Proceedings of the 10th International Conference "Reliability and Statistics in Transportation and Communication" (RelStat'10), 20–23 October 2010, Riga, Latvia, p. 152-162. ISBN 978-9984-818-34-4
Transport and Telecommunication Institute, Lomonosova 1, LV-1019, Riga, Latvia

USING THE R ENVIRONMENT FOR PROCESSING STATISTICAL DATA AND CONSTRUCTION REGRESSION DEPENDENCES

Aleksandr Krivchenkov, Victor Lyumkis

Transport and Telecommunication institute
Lomonosova str. 1, Riga, LV-1019, Latvia
E-mail: aak@tsi.lv, vlyumkis@yahoo.com,

The considerable quantity of publications about significant opportunities statistical and free-of-charge R environment makes to pay to it the most steadfast attention. In the present work the review of the user opportunities of language and R environment is done. The explanatory of structure of environment and detailed elaboration of some opportunities will allow the user to get more deeply and more quickly skills of work in environment R. From the other side, the basic applicability of R environment is reduced to realization of classical and modern statistical methods. The problem of construction nonlinear regression dependences of dependent variable Y on independent variable X remains topical. On an example of this class of problems we show opportunities of language means of R environment for their solution. In work the so-called nuclear construction tools of nonlinear regresses are discussed, their comparison with usual estimations of a method of the least squares is done. For realization of such nuclear estimations modern means of environment R, which are reduced to an opportunity of a choice at a program level the so-called "width of a window" – bandwidths, which allows receiving nuclear regression estimation in the "best" image, are used.

Keywords: *R environment, regression analysis, nuclear estimations*

1. Introduction

Now the significant amount of publications [1, 2, 3] about opportunities of language and software of R environment for the statistical analysis of data forces to pay to it the most steadfast attention. R environment is freely distributed according to license GNU (General Public Licence), it is developed and accompanied by statisticians group, known as R Development Core Team [5]. R is a system for statistical analyses and graphics. R is both software and a language considered as a dialect of the S language created by the AT&T Bell Laboratories. S is available as the software S-PLUS commercialised by Insightful [7]. There are important differences in the designs of R and S. But R users may benefit from a large number of programs written for S and available on the internet [8].

R has many functions for statistical analyses and graphics; the latter are visualized immediately in their own window and can be saved in various formats (jpg, png, bmp, ps, pdf, emf, pictex, xfig; the available formats may depend on the operating system). The results from a statistical analysis are displayed on the screen; some intermediate results (P-values, regression coefficients, residuals and etc.) can be saved, written in a file, or used in subsequent analyses. The power of R environment is supported with such objects as functions and packages (group of functions) and we consider how it works for solution statistical problems.

2. The Structure and Usability of R Environment

When R is running, variables, data, functions, results, etc, are stored in the active memory of the computer in the form of objects. The user can do actions on these objects with operators (arithmetic, logical, comparison and etc.) and functions (which are themselves objects). All the actions of R are done on objects stored in the active memory of the computer: no temporary files are used (Fig. 1). The readings and writings of files are used for input and output of data and results. The user executes the functions via some commands. The results are displayed directly on the screen, stored in an object, or written on the

disk (particularly for graphics). Since the results are themselves objects, they can be considered as data and analysed as such. Data files can be read from the local disk or from a remote server through Internet.

The basic applicability of R environment is the realization of classical and modern statistical methods. Some of these means are built in basic set R, but many of them are delivered as packages using Comprehensive R Archive Network (CRAN) [6], – unique enough resource in the Internet. Now it is known more than 2400 packages delivered with R which contain modern means of the decision of various statistical problems.

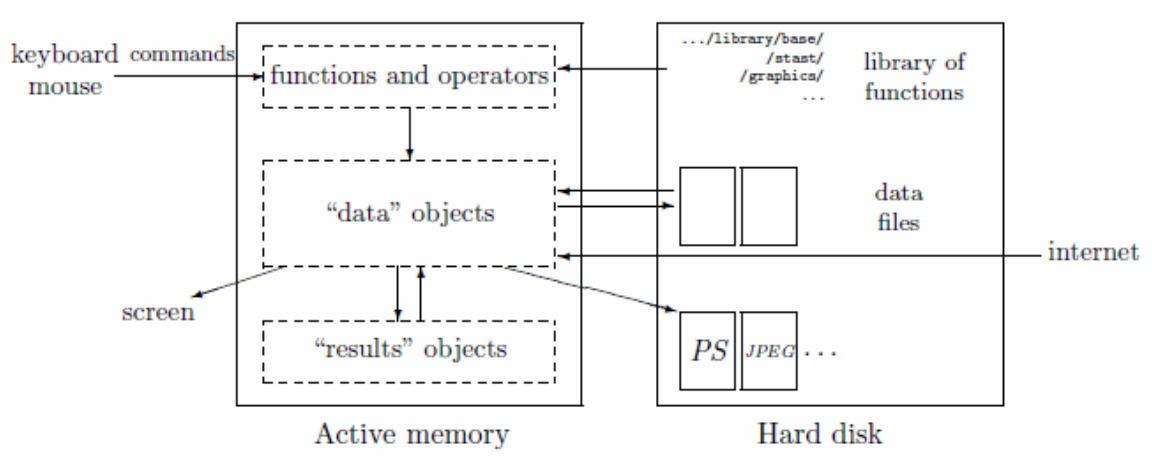

Figure 1. A schematic view of how R works

The following screen (Fig. 2) lists the standard packages which are distributed with a base installation of R. Some of them are loaded in memory when R starts; this can be displayed with the function *search()*, the other packages may be used after being loaded (for an example package *grid* with command *library(grid)*), the list of the functions in a package can be displayed with *library(help = grid)* or by browsing the help.

