CLADAG2017

Book of Short Papers

Editors: Francesca Greselin,
Francesco Mola and Mariangela Zenga

via Sottoriva, 9 - 46100 Mantova (MN)
P. IVA 02346110204
E-book (PDF version) published in September 2017
ISBN 978-88-99459-71-0

This book is the collection of the Abstract / Short Papers submitted by the authors of the International Conference of The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS), held in Milan (Italy), University of Milano-Bicocca, September 13-15, 2017.

Euro 9,00

Keynotes

Statistical models for complex extremes

Antony Davison,
Institute of Mathematics,
Ecole Polytechnique Federale de Lausanne, Switzerland

Classified Mixed Model Prediction

J. Sunil Rao,
Division of Biostatistics,
Department of Public Health Sciences, University of Miami, Florida

An URV approach to cluster ordinal data

Roberto Rocci,
Dipartimento di Economia e Finanza,
Università degli studi di Tor Vergata, Rome, Italy

Invited sessions

Clustering and Dimensionality Reduction

Mixture models for simultaneous classification and reduction of three-way data

Roberto Rocci, Maurizio Vichi, *Monia Ranalli*

High-dimensional Clustering via Random Projections

Laura Anderlucci, *Francesca Fortunato*, Angela Montanari

Clustering and Structural Equation Modeling

Mario Fordellone, Maurizio Vichi

Hidden Markov Models for Longitudinal Data

Package LMest for Latent Markov Analysis of longitudinal categorical data

Francesco Bartolucci, Silvia Pandolfi, Fulvia Pennoni

Dynamic sequential analysis of careers

Fulvia Pennoni, Raffaella Piccarreta

Multivariate hidden Markov regression models with random covariates

Antonio Punzo, Salvatore Ingrassia, *Antonello Maruotti*

Analysis of partially ordered data in socio-economics

Comparing three methodological approaches for synthesizing complex phenomena

Katia Iglesias, Christian Suter, Tugce Beycan, B.P. Vani

New posetic tools for the evaluation of financial literacy

Marco Fattore, Mariangela Zenga

Poset theory and policy making: three case studies

Enrico di Bella

Classification models in Economics and Business

Poland on Global Consumer Markets – Multilevel Segmentation of Countries on the basis of Market Potential Index

Adam Sagan, Eugene Kąciak

Hidden Variable Models for Market Basket Data

Harald Hruschka

Sensitivity Analysis in Corporate Bankruptcy Prediction

Barbara Pawełek, Jozef Pociecha

Advances in Functional data analysis

Growth processes in forensic entomology: a functional data perspective

Davide Pigoli, John Aston, Frédéric Ferraty

Density based classification methods for functional data

Enea Bongiorno, Aldo Goia

Permutation methods for multi-aspect local inference on functional data

Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti

New results in Robust estimation

A proposal for robust functional clustering based on trimming and constraints

Luis Angel García-Escudero, Diego Rivera-García, Joaquín Ortega and Agustin Mayo-Iscar

Trimming in probabilistic clustering

Gunter Ritter

Covariance matrices of robust estimators in regression

Silvia Salini, Fabrizio Laurini, Marco Riani, Andrea Cerioli

Innovative applications of multidimensional scaling

Preference judgments of curvature and angularity in architectural façades

Giuseppe Bove, Nicole Ruta, Stefano Mastandrea

Changes in couples' breadwinning patterns and wife's economic role in Japan

Miki Nakai

Individual differences in brand switching

Akinori Okada, Hiroyuki Tsurumi

Advances in Robust methods

Weighted likelihood estimation of multivariate location and scatter

Luca Greco, Claudio Agostinelli

Efficient robust methods for multivariate data via monitoring

Antony C. Atkinson

A new robust estimator of multilevel models based on the forward search approach

Aldo Corbellini, *Luigi Grossi,* Fabrizio Laurini

Big Data - Big Knowledge

Flexible Inference for FMRI Data
Aldo Solari

Opinion Mining and City Branding
Federico Neri, Roberto Grandi - Integris

From predictive to reactive approach: how not to be biased from the past in volatile contexts
Federico Stefanato, Marco Cagna - Waterdata

