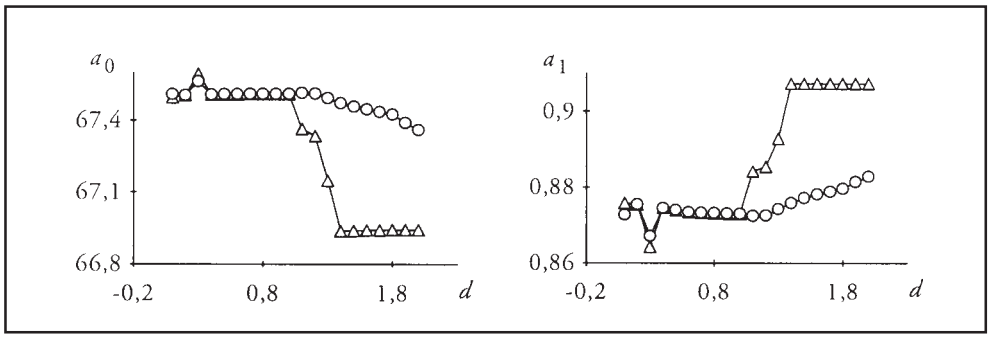
β) Let in (5) Andrews function φ(*r*) {φ(*r*) = *d*(1–cos(*r*/*d*)) if $|r| \le d\pi$ and φ(*r*) =0 if $|r| > d\pi$} be applied.

It is articulate in Figure 4 that the values of $a_0$ and $a_1$ parameters of the linear model, fitted D. I. Mendeleev data array, depend on

a) values of internal parameter *d* of robust Andrews estimator φ(*r*);

b) type of the minimisation robust regression problems, mentioned in (5) (solutions of the first and second minimisation problem of (5) are marked respectively by triangles and circles in Figure 4).

***Figure 4. Dependences of values of $a_0$ and $a_1$ parameters of the linear model, fitted D. I. Mendeleev data array, on values internal parameter d of robust Andrews estimator and the type of the minimisation robust regression problems, mentioned in (5).***




Thus, in this case the main difficulty is exhibited in a fact, that

The robust estimates may be not robust in actual practice.

## 3. MAIN STATEMENTS AND METHODS OF ADAPTIVE REGRESSION ANALYSIS

As has been mentioned in Section 1 and 2

a) The main difficulty of the modern regression analysis theory consists in disagreement between statements of this theory, which ensure uniqueness of solutions for fitting problems, and multiplicity of solutions, existed in actual practice.

b) If the regression analysis of dependence $\{y_n, \mathbf{X}_n\}$ describing the heterogeneous system behaviour is performed then three various situations can be met: postulated fitting model $F(\mathbf{A}, \mathbf{X})$ may be (*i*) adequate, (*ii*) removable (local) inadequate and (*iii*) irremovable (global) inadequate.

c) Situation (*ii*) has a wide distribution in practice. And if it occurs then it means that the data array under consideration contains some number of outliers.

It is proved in [1, 2] that in situation (*ii*) one may obtain the correct solution of the problem on revealing outliers in the data array $\{y_n, \mathbf{X}_n\}$ by means of using a set of methods of adaptive regression analysis.

In this Section we consider the main statements and methods of adaptive regression analysis.

1. The relation

$$y_n = g(F(\mathbf{A}, \mathbf{X}_n) + e_n) \qquad \textbf{[6]}$$

is the main data analysis *model* of adaptive regression analysis, where $y_n$ is $n$-th value of dependent variable; $n = 1, 2, \ldots, N$; $F(\mathbf{A}, \mathbf{X})$ is a fitting function; $e_n$ is $n$-th value of random variable; $g$ is a function, allowing to describe the way of dependent variable measurement.

We add

*i*) If the form of measurement function $g$ in (6) is unknown then it is usually assumed that $g = g_a$, where $g_a$ is the truncation function:

$$g_\alpha(y) = 2\alpha[y/(2\alpha)] + 2\alpha \text{ if } |y - 2\alpha\,[y/(2\alpha)]| \geq \alpha, \text{ else } g_a(y) = 2\alpha\,[y/(2\alpha)]; \quad [7]$$

*ii*) For the sake of simplicity we will assume further that in (6) $g = g_\alpha$ in all cases.

