Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 267–286
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Least squares estimation of a linear regression model with LR fuzzy response

Renato Coppi[a,*], Pierpaolo D'Urso[b], Paolo Giordani[a], Adriana Santoro[b]

[a]*Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma "La Sapienza", P.le A. Moro, 5 - 00185 Roma, Italy*
[b]*Dipartimento di Scienze Economiche, Gestionali e Sociali, Università degli Studi del Molise, Via De Sanctis, 86100 Campobasso, Italy*

Available online 24 May 2006

Abstract

The problem of regression analysis in a fuzzy setting is discussed. A general linear regression model for studying the dependence of a LR fuzzy response variable on a set of crisp explanatory variables, along with a suitable iterative least squares estimation procedure, is introduced. This model is then framed within a wider strategy of analysis, capable to manage various types of uncertainty. These include the imprecision of the regression coefficients and the choice of a specific parametric model within a given class of models. The first source of uncertainty is dealt with by exploiting the implicit fuzzy arithmetic relationships between the spreads of the regression coefficients and the spreads of the response variable. Concerning the second kind of uncertainty, a suitable selection procedure is illustrated. This consists in maximizing an appropriately introduced goodness of fit index, within the given class of parametric models. The above strategy is illustrated in detail, with reference to an application to real data collected in the framework of an environmental study. In the final remarks, some critical points are underlined, along with a few indications for future research in this field.

Keywords: Multiple linear regression; Least-squares approach; LR fuzzy response variable; Analysis of uncertainty; Goodness of fit

1. Introduction

The study of relationships between real world phenomena is one of the basic aims in science, and plays a fundamental role in decision making in everyday life. Statistical regression analysis is a powerful tool in this domain. It concerns the analysis of the statistical link between a "response" variable (say Y) and a set of "explanatory" or "predictive" variables (say $X_1, \ldots, X_m$), on the basis of a set of observations of the joint behavior of these variables.

Several sources of uncertainty may affect this kind of study, including: (a) the sampling effects due to the selection of the specific set of statistical units on which the analysis is carried out; (b) the ignorance concerning the type of model expressing the dependence relationship of Y on the X_j's ($j = 1, \ldots, m$); (c) the ignorance about the specific model

* Corresponding author. Tel.: +390649910731; fax: +39064959241.
E-mail addresses: renato.coppi@uniroma1.it (R. Coppi), durso@unimol.it, pierpaolo.durso@uniroma1.it (P. D'Urso), paolo.giordani@uniroma1.it (P. Giordani), santoro@unimol.it (A. Santoro).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.04.036

to be selected within a given class of regression models; (d) the imprecision of the mechanism ruling the dependence relationship, whatever the type of model considered; (e) the imprecision/vagueness of the observed data.

In the literature, the uncertainty stemming from sources (a) and (c) has been widely explored. The classical theory of Linear Models (e.g. Graybill, 1961; Neter et al., 1996) provides the most relevant piece of methodology in this respect. More recent developments enlarge the scope of the traditional approach to non-parametric methods and statistical learning techniques (e.g. Hastie et al., 2001). To a certain extent, also uncertainty of type (b) is dealt with in the latter case. Bayesian analysis provides a different viewpoint in coping with uncertainty of types (a), (b) and (c), making a "full" use of probability in jointly managing the randomness of the data and the uncertainty concerning the models and their parameters (see, e.g., Gelman et al., 1995). In the latter case, the introduction of prior probability distributions over the set of regression parameters represents a way of dealing with the "imprecision" of the regression mechanism (source (d)).

However, in the traditional framework, uncertainty of type (e) is not envisaged. The data are considered as "crisp" empirical information to be fed into the "Statistical Reasoning" process, which may be affected only by other sources of uncertainty.

In this paper, we introduce and develop a statistical regression model enabling us to manage imprecise (fuzzy) data, with particular reference to the response variable, which, in this case, will be denoted by $\tilde{Y}$. The fuzziness of $\tilde{Y}$ may stem from various sources: (i) imprecision in measuring the empirical phenomenon represented by Y; (ii) vagueness of Y when this is expressed in linguistic terms; (iii) partial or total ignorance concerning the values taken by Y on specific observational instances; (iv) "granularity" of the Y-variable, with reference to the way it is defined and used in the analysis (e.g. the age of a person may be described in terms of 5-year intervals, or just as "young", "middle age", "old"; to each of these "granulations" there is associated a different amount of uncertainty; see Zadeh, 2005). We argue that, in the above mentioned situations, an appropriate fuzzification of Y may exploit the available information in a more complete and efficient way, than just reducing it to a single value (a number, or a category).

Several approaches to regression analysis for fuzzy data have been developed, starting from the pioneering works by Tanaka et al. (1982), Celminš (1987), Diamond (1988), based respectively on possibilistic (the first one) and least squares principles. A brief overview of the various proposals in this domain will be given in Section 2.2.

The present work focuses on the observational situation where the response variable is fuzzy and the explanatory variables are crisp quantitative characters. In this context, we set up a general linear regression model, assuming that the membership function of the response variable belongs to the LR family. This is illustrated in Section 2.3. Then, in Section 3, the estimation procedure is described. This is based on the least squares (LS) principle. In this connection, an appropriate distance function for LR fuzzy variables is introduced and the corresponding LS objective function is defined (Section 3.1). An iterative LS solution is shown in Section 3.2 and some relevant properties of this solution are proved in Section 3.3, while in Section 3.4 specific methodological aspects related to the estimation procedure are discussed. In Section 4, a procedure for assessing the imprecision associated with the estimates of the regression coefficients obtained by the proposed model, is illustrated. This involves the use of an implicit fuzzy regression model with LR fuzzy coefficients, whose parameters (centers and spreads) are estimated by means of independent LS equations with input given by the estimates obtained by the basic regression model (exploiting fuzzy arithmetic relationships).

Next, in Section 5, the problem of model selection is faced. The corresponding source of uncertainty (of the above mentioned type (c)), is represented by the selection of an "optimal" regression function from among a suitable class of parametric models (Section 5.1). In Section 5.2, a selection tool is suggested, based on an appropriate decomposition of the sum of squares of the response variable and the construction of a multiple determination coefficient. This is used in setting up a suitable selection procedure in the above mentioned parametric class. In Section 6, an application to real world data collected in the framework of an environmental study is utilized for showing the informational capability of the proposed strategy of regression analysis. In this connection, another source of uncertainty is discussed, namely the one related to the sampling variation (source (a)). A possible way for coping with it is suggested, based on a bootstrap procedure for estimating the standard errors pertaining to the estimates of the various parameters introduced in the regression model. Finally, in Section 7, we make a few concluding remarks concerning some critical points of the proposed strategy of analysis, and outline possible perspectives of future research in this domain.