Big data and Design of experiments

Passive and active observation: experimental design issues in big data
Henry Wynn

Optimal design of experiments in the presence of covariate information
Peter Goos

Is it possible a design of experiment with puzzling dynamic data?
Giacomo Aletti

Robust Clustering

Robust clustering tools based on optimal transportation
Eustasio del Barrio

Advances in robust clustering for regression structures
Domenico Perrotta, Francesca Torti, Andrea Cerioli, Marco Riani

Robustness aspects of DD-classifiers for directional data
Giuseppe Pandolfo

Classification and Visualization

Explorative visualization techniques for imbalanced classification tasks
Adalbert Wilhelm

Calibrated cluster validity for comparing the quality of clusterings
Christian Hennig

Visual Tools for Interactive Clustering of UE State Members via Metabolic Patterns
Massimo Aria, Carmela Iorio, Roberta Siciliano, *Michele Staiano*

Designing clinical trials

The rise of early phase clinical trials

Nancy Flournoy

Adaptive dose-finding designs to identify multiple doses that achieve multiple response targets

Adrian Mander

A new design strategy for hypothesis testing under response adaptive randomization

Maroussa Zagoraiou, Alessandro Baldi Antognini, Alessandro Vagheggin

Advances in Credit Risk modelling

Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards

Viani Djeundje, *Jonathan Crook*

Scoring models for P2P lending platforms: an evaluation of predictive performance

Paolo Giudici, *Branka Hadij-Misheva*

Advances in risk measurement in a distressed banks scenario

Mauro Bernardi, Roy Cerqueti, Arsen Palestini

Heterogeneity and new statistical models

Fitting Cluster-Weighted Models in R

Angelo Mazza, *Antonio Punzo,* Salvatore Ingrassia

Outcome evaluation in healthcare: The Multilevel Logistic Cluster Weighted Model

Paolo Berta, Fulvia Pennoni, Veronica Vinciotti

Mixture model under overlapping clusters: an application to network data

Saverio Ranciati, *Veronica Vinciotti,* Ernst Wit

A world of data

Active and passive measurement: a paradigm change

Giorgio Licastro - GFK Eurisko

Data Science approach and challenges in private sectors

Rocco Michele Lancellotti - Data Reply

Analytics Data LAB: The power of Big Data Investigation and Advanced Analytics to maximize the Data Capital

Roberto Falcinelli - Oracle

Advances in Biostatistics

Regression models for the restricted residual mean time for right-censored and left-truncated data

Giuliana Cortese, Thomas Scheike

Estimating mediation effects in epigenomic studies

Vera Djordjilovic

Statistical challenges in single-cell RNA sequencing

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, Jean-Philippe Vert

Advances in Ordinal and Preference data

Zero inflated CUB models for the evaluation of leisure time activities

Maria Iannario, Rosaria Simone

Constrained consensus bucket order

Antonio D'Ambrosio, Carmela Iorio, Roberta Siciliano

Ensemble methods for Ranking data

Antonella Plaia*, Mariangela Sciandra,* Roberta Murò

Inequality indices and their decomposition

Contributions from macro-regions and from income components to the Zenga Index I(Y): an application to data from Poland and Italy

Michele Zenga, *Alina Jedrzejczak,* Igor Valli

Joint decomposition by subpopulations and sources of the point and synthetic Bonferroni inequality measures

Michele Zenga, *Igor Valli*

Transfers between sources and units in Zenga's inequality index decomposition

Alberto Arcagni

Advances in Classification and Clustering of complex Data

Combined methods in multi-label classification algorithms

Luca Frigau, Claudio Conversano, Francesco Mola

Validation of Experiments Involving Image Segmentation of Botanic Seeds

Jaromir Antoch, Claudio Conversano, Luca Frigau, Francesco Mola

Time Series Clustering for Portfolio Selection

Carmela Iorio, Antonio D'Ambrosio

Preferences in freshly graduates recruiting

Academic disciplines as perceived by entrepreneurs
Luigi Fabbris, *Manuela Scioni*

University and tourism. Graduates' profiles for the tourism sector
Antonio Giusti, Laura Grassini, Manuela Scioni

