SINAPE 2022

GRAMADO

Livro de Resumos

The electronic version of this booklet can be found at:
https://sinape2022.eventize.com.br/

Contents

APRESENTAÇÃO

SINAPE - Simpósio Nacional de Probabilidade e Estatística é a principal reunião científica da comunidade estatística brasileira, sendo organizado pela Associação Brasileira de Estatística (ABE). É um fórum único para a difusão no Brasil dos avanços da estatística mundial, tanto em termos teóricos e metodológicos, quanto em termos de sua interação e aplicação nas diversas áreas do conhecimento. Há uma sólida participação de pesquisadores nacionais, internacionais, alunos e profissionais do mercado.

Comissão Organizadora

Geral (ABE)

Marcos Oliveira Prates (Coordenador)
Celso Rômulo Barbosa Cabral
Larissa Ávila Matos

Local (UFRGS)

Flávio Ziegelmann (Coordenador)
Danilo Marcondes Filho
Hudson da Silva Torrent
Márcia Barbian
Patrícia Klarmann Ziegelmann
Silvana Schneider

Equipe Técnica

Geral (ABE)

Marcos Oliveira Prates

Local (UFRGS)

Alexandra Bautista
Andressa Dornelles
Elisa Fink
Fernanda Bianchi
Martha Reichel Reus
Raquel Pereira

Comissão Científica

Clarice Demétrio (ESALQ, Brasil – Coordenador)
Dani Gamerman (UFRJ e UFMG, Brasil)
Elias T. Krainski (KAUST, Arabia Saudita)
Flávio Ziegelmann (UFRGS, Brasil)
Francisco Cribari (UFPE, Brasil)
Jalmar Carrasco (UFBA, Brasil)
Jeremias Leão (UFAM, Brasil)
Marcelo Fernandes (FGV-SP, Brasil)
Marcia Branco (USP, Brasil)
Nancy Garcia (UNICAMP, Brasil)
Pedro Silva (ENCE-IBGE, Brasil)
Rosangela Loschi (UFMG, Brasil)

CONFERÊNCIAS PLENÁRIAS

C1

Coordenador: Renato Assunção (UFMG)

21st Century Teaching and Learning - What & How Should We Be Teaching and Learning?

Jo Boaler[1].

Recent years have seen an explosion of scientific evidence showing that there is a different way to learn, lead and live, available to us all. When people take a limitless approach to learning – in mathematics and in life – different pathways open up, leading to higher, more equitable and more enjoyable achievement. In this session we will consider what this different approach is, thinking about the ways we can teach students to increase equity, engagement, and achievement. We will also consider the nature of the content we are teaching. We are in an exciting time in terms of the knowledge we can all access, and the ways knowledge is being communicated. Our world is filled with data and data visualizations and a new, important goal for our teaching is to help students become data literate, learning to make sense of data in their lives and separate fact from fiction. All teachers can teach with a data perspective, integrating the ideas from data science into their teaching. This session will invite you to think about the ways we can teach and learn with a 21st century perspective.

[1]Stanford University

C2

Coordenador: Marcos Prates (UFMG)

Beyond Gaussian Processes: Flexible Bayesian Modeling and Inference for Geostatistical Processes

Flavio Bambirra Gonçalves[1].

In this talk, I will present a novel family of geostatistical models to account for features that cannot be properly accommodated by traditional Gaussian processes. The family is specified hierarchically, through a latent Poisson process, and combines the infinite-dimensional dynamics of Gaussian processes with that of any multivariate continuous distribution. The resulting process is called the Poisson-Gaussian Mixture Process - POGAMP. Whilst the attempt of defining geostatistical processes by assigning some arbitrary continuous distribution to be the finite-dimension distributions usually leads to non-valid processes, the finite-dimensional distributions of the POGAMP can be arbitrarily close to any continuous distribution and still define a valid process. Formal results to establish the existence and some important properties of the POGAMP, such as absolute continuity with respect to a Gaussian process measure, are provided. Also, an MCMC algorithm is carefully devised to perform Bayesian inference when the POGAMP is discretely observed in some space domain.

[1]UFMG

CONFERÊNCIAS PLENÁRIAS

C3

Coordenador: Clarice Garcia Borges Demétrio (USP/ESALQ)

A Estrutura Dual na Redução de Dimensionalidade: Aplicação em Dados Multiômicos

Julia Pavan Soler[1].

