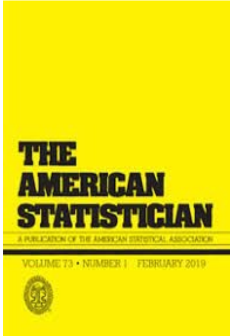

The American Statistician, ISSN 0003-1305
Volume 77, number 4 (november 2023)

Distribution-Free Location-Scale Regression

P. 345-356

Sandra Siegfried, Lucas Kook & Torsten Hothorn

Abstract

We introduce a generalized additive model for location, scale, and shape (GAMLSS) next of kin aiming at distribution-free and parsimonious regression modeling for arbitrary outcomes. We replace the strict parametric distribution formulating such a model by a transformation function, which in turn is estimated from data. Doing so not only makes the model distribution-free but also allows to limit the number of linear or smooth model terms to a pair of location-scale predictor functions. We derive the likelihood for continuous, discrete, and randomly censored observations, along with corresponding score functions. A plethora of existing algorithms is leveraged for model estimation, including constrained maximum-likelihood, the original GAMLSS algorithm, and transformation trees. Parameter interpretability in the resulting models is closely connected to model selection. We propose the application of a novel best subset selection procedure to achieve especially simple ways of interpretation. All techniques are motivated and illustrated by a collection of applications from different domains, including crossing and partial proportional hazards, complex count regression, nonlinear ordinal regression, and growth curves. All analyses are reproducible with the help of the tram add-on package to the R system for statistical computing and graphics.

Hypothesis Testing for Matched Pairs with Missing Data by Maximum Mean Discrepancy: An Application to Continuous Glucose Monitoring

P. 357-369

Marcos Matabuena, Paulo Félix, Marc Ditzhaus, Juan Vidal & Francisco Gude

Abstract

A frequent problem in statistical science is how to properly handle missing data in matched paired observations. There is a large body of literature coping with the univariate case. Yet, the ongoing technological progress in measuring biological systems raises the need for addressing more complex data, for example, graphs, strings, and probability distributions. To fill this gap, this article proposes new estimators of the maximum mean discrepancy (MMD) to handle complex matched pairs with missing data. These estimators can detect differences in data distributions under different missingness assumptions. The validity of this approach is proven and further studied in an extensive simulation study, and statistical consistency results are provided. Data obtained from continuous glucose monitoring in a longitudinal population-based diabetes study are used to illustrate the application of this approach. By employing new distributional representations along with cluster analysis, new clinical criteria on how glucose changes vary at the distributional level over 5 years can be explored.

Estimating the Performance of Entity Resolution Algorithms: Lessons Learned Through PatentsView.org

P. 370-380

Olivier Binette, Sokhna A York, Emma Hickerson, Youngsoo Baek, Sarvo Madhavan & Christina Jones

Abstract

This article introduces a novel evaluation methodology for entity resolution algorithms. It is motivated by PatentsView.org, a public-use patent data exploration platform that disambiguates patent inventors using an entity resolution algorithm. We provide a data collection methodology and tailored performance estimators that account for sampling biases. Our approach is simple, practical, and principled—key characteristics that allow us to paint the first representative picture of PatentsView's disambiguation performance. The results are used to inform PatentsView's users of the reliability of the data and to allow the comparison of competing disambiguation algorithms.

MOVER-R and Penalized MOVER-R Confidence Intervals for the Ratio of Two Quantities

P. 381-389

Peng Wang, Yilei Ma, Siqi Xu, Yi-Xin Wang, Yu Zhang, Xiangyang Lou, Ming Li, Baolin Wu, Guimin Gao, Ping Yin & Nianjun Liu

Abstract

Developing a confidence interval for the ratio of two quantities is an important task in statistics because of its omnipresence in real world applications. For such a problem, the MOVER-R (method of variance recovery for the ratio) technique, which is based on the recovery of variance estimates from confidence limits of the numerator and the denominator separately, was proposed as a useful and efficient approach. However, this method implicitly assumes that the confidence interval for the denominator never includes zero, which might be violated in practice. In this article, we first use a new framework to derive the MOVER-R confidence interval, which does not require the above assumption and covers the whole parameter space. We find that MOVER-R can produce an unbounded confidence interval, just like the well-known Fieller method. To overcome this issue, we further propose the penalized MOVER-R. We prove that the new method differs from MOVER-R only at the second order. It, however, always gives a bounded and analytic confidence interval. Through simulation studies and a real data application, we show that the penalized MOVER-R generally provides a better confidence interval than MOVER-R in terms of controlling the coverage probability and the median width.

Hierarchical Spatio-Temporal Change-Point Detection

P. 390-400

Mehdi Moradi, Ottmar Cronie, Unai Pérez-Goya & Jorge Mateu

Abstract

Detecting change-points in multivariate settings is usually carried out by analyzing all marginals either independently, via univariate methods, or jointly, through multivariate approaches. The former discards any inherent dependencies between different marginals and the latter may suffer from domination/masking among different change-points of distinct marginals. As a remedy, we propose an approach which groups marginals with similar temporal behaviors, and then performs group-wise multivariate change-point detection. Our approach groups marginals based on hierarchical clustering using distances which adjust for inherent dependencies. Through a simulation study we show that our approach, by preventing domination/masking, significantly enhances the general performance of the employed multivariate change-point detection method. Finally, we apply our approach to two datasets: (i) Land Surface Temperature in Spain, during the years 2000–2021, and (ii) The WikiLeaks Afghan War Diary data.

