SiDOR
Statistical Inference
Decision & Operations Research Group
Universidade de Vigo

Dep. de Estatística
e Investigación Operativa

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"

Universidade de Vigo
cinbio

Workshop

'Methodological and computational challenges in statistical research'

(RETOS METODOLÓGICOS Y COMPUTACIONALES EN LA INVESTIGACIÓN ESTADÍSTICA)

November, 19-20th 2015

Facultad de CC. Económicas y Empresariales de Vigo, Aula-seminario 7

Program Overview

Thursday, November 19th

12:00 - 12:45 hrs. "Added prognostic value of omic predictors in survival analysis" – Mar Rodríguez Girondo (Leiden University Medical Center -LUMC, The Netherlands)

12:45 - 13:30 hrs. "From MTM2005 to MTM2014: Some achievements and some open problems" – Jacobo de Uña-Álvarez (SiDOR Research Group, Universidade de Vigo)

Lunch break

15:15 - 16:00 hrs. "SOP algorithm: Smoothing Parameter Selection in P-splines with Overlapping Penalties" – María Xosé Rodríguez Álvarez (SiDOR Research Group, Universidade de Vigo)

16:00 - 16:45 hrs. "Registro de pacientes con Lupus eritematoso sistémico de la Sociedad Española de Reumatología (RELESSER). Situación actual" - José María Pego Reigosa (División de Reumatología, Hospital Meixoeiro, Vigo)

Friday, November 20th

10:30 - 11:15 hrs. "Forecasting SO2 pollution incidents by means of multivariate quantile regression" – Javier Roca Pardiñas (SiDOR Research Group, Universidade de Vigo)

Coffee break

12:00 - 12:45 hrs. "Robust testing for superiority between two regression curves" – Juan Carlos Pardo-Fernández (SiDOR Research Group, Universidade de Vigo)

12:45 - 13:30 hrs. "Generalizations of the ROC curve" – Pablo Martínez-Camblor (Hospital Universitario Central de Asturias – HUCA, and Universidad de Oviedo)

Lunch break

15:15 - 16:00 hrs. "La estadística en la empresa: data mining" – Irene Castro Conde (Optare Solutions, S.L.)

ABSTRACTS

Thursday, November 19th

12:00 – 12.45 hrs. ***Title:*** **Added prognostic value of omic predictors in survival analysis – Mar Rodríguez Girondo (Leiden University Medical Center -LUMC, The Netherlands)**

Abstract: Augmentation of previously established high-dimensional survival models with new molecular markers is an important task in biomedicine. We introduce a new approach for the assessment of the augmented predictive value of omics predictors when considering a time-to-event (survival) outcome, potentially subject to left-truncation and right censoring. Our procedure is based on sequentially fitting penalized regression analysis of survival data using pseudo-observations as outcome. We propose several performance indices to summarize the two-stage prediction procedure and a permutation test to formally assess the augmented predictive value of a second omic set of predictors over a primary source. The performance of the test is investigated through simulations and illustrated through the analysis of added value of glycomics and metabolomics to predict survival in two different European populations.

12:45 – 13:30 hrs. ***Title:*** **From MTM2005 to MTM2014: Some achievements and some open problems – Jacobo de Uña-Álvarez (SiDOR Research Group, Universidade de Vigo)**

Abstract: En octubre de 2005, un equipo formado por cuatro doctores del Departamento de Estadística e Investigación Operativa de la Universidad de Vigo conseguía, por primera vez en esta institución, financiación ministerial para un proyecto de investigación en el ámbito de la Estadística No Paramétrica y Semiparamétrica. Aquél fue el origen de más de una década de investigación competitiva, culminada en los últimos años con la participación en diversas redes internacionales y nueva financiación ministerial para el período 2015-2017. A lo largo de este decenio, la masa crítica de este equipo investigador se ha incrementado notablemente; sus diez doctores actuales han dirigido una veintena de tesis doctorales, y han contribuido con más de 300 publicaciones en revistas de impacto. La financiación ministerial concedida desde el proyecto semilla MTM2005 hasta el actualmente en curso MTM2014 ha posibilitado plantear y resolver problemas estadísticos emergentes en distintas áreas (biomedicina, economía, medioambiente, ingeniería, bioinformática, deporte…) En esta charla se presentarán algunos de los logros alcanzados, y parte de los objetivos proyectados para el trienio en curso.