The effect of the firm size in the selection of recruitment for new graduate
Franca Crippa, Paolo Mariani, Andrea Marletta, *Mariangela Zenga*

Network Analysis with applications on biological, financial and social networks

Co-authorship Network in Statistics: methodological issues and empirical results
Susanna Zaccarin, Maria Prosperina Vitale, Domenico De Stefano

Network inference in genomics
Ernst Wit

Spatial modeling of brain connectivity data
Daniele Durante, Emanuele Aliverti

Contributed sessions

Classification of Multiway and Functional Data

A generalized Mahalanobis distance for the classification of functional data

Andrea Ghiglietti, Francesca Ieva, Anna Maria Paganoni

Classification methods for multivariate functional data with applications to biomedical signals

Andrea Martino, Andrea Ghiglietti, Anna M. Paganoni

A new Biclustering method for functional data: theory and applications

Jacopo Di Iorio, Simone Vantini

A leap into functional Hilbert spaces with Harold Hotelling

Alessia Pini, Aymeric Stamm, *Simone Vantini*

Sampling Designs and Stochastic models

Statistical matching under informative probability sampling

Daniela Marella, Danny Pfeffermann

Goodness-of-fit test for discrete distributions under complex sampling design

Pier Luigi Conti

Structural learning for complex survey data

Daniela Marella*, Paola Vicard*

The size distribution of Italian firms: an empirical analysis

Anna Maria Fiori, Anna Motta

Robust statistical methods

New proposal for clustering based on trimming and restrictions

Luis Angel Garcìa Escudero, Francesca Greselin, *Agustin Mayo Iscar*

Wine authenticity assessed via trimming

Andrea Cappozzo, Francesca Greselin

Robust and sparse clustering for high-dimensional data

Sarka Brodinova, Peter Filzmoser, Thomas Ortner, Maia Zaharieva, Christian Breiteneder

M-quantile regression for multivariate longitudinal data

Marco Alfo', *Maria Francesca Marino,* Maria Giovanna Ranalli, Nicola Salvati, Nikos Tzavidis

New proposals in Clustering methods

Reduced K-means Principal Component Multinomial Regression for studying the relationships between spectrometry and soil texture

Pietro Amenta, *Antonio Lucadamo*, Antonio Pasquale Leone

Comparing clusterings by copula information based distance

Marta Nai Ruscone

Fuzzy methods for the analysis of psychometric data

Isabella Morlini

Inverse clustering: the paradigm, its meaning, and illustrative examples

Jan W. Owsinski, Jaroslaw Stanczak, Karol Opara, Slawomir Zadrozny

Big data mining and classification

The importance of the minorities' viewpoints: Rare Event Sampling Technique on Sentiment analysis supervised algorithm

Marika Arena, *Anna Calissano,* Simone Vantini

A generalized K-means algorithm for multivariate big data with correlated components

Giacomo Aletti, *Alessandra Micheletti*

Big data process analysis: from data mining to process mining

Massimiliano Giacalone, Carlo Cusatelli, Roberto Casadei, Angelo Romano, Vito Santarcangelo

Semiparametric estimation of large conditional variance-covariance and correlation matrices with an application to financial data

Claudio Morana

Advances in model-based clustering

Probabilistic Distance Algorithm generalization to Student's t mixtures

Christopher Rainey, Cristina Tortora, *Francesco Palumbo*

Model-based Clustering of Data with Measurement Errors

Michael Fop, Thomas Brendan Murphy, Lorraine Hanlon

Gaussian Mixture Modeling Under Measurement Uncertainty

Volodymyr Melnykov, Shuchismita Sarkar, Rong Zhengi

A dynamic model-based approach to detect the trend of Statistics from 1970 to 2015

Laura Anderlucci, Angela Montanari, Cinzia Viroli

Bayesian methods and networks

Non parametric Bayesian Networks for measurement error detection
Daniela Marella, Paola Vicard, *Vincenzina Vitale*

Sparse Naïve Bayes Classification
Rafael Blanquero, Emilio Carrizosa, Pepa Ramírez-Cobo, *M. Remedios Sillero-Denamiel*

A Constraint-based Algorithm for Nonparanormal Data
Flaminia Musella, *Paola Vicard,* Vincenzina Vitale