2. The main computational *problems* of adaptive regression analysis are

a) Finding such minimum value $\alpha = \alpha_{\min}$, that, for all experiment realisations of $\{\mathbf{X}_U\}$, containing all possible $U$ subsamples from $N, N-1, \ldots, N-n_0$ readings, the inequality

$$|y_n - g_a(y_n')| \leq \alpha_{\min} \qquad [8]$$

is fulfilled, where $N$ is the dimension of the initial data array $\{y_n, \mathbf{X}_n\}$; $y_n'$ is an estimate of $n$-th value of the dependent variable; $n_0$ is a given integer number, which assigns a maximum level of truncating the initial data array $\{y_n, \mathbf{X}_n\}$.

The value of $\alpha_{\min}$ is found as a solution of the following extreme problem

$$\min_{\alpha} \max_{U} \max_{n} |y_n - g_\alpha(y_n')|, \qquad \textbf{[9]}$$

where the maximum on $U$ means finding the solution on all $U$ sets of $\{\mathbf{X}_U\}$, composed of $N, N-1, \ldots, N-n_0$ readings;

b) Finding a general analytical solution, which is a set of the equivalent analytical functions $g_\alpha(F((\mathbf{C}_i', \xi), \mathbf{X}_n))$, where $(\mathbf{C}_i', \xi)$ are some multinomials of degree $m_i$, where $\xi$ is a variable $(-1 \leq \xi \leq 1)$ and $\mathbf{C}_i$ is a vector parameter.

The family of equivalent analytical formulae is constructed by replacing the vector parameters $\mathbf{A}$ of functions $g_\alpha(F(\mathbf{A}, \mathbf{X}))$ by $\mathbf{A} = \{(\mathbf{C}_i, \xi)\}$ and determining the minimum values of degrees and the estimates of coefficients in multinomials $(\boldsymbol{C}_i, \xi)$.

For example, if $F(\mathbf{A}, \mathbf{X})$ is a linear model, $\alpha$=0.94, $-1 \le \xi \le 1$ and the family of equivalent analytical formulae has the form

$$y'(\xi, x) = g_{\alpha}( 67.77 - 0.35\xi + (0.86457 + 0.00805\xi)\, x ), \qquad \textbf{[10]}$$

then $| y_n - y'(\xi, x_n) | \le 1.73$ for D. I. Mendeleev data array adduced in Table 1, where the value 1.73 is the limiting one for the confidence interval $\delta(y - y')$ {see point ($\alpha$) of Section 2.1}.

It may be computed from (10)

*i*) If $S = \sum_{n=1}^{N} (y_n - y'(\xi, x_n))^2$, then dependence of $S$ on x has the following form

$$S(\xi) = 6.69 - 0.661\xi + 0.441\xi \qquad \textbf{[11]}$$

for D. I. Mendeleev data array. Consequently, $S(\xi)$ has least value in (11) when $\xi$= 0.661/(2×0.441) = 0.749. But if $\xi$ = 0.749, then standard LS-solutions $\{\mathbf{A}'_{LS}= (67.51; 0.871); S_{LS}= 6.44\}$ are obtained from (10) and (11) for D. I. Mendeleev data array;

*ii*) If $D = \max_n | y_n - y'(\xi, x_n) |$, then dependence of $D$ on $\xi$ has the following form:

$$\textit{If } \xi \le 1.07 \text{ then } D(\xi) = 1.7369 - 0.0606\xi;$$
$$\textit{If } x > 1.07 \text{ then } D(\xi) = 1.4608 + 0.197\xi \qquad \textbf{[12]}$$

for D. I. Mendeleev data array. Consequently, $D(\xi)$ has least value =1.672 in (12) when $\xi$= 1.07. But this value of is obtained when $\xi > 1$ and so it is not solution for the problem under consideration. Thus, the correct reply for D. I. Mendeleev data array is $D_{\min}$ =1.676 ($\xi$= 1.00).