Package	Description
base	base R functions
datasets	base R datasets
grDevices	graphics devices for base and grid graphics
graphics	base graphics
grid	grid graphics
methods	definition of methods and classes for R objects and programming tools
splines	regression spline functions and classes
stats	statistical functions
stats4	statistical functions using S4 classes
tcltk	functions to interface R with Tcl/Tk graphical user interface elements
tools	tools for package development and administration
utils	R utility functions

Figure 2. The basic R packages

Many contributed packages add to the list of statistical methods available in R. They are distributed separately, and must be installed and loaded in R. A complete list of the contributed packages, with descriptions, is on the CRAN Web site. Several of these packages are recommended since they cover

statistical methods often used in data analysis. The recommended packages are often distributed with a base installation of R. They are briefly described (Fig. 3).

It is the main advantage of the R environment – placing functions in the libraries, which are collected in packages and are automatically available through the Internet in a global network of CRAN servers. There are two other main repositories of R packages: the Omegahat Project for Statistical Computing [9] which focuses on web-based applications and interfaces between software and languages, and the Bioconductor Project [10] specialized in bioinformatic applications (particularly for the analysis of microarray data).

Package	Description
boot	resampling and bootstraping methods
class	classification methods
cluster	clustering methods
foreign	functions for reading data stored in various formats (S3, Stata, SAS, Minitab, SPSS, Epi Info)
KernSmooth	methods for kernel smoothing and density estimation (including bivariate kernels)
lattice	Lattice (Trellis) graphics
MASS	contains many functions, tools and data sets from the libraries of "Modern Applied Statistics with S" by Venables & Ripley
mgcv	generalized additive models
nlme	linear and non-linear mixed-effects models
nnet	neural networks and multinomial log-linear models
rpart	recursive partitioning
spatial	spatial analyses ("kriging", spatial covariance, ...)
survival	survival analyses

Figure 3. Recommended R packages

R could seem too complex for a non-specialist. This may not be true actually. But the following table gives an overview of the type of objects representing data.

Object	**Mode**
vector	Numeric, character, complex, logical
factor	Numeric, character
array	Numeric, character, complex, logical
matrix	Numeric, character, complex, logical
data frame	Numeric, character, complex, logical
time set	Numeric, character, complex, logical
list	Numeric, character, complex, logical, function, expression

It is clear that manipulation of such objects requires detailed knowledge of the language and enough programmer skill.

The initial organization of environment in which operators of language are carried out in the lines, from a command line, looks rather unusually. There are many publications [3] where updating

environment the various routine operations are automated and graphic environments are offered. One of such GUI (Graphical User Interface) for R is R Commander [11]. Only those features are simplified:

- The R Commander provides several ways to get data into R. To enter data directly via New data set. This is a reasonable choice for a very small data set. To import data from a plain-text ("ascii") file or the clipboard, from another statistical package (Minitab, SPSS, or Stata), or from an Excel, Access, or dBase data set. To read a data set that is included in an R package, either typing the name of the data set (if you know it), or selecting the data set in a dialog box.
- The R Commander menus can be used to produce a variety of numerical summaries and graphs.
- Several kinds of statistical models can be fit in the R Commander using menu items. Models: linear models, generalized linear models, multinomial logit models, and proportional-odds models. Double-clicking on a variable in the variable-list box copies it to the model formula. The row of buttons above the formula can be used to enter operators and parentheses into the formula. You can also type directly into the formula fields, and indeed have to do so, for example, to put a term such as *log(income)* into the formula. You can type an R expression into the box labelled Subset expression; if supplied, this is passed to the subset argument of the *lm* function, and is used to fit the model to a subset of the observations in the data set.

On the next screenshot (Fig. 4) you can see how the work in R environment in some cases can be simplified using R commander.

