SiDOR
Statistical Inference, Decision
& Operations Research Group
Universidade Vigo

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN

Workshop on Statistics

May 12th, 2025

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

General Schedule

10:30–10:55 Testing for the zero-altered Poisson distribution with positive data. María Dolores Jiménez-Gamero (Universidad de Sevilla)

10:55–11:20 Modelos de regresión con datos composicionales. Ana Pérez González (Universidade de Vigo)

11:20–12:00 Coffee break

12:00–12:25 Semiparametric estimator for the covariate-specific ROC curve. Juan Carlos Pardo-Fernández (Universidade de Vigo)

12:25–12:50 Penalised spline estimation of covariate-specific time-dependent ROC curves. María Xosé Rodríguez-Álvarez (Universidade de Vigo)

12:50–13:15 Confidence intervals for the AUC with missing data. Susana Martins (Universidade de Vigo)

13:30–15:30 Lunch break

15:30–15:55 Goodness-of-fit testing with survival data. Jacobo de Uña-Álvarez (Universidade de Vigo)

15:55–16:20 The k-sample problem with left-truncated and right-censored data. Adrián Lago (Universidade de Vigo)

16:20–16:45 Robust estimation of conditional regions using quantile regression. Application in diabetes research. Sara Rodríguez Pastoriza (Universidade de Vigo)

Abstracts

Testing for the zero-altered Poisson distribution with positive data

María Dolores Jiménez-Gamero (Universidad de Sevilla)

Abstract: This talk proposes three new goodness-of-fit tests for the zero-altered Poisson distribution, or equivalently for the positive Poisson distribution, based on positive data, that is, data truncated at 0, whose test statistic is built using a characterization of this law. It is shown that the novel tests are consistent against any fixed alternative, and that the parametric bootstrap method can accurately approximate their null distributions. The power of these tests is investigated through a large simulation study, where it is also compared with some existing tests, showing a very competitive behavior. Several applications to real datasets illustrate the usefulness of the tests.

This is joint work with P. Puig and J. Ngatchou-Wandji.

Modelos de regresión con datos composicionales

Ana Pérez González (Universidade de Vigo)

Abstract: Los datos composicionales son un tipo de datos que por su naturaleza, suma constante, requieren de un tratamiento estadístico adecuado. En particular en los modelos de regresión esta característica pueden aparecer tanto en la variable dependiente, como en la independiente o incluso en ambas. En esta presentación se contextualizan los modelos de regresión con respuesta escalar y covariable composicional y se presenta un trabajo reciente sobre la estimación no paramétrica robusta de la función de regresión. Finalmente, se muestran las primeras ideas en las que estamos trabajando sobre la estimación en regresión modal con covariable composicional.

Semiparametric estimator for the covariate-specific ROC curve

Juan Carlos Pardo-Fernández (Universidade de Vigo)

Abstract: The ROC curve is routinely used for evaluating the performance of a continuous marker as diagnostic tool in a binary classification problem. In many practical applications, covariates related to the marker may be available. Under these circumstances, it is of interest to evaluate the influence that those covariates might have in the performance of the marker in terms of classification ability. Two extensions of the ROC are commonly used: the covariate-specific ROC curve and the covariate-adjusted ROC curve. In this talk we will review these

concepts. Since they are strongly related with the conditional distribution of the marker, the use of proportional hazard regression models arises in a very natural way. We will explore the use of flexible proportional hazard Cox regression models for estimating the covariate-specific and the covariate-adjusted ROC curves. We study their large- and finite-sample properties and apply the proposed estimators to a real-world problem.

This is joint work with Pablo Martínez-Camblor.