We may conclude

*The family of equivalent fitting models may be used for obtaining such correct standard and alternative solutions of modern regression analysis theory as minimised different criteria.*

c) Finding the value of variance $\delta\mathbf{A}'$.

The value of variance $\delta\mathbf{A}'$ is computed by setting extreme values of parameter $\xi$ into the general analytical solution $g_{\alpha}(F((\mathbf{C}_i', \xi), \mathbf{X}_n))$.

For example

*i*) If values of $\xi = \pm 1$ are inserted into (10), then the exact limit of the variation of parameters $a_0$ and $a_1$ may be determined: $a_0 = 67.77 \pm 0.35$ and $a_1 = 0.865 \pm 0.008$ for function (10), fitted D. I. Mendeleev data array;

*ii*) If the set of readings with numbers 1, 7, 8 and 9 deletes from D. I. Mendeleev data array, then the family of equivalent analytical formulae must have the following form

$$y'(\xi, x) = g_{\alpha}( 67.521 - 0{,}045\xi + (0.872243 + 0.002196x)\, x ) \qquad \textbf{[13]}$$

for the truncated data array $\{y_n, x_n\}^*$ in order to keep the difference of $|y_n - y_n'|$ within the limit of the error $\varepsilon = 0.1$, where $\alpha = 0.07$, $-1 \le \xi \le 1$. Consequently, $a_0 = 67.52 \pm 0.045$ and $a_1 = 0.872 \pm 0.002$ in this case.

If solutions (10) and (13) are compared with the standard LS-solution $\mathbf{A}' = (67.5 \pm 0.5;\ 0.87 \pm 0.2)$, then the following conclusion may be obtained

*Values of variances $\delta a_0'$ and $\delta a_1'$, computed by standard method (1), disagree with exact values determined by (10) and (13).*

## 4. APPLICATION OF METHODS OF ADAPTIVE REGRESSION ANALYSIS IN ECONOMETRICS

In this Section we adduce such solutions of fitting problems for three econometric heterogeneous dependences as obtained by means of the techniques of adaptive regression analysis and compare these solutions with standard ones computed from formulae (1).

Problem 1. As stated in [8, P.49–52], standard solutions (1) have the following form

$$\mathbf{A}' = (31.67 \pm 2.25;\ 1.279 \pm 0.003), \qquad \textbf{[14]}$$

for the linear model $y = a_0 + a_1 x_1$, fitted data array $\{y_n, x_n\}$ of Table 2, where $x_n$ is operation time (months) for $n$-th car out of service and $y_n$ is one year upkeep (hundreds of francs) for the same car.

**Table 2. The data array for Problem 1**

| $n$ | $x_n$ | $y_n$ | $n$ | $x_n$ | $y_n$ | $n$ | $x_n$ | $y_n$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 48 | 6 | 8 | 40 | 11 | 32 | 71 |
| 2 | 8 | 43 | 7 | 21 | 56 | 12 | 17 | 58 |
| 3 | 36 | 77 | 8 | 21 | 62 | 13 | 58 | 102 |
| 4 | 41 | 89 | 9 | 53 | 100 | 14 | 6 | 35 |
| 5 | 16 | 50 | 10 | 10 | 47 | 15 | 20 | 60 |

The techniques of adaptive regression analysis give following solutions for the fitting problem under consideration:

*i*) The family of equivalent analytical formulae has the form

$$y'(\xi,\ x) = g_\alpha(31.284 + 0.126\xi + (1.2901 - 0.0036\xi)\, x), \qquad \textbf{[15]}$$

where $\alpha = 4.85$, $-1 \le \xi \le 1$. Consequently, $a_0 = 31.28 \pm 0.13$ and $a_1 = 1.2901 \pm 0.0036$;

*ii*) If $S = \sum_{n=1}^{N} \left(y_n - y'(\xi, x_n)\right)^2$, then dependence of $S$ on $\xi$ has the following form

$$S(\xi) = 132.7 - 0.4427\xi + 0.0716\xi. \qquad [16]$$

Consequently, $S(\xi)$ has least value in (16) when $\xi = 0.4427/(2\times0.0716) = 3.09$. If $\xi = 3.09$, then standard LS-solutions $\{\mathbf{A}'_{LS} = (31.67; 1.279); S_{LS} = 132.015\}$ are obtained from (15) and (16). But the value $\xi = 3.09 >> 1$ and so standard LS-solutions (14) are no part of correct ones.