Interventional data and Markov equivalence classes of DAGs
Federico Castelletti, Guido Consonni

Categorical data analysis

Study of context-specific independencies through Chain Stratified Graph Models for categorical variables
Federica Nicolussi, Manuela Cazzaro

Redundancy Analysis Models with Categorical Endogenous Variables: A New Estimation Technique
Gianmarco Vaccà

Mixture of copulae based approach for defining the subjects distance in cluster analysis
Andrea Bonanomi, Marta Nai Ruscone, Silvia Angela Osmetti

Dissimilarity profile analysis for assessing the quality of imputation in cardiovascular risk studies
Nadia Solaro

Data Analysis

Measuring vulnerability: a Structural Equation Modelling approach

Ambra Altimari, *Simona Balzano*, Gennaro Zezza

On the turning point detection in financial time series

Riccardo Bramante, *Silvia Facchinetti*

Optimization of the Listwise Deletion Method

Graziano Vernizzi, Miki Nakai

Discretization of measures: an IRT approach

Silvia Golia

Mixture and Latent Class Models for Clustering

Analysis of university teaching quality merging student ratings with professor characteristics and opinions

Francesca Bassi, Leonardo Grilli, Omar Paccagnella, *Carla Rampichini,* Roberta Varriale

Clustering technique for grouped survival data with a nonparametric frailty term

Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples

A latent trajectory model for migrants' remittances: an application to the German Socio-Economic Panel data

Silvia Bacci, Francesco Bartolucci, Giulia Bettin, *Claudia Pigini*

Stepwise latent Markov modelling with covariates in presence of direct effects

Roberto Di Mari, Zsuzsa Bakk

Network analysis

Non-parametric inference for network-valued data
Ilenia Lovato, Alessia Pini, Aymeric Stamm, Simone Vantini

Applying network analysis to online news big data
Giovanni Giuffrida, Simona Gozzo and Francesco Mazzeo Rinaldi, *Venera Tomaselli*

Interval Regression Analysis for the Representation of the Core-Periphery Structure on Large Networks
Carlo Drago

A Latent Space Model for Multidimensional Networks
Silvia D'Angelo, Thomas Brendan Murphy and Marco Alfò

Advances in LMs, GLMs and PCA

Bootstrap prediction intervals in linear models
Davide Passaretti, Domenico Vistocco

Bayesian Variable Selection in Linear Regression Models with non-normal Errors
Saverio Ranciati, Giuliano Galimberti and Gabriele Soffritti

Principal Component Analysis: the Gini Approach
Stéphane Mussard*, Téa Laurent Jérome Akeywidi Ouraga*

On Proportional Odds Modelling and Marginal Effects of Sardinian Hotels data
Giulia Contu, Claudio Conversano, Thomas W. Yee

Advances in Classification

Classification of Textual Data

Evaluation in Education

Nonparametric mixed-effects model for unsupervised classification in the Italian education system

Chiara Masci, Francesca Ieva, Anna Maria Paganoni, Tommaso Agasisti

Multivariate mixed models for assessing equity and efficacy in education. An analysis over time using EU15 PISA data

Isabella Sulis, Francesca Giambona, Mariano Porcu

A zero-inflated beta regression model for predicting first-year performance in university career

Matilde Bini, Lucio Masserini

Students' satisfaction in higher education: how to identify courses with low-quality teaching

Marco Guerra, *Francesca Bassi,* José G. Dias

Statistical models for complex data

Spatial Survival Models for Analysis of Exocytotic Events on Human beta-cells Recorded by TIRF Imaging

Thi Huong Phan, Giuliana Cortese

Testing different structures of spatial dynamic panel data models

Francesco Giordano, Massimo Pacella, *Maria Lucia Parrella*

Identification of earthquake clusters through a new space-time-magnitude metric

Renata Rotondi, Antonella Peresan, Stefania Gentili, Elisa Varin

A circular density strip plot

Davide Buttarazzi, Giovanni Camillo Porzio

Mixture Models

A special Dirichlet mixture model for multivariate bounded responses

Agnese Maria Di Brisco, Sonia Migliorati

Cluster-Weighted Beta Regression

Marco Alfò, *Luciano Nieddu,* Cecilia Vitiello

A Special Dirichlet Mixture Model in a Bayesian Perspective

Roberto Ascari, Sonia Migliorati, Andrea Ongaro

Advances in data Analysis

Assessing Heterogeneity in a Matching Estimation of Endogenous Treatment Effect

Maria Gabriella Campolo, *Antonino Di Pino,* Edoardo Otranto

Template matching for hospital comparison: an application to birth event data in Italy