Penalised spline estimation of covariate-specific time-dependent ROC curves

María Xosé Rodríguez-Álvarez (Universidade de Vigo)

Abstract: The identification of biomarkers with high predictive accuracy is a crucial task in medical research, as it may help clinicians make early decisions, thus reducing morbidity and mortality in high-risk populations. Time-dependent receiver operating characteristic (ROC) curves are the main tool used to assess the accuracy of prognostic biomarkers for outcomes that evolve over time. Recognising the need to account for patient heterogeneity when evaluating the accuracy of a prognostic biomarker, we introduce a novel penalised spline-based estimator of the time-dependent ROC curve that accommodates the possible modifying effect of covariates. Building on previous proposals, we adopt a modelling approach that considers flexible models for both the conditional hazard function and the biomarker, enabling the accommodation of non-proportional hazards and nonlinear effects through penalised splines, thus overcoming the limitations of earlier methods. A simulation study demonstrates that our approach successfully recovers the true functional form of the covariate-specific time-dependent ROC curve and the corresponding area under the curve across a variety of scenarios. Comparisons with existing methods further show that our approach performs favourably in multiple settings. We illustrate the utility of our method by assessing the ability of the Global Registry of Acute Coronary Events (GRACE) risk score to predict post-discharge mortality in patients with acute coronary syndrome, and how this predictive ability may vary with left ventricular ejection fraction.

This is joint work with Vanda Inácio.

Confidence intervals for the AUC with missing data

Susana Martins (Universidade de Vigo)

Abstract: The Area Under the ROC Curve (AUC) plays an important role in several fields, such as biomarker identification or the study of the predictive capacity of regression models. In this work, we consider the construction of confidence intervals for the AUC in the presence of missing data. Several approaches to construct confidence intervals from the empirical (nonparametric) AUC are explored: the Newcombe variance-based method, the percentile bootstrap method, and the normal quantile bootstrap method. To deal with missing data, we use complete case

analysis and multiple imputation. The relative performance of the methods is investigated in simulated missing completely at random and missing at random scenarios, considering a setting with a fully observed univariate biomarker and a potentially missing binary outcome. The methods are compared in terms of empirical coverage and interval width, providing practical recommendations for selecting the most appropriate procedure for AUC inference in the presence of incomplete data.

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

Goodness-of-fit testing with survival data

Jacobo de Uña-Álvarez (Universidade de Vigo)

Abstract: In this talk I will present a new general strategy for goodness-of-fit testing with survival data. The setting is that of testing for a parametric family of distribution functions when the data are deteriorated due to random censoring and/or random truncation. A key step is the characterization of the null hypothesis through a moment equation which involves the estimation of the observable distribution under both the null and the alternative. An omnibus test based on a maximum mean discrepancy principle will be proposed, and its theoretical properties will be presented. The finite sample performance of the proposed test will be investigated through simulations. Illustrative real data applications will be given.

This is joint work with Juan Carlos Escanciano.

The k-sample problem with left-truncated and right-censored data

Adrián Lago (Universidade de Vigo)

Abstract: In this talk, we propose k-sample versions of the Kolmogorov-Smirnov and Cramér-von Mises tests for data subject to left truncation and right censoring. We study their asymptotic behaviour under both the null and alternative hypotheses, and introduce a bootstrap resampling scheme to approximate the null distribution of the proposed statistics. The methodology is validated through Monte Carlo. Its finite-sample performance is assessed in a simulation study, which also includes the classical log-rank test for comparison. The performance of the new tests is illustrated with a real dataset on unemployment duration.

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

Robust estimation of conditional regions using quantile regression. Application in diabetes research

Sara Rodríguez Pastoriza (Universidade de Vigo)

Abstract: In clinical practice, the interpretation of continuous diagnostic markers requires reference intervals. For a single marker, these intervals define a range that captures 95% of the results from healthy individuals. However, for diseases involving two correlated markers, such as diabetes, a more powerful approach is to analyze them jointly using reference regions. These regions contain 95% of the results from healthy individuals, taking into account the joint distribution of both markers. My doctoral thesis explores and proposes advanced methods for the robust estimation of conditional reference regions, which I present in this article. We propose a new bivariate regression model based on quantile regression that estimates conditional bivariate regions. This method offers several advantages over previous approaches, including flexibility in modeling covariate effects, robustness to outliers, the ability to capture nonlinear covariate effects using cyclic spline functions, and the property that the estimated region leaves the same percentage of data outside in all possible directions. We validate our model through simulations, demonstrating its accuracy and robustness in the presence of outliers, and we explore its behavior on both Gaussian and non-Gaussian data. Finally, we apply it to a real study related to diabetes, modeling the joint distribution of two blood sugar markers and analyzing how age affects their distribution.

This is joint work with Javier Roca-Pardiñas and Óscar Lado-Baleato.