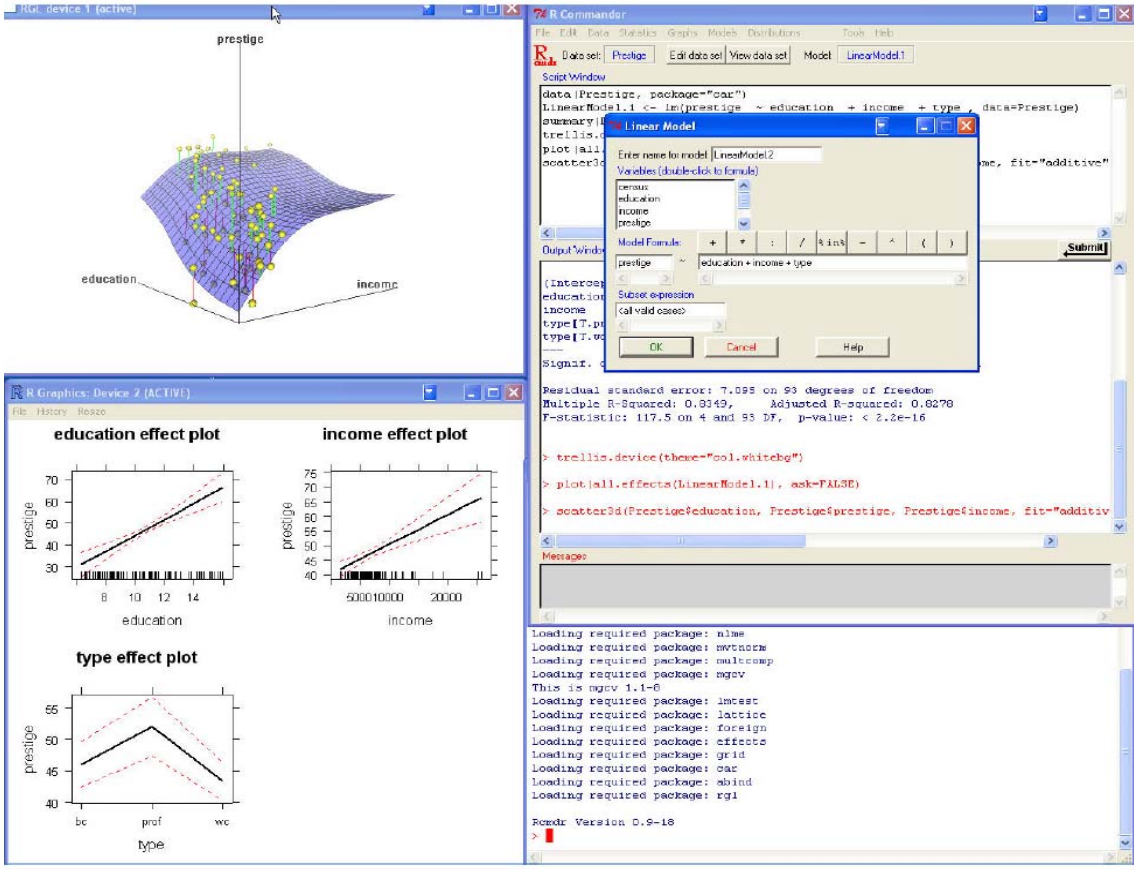

Figure 4. A view of how R Commander works

The next GUI for R is Java GUI for R – speak Jaguar (JGR) [11]. JGR provides only those features (Fig. 5): Console with several not valuable conveniences, Editor, Objects browser, Package manager, Spread sheet for manipulation with data and Help system.