Problem 2. As stated in [9, 10], for the data array $\{y_n, \mathbf{X}_n\}$ shown in the Table 3, the multivariate model $Y_3 = a_0+a_1x_1+a_2x_2$ demonstrates well some effects of multicollinearity and so fitting models $Y_1 = b_0+b_1x_1$ and $Y_2 = d_0+d_1x_2$ are more preferable.

**Table 3. The econometric data $\{y_n, \mathbf{X}_n\}$ on investigating the import turnover y (billion dollars) on gross national product $x_1$ (billion dollars) and consumer price index $x_2$ in USA**

| *Year* | *Y* | $x_1$ | $x_2$ |
|---|---|---|---|
| 1964 | 28.4 | 635.7 | 92.9 |
| 1965 | 32.0 | 688.1 | 94.5 |
| 1966 | 37.7 | 753.0 | 97.2 |
| 1967 | 40.6 | 796.3 | 100.0 |
| 1968 | 47.7 | 868.5 | 104.2 |
| 1969 | 52.9 | 935.5 | 109.8 |
| 1970 | 58.5 | 982.4 | 116.3 |
| 1971 | 64.0 | 1063.4 | 121.3 |

| *Year* | *Y* | $x_1$ | $x_2$ |
|---|---|---|---|
| 1972 | 75.9 | 1171.1 | 125.3 |
| 1973 | 94.4 | 1306.6 | 133.1 |
| 1974 | 131.9 | 1412.9 | 147.7 |
| 1975 | 126.9 | 1528.8 | 161.2 |
| 1976 | 155.4 | 1702.2 | 170.5 |
| 1977 | 185.8 | 1899.5 | 181.5 |
| 1978 | 217.5 | 2127.6 | 195.4 |
| 1979 | 260.9 | 2368.5 | 217.4 |

**Table 4. Computation results**

a) The function $g_a(Y_i)$ (all readings)

| $i=1$ | | $i=2$ | | $i=3$ | |
|---|---|---|---|---|---|
| $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ |
| 0 | 37.7 | 0 | 40.6 | 0 | 47.7 |
| 1 | 45.6 | 1 | 55.6 | 1 | 56.6 |

b) The function $g_a(Y_i)$ (the first ten readings)

| $i=1$ | | $i=2$ | | $i=3$ | |
|---|---|---|---|---|---|
| $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ |
| 0 | 2.7 | 0 | 6.4 | 0 | 2.7 |
| 1 | 5.7 | 1 | 11.1 | 1 | 5.7 |

c) The function $g_a(Y_i)$ (the last six readings)

| $i=1$ | | $i=2$ | | $i=3$ | |
|---|---|---|---|---|---|
| $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ | $n_0$ | $\alpha_{min}$ |
| 0 | 9.0 | 0 | 16.0 | 0 | 8.0 |
| 1 | 15.5 | 1 | 22.2 | 1 | 29.0 |

Let us find the best fitting model $g_\alpha(Y_i)$ ($i$ = 1, 2, 3) for the data array of Table 3. Our calculations give the results presented in points (a – c) of Table 4. Analysing these results we may conclude that

*a*) for case (a) of Table 4

*i*) The model $g_\alpha(Y_1)$ has the minimal value $\alpha_{\min}$ and so it is the best fitting model among all tested models $g_a(Y_i)$;

*ii*) The econometric system under analysis is heterogeneous for all tested models since even the minimal value $\alpha_{\min}$ exceeds appreciably the (measurement) error of the dependent variable $y$ (e = 1);

*b*) for cases (b) and (c) of Table 4

*i*) In order to reduce fitting errors of models $g_\alpha(Y_i)$, the econometric data of Table 3 are to divide into two data arrays, contained respectively the first ten (A1) and the last six (A2) readings;

*ii*) for arrays A1 and A2, the model $g_\alpha(Y_1)$ is the best fitting model and so the value of $y'$ may be calculated by the formula

$$y' = g_{2.7}(-34.7 + 0.09553\, x_1) \qquad \textbf{[17]}$$

for readings of array A1 and by the formula

$$y' = g_{9.0}(-81.8 + 0.14212\, x_1) \qquad \textbf{[18]}$$

for readings of array A2.