Massimo Cannas, Paolo Berta, Francesco Mola

On variability analysis of evolutionary algorithm-based estimation

Manuel Rizzo

Poster session

Accounting for Model Uncertainty in Individualized Designs for Discrete Choice Experiments

Eleonora Saggini, Laura Deldossi, Guido Consonni

Financial-literacy: Socio-demographic variables versus environment

Doriana Cuccinelli, Paolo Trivellato, Mariangela Zenga

Joint models for survival and bivariate longitudinal data: a likelihood formulation

Marcella Mazzoleni, Mariangela Zenga

A Spatial and model-based approach to identify the effect of cultural capital on high school dropout. The Italian case.

Stefano Barberis, Enrico Ripamonti

M-quantile Regression in small area estimation: estimation and testing

Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati, Nikos Tzavisis

A ZERO-INFLATED BETA REGRESSION MODEL FOR PREDICTING FIRST-YEAR PERFORMANCE IN UNIVERSITY CAREER

Matilde Bini[1], Lucio Masserini[2]

[1] Department of Human Sciences, European University of Rome, e-mail: matilde.bini@unier.it

[2] Statistical Observatory, University of Pisa; e-mail: lucio.masserini@unipi.it

ABSTRACT: The background preparation of students entering the university system is checked through evaluation tests in Italy. The test is non-selective in most degree programmes, as it does not preclude the possibility of enrolling in the student's chosen program. However, the initial preparation and attitude of the students seem to be key issues in explaining their performance and predicting the performance outcome of their first-year in university. The evaluation test results are used to predict the students' performance at the end of the first year by a zero inflated beta regression model. The analysis was conducted on the evaluation test carried out in 2013 with students at the Department of Economics and Management, University of Pisa.

KEYWORDS: non-selective evaluation test, first-year university career, zero-inflated beta regression.

1 Introduction

The background preparation of students entering the university system is a fairly recent issue in Italy. The Ministerial Decree 370/04 introduced the need to evaluate their preparation in degree programmes with open access, and universities are required to organize evaluation tests. These evaluation procedures aim at identifying students potentially able to undertake university studies and informing others that the gaps in their initial preparation could make their future university career problematic. The test provides a measurement of students' attitude, and is non-selective, as it does not preclude the possibility of enrolling in the student's chosen program. However, students having a low score could have to bear formative debts by taking special courses organized by the relevant department. This study aims at collecting evidence on the Italian university system and has two goals: i) to investigate whether a non-selective evaluation test can be effective for predicting the performance outcome of a student's first-year university career and ii) to detect what other factors affect the student's probability of earning (or not) credits and their career progression within their first year of study, among those variables available at the time of the enrolment. The analysis refers to the students enrolled in the

Department of Economics and Management of the University of Pisa who carried out a non-selective evaluation test in September 2013.

2 Measure of students' academic performance

In many research studies, grade point average (GPA) and cumulative grade point average (CGPA) are commonly employed as measures of students' academic performance (Park and Kerr 1990; Broh 2000; Darling 2005; Rienties et al. 2012). Both GPA and CGPA are computed as the weighted mean of grade points earned by students from the time of enrolment and, as such, reflect the teacher's judgment of a student's academic achievement. Specifically, GPA is obtained by dividing the total number of a student's grade points earned in a given period of time (usually a particular semester or one year) by the total amount of credit hours attempted; alternatively, CGPA is given by the average of all grades for a complete educational career. In this study academic performance is defined by computing a composite indicator, with the aim of measuring both the students' achievement and advancement in their university career during their first year. This indicator, defined as *Career Progression Index* (CPI), is given as follows:

$$CPI_i = \begin{cases} 0 & \text{if} \quad CFU_j = 0 \\ (0,1] & \text{if} \quad \dfrac{\sum_j CFU_j \times v_{ij}}{\max\left(\sum_j CFU_j \times v_{ij}\right)} \end{cases}$$

where CFU_j is the number of credits in the first year for exam *j* and υ_{ij} is the corresponding grade for student *i*. The proposed CPI measure generates a continuous and doubly bounded random variable, defined in the unit interval (0,1]. Here, the quantity at the denominator is constant and equal to 1710, given by the product between 57, the number of the most credits a student can earn during the first year, and 30, the maximum grade for each exam (in the Italian university system, grades range from 18 to 30). However, in this case CPI starts from 0.06 which represents the value obtained considering the minimum number of credits multiplied by the lowest grade. As a consequence, it can be considered as a normalized indicator of career progression, since the value obtained by each student is compared to its maximum, representing the hypothetical limit of a student's level of academic advancement.

3 Data and variables

The data include information from two different sources, the administrative archive of the University of Pisa, where information on the university careers and the main characteristics of students are recorded, and the database of the evaluation test results. Specifically, among the 867 participants enrolled in a degree programme at the Department of Economics and Management, who undertook the evaluation test carried out in September 2013, the number of students who decided to enroll was 709 (81.8%). The test was made up of 40 multiple choice questions concerning three different areas: Logic, Reading comprehension and Mathematics. For each question, one out of five answers was correct. A score of 1 was assigned to the correct answer, -0.25 for the wrong answer and 0 for a non-response. The total score was given by the sum of the scores in each area and ranged from -10 to 40. Despite the fact that the evaluation test was compulsory, the value of the total score did not affect the possibility of enrolling in a degree programme. A set of variables was also collected from those available in the administrative archive: Gender (Female = 0; Male = 1), High school diploma (lyceum, commercial and technical institute, vocational, other), High school grade (60-74; 75-90; 91-100), Diploma after the age of 19 years (No = 0; Yes = 1), Province of residence (Other = 0; 1 = neighbouring provinces of Livorno, Lucca and Pisa). The descriptive statistics for some interesting outcomes (proportion of students with 0 CFU, average number of CFU and the CPI value) haven carried out but not here presented for lack of space. Test results variables were considered as standardized test scores for each area to eliminate the effect due to different magnitudes and variability, and defined as *zlogic*, *zread* and *zmath*.

4 Zero-inflated beta regression model

Dependent variables measured in the social sciences often assume a limited range of values. Classical examples are categorical or binary responses, but continuous variables in the open interval (0,1) with boundaries such as fractions or proportions are also popular. Beta regression (Ferrari and Cribari-Neto 2004; Smithson and Merkle 2013) is suitable for modelling beta-distributed continuous dependent variables defined over the interval (0,1), but it is not applicable when data include observations at the boundaries, since it does not allow positive probability masses at the extremes. For observed data which include one or both the boundaries, such as [0,1), (0,1] or [0,1], Ospina and Ferrari (2010, 2012) proposed a mixture of a continuous distribution on (0,1) and a degenerate distribution that assigns a non-negative probability to 0 or 1. Following this approach, the probability density function of the response variable y (iid), with respect to the measure generated by the mixture is given by:

$$BI_c(y;\alpha,\mu,\phi)=\begin{cases}\alpha & \text{if} \quad y=c\\(1-\alpha)f(y;\mu,\phi) & \text{if} \quad y\in(0,1)\end{cases},$$

where $0 < \alpha < 1$ is the mixture parameter and $f(y; \mu, \phi)$ is the beta density function:

$$f(y;\mu,\phi)=\frac{\Gamma\phi}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}.$$

where $y \in (0,1)$ is the beta-distributed random variable, indicated by $y \sim \text{Be}(\mu, \phi)$; $E(y) = \mu$ with $0 < \mu < 1$, and $\text{Var}(y) = \mu(1-\mu)/(\phi+1)$. More specifically, the parameter