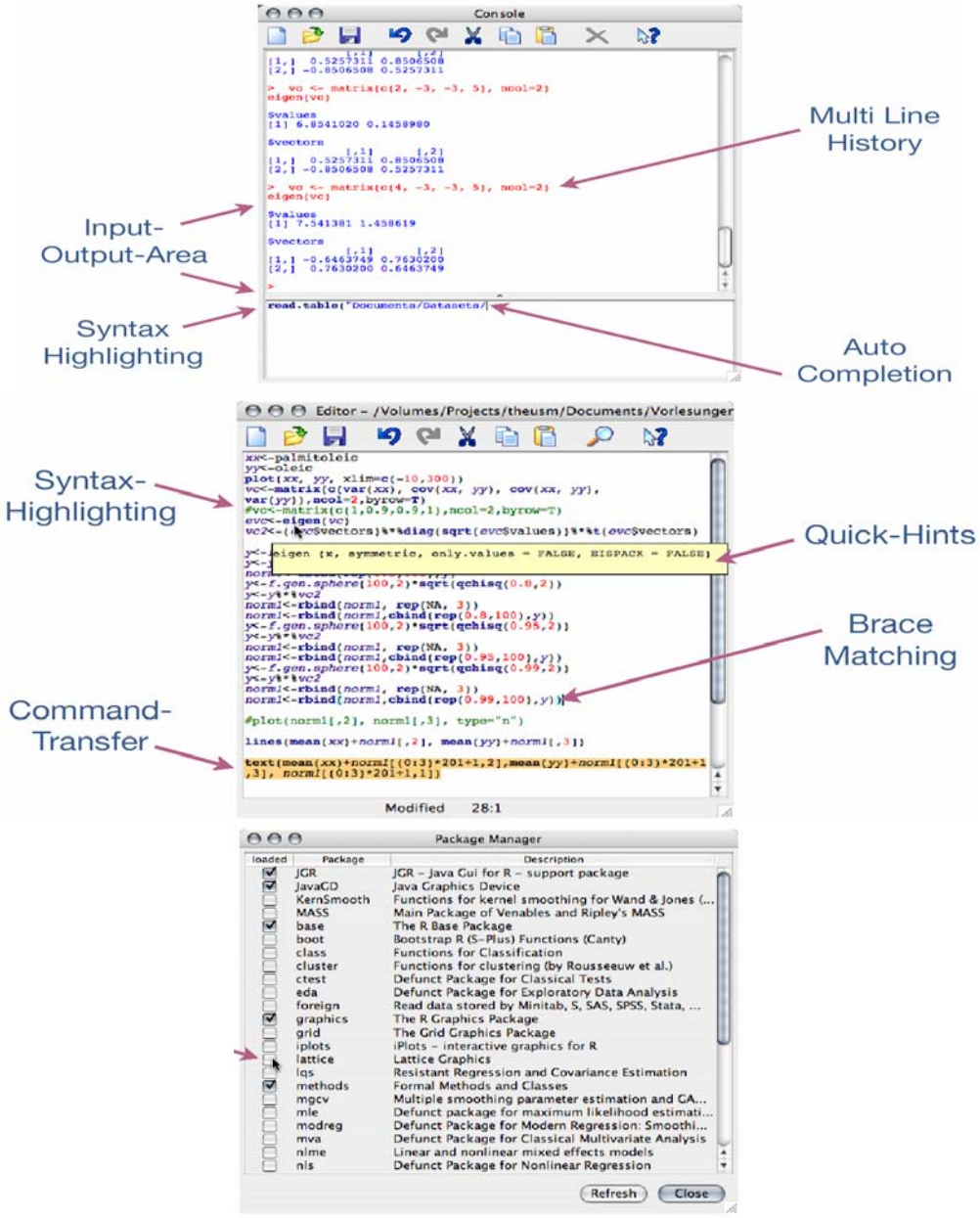

Figure 5. Screenshots of JGR user interface

So the functionality of the JGR package expands R environment and simplifies the packages control but the necessity of elaborate programming is still here. The GUI interfaces for R seems still have problem oriented nature and it will be so in the future.

3. Application of Language R for the Decision of Statistical Problems, Construction of Nonlinear Regression Dependences

As it was already mentioned, the basic applicability of R environment is reduced to realization of classical and modern statistical methods. In this connection it is possible to remind, that the problem of construction nonlinear regression dependences of dependent variable Y on independent variable X remains rather actual [13,14] and is one of central in a number of statistical problems. On an example of this class of problems we will show opportunities of language means of R environment of the solution of similar problems. In work the so-called nuclear construction tools of nonlinear regresses [4,14] are discussed, their comparison with usual estimations of a method of the least squares is done. For realization of such nuclear estimations modern means of environment R which are reduced, in particular,

to an opportunity of a choice at a program level the so-called " width of a window ", bandwidth – sizes, which allow receiving regression nuclear estimation in the "best" image, are used. So, the further results are based on consideration of a concrete example of nonlinear dependence Y from independent variable X, thus we were limited to detailed consideration only one example by virtue of the limited opportunities of present clause.

Let's notice further, that quite often it is possible to postulate the so-called regression equation of relative between dependent variable Y and independent variable X of a kind

$$Y = f(X,\theta) + \varepsilon, \tag{1}$$

where ε – some random variable having at everyone X=x distribution with a mathematical expectation $M(\varepsilon) = 0$ and a variation $D(\varepsilon) = \sigma^2$, $f(X,\theta)$ – some known smooth function depending on a vector of explaining variables X = ($x_1, x_{2,} ... x_k$) and a vector of parameters $\theta = (\theta_1, \theta_2, ..., \theta_p)$.

The standard way of a finding of a vector of unknown parameters, as is known [13], is reduced to minimization of the sum of squares of discrepancies – to use of a method of the least squares (MLS) that is reduced to minimization the function of a kind

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f(x_i,\theta)^2 , \tag{2}$$

where *n* – number of observations of dependent variable Y from a vector of explaining variables.

To find estimation $\hat{\theta}$ on MLS we should differentiate (2) on θ and equate the left parts to 0 for reception of the so-called system p normal equations. Difficulties of the decision of received system of the normal equations are well-known, and for its decision it is necessary to use iterative methods, moreover difficulties are aggravated with that there can be a set of stationary values of function $S(\theta)$. As opposed to the classical method of the least squares, considered above, it is possible to paraphrase a little in some cases required regression equation of relationship between dependent variable Y and a vector of explaining variables X and to reduce to model of a kind

$$Y = m(X) + \varepsilon, \tag{3}$$

where ε – some random variable having at everyone X=x distribution with a mathematical expectation $M(\varepsilon) = 0$ and a variation $D(\varepsilon) = \sigma^2$, m(X) – some smooth function.

By definition the conditional average of a continuous random variable can be presented thus

$$M(Y|X = x) = \int yg(y/x)dy = \int y\frac{f(x,y)}{f(x)}dy = m(x), \tag{4}$$

where g (y/x) – conditional density of random variable Y at fixed value X=x, and f (x, y) joint density of random variables X, Y.

