STATISTICAL ANALYSIS AND DATA MINING

Original Article

Linear regression model with histogram - valued variables

Sónia Dias, Paula Brito

First published: 22 April 2015
https://doi.org/10.1002/sam.11260
Cited by: 3

Abstract

Histogram - valued variables are a particular kind of variables studied in *Symbolic Data Analysis* where to each entity under analysis corresponds a distribution that may be represented by a histogram or by a quantile function. Linear regression models for this type of data are necessarily more complex than a simple generalization of the classical model: the parameters cannot be negative; still the linear relation between the variables must be allowed to be either direct or inverse. In this work, we propose a new linear regression model for histogram - valued variables that solves this problem, named *Distribution and Symmetric Distribution Regression Model*. To determine the parameters of this model, it is necessary to solve a quadratic optimization problem, subject to non - negativity constraints on the unknowns; the error measure between the predicted and observed distributions uses the Mallows distance. As in classical analysis, the model is associated with a goodness - of - fit measure whose values range between 0 and 1. Using the proposed model, applications with real and simulated data are presented.