Regression models with responses on the unit interval: specification, estimation and comparison

Wagner Bonat

PhD student from Universidade Federal do Paraná, Brazil, and University of Southern Denmark

March 11th, 2014

Abstract

Regression models are widely used in a diversity of application areas to describe associations between explanatory and response variables. The initially and frequently adopted Gaussian linear model was gradually extended to accommodate different kinds of response variables. These models were later described as particular cases of the generalized linear models (GLM). The GLM family allows for a diversity of formats for the response variable and functions linking the parameters of the distribution to a linear predictor. This model structure became a benchmark for several further extensions and developments in statistical modelling such as generalized additive, overdispersed, zero inflated, and other models. Response variables with values restricted to an interval, often (0,1), are common in social sciences, agronomy, psychometrics and other areas. Beta or Simplex distributions are often used, although other options are available in the literature. In this seminar, a generic structure is presented to define a set of regression models for response variables on (0,1), not only including the usual formats but allowing for a wider range of models. Individual models are defined by choosing three components: the probability distribution for the response; the function linking the parameter of the distribution of choice with the linear predictor; and the transformation function for the response. We report results of the analysis of four different data sets using Beta, Simplex, Kumaraswamy and Gaussian distributions. For the link and transformation functions the logit, probit, complementary log-log, log-log, Cauchit and Aranda-Ordaz are considered. Likelihood-based analysis for model fitting, comparison and model choice are carried out in a unified way and computer code is available. Our results show the there is no overall best model within this class, highlighting the importance of investigating a wide range of models for each problem at hand.