# Multiple Linear Regression for Histogram Data using Least Squares of Quantile Functions: a Two-components model.

Rosanna Verde*, Antonio Irpino*

* Dipartimento di Scienze Politiche "J. Monnet"
Seconda Universitá degli Studi di Napoli
Viale Eliitico 31, Caserta, Italy
rosanna.verde@unina2.it, antonio.irpino@unina2.it

**Abstract.** Histograms are commonly used for representing summaries of observed data and they can be considered non parametric estimates of probability distributions. Symbolic Data Analysis formalized the concept of *histogram symbolic variable*, as a variable which allows to describe statistical units by histograms instead of single values. In this paper we present a linear regression model for multivariate histogram variables. We use a Least Square estimation method where the sum of squared errors is defined according to the $\ell_2$ Wasserstein metric between the observed and the predicted histogram data. Consistently with the $l_2$ Wasserstein metric, we solve the Least Square computational problem by introducing a suitable inner product between two vectors of histogram data. Finally, measures of goodness of fit are discussed and an application on real data shows some interpretative advantages of the proposed method.

## 1 Introduction

Symbolic Data Analysis (SDA) is a relatively new statistical approach concerning the analysis of *higher level individuals* (like typologies, classes or concepts) that are described by *multi-valued variables* (Bock and Diday (2000); Diday and Noirhomme-Fraiture (2008)).
The term *symbolic variable* was coined in order to introduce such new set-valued descriptions. In a classic data table ($n \times p$ individuals per variables) each individual is described by a vector of values, similarly, in a *symbolic data table* each individual is described by a vector of set-valued descriptions (like intervals of values, histograms, set of numbers or of categories, sometimes equipped with weights, probabilities, frequencies, an so on). According to the taxonomy of symbolic variables presented in Bock and Diday (2000) we may consider as numerical symbolic variables all those symbolic variables whose support is numeric. The main quantitative and multi-valued symbolic variable types are *interval* and *modal numeric symbolic variable*.
Linear regression models allow to modeling the linear relationship between a quantitative response *symbolic* variable and a set of independent or explicative quantitative *symbolic* variables of the same type. Regression models for interval data extend the classic linear model to interval variables (see Afonso et al. (2008) and Lima Neto and de Carvalho (2010)) and the therein