-Lunch break-

15:15 – 16:00 hrs. ***Title:*** **SOP algorithm: Smoothing Parameter Selection in P-splines with Overlapping Penalties – María Xosé Rodríguez Álvarez (SiDOR Research Group, Universidade de Vigo)**

Abstract: Roughness penalty smoothing has become the most popular method for performing non-parametric regression. However, this methodology depends on a key step: the selection of the smoothing parameter, which controls the trade off between fidelity to the data and smoothing. There are two main approaches to smoothing parameter selection: one based on the optimization of some criteria such as Akaike Information Criterion (AIC) or Generalized cross-validation (GCV), and one in which the smooth function is treated as random, and the smoothing parameters (or variance components) estimated by maximum likelihood (ML), or restricted maximum likelihood (REML). When it came to extending the aforementioned approaches to the estimation of multidimensional interaction surfaces, anisotropic low-rank tensor product smoothers have become the general approach (Eilers and Marx 2003, Wood 2006). However, for the REML/ML-based estimation approaches one is faced with the fact that estimation of the variance components cannot be accommodated using standard mixed model software, since the covariance matrix of the random effects has a non-standard form, with a block involving several variance components. Although estimation can be done by numerical maximization of the (restricted) log - likelihood (Wood 2006; 2011), it has the drawback of being computationally demanding, especially for large datasets. We present a fast and stable computational algorithm for estimating the smoothing parameters of a P-spline mixed model with overlapping penalties. The algorithm can be used whenever the penalty matrix of the P-spline model can be specified as a linear combination defined over the smoothing parameters. This includes, as particular cases, the multidimensional P-spline model with anisotropic penalty, as well as its adaptive counterpart (as presented in Wood 2011). Closed form expressions for the estimates of each variance component have been derived, which are then embedded into an iterative procedure. We call the algorithm SOP (Separation of Overlapping Penalties). The algorithm is a generalization to more complex situations of the SAP (Separation of Anisotropic Penalties) algorithm recently proposed by Rodríguez-Álvarez et al. (2015). This is a joint work with Dae-Jin Lee (Basque Center for Applied Mathematics, Spain), Thomas Kneib (Georg-August-Universität Göttingen, Germany), María Durbán (Carlos III University of Madrid, Spain) and Paul Eilers (Erasmus University Medical Centre, Rotterdam, The Netherlands)

16:00 – 16:45 hrs. ***Title:*** **Registro de pacientes con Lupus eritematoso sistémico de la Sociedad Española de Reumatología (RELESSER). Situación actual - José María Pego Reigosa (División de Reumatología, Hospital Meixoeiro, Vigo)**

Abstract: Systemic Lupus Erythematosus (SLE) is a relatively uncommon rheumatic disease. It can affect every organ system and has an autoimmune ethiology. As its prevalence is low, multicentric projects are necessary to get large samples of SLE patients that can be analyzed in order to get reliable results. The Spanish Society of Rheumatology recently created the Spanish Registry of SLE Patients (RELESSER) to have useful and reliable information about this disease in Spain. RELESSER consists of two stages: a) the cross-sectional one (already finished with several articles published in international journals) and b) the prospective one, more ambicious that is being developed at present: a follow-up study of different subgroups of the whole cohort of more than 4,000 patients. Information about the results of the analysis of RELESSER data and the current situation of the Registry are presented.

Friday, November 20th

10:30 – 11:15 hrs. ***Title:*** **Forecasting SO2 pollution incidents by means of multivariate quantile regression – Javier Roca Pardiñas (Universidad de Vigo)**

Abstract: More than 90% of the sulfur dioxide in the air comes from human sources. Due to the adverse health effects of high levels of sulfur dioxide, specific regulations have been adopted to manage and reduce the amount of sulfur dioxide produced. However, some SO2 emission incidents (i.e., emission exceeding the limits established by law) still ocurr. The aim of this paper is to predict time series of SO2 concentrations in order to estimate in advance high emission episodes and analyze the influence of previous series in the prediction. Previous studies aimed to forecast SO2 pollution incidents are based on estimating mean values. Instead, we propose the use of quantile regression models as they provide not only the mean but also the whole distribution of the pollution levels. A backfitting algorithm with local polynomial kernel smoothers was used to estimate the model, and critical values of the hypothesis test were obtained by means of bootstrapping. The performance of the method was evaluated using simulated data as well as real data drawn from an SO2 time series of a coal-fired power station located in Northern Spain. This is joint work with Isabel Martínez Silva and Celestino Ordóñez Galán.

-Coffee break-

12:00 – 12:45 hrs. ***Title:*** **Robust testing for superiority between two regression curves – Juan Carlos Pardo-Fernández (Universidad de Vigo)**

Abstract: In this talk we will focus on the problem of testing the null hypothesis that the regression functions of two populations are equal versus one-sided alternatives under a general nonparametric homoscedastic regression model. To protect against atypical observations, the test statistic is based on the residuals obtained by using a robust estimate for the regression function under the null hypothesis. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root-n local alternatives. A Monte Carlo study is performed to compare the finite sample behaviour of the proposed tests with the classical one obtained using local averages. A sensitivity analysis is carried on a real data set. This is joint work with Graciela Boente (Universidad de Buenos Aires).