The estimation of a conditional average turns out replacement of unknown functions of density joint and marginal distributions, f (x, y) and f (x), their "good" estimations. Thus "good" estimation of a conditional average is the nuclear estimation (5) which detailed design is described, for example, in works [4, 14]. It allows reducing an estimation of conditional average Y at everyone x to an estimation of a kind:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h})}{\sum_{i=1}^{n} K(\frac{X_i - x}{h})}, \tag{5}$$

where K(z) – value of function of density of standard normal distribution in a point z.

In table 1 the given sample dependences of variable Y from x are presented in the assumption that true dependence looks like

$$Y = \frac{\theta_1 x}{(\theta_2 + x)} + \varepsilon . \tag{6}$$

Table 1. Dependence Y from x for model (6)

x	2	2	0.667	0.667	0.4	0.4	0.286	0.286	0.222	0.222	0.2	0.2
Y	0.0615	0.0527	0.0334	0.0258	0.0138	0.0258	0.0129	0.0183	0.0083	0.0169	0.0129	0.0087

It is necessary by means of method of the least squares to receive an estimation of parameters of this nonlinear model and it is possible to emphasize, that search of parameters of model is carried out on small sample and that is characteristic for some practical problems. Methodologically, first of all, there arises a problem of choice initial approaches for a correct finding of estimations of parameters on a method of the least squares. In this connection, rejecting ε in model 6, we can receive system of the equations of a kind

$$Y_i x_i = \theta_1 x_i - Y_i \theta_2,\ i = 1, 2, \ldots n,\ n = 12.$$

Designating a required vector $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ and using the solution of plural linear regression, we shall receive $\theta = (Z^T Z)^{-1} Z^T U = \begin{pmatrix} 0.084 \\ 1.031 \end{pmatrix}$, where $Z = \begin{pmatrix} x_1 & -Y_1 \\ \ldots & \ldots \\ x_n & -Y_n \end{pmatrix}, U = \begin{pmatrix} Y_1 x_1 \\ \ldots \\ Y_n x_n \end{pmatrix}$.

Having received initial approach, we use the standard method of search of such estimations of parameters of model (6) which minimize the value of discrepancy (2). Taking private derivatives on parameters in (2) and having equated them to zero, we shall receive corresponding system of the equations of a kind

$$\begin{cases} \sum\limits_{i=1}^{n} \left(\frac{x_i y_i}{x_i + \theta_2}\right) - \theta_1 \sum\limits_{i=1}^{n} \frac{x_i^2}{(x_i + \theta_2)^2} = 0 \\ \sum\limits_{i=1}^{n} \left(\frac{x_i y_i}{(x_i + \theta_2)^2}\right) - \theta_1 \sum\limits_{i=1}^{n} \frac{x_i^2}{(x_i + \theta_2)^3} = 0 \end{cases}. \quad (7)$$

The numerical solution of this system of the equations in R at presence initial approaches $\theta = \begin{pmatrix} 0.084 \\ 1.031 \end{pmatrix}$ gives the final solution of a kind $\theta = \begin{pmatrix} 0.10564 \\ 1.70269 \end{pmatrix}$. The list of R commands for the solution of system of the equations (7) is as follows:

```
library("rootSolve")
xt<-c(2,2,0.667,0.667,0.4,0.4,0.286,0.286,0.222,0.222,0.2,0.2)
yt<-c(0.0615,0.0527,0.0334,0.0258,0.0138,0.0258,0.0129,0.0183,0.0083,0.0169,0.0129,0.0087)
model<-function(o)  c(
F1=-sum(xt*yt/(xt+o[2])) + sum(o[1]*(xt^2)/((xt+o[2])^2)),
F2=-sum(xt*yt/((xt+o[2])^2)) + sum(o[1]*(xt^2)/((xt+o[2])^3)))
MRres<-multiroot(model,c(0.08,1))
yregr <- function(x){
MRres$root[1]*(x/(MRres$root[2]+x))}
plot(xt,yt)
points(xt,yt, col='red',bg='red',pch=21)
lines(xt,yregr(xt), col = 'blue', lty = "dotted")
```

Classical regression dependence can be presented now as $YR(x) = 0.1056 \frac{x}{(1.702 + x)}$.

On Figure 6 the corresponding regression smoothing for the given correlation field is presented.

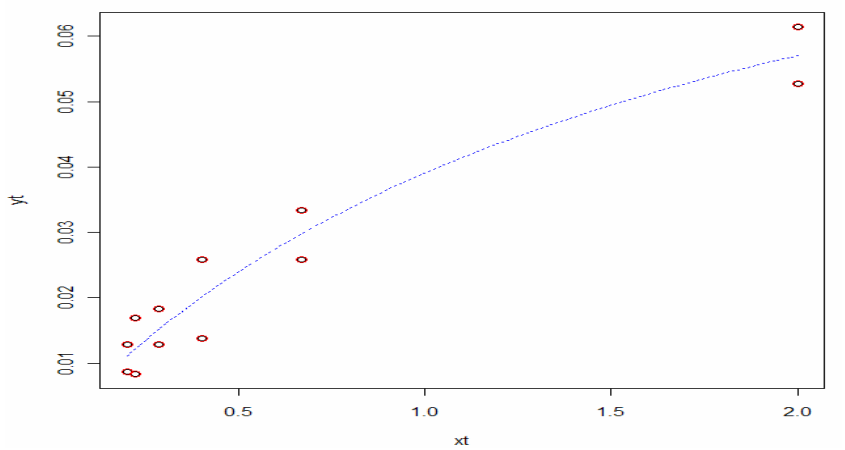

Figure 6. A field of variability and corresponding classical regression for the model (6).

