"Workshop on Statistical Modeling and Data Science UC"
Department of Statistics UC

November 06-07, 2023
Sala Multiusos N°2, Segundo Piso
Edificio Felipe Villanueva, Facultad de Matemáticas UC

Organizer: Mauricio Castro

Program Day 1

13.45 **Registration/Opening**

14.00 **Rosangela Loschi** (Department of Statistics, Universidade Federal de Minas Gerais, Brasil):
Flexible Bayesian Modelling in Dichotomous Item Response Theory Using Mixtures of Skewed Item Curves

15.00 **Anuradha Roy** (Department of Management Science and Statistics, The University of Texas at San Antonio, USA):
Linear Models for Multivariate Repeated Measures Data with Block Exchangeable Covariance Structure

16.00 **Coffee Break**

16.30 **Jorge Bazán** (Department of Applied Mathematics and Statistics, Universidade de Saõ Paulo, Brasil):
Beyond Beta Distribution: A Personal Journey

17.30 **Carlos Sing-Long** (Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile, Chile):
TBA

18.30 **Welcome Reception**

Program Day 2

09.00 **Richard Warr** (Department of Statistics, Brigham Young University, USA):
A Fiducial-based Confidence Interval for the Linear Combination of Multinomial Probabilities

10.00 **Daniela Castro** (School of Mathematics and Statistics, University of Glasgow, UK):
Probabilistic Forecasting of Weather-Driven Faults on Electricity Distribution Networks

11.00 **Coffee Break**

11.30 **Armin Schwartzman** (Division of Biostatistics, University of California San Diego, USA)
TBA

12.30 **Lunch**

14.30 **Alejandro Murúa** (Department of Mathematics and Statistics, Université de Montréal, Canada):
Qini Based Uplift Regression

15.30 **Closing Ceremony**

Practical information

Prof. Mauricio Castro
E-mail: lmcastro@uc.cl

Rosangela Loschi

Flexible Bayesian Modelling in Dichotomous Item Response Theory Using Mixtures of Skewed Item Curves

Abstract

Most Item Response Theory (IRT) models for dichotomous responses are based on probit or logit link functions which assume a symmetric relationship between the probability of a correct response and the latent traits of individuals submitted to a test. This assumption restricts the use of those models to the case in which all items have a symmetric behaviour. On the other hand, asymmetric models proposed in the literature impose that all the items in a test have an asymmetric behaviour. This assumption is inappropriate for great part of the tests which are, in general, composed by both symmetric and asymmetric items. Furthermore, a straightforward extension of the existing models in the literature would require a prior selection of the items' symmetry/asymmetry status. This paper proposes a Bayesian IRT model that accounts for symmetric and asymmetric items in a flexible though parsimonious way. That is achieved by assigning a finite mixture prior to the skewness parameter, with one of the mixture components being a point-mass at zero. This allows for analyses under both model selection and model averaging approaches. Asymmetric item curves are designed through the centered skew normal distribution, which has a particularly appealing parameterisation in terms of parameter interpretation and computational efficiency. An efficient MCMC algorithm is proposed to perform Bayesian inference and its performance is investigated in some simulated examples. Finally, the proposed methodology is applied to a data set from a large scale educational exam in Brazil.

Anuradha Roy

Linear Models for Multivariate Repeated Measures Data with Block Exchangeable Covariance Structure

Abstract

Multivariate repeated measures data, where observations are made on p response variables and each response variable is measured over n sites or time points, construct matrix-valued response variable, and arise across a wide range of disciplines, including medical, environmental and agricultural studies. The popularity of the classical general linear model (CGLM) is mostly due to the ease of modeling and authentication of the appropriateness of the model. However, CGLM is not appropriate for correlated multivariate repeated measures data. We propose an extension of CGLM for multivariate repeated measures data with exchangeably distributed errors for multiple observations. Maximum likelihood estimates of the matrix parameters of the intercept, slope and the eigenblocks of the exchangeable error matrix are derived. The distributions of these estimators are also derived. The practical implications of the methodological aspects of the proposed extended model for multivariate repeated measures data are demonstrated using two medical datasets.

Jorge Bazán

Beyond Beta Distribution: A Personal Journey

Abstract

We will show a state of art about new development and applications for bounded response. We show our different contributions with students and collaborators proposing different regression analysis and mixed models considering alternative distributions to the beta distribution over the last 11 years. An application showing the advantages of considering alternatives to Beta mixed regression models is presented and some proposal to future works are discussed.

Richard L. Warr

A Fiducial-based Confidence Interval for the Linear Combination of Multinomial Probabilities

Abstract

Across a broad set of applications, system outcomes may be summarized as probabilities in confusion matrices or contingency tables. In settings with more than two outcomes (e.g., stages of cancer), these outcomes represent multinomial experiments. Measures to summarize system performance have been presented as linear combinations of the resulting multinomial probabilities. Statistical inference on the linear combination of multinomial probabilities has been focused on large-sample and parametric settings and not small-sample settings. Such inference is valuable, however, especially in settings such as those resulting from pilot or low-cost studies. To address this gap, we leverage the fiducial approach to derive confidence intervals around the linear combination of multinomial parameters with desirable frequentist properties. One of the original arguments against the fiducial approach was its inability to extend to multiparameter settings. Therefore, the great novelty of this work is both the derived interval and the logical framework for applying the fiducial approach in multiparameter settings. Through simulation, we demonstrate that the proposed method maintains a minimum coverage of $1 - \alpha$, unlike the bootstrap and large-sample methods, at comparable interval lengths. Finally, we illustrate its use in a medical problem of selecting classifiers for diagnosing chronic allograph nephropathy in post kidney transplant patients.

Daniela Castro-Camilo

Probabilistic Forecasting of Weather-Driven Faults on Electricity Distribution Networks

Abstract

Electricity networks are exposed to the weather, and severe weather may cause faults that result in power cuts. Predicting the occurrence of faults in a region on time scales from hours to days ahead can increase preparedness, accelerate the response to weather-related faults, and ultimately reduce the duration of power cuts. Furthermore, these predictions should quantify uncertainty so that planners may assess risk and distribute limited resources accordingly. We present a method for probabilistic fault prediction that leverages ensemble numerical weather prediction and methods from extreme value theory for discrete processes. Data describing network topology and vulnerability, such as elevation and proximity to vegetation, are combined with meteorological data to model the occurrence of stochastic faults, which may be heavy-tailed. In addition, forecasts of future weather conditions are required, and associated uncertainty is quantified via ensemble numerical weather prediction, which requires statistical post-processing. Finally, we discuss the communication of the resulting complex forecast information to decision-makers.

Alejandro Murúa

Qini Based Uplift Regression

Abstract

Models for uplift are commonly used to isolate the marketing effect of a campaign. For customer churn reduction, uplift models identify customers who are likely to answer positively to a retention activity only if explicitly targeted. They are also used to avoid wasting resources on customers that are very likely to switch companies. In practice, the models' performance is measured with the Qini coefficient. We introduce a Qini-based uplift regression model to analyze a large insurance company's retention marketing campaign. Our approach is based on logistic regression. We show that a Qini-optimized uplift model acts as a regularization in uplift models, yielding interpretable models with few relevant explanatory variables. Our results also show that the parameter estimation based on our Qini-optimized regression significantly improves the Qini prediction performance of uplift models.