MINISTERIO DE CIENCIA E INNOVACIÓN

AGENCIA ESTATAL DE INVESTIGACIÓN

SiDOR
Statistical Inference
Decision & Operations Research Group

Universidade Vigo

Workshop on Statistics

June 14th, 2024

Facultade de Ciencias Económicas e Empresariais
– Aula-Seminario 8 –
Universidade de Vigo

General Schedule

10:00–10:25 Two density-based tests for the $k-$sample problem with left-truncated data. Adrián Lago (Universidade de Vigo)

10:25–10:50 Measuring the predictive capacity in logistic regression with missing data. Susana Martins (Universidade de Vigo)

10:50–11:15 Aproximación a los intervalos de confianza para una combinación lineal de proporciones en presencia de una covariable. María Álvarez Hernández (Centro Universitario da Defensa)

11:15–11:45 Coffee break

11:45–12:10 Clustering curves in a competing risk framework. Marta Sestelo (Universidade de Vigo)

12:10–12:35 Modal regression with missing response data. Tomás Cotos Yáñez (Universidade de Vigo)

12:35–13:00 Eigenvalues Approximation of integral covariance operators with applications to weighted L^2 statistics. María Dolores Jiménez-Gamero (Universidad de Sevilla)

Abstracts

Two density-based tests for the $k-$sample problem with left-truncated data

Adrián Lago (Universidade de Vigo)

Abstract: The comparison of populations can be addressed in many different ways, depending on the interests of the researcher or the a priori information that one has about the target distribution. One can, for example, employ the well-known t-test or ANOVA test for to compare means under normality. If one is interested in any kind of differences between the populations, distinct functions related to a random variable can be employed. In this case, we will focus on tests based on estimators of the density function, which are known to be more powerful than the distribution-based tests in certain situations.

On the other hand, there exist situations in which the observation of individuals is partially hidden by the presence of another random variable; this is called truncation. Literature referring to hypothesis contrasts under truncation is, until now, vaguely developed, being the log-rank and generalizations of it the most employed tests. From the estimation of the cumulative distribution function, one can define a proper estimator of the density function under one-sided truncation and study tests based on it.

In this talk, two different tests based on such estimator will be proposed to address the $k-$sample problem with left-truncated data. Their asymptotic distributions will be studied and, due to the impossibility of its application in practice, two slightly different bootstrap resampling plans will be proposed to approximate the null distributions of the tests statistics. Both the validation of these methods and the choice of the smoothing parameter will be addressed via Monte Carlo simulations. Moreover, both tests will be compared to existing tests in the left-truncation framework, such as the Kolmogorov-Smirnov and the log-rank tests. Two real datasets regarding pregnancy and unemployment times will be employed to exemplify the performance of the proposed tests.

This is joint work with Ingrid Van Keilegom, Jacobo de Uña-Álvarez and Juan Carlos Pardo-Fernández.

Measuring the predictive capacity in logistic regression with missing data

Susana Martins (Universidade de Vigo)

Abstract: In some studies the goal is to relate a main outcome of interest to a number of variables, the so-called regression setup. Missing information complicates the analysis. Our research focuses on logistic regression with missing data. In particular, we are interested in measuring the predictive capacity of the logistic regression model, where the Area Under the ROC Curve (AUC) plays an outstanding role. Estimation of the AUC with missing data can be performed

through several methods: complete case analysis, inverse probability weighting or multiple imputation. Since the apparent AUC overestimates the true AUC, in this work we consider several approaches to correct for this overestimation: Split-Sample, K-fold cross-validation and Leave-one-out cross-validation, all of them adapted to missing data. We conduct a simulation study to evaluate performance of the correction methods in the presence of missing data. Although no method uniformly dominates along all the simulated scenarios, the approaches based on multiple imputation exhibit the best overall performance.

This is joint work with María del Carmen Iglesias-Pérez and Jacobo de Uña-Álvarez.

Aproximación a los intervalos de confianza para una combinación lineal de proporciones en presencia de una covariable

María Álvarez Hernández (Centro Universitario da Defensa)

Abstract: El estudio de nuevos procedimientos para efectuar inferencias sobre una combinación lineal (L) de K proporciones independientes ha recibido bastante atención en los últimos años dada su gran importancia práctica en el ámbito clínico. Es habitual que este parámetro L sea estudiado en presencia de otra variable, afectando a las estimaciones individuales de cada proporción y por ello a la estimación objetivo. Además, la construcción de intervalos de confianza (IC) para una combinación lineal de proporciones, planteada inicialmente desde una perspectiva aproximado, ha sido tratada recientemente desde un punto de vista exacto, basándose en ambos casos en el método score (estimador de L de máxima verosimilitud bajo la hipótesis nula). Por ello, en esta investigación se pretende hacer una acercamiento a la obtención de IC para el parámetro L en presencia de una covariable, desde un punto de vista semiparamétrico, tratando de adaptar los últimos estudios que existen al respecto en la literatura.

Clustering curves in a competing risk framework

Marta Sestelo (Universidade de Vigo)

Abstract: The cumulative incidence function is the standard method for estimating the marginal probability of a given event in the presence of competing risks. One basic but important goal in the analysis of competing risk data is the comparison of these curves, for which limited literature exists. We proposed a new procedure that lets us not only test the equality of these curves but also group them if they are not equal. The proposed method allows determining the composition of the groups as well as an automatic selection of their number. Simulation studies show the good numerical behavior of the proposed methods for finite sample size. The applicability of the proposed method is illustrated using real data.

Modal regression with missing response data

Tomás Cotos Yáñez (Universidade de Vigo)

Abstract: Modal regression estimates local modes of the conditional distribution of a response variable Y conditional on X=x. It is an alternative method to the classic mean regression that has gained much importance in recent decades due to its suitability, for example, when the conditional distribution has heavy tails, is not symmetrical or has more than one mode. On the other hand, we often encounter incomplete samples where the study of the behavior of any estimator under missing information is crucial to make decisions about the different missing data methodologies that can be applied. In this work, we adapt some methodologies used in modal regression to the context of missing response data: using just complete observations incorporating weights to the estimator, or imputing the missing response (simple or multiple imputation). The performance of the different estimated estimators is analysed in an extensive simulation study and an application to real data is also included.

This is joint work with Ana Pérez González and Rosa M. Crujeiras.

Eigenvalues Approximation of integral covariance operators with applications to weighted L^2 statistics

María Dolores Jiménez-Gamero (Universidad de Sevilla)

Abstract: Finding the eigenvalues connected to the covariance operator of a centred Hilbert-space valued Gaussian process is genuinely considered a hard problem in several mathematical disciplines. In statistics this problem arises for instance in the asymptotic null distribution of goodness-of-fit test statistics of weighted L^2-type. For this problem we present the Rayleigh-Ritz method to approximate the eigenvalues. The usefulness of these approximations is shown by high lightening implications such as critical value approximation and theoretical comparison of test statistics by means of Bahadur efficiencies.

This is joint work with Bruno Ebner and Bojana Milošević.