As opposed to classical regression nuclear regression as it was marked earlier, is reduced to realization of the formula (5) where all is simple enough except for a choice of a step h. A key to carrying out of the further researches connected with nonparametric nuclear estimation is the choice of size h – width of a window (bandwidth). There were many works connected with an optimum choice of width of a window from which, certainly, it is possible to allocate works [4, 14]. In them authors offer various algorithms of a choice h, thus concentrate the attention to algorithm cross-validation. The merit of authors of the specified works is reduced to that they bring theoretical algorithms of a choice of width of a window to application of convenient programs in the practical plan. All these programs are realized in the package *np*. We shall give an example applications of the key function *npregbw()* (Nonparametric Regression Bandwidth Selection) from the package, providing a choice optimum size of width of a window by consideration of our model (6):

```
library(np)
x1<-c(2,2,0.667,0.667,0.4,0.4,0.286,0.286,0.222,0.222,0.2,0.2)
y1<c(0.0615,0.0527,0.0334,0.0258,0.0138,0.0258,0.0129,0.0183,0.0083,0.0169,0.0129,0.0087)
m1<-npregbw(formula=y1~x1,tol=0.1,ftol=0.1)
m1
Regression Data (12 observations, 1 variable(s)):x1
Bandwidth(s): 0.1379053
Regression Type: Local-Constant
Bandwidth Selection Method: Least Squares Cross-Validation
Formula: y1 ~ x1
Bandwidth Type: Fixed
plot(m1,xlim=c(0,2),ylim=c(0,0.1))
points(x1,y1,cex=1.25,col="red")
```

On the following Figure 7 the curve of nuclear smoothing for data of table 1 with optimum choice of a step h=0.138 is presented.

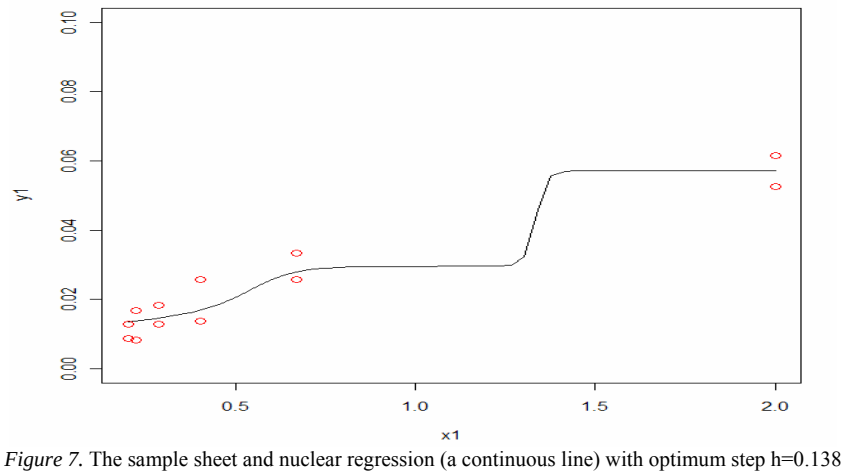

Figure 7. The sample sheet and nuclear regression (a continuous line) with optimum step h=0.138

In our work some analysis of the chosen smoothing dependences connected both with a choice of value of a step h, and with research of quality of smoothing at increase size of sample is done. Let we in the beginning have a sample of volume n = 20 from model 6, simulated at true parameters $\theta_1 = 1, \theta_2 = 2$. On Figure 8 the corresponding correlation field and smoothing regression dependence on a base MLS is presented.

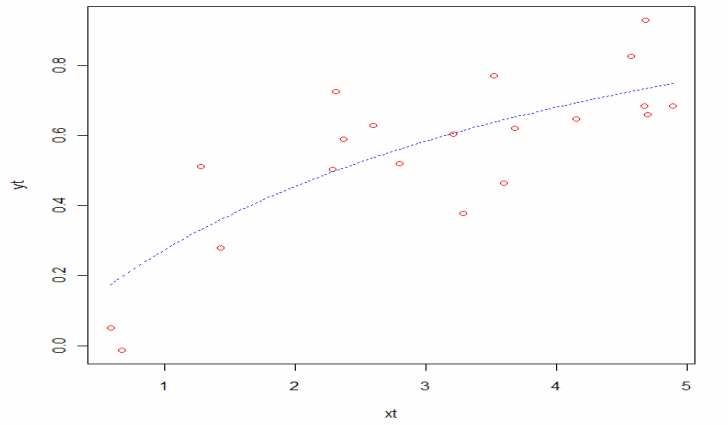

Figure 8. A field of variability and corresponding classical regression for model (6) at volume of sample n=20

On Figure 9 the nuclear smoothing curve received at use of the formula (5) and optimum chosen size of a step h = 0.508 is presented.

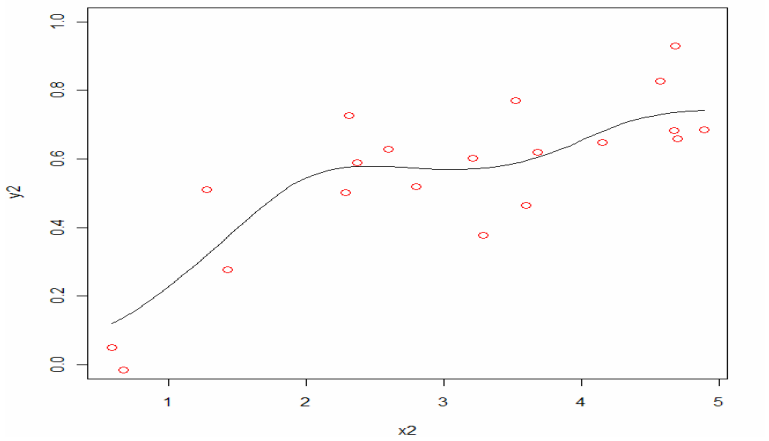

Figure 9. A field of variability and corresponding nuclear regression for model (6) at volume of sample n=20 and an optimum step h=0.508

Let's notice, that even visually the optimality of a choice of a step h=0.508 is visible.

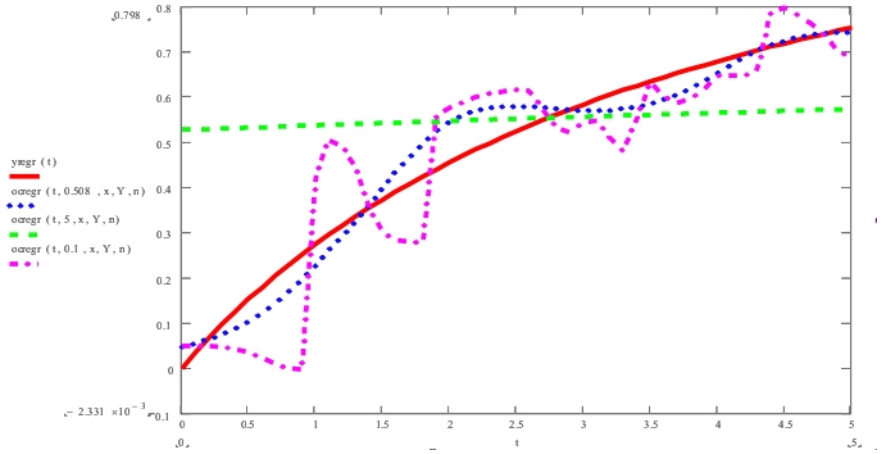

Figure 10. Smoothing curves: red continuous – regression on MLS, a dark blue dotted line – nuclear smoothing at an optimum step h=0.508, a green dotted line – nuclear excessive smoothing at h=5, an orange dotted line-insufficient nuclear smoothing at h=0.1

From the above-stated it is possible to make a conclusion, that in conditions when a method of the least squares to realize difficultly, the method of nuclear smoothing in some cases can be not the worst alternative. Apparently, especially evidently it will be shown at rather great volume of sample observations. On Figure 11 the simulated sample in volume n=200 for model (6) is presented and received corresponding classical regression dependence.

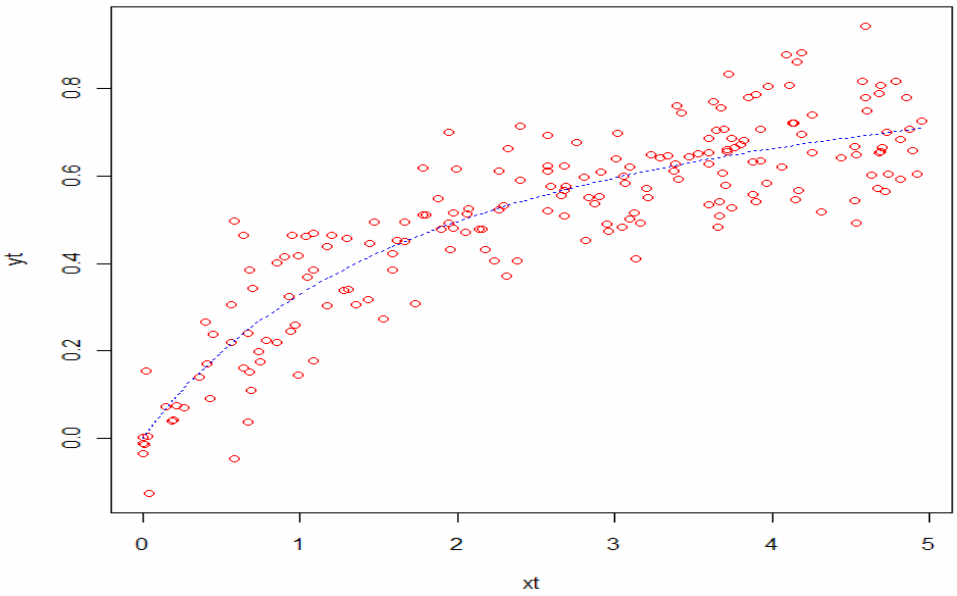

Figure 11. A field of variability and corresponding classical regression for model (6) at volume of sample n=200

On Figure 12 corresponding nuclear smoothing dependence for the same data is presented at volume of sample n=200.