We add that fitting errors may be reduced by means of replacing the model $y = g_\alpha(a_0 + a_1 x_1)$ with $\ln y = g_\alpha(a_0 + a_1 \ln x_1)$ and eliminating data of 1974 year from the initial dependence $\{y_n, \mathbf{X}_n\}$ adduced in Table 3. In this case the family of equivalent analytical formulae has the form

$$\ln y'(\xi,\ x) = g_a(-7.675 - 0.610\xi + (1.70465+0.08625\xi)\ln x_1), \qquad \textbf{[19]}$$

where $\alpha$=0.072, $-1 \le \xi \le 1$.

It may be computed from (19)

*i*) If $S = \sum_{n=1}^{N} \left(y_n - y'(\xi, x_n)\right)^2$, then dependence of $S$ on $\xi$ has the following form

$$S(\xi) = 97.57 - 80.475\xi + 471.43\xi + 24.042\xi. \qquad \textbf{[20]}$$

Consequently, $S(\xi)$ has least value in (20) when $\xi$= 0.0848. But if $\xi$=0.0848, then standard LS-solutions $\{\mathbf{A}'_{LS}= (-7.7267;\ 1.7119);\ S_{LS}= 94.18\}$ are obtained from (19) and (20) for the model $y = \exp(a_0 + a_1 \ln x_1)$ fitted the data array adduced in Table 3;

*ii*) If $D = \max_n |y_n - y'(\xi, x_n)|$, then dependence of $D$ on $\xi$ has the minimum value $D_{\min} = 4.935$ when $\xi = 0.211$ and the best fitting model (for this value of $\xi$) has the following form

$$y' = \exp(-7.8037 + 1.7228 \times \ln x_1). \qquad \textbf{[21]}$$

Problem 3. Let us discuss some regression solutions adduced in [11] for the data array $\{y_n, x_n\}$ shown in the Table 5.

**Table 5. The econometric data $\{y_n, x_n\}$ on full (y) and stated (x) cost of bank shares in million pesetas for 20 Spanish banks and classification of these banks on values of $|y_n - y_n'|$ and $4x_n/y_n$.**

| $n$ | Bank name | $x_n$ | $y_n$ | $y_n - y_n'$ | $4x_n/y_n$ |
|---|---|---|---|---|---|
| | | ***First group*** | | | |
| 1 | Alicante | 8.018 | 29.988 | –9.21 | 1.07 |
| | | ***Second group*** | | | |
| 2 | Andalucía | 39.958 | 89.905 | 11.77 | 1.78 |
| 3 | Galicia | 16.538 | 34.233 | –15.35 | 1.93 |
| 4 | Mapfre | 84.128 | 173.303 | 41.33 | 1.94 |
| 5 | Popular | 274.906 | 558.059 | 193.52 | 1.97 |
| 6 | Castilla | 24.122 | 47.038 | –11.79 | 2.05 |
| 7 | Vasconia | 8.261 | 15.200 | –24.29 | 2.17 |
| 8 | Santander | 327.498 | 582.947 | 154.29 | 2.25 |
| 9 | Bankinter | 99.526 | 172.180 | 21.43 | 2.31 |
| | | ***Third group*** | | | |
| 10 | BBV | 568.672 | 892.815 | 170.16 | 2.55 |
| 11 | Credito Bale | 7.165 | 10.819 | –27.34 | 2.65 |
| 12 | Guipuzcuano | 20.513 | 30.974 | –23.45 | 2.65 |
| 13 | Valencia | 24.869 | 37.552 | –22.18 | 2.65 |
| 14 | Exterior | 232.184 | 325.057 | 12.60 | 2.86 |
| 15 | Atlantico | 47.899 | 64.664 | –23.15 | 2.96 |
| 16 | Argentaria | 474.760 | 612.440 | 4.27 | 3.10 |
| | | ***Fourth group*** | | | |
| 17 | Zaragozano | 40.065 | 42.068 | –36.19 | 3.81 |
| 18 | Pastor | 55.923 | 54.245 | –43.35 | 4.12 |
| | | ***Fifth group*** | | | |
| - | - | - | - | - | - |
| | | ***Sixth group*** | | | |
| 19 | Central Hisp | 604.973 | 429.531 | –337.37 | 5.63 |
| 20 | Vitoria | 11.239 | 7.418 | –35.70 | 6.06 |