Revisiting the Name Variant of the Two-Children Problem

P. 401-405

Davy Paindaveine & Philippe Spindel

Abstract

Initially proposed by Martin Gardner in the 1950s, the famous two-children problem is often presented as a paradox in probability theory. A relatively recent variant of this paradox states that, while in a two-children family for which at least one child is a girl, the probability that the other child is a boy is 2/3, this probability becomes 1/2 if the first name of the girl is disclosed (provided that two sisters may not be given the same first name). We revisit this variant of the problem and show that, if one adopts a natural model for the way first names are given to girls, then the probability that the other child is a boy may take *any* value in (0,2/3). By exploiting the concept of Schur-concavity, we study how this probability depends on model parameters

RafterNet: Probabilistic Predictions in Multi-Response Regression

P. 406-416

Marius Hofert, Avinash Prasad & Mu Zhu

Abstract

A fully nonparametric approach for making probabilistic predictions in multi-response regression problems is introduced. Random forests are used as marginal models for each response variable and, as novel contribution of the present work, the dependence between the multiple response variables is modeled by a generative neural network. This combined modeling approach of random forests, corresponding empirical marginal residual distributions and a generative neural network is referred to as RafterNet. Multiple datasets serve as examples to demonstrate the flexibility of the approach and its impact for making probabilistic forecasts.

Mapping Life Expectancy Loss in Barcelona in 2020

P. 417-424

Xavier Puig & Josep Ginebra

Abstract

We use a Bayesian spatio-temporal model, first to smooth small-area initial life expectancy estimates in Barcelona for 2020, and second to predict what small-area life expectancy would have been in 2020 in absence of covid-19 using mortality data from 2007 to 2019. This allows us to estimate and map the small-area life expectancy loss, which can be used to assess how the impact of covid-19 varies spatially, and to explore whether that loss relates to underlying factors, such as population density, educational level, or proportion of older individuals living alone. We find that the small-area life expectancy loss for men and for women have similar distributions, and are spatially uncorrelated but positively correlated with population density and among themselves. On average, we estimate that the life expectancy loss in Barcelona in 2020 was of 2.01 years for men, falling back to 2011 levels, and of 2.11 years for women, falling back to 2006 levels.

Global Simulation Envelopes for Diagnostic Plots in Regression Models

P. 425-431

David I. Warton

Abstract

Residual plots are often used to interrogate regression model assumptions, but interpreting them requires an understanding of how much sampling variation to expect when assumptions are satisfied. In this article, we propose constructing global envelopes around data (or around trends fitted to data) on residual plots, exploiting recent advances that enable construction of global envelopes around functions by simulation. While the proposed tools are primarily intended as a graphical aid, they can be interpreted as formal tests of model assumptions, which enables the study of their properties via simulation experiments. We considered three model scenarios—fitting a linear model, generalized linear model or generalized linear mixed model—and explored the power of global simulation envelope tests constructed around data on quantile-quantile plots, or around trend lines on residual versus fits plots or scale-location plots. Global envelope tests compared favorably to commonly used tests of assumptions at detecting violations of distributional and linearity assumptions. Freely available R software (ecostats::plotenvelope) enables application of these tools to any fitted model that has methods for the simulate, residuals and predict functions.

Improved Approximation and Visualization of the Correlation Matrix

P. 432-442

Jan Graffelman & Jan de Leeuw

Abstract

The graphical representation of the correlation matrix by means of different multivariate statistical methods is reviewed, a comparison of the different procedures is presented with the use of an example dataset, and an improved representation with better fit is proposed. Principal component analysis is widely used for making pictures of correlation structure, though as shown a weighted alternating least squares approach that avoids the fitting of the diagonal of the correlation matrix outperforms both principal component analysis and principal factor analysis in approximating a correlation matrix. Weighted alternating least squares is a very strong competitor for principal component analysis, in particular if the correlation matrix is the focus of the study, because it improves the representation of the correlation matrix, often at the expense of only a minor percentage of explained variance for the original data matrix, if the latter is mapped onto the correlation biplot by regression. In this article, we propose to combine weighted alternating least squares with an additive adjustment of the correlation matrix, and this is seen to lead to further improved approximation of the correlation matrix.

The Wald Confidence Interval for a Binomial *p* as an Illuminating "Bad" Example

P. 443-448

Per Gösta Andersson

Abstract

When teaching we usually not only demonstrate/discuss *how* a certain method works, but, not less important, *why* it works. In contrast, the Wald confidence interval for a binomial *p* constitutes an excellent example of a case where we might be interested in why a method does *not* work. It has been in use for many years and, sadly enough, it is still to be found in many textbooks in mathematical statistics/statistics. The reasons for not using this interval are plentiful and this fact gives us a good opportunity to discuss all of its deficiencies and draw conclusions which are of more general interest. We will mostly use already known results and bring them together in a manner appropriate to the teaching situation. The main purpose of this article is to show how to stimulate students to take a more critical view of simplifications and approximations. We primarily aim for master's students who previously have been confronted with the Wilson (score) interval, but parts of the presentation may as well be suitable for bachelor's students.