Comparing Figures 11 and 12, we see comprehensible enough coincidence regression curve MLS with corresponding nuclear smoothing.

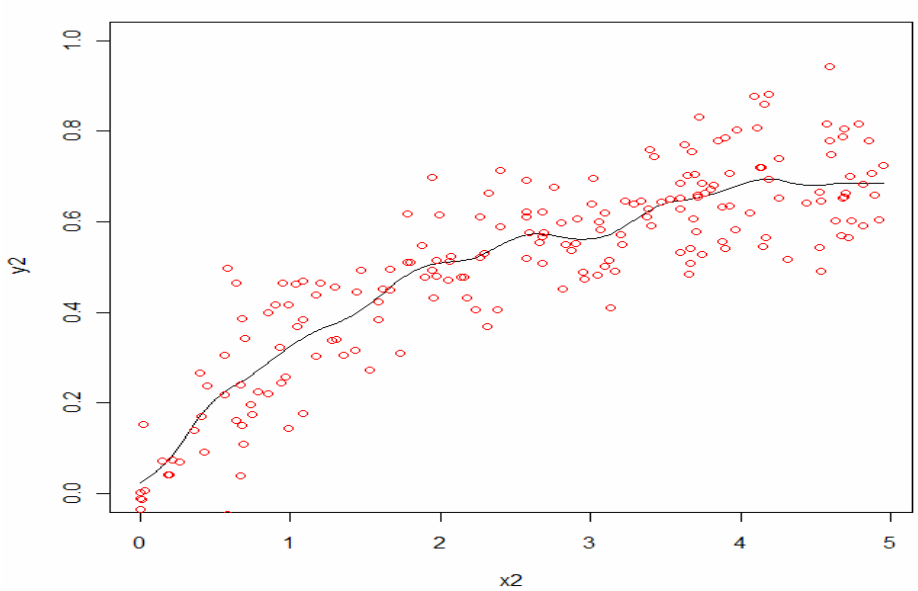

Figure 12. A field of variability and corresponding nuclear regression for model (6) at volume of sample n=200

4. Conclusions

The use of various statistical procedures and methods in practical activities by many engineers demands universal use of qualitative software. Such characteristic is deserved with R environment. In the present work authors bring to a focus to the user opportunities and structure of R environment suitable for the solution of a wide class of statistical problems. It is paid attention to that as function environment, on

every possible modifications of environment, what in whole can facilitate the user work. From the other hand, in work the concrete approach to the solution of an actual statistical problem – construction on sample data nonlinear regression model by means of R environment is presented. On a concrete example the comparative analysis of construction of models of nonlinear regression, as by means of the classical regression and with the use of nuclear regression, is shown. In conditions of nonparametric models the use of the nuclear approach obviously is reasonable. On the other hand even at presence of parametrical model it is possible to use nuclear regression, if difficulties of realization of parametrical model (for example, a choice initial approach) take place. The difference between applications of both approaches is levelled, as shown in the study, at increase volume of sample data. The use of R for the solution of similar problems is difficult to overestimate. Convenient software of a choice of an optimum weight step h in the formula of nuclear smoothing and the modern graphic interface allow the user to reach the demanded result of data smoothing.

References

1. Crawley, Michael J. *The R Book*, John Wiley & Sons, 2007. 939 p.
2. Dalgaard, P. *Introductory Statistics with R*. New York: Springer, 2002. 267 p.
3. Heiberger, Richard M., Neuwirth, Erich. *R Through Excel*. New York: Springer, 2009. 340 p.
4. Racine, Jeff & Li. Nonparametric estimation of regression functions with both categorical and continuous data, *Journal of Econometrics,* Vol. 119(1), 2004. pp. 99-130,
5. R Development Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,* Vienna, Austria, 2005. ISBN 3-900051-07-0, URL: http://www.R-project.org
6. http://cran.r-project.org/
7. http://www.insightful.com/products/splus/default.asp
8. http://stat.cmu.edu/S/
9. http://www.omegahat.org/R/
10. http://www.bioconductor.org/
11. http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/index.html
12. http://rforge.net/JGR/
13. Draper, N. and Smith, H. *Applied Regression Analysis. Third ed.* New York: John Wiley and Sons, 1998. 706 p.
14. Härdle, W., Muller, M., Sperlich, S, Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin-Heidelberg-New York: Springer, 2004. 328 p.