Dados multiômicos (coletados do Genoma, Transcriptoma, Proteoma, Fenoma, etc.) inauguram uma nova fase de pesquisa nas diferentes áreas factuais, como Medicina, Agropecuária e até mesmo no Marketing. Os desafios impostos se estendem desde o planejamento do estudo, coleta, armazenamento e análise dos dados, até a validação dos resultados, mediados por dificuldades decorrentes da natureza multidisciplinar envolvida. Além disso, o estrondo provocado pelo big-data também tem produzido ecos na pesquisa multiômica, sedenta por representações apropriadas, em baixa dimensão, de dados em altíssima dimensão (especificamente, baixo tamanho amostral realtivamente ao número de variáveis, isto é, n«p). Nesse contexto, muitos métodos de redução de dimensionalidade têm se apoiado nas propriedades de espaços duais, sob as quais, a análise de matrizes retangulares (nxp) de dados é realizada com base nas (co)variâncias (espaço pxp) ou nas distâncias (espaço nxn), como é o caso de Componentes Principais e Coordenadas Principais, que permitem também o uso de soluções penalizadas. Adicionalmente, levando em conta o efeito de preditores ou do desenho do estudo, a Regressão Multivariada (ou a MANOVA) também tem sua formulação clássica via (co)variâncias estendida para a análise de componentes simultâneos (ASCA) ou ainda para as correspondentes matrizes de distância, como é o caso da estatística de Mantel. Na literatura, há diferentes propostas de análises gravitando entre as representações retangulares de dados e suas formas quadráticas decorrentes, sendo imperativo levar em conta a estrutura dos dados para se alcançar os resultados pretendidos de modo apropriado e em baixa dimensão. Neste Encontro, para atender à finalidade de redução de dimensionalidade, integração de bancos de dados, bem como para o aprendizado de estruturas causais, farei considerações sobre como estender propriedades da estrutura dual em dados independentes para a modelagem de dados com dependência entre indivíduos. A motivação vem da análise de dados multiômicos avaliados em indivíduos e em seus familiares amostrados da população brasileira. Este trabalho tem a colaboração da pesquisadora Adèle Ribeiro (Pósdoutoranda na Columbia University) e é parcialmente financiado pela FAPESP (Projeto 17/05125-7)

[1]IME/USP

C4

Coordenador: Rosangela Loschi (UFMG)

The Statistical Analysis of Replication Success

Leonhard Held[1].

Although replication has long been a central part of the scientific method in many fields, the so-called replication crisis has led to increased interest in replication studies over the last decade. These developments eventually culminated in large-scale replication projects that were conducted in various fields. Declaring a replication as successful is, however, not a straightforward task, and diferente statistical methods are currently being used, including significance of both the original and replication study, compatibility of their effect estimates based on the Q-test, and computation of a meta-analytic combined effect estimate with confidence interval. I will propose a unified statistical framework for replicability which provides an assessment of significance, an assessment of compatibility and a combined confidence interval. The approach is based on the sceptical p-value, a method that combines reverse-Bayes methodology with a prior-predictive conflict assessment for the analysis of replication success. I will show how the method can be recalibrated to obtain a valid p-value with exact overall Type-I error control. The resulting "controlled" sceptical p-value avoids the dichotomisation for significance of the two studies and has larger project power to detect true existing effects. It gives rise to a p-value function which can be used to compute a confidence interval for the underlying true effect. If the effect estimates are compatible, the esulting confidence interval is similar to the meta-analytic one, but in the presence of conflict, the confidence interval splits into two disjoint intervals. Whether or not a split occurs is directly related to the result of the Q-test.

[1]University of Zurich

CONFERÊNCIAS PLENÁRIAS

C5

Coordenador: Elias T. Krainski (KAUST)

Food Security in Brazil: a Bayesian Network Model for Discrete Multivariate Time Series

Thais Cristina Oliveira da Fonseca[1]; Luiz Eduardo S. Gomes[1].