ϕ is known as a 'precision' parameter, since, for fixed μ, the larger ϕ is, the smaller the variance of the response variable y. Moreover, α is the probability mass at c and represents the probability of observing zero ($c = 0$). If $c = 0$, the distribution is called the inflated beta distribution at zero (BEZI) and is indicated with $y \sim \mathrm{BEZI}(\alpha, \mu, \phi)$; Here, the mean and variance of y are, respectively: $E(y) = \alpha c + (1-\alpha)\mu$ and $\mathrm{Var}(y) = (1-\alpha)(\mu(1-\mu))/(\phi+1) + \alpha(1-\alpha)(c-\mu)^2$. A general class of zero-or-one inflated beta regression models can be defined by assuming the following relations for the conditional mean, the mixture parameter and the precision parameter (Ospina and Ferrari 2012):

$$h(\alpha_t) = \sum_{i=1}^{M} z_{ti}\gamma_i = \zeta_t, \quad g(\mu_t) = \sum_{i=1}^{m} x_{ti}\beta_i = \eta_t \text{ and } b(\phi_t) = \sum_{i=1}^{M} s_{ti}\lambda_i = \kappa_t.$$

where $\gamma = (\gamma_1, \ldots, \gamma_M)^T$, $\beta = (\beta_1, \ldots, \beta_m)^T$ and $\lambda = (\lambda_1, \ldots, \lambda_q)^T$ are vectors of unknown parameters, with $(M + m + q < n)$; ζ_t, η_t and κ_t are linear predictors; and $z_{t1}, \ldots, z_{tM}$, $x_{t1}, \ldots, x_{tm}$, and $s_{t1}, \ldots, s_{tq}$ are fixed and known covariates which may be identical or partly overlapping. The previous equations define the sub-models for the inflated beta regression, The link function for $g(\cdot)$ and $h(\cdot)$ is the logit and the resulting regression parameters are interpretable in terms of log-odds. The link for $b(\cdot)$ is the log, because the precision parameter ϕ must be positive since a variance cannot be negative. Parameter estimation is performed by maximum likelihood ML (see for details Ospina and Ferrari, 2012) by using the GAMLSS framework, implemented in the R package 'gamlss' (Stasinopoulos and Rigby 2007).

5 Results

For predicting the outcome of the university career at the end of the first academic year, as measured by the CPI, the same set of explanatory variables was used for the first two sub-models (zero-inflated and proportion) and includes student demographic characteristics, high school experience before enrolment and evaluation test results. Parameter estimates together with the corresponding standard errors, p-values and 95% c.i. are summarized in Table 1. As a global goodness-of-fit measure, a pseudo R-square is computed by the square of the correlation coefficient between the response variable and the corresponding predicted values. Its value is equal to 0.369 and can be considered indicative of good model fit.

Table 1. Parameter estimates, standard errors, p-values and 95% confident interval

α (zero-inflated)	Estimate	S.e.	p
constant	0.27	0.19	0.159
zread	-0.23	0.10	0.019
zmath	-0.37	0.11	0.001
lyceum	-0.43	0.19	0.026
vocational	1.13	0.49	0.023
grade (75–90)	-0.89	0.20	0.000
grade (91–100)	-1.70	0.33	0.000
later diploma	0.82	0.20	0.001

μ (proportion)	Estimate	S.e.	p
constant	-0.38	0.08	0.000
zmath	0.19	0.04	0.000
commercial and technical	-0.17	0.09	0.047
grade (75–90)	0.53	0.09	0.000
grade (91–100)	1.15	0.12	0.000
later diploma	-0.38	0.11	0.000
ϕ (precision)	Estimate	S.e.	p
constant	1.74	0.11	0.000
grade (75–90)	-0.29	0.14	0.040
grade (91–100)	-0.51	0.17	0.003