12:45 – 13:30 hrs. ***Title:*** **Generalizations of the ROC curve – Pablo Martínez-Camblor (HUCA and Universidad de Oviedo)**

Abstract: The receiver operating characteristic curve is a popular graphical method frequently used in order to study the diagnostic capacity of continuous markers. It represents in a plot true-positive rates against the false-positive ones. Both the practical and theoretical aspects of the receiver operating characteristic curve have been extensively studied. Conventionally, it is assumed that the considered marker has a monotone relationship with the studied characteristic; i.e., the upper (lower) values of the (bio)marker are associated with a higher probability of a positive result. However, there exist real situations where both the lower and the upper values of the marker are associated with higher probability of a positive result. We propose a receiver operating characteristic curve generalization, Rg, useful in this context. All pairs of possible cut-off points, one for the lower and another one for the upper marker values, are taken into account and the best of them are selected. The natural empirical estimator for the Rg curve is considered and its uniform consistency and asymptotic distribution are derived. Finally, two real-world applications are studied.

-Lunch break-

15:15 – 16:00 hrs. ***Title:*** **La estadística en la empresa: data mining – Irene Castro Conde (Optare Solutions, S.L.)**

Abstract: Los operadores del sector de las telecomunicaciones disponen de grandes volúmenes de datos que no se aprovechan completamente y que, por tanto, necesitan de análisis avanzados que aumenten el valor de esta información para conseguir objetivos como incrementar los ingresos, gestionar la base de clientes, optimizar los procesos y generar nuevos modelos de negocio. Recientemente, la consultora tecnológica Optare Solutions S.L. ha participado en el proyecto BDA4T (Big Data Analytics for Telecoms), un proyecto de I+D en Cooperación Nacional, financiado por el CDTI (Centro para el Desarrollo Tecnológico Industrial) y que cuenta con el apoyo de la Universidad de Vigo como organismo de investigación. Uno de los objetivos de este proyecto consistía en el desarrollo de un sistema de análisis predictivo de bajas (churn) en los operadores de telecomunicaciones, con datos proporcionados por sus sistemas de gestión relacionados con el negocio e interacción con el cliente. El anglicismo churn describe la tasa de abandono de los clientes, es decir, aquellos clientes que dejan la compañía o el proveedor de un servicio durante un periodo de tiempo determinado. Este es uno de los problemas que resulta de mayor interés dentro del sector de las telecomunicaciones ya que resulta entre cinco y quince veces más caro captar nuevos clientes que retener a los actuales. Esto hace que la predicción del churn sea un arma muy potente para conocer de forma temprana qué clientes abandonarán la compañía. El objetivo final es poder construir una estrategia sólida para aumentar el grado de retención y ahorrar costes de adquisición para ser más eficientes y competitivos. Para llevar a cabo este cometido se realizó un análisis exhaustivo de diferentes herramientas de analítica: R, Weka, RapidMiner, KNIME, etc., y algoritmos de Data Mining para clasificación supervisada como Árboles de Decisión, Naive Bayes, etc., con el fin de decidir qué era lo más óptimo para resolver el problema en cuestión. Hecho esto se desarrolló el caso de uso del churn siguiendo el modelo CRISP-DM y obteniendo unos resultados muy satisfactorios. Uno de los siguientes pasos a seguir será estudiar la viabilidad del uso de técnicas Deep Learning en la predicción del churn. Este nuevo proyecto está financiado por una Ayuda Torres Quevedo.

REGISTRATION/INSCRIPCIÓN:

There is no registration fee. However, due to organizational issues, sending an e-mail to chuslonga@uvigo.es, with subject: **Registration Workshop 19-20N** before Thursday, November 12th at 14:00 is mandatory. Please indicate full name and affiliation (students at the Galician interuniversity Statistics Master Program or at the Galician interuniversity Statistics and OR Doctorate Program are particularly welcome).

La inscripción es gratuita. Por motivos de organización, para inscribirse es imprescindible enviar un e-mail antes del jueves, 12 de noviembre, a las 14:00 hrs., a chuslonga@uvigo.es, con el asunto: **Inscripción Workshop 19-20N,** indicando nombre completo y afiliación (los estudiantes del Máster interuniversitario en Técnicas Estadísticas y del Programa de Doctorado interuniversitario en Estadística e Investigación Operativa son especialmente bienvenidos).