As stated in [11], for the data array $\{y_n, x_n\}$ shown in the Table 5, the linear model $y = a_0 + a_1 x$ has maximum fitting errors (maximum values of the difference $|y_n - y_n'|$) for four Spanish banks (Popular, Santander, BBV and Central Hisp). And this fact connects only with big size of these banks.

Our opinion is that

*i*) Regression solutions obtained in [11] do not agree with big degree of the heterogeneity of dependence $\{y_n, x_n\}$ shown in Table 5;

*ii*) Correct classification of 20 Spanish banks may be performed on values of $4x_n/y_n$: the number $K_n$ of classification group for readings $(y_n, x_n)$ is determined by formula

$$K_n = g^+(4x_n / y_n), \qquad \textbf{[22]}$$

where function is "the nearest integer to $z$" (for example, $g^+(1.07)=1$, $g^+(1.78)=2$).

## 5. CONCLUSION

In this paper the emergence of computational difficulties of regression analysing heterogeneous dependences is explained by inadequacy of the fitting function $F(\mathbf{A}, \mathbf{X})$. It is stated that the following criteria and methods are to be constructed for bypassing the computational difficulties:

a) Objective criteria allowing on the given dependence $\{y_n, \mathbf{X}_n\}$ to determine whether the heterogeneity has effect on the form of regression solutions;

b) Methods of solving regression problems for inadequate fitting models.

It is demonstrate in Sections 3 and 4 that the problem (a) is solved completely by means of techniques of adaptive regression analysis. Thus in order to carry the theory of adaptive estimation of parameters to completion it is necessary to elaborate advanced estimation method for the case, when the fitting model is irremovable (global) inadequate. This problem is theme of our further studies. Some preliminary results on solving this problem may be found in [1, 12].

## REFERENCES

Y. V. CHEBRAKOV, *Parameters Estimation Theory in Measuring Experiments* (St.–Petersburg State Univ. Press, 1997).

Y. V. CHEBRAKOV & V. V. SHMAGIN, *Regression Data Analysis for Physicists and Chemists* (St.–Petersburg State Univ. Press, 1998).

G. PEREZ, Y.V. CHEBRAKOV y V.V. SHMAGUIN, *Analisis de datos: dificultades y metodos alternativos* (Universidad de Almeria, 1999).

K. R. DRAPER & H. SMITH, *Applied Regression Analysis* (Wiley & Sons, 1981).

Y.V. LINNIK, *Least squares method and foundations of mathematical–statistical theory of reduction of observations* (Moscow Publ. H. of Phys.–Math. Lit., 1958).

P. J. HUBER, *Robust Statistics* (Wiley & Sons, 1981).

P. J. ROUSSEEUW & A. M. LEROY, *Robust Regression and Outlier Detection* (Wiley & Sons, 1987).

R. GIRAUD & N. CHAIX, *Econometric* (Presses Univ. de France, 1989).

D. SALVATORE, *Statistics and Econometrics* (McGraw–Hill, 1982).

P. K. KATYSHEV, Y. R. MAGNUS & A. A. PERESECTSKIY, *Econometrics for beginners: Problems and solutions* (Moscow Publ. H. of Delo, 2002).

E. J. JIMÉNEZ & J. G. ROSAT, *Econometría aplicada* (Editorial AC, 1997).

J. SACKS & D. YLVISAKER, Linear estimation for approximately linear models, *Annals. Statist.*, 6(5), 1978, 1122–1137.