Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 160 (2009) 357 – 370
FUZZY sets and systems
www.elsevier.com/locate/fss

Estimation of a simple linear regression model for fuzzy random variables☆

Gil González-Rodríguez[a], Ángela Blanco[b], Ana Colubi[b], M. Asunción Lubiano[b,*]

[a] Research Unit on Intelligent Data Analysis and Graphical Models, European Centre for Soft Computing, 33600 Mieres, Spain
[b] Dpto. de Estadística e I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain

Available online 23 July 2008

Abstract

A generalized simple linear regression statistical/probabilistic model in which both input and output data can be fuzzy subsets of $\mathbb{R}^p$ is dealt with. The regression model is based on a fuzzy-arithmetic approach and it considers the possibility of fuzzy-valued random errors. Specifically, the least-squares estimation problem in terms of a versatile metric is addressed. The solutions are established in terms of the moments of the involved random elements by employing the concept of support function of a fuzzy set. Some considerations concerning the applicability of the model are made.

Keywords: Linear regression model; Fuzzy random variable; Fuzzy-arithmetic approach; Support function; Point estimation; Least-squares method; Minimization problem

1. Introduction

Consider the situation in which the value assigned to each and every possible outcome in the performance of a random experiment can be described by means of a fuzzy set. Fuzzy random variables (hereafter FRVs for short) in Puri and Ralescu's sense [31] were introduced as an extension of random sets to model this kind of situation, and they can be employed in many contexts (see, for instance, examples in [3,15,11]).

A basic problem concerning random elements/variables is focussed on estimating the functional dependence of a response variable on a set of explanatory variables, on the basis of their joint observation on n statistical units. In this respect, descriptive and inferential regression and correlation analyses have been widely studied in the literature.

When dealing with FRVs, regression and correlation analyses become more complex due on one hand to the interpretation of the relationships and, on the other hand, to the difficulties in finding optimal solutions and measures of their adequacy.

☆ The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2006-07501 and by the Govern of Asturias through the Innovation, Technology and Science Plan (PCTI) 2006–2009, Grant BP06-010. Their financial support is gratefully acknowledged.

* Corresponding author. Tel.: +34 985103190; fax: +34 985103356.
E-mail addresses: gil.gonzalez@softcomputing.es (G. González-Rodríguez), blancoangela.uo@uniovi.es (Á. Blanco), colubi@uniovi.es (A. Colubi), lubiano@uniovi.es (M.A. Lubiano).

0165-0114/$ - see front matter
doi:10.1016/j.fss.2008.07.007

The regression problem with fuzzy data has been previously treated in the literature from different points of view and by considering different kinds of input/output data. In this sense, Näther [29] and Coppi et al. [4] provide extensive reviews of the main approaches to regression in a fuzzy setting and may serve as general references.

In the context of inferential regression analysis, two main problems come up, namely, the model specification and the estimation of a given model. The first problem is focused on establishing the suitability of a given model for a set of data, and it is traditionally addressed *a posteriori*, in order to take advantage of the estimators found for the second one. Thus, as a first step, a new simple linear model for FRVs in $\mathbb{R}^p$ is formulated in this paper in terms of the usual fuzzy arithmetic. This linear model has been analyzed in the particular case of convex compact sets in $\mathbb{R}^p$ by González-Rodríguez et al. [16] as a broad alternative to that of Gil et al. [13]. As a consequence, this paper means a generalization of the studies carried out by Gil et al. [12–14] and González-Rodríguez et al. [17,16]. The estimation of the linear regression (i.e., the conditional expectation) is addressed. As in many previous works involving fuzzy data, the least-squares technique is employed.

The model for FRVs considered in this work agrees with those in Diamond [5,7], Diamond and Körner [9], Körner and Näther [19], Näther [28], Wünsche and Näther [32] and Krätschmer [21], in the sense of involving the same regression function under particular conditions. The objective in those works has been twofold. On one hand, the least-squares solution of the fitting or approximation problem under particular conditions, or by considering a particular kind of fuzzy numbers (such as the so-called LR fuzzy numbers), has been addressed (see, for instance [5,7,9,32]). On the other hand, some theoretical properties of general estimators have been also established (see [19,28,21–23]).

The main contribution of the current paper w.r.t. the previous works in the literature is the obtaining of explicit expressions for suitable estimators of the model under general conditions and for a wide class of fuzzy data generated from the model. It should be noted that, in contrast to what happens in the real case, the numerical fitting problem and the statistical estimation problem for the linear regression are different in the fuzzy case. The reason is that the lack of linearity of the space of fuzzy data makes that considering or not the data generation process lead to different restrictions in the minimization problem. This paper concerns the estimation problem. To be precise, we consider a problem in which two FRVs $\mathcal{X}, \mathcal{Y}$ are related through a probabilistic model involving a linear function (namely, $\mathcal{Y} = a\mathcal{X} + \varepsilon$, with $E(\varepsilon) = B$, see Section 3 for details), and the aim is to estimate the parameters of the model (namely, a and B) on the basis of a random sample obtained from the observation of $\mathcal{X}$ and $\mathcal{Y}$ on n statistical units taking into account that the data are generated from the model (namely, $\mathcal{Y}_i = a\mathcal{X}_i + \varepsilon_i$ for all $i = 1, \ldots, n$).

It is important to remark that the model considered in this paper (and all those in [5,7,9,19,22,23,29,32]) establishes the relationship between $\mathcal{X}$ and $\mathcal{Y}$ through only two parameters a and B. However, fuzzy sets are complex data (not points of an Euclidean space, but functions), and for this reason the applicability of the model is limited, since only two parameters are frequently not enough to gather the expressiveness of the class of fuzzy sets. To overcome this inconvenience particular models applicable for concrete families of fuzzy sets may be considered. For instance, if we restrict to the class of LR fuzzy numbers, separated models for the center and the spreads can be analyzed. The separated models can fit better many practical situations, because six parameters (or more if cross-combinations are allowed) may be considered. The inconvenience of these models is that they imply non-negativity restrictions, which are very difficult to handle in the inferential context (see, for instance [10,24]). Note that it is not possible to consider separated models for general fuzzy sets in families not determined by a finite quantity of real numbers.

The paper is organized as follows: In Section 2 some notations and preliminary concepts concerning FRVs will be presented. The considered simple linear model is stated in Section 3, and the least-squares estimators of the linear regression are obtained in Section 4. In Section 5, a real-life example to illustrate the results is included. Finally, in Section 6, some relevant conclusions and future directions related to the study developed in this paper will be commented.

2. Preliminaries

Let $\mathcal{K}_c(\mathbb{R}^p)$ be the class of the non-empty compact convex subsets of $\mathbb{R}^p$ and let $\mathcal{F}_c(\mathbb{R}^p)$ denote the class of normal and convex upper semicontinuous fuzzy sets of $\mathbb{R}^p$ with bounded closure of the support, that is,

$$\mathcal{F}_c(\mathbb{R}^p) = \{U : \mathbb{R}^p \to [0,1] | U_\alpha \in \mathcal{K}_c(\mathbb{R}^p) \text{ for all } \alpha \in [0,1]\},$$

where U_α is the α-level set of U if $\alpha \in (0,1]$, and $U_0 = \text{cl}(\{x \in \mathbb{R}^p | U(x) > 0\}$. An example of fuzzy set on $\mathcal{F}_c(\mathbb{R})$ we will consider is the so-called *trapezoidal* fuzzy number $A = \text{Tra}(A^c, A^s, A^l, A^r)$ with the center $A^c \in \mathbb{R}$ and the