Table 1 shows that only a subset of the observed covariates are related to the probability of having zero career progression by the end of the first academic year. The model demonstrates that there are no differences in the demographic characteristics. Instead, with regard the standardized test scores, a significant effect is observed for *zread* (-0.23, p=0.019) and *zmath* (-0.37, p=0.001) but not for *zlogic*. Since the coefficients are negative, students who achieve higher scores in reading and mathematics have a lower probability of having zero career progression. In particular, if the effect of *zmath* is not surprising, such as the fact that during the first year three out of five exams have mathematical contents, the relationship with *zread* seems more interesting. Substantial considerations also arise from the variables related to high school experience. Indeed, the probability of having zero career progression is significantly lower for students coming from *lyceum* (-0.43, p=0.026). In particular, for these students the odds of not starting their university career is about 1,5 times lower than that of students from other institutes. Furthermore, this probability varies across the categories of *high school grade*: the odds for students with a high school grade of 75–90 is estimated to be 2.4 times less than 60–74, whereas that of students with a high school grade of 91–100 is about 5.5 times lower. A significant effect is also observed for the variable *later diploma*, which shows that students who achieved their diploma after the age of 19 years have an odds of not having started their university career 2.3 times higher than that of those who have a regular course. Finally, the probability of zero career progression was computed for two hypothetical students after identifying their corresponding profiles. Such profiles (referred to as profile 1 and profile 2) were defined to have a high or low probability of being in the status of zero career progression, respectively, by taking parameters values into account. In particular, profile 1 has a probability of having a zero career progression equal to 0.09 and describes students with the following characteristics: coming from a lyceum, high school grade of 91–100 and a standardized test score equal to the third quartile in each area. Alternatively, profile 2 has a probability of having a zero career progression equal to 0.935 and describes students with the following characteristics: coming from a vocational institute, high school grade of 60–74 and a standardized test score equal to the first quartile in each area. The comparison between these two opposite profiles, as well as any intermediate profiles corresponding to real situations, can be of great help for identifying students with potentially lower or higher risk of not starting their academic career within the first year of study. The proportion of career progression is modelled with a beta distribution as specified in the sub-model which defines the continuous component of the mixture, and the results are only partly consistent with those of the zero-inflated sub-model. In particular, demographic characteristics are not significant, either, in explaining career progression. Instead, differences from the zero-inflated sub-model were found regarding the standardized test scores, since a significant effect is observed only for *zmath* (+0.19, $p<0.001$) but not for *zlogic* and *zread*. Given that its coefficient is positive, students who achieve higher scores in mathematics have on average a more advanced career progression, while holding constant the other covariates. Moreover, further differences with the zero-inflated sub-model were found as concern the variables related to high school experience. More specifically, having attended *lyceum* seems not to favour career progression, while those students coming from commercial and technical institutes have a lower progression. On the other hand, significant differences are confirmed across the categories of high school grades. In fact, students with higher grades are more likely to have on average a more advanced career progression, with grades 75–90 slightly less (+0.53, $p<0.001$) than 91–100 (+1.15, $p<0.001$), respectively. Furthermore, a confirmation is also found for the variable *later diploma*, since students who achieved their diploma after the age of 19 years have a career progression significantly less advanced than students with a regular course (-0.38, $p<0.001$).

References

BROH, B.A. 2002. Linking extracurricular programming to academic achievement: Who benefits and why? *Sociology of Education*, **75**(1), 69–95.

DARLING, N. 2005. Participation in extracurricular activities and adolescent adjustment: Cross-sectional and longitudinal findings. *Journal of Youth and Adolescence*, **34**, 493–505.

FERRARI, S.L.P., CRIBARI-NETO, F. 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **7**, 799–815.

OSPINA, R., FERRARI, S.L.P. 2010. Inflated beta distributions. *Statistical Papers*, **51**, 111–126.

OSPINA, R., FERRARI, S.L.P. 2012. A general class of zero-or-one inflated beta regression models. *Comput. Stat. Data Anal.*, **56**, 1609–1623.

PARK, K. H., KERR, P.M. 1990. Determinants of academic performance: a multinomial logit approach. *Journal of Economic Education*, **21**(2), 101–111.

RIENTIES, B., BEAUSAERT, S., GROHNERT, T., NIEMANTSVERDRIET, S., KOMMERS, P. 2012. Understanding academic performance of international students: the role of ethnicity, academic and social integration. *Higher Education*, **63**, 685–700.

SMITHSON, M., MERKLE, E.C. 2013. *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. CRC Press.

STASINOPOULOS, D.M., RIGBY, R.A. 2007. Generalized additive models for location scale and shape (GAMLSS). R. *Journal of Statistical Software*, **23**, 1–43.