Chronic food insecurity represents one of the main obstacles to economical and social development in many countries. In ever-larger dynamic systems, such as the food system, it is increasingly difficult for decision-makers to effectively account for all the variables within the system that may influence the outcomes of interest under enactments of various candidate policies. For instance, the food system is usually influenced by food prices, availability, income, unemployment and demographic factors to cite a few. Each of the influencing variables is likely to be dynamic sub-systems with expert domains supported by sophisticated probabilistic models. In this talk, I present a graphical modelling approach that decouples a large multivariate system into a sub-system of smaller dimensions and recouples the original system for prediction and decision making. The model integrates historical data and expert knowledge within the framework of graphical models to allow for modelling and predicting food security over time. The proposal is based on Dynamic Bayesian Networks for discrete multivariate series which are computational efficient providing fast risk evaluation for several competing policies. The Dynamic Dirichlet process is considered to evolve the conditional probability tables (CPTs) over time using filtering and smoothing equations. An analysis of food security in Brazil is presented at the household level using the PNAD data.

[1]UFRJ

C6

Coordenador: Rosangela Loschi (UFMG)

Recent Advances in Adversarial Risk Analysis

Fabrizio Ruggeri[1],

At SINAPE 2016 I presented my first work in the field of Adversarial Risk Analysis (ARA). The talk was about Adversarial Hypothesis Testing in which an agent, called Defender, wants to ascertain which of several hypothesis holds, based on observations from a source that might be perturbed by another agent, whom we will call Attacker. A lot of work has been done about ARA since then, including the now popular field of Adversarial Machine Learning. I will present my recent and current works in the field, especially related to cybersecurity and industrial applications. For the former, I will present results on adversarial classification when the Attacker tries to manipulate data to induce the Defender to classify an item incorrectly. An example is a spam e-mail modified so that the Defender can classify it as a legitimate. For the latter, I will discuss about ARA for batch acceptance where a customer (Defender) must decide if to purchase or not a lot of certain items based on a sample which could be maliciously modified by the seller (Attacker). Finally, I will discuss about the optimal release policy, in the ARA framework, of a software when there are two producers and one buyer willing to purchase software from at most one of the two suppliers.

[1]CNR - Istituto Di Matematica Aplicata e Tecnologie Informatiche

CONFERÊNCIAS PLENÁRIAS

C7

Coordenador: Pedro Silva (ENCE/IBGE)

Experiências com Cursos para Amplas Audiências de Graduação e Pós-Graduação

Paulo Justiniano Ribeiro Jr[1].

Os fundamentos e métodos estatísticos são instrumentos frequentemente utilizados nas mais diferentes áreas do conhecimento, seja em estudos científicos, aplicações industriais, estatísticas oficiais ou outras. Isto faz com que profissionais de diversas áreas busquem, em algum nível, por proficiência na área. Além disso, conceitos fundamentais são necessários mesmo para o exercício da vida em sociedade e cidadania em tempos de informação abundantes. No ambiente acadêmico, estudantes tanto de graduação quanto de pós-graduação frequentemente possuem em sua formação disciplinas de estatística, seja obrigatória, optativa ou eletiva. A necessidade de conhecimento leva autores a produzirem materiais como cursos, livros, vídeos com diferentes ênfases e cada vez mais facilmente acessíveis. Seja com foco em fundamentos, em uso de recursos computacionais, de temas gerais ou específicos de uma área em particular, as iniciativas se multiplicam gerando uma saudável e rica diversidade de opções. Nesta apresentação iremos discutir duas iniciativas conduzidas no âmbito do Departamento de Estatística da Universidade Federal do Paraná (UFPR), uma para pós-graduação e outra para graduação. Serão detalhadas as motivações, os desafios, a construção e implementação de cursos voltados para amplas audiências. A iniciativa da pós-graduação teve início em 2018 na programação mais geral de disciplinas transversais da UFPR, enquanto a da graduação se iniciou em 2020 diante da situação de pandemia. Além de discorrer sobre concepções e recursos utilizados, a apresentação convida a uma reflexão sobre ambientes de troca de informações bem como de necessidades, possibilidades e oportunidades para iniciativas de unidades acadêmicas e associações da área.

[1]UFPR

C8

Coordenador: Elias T. Krainski (KAUST)

High-Dimensional Extreme Quantile Regression Using Partially-Interpretable Neural Networks

Raphaël Huser[1],

Risk management for extreme wildfires requires an understanding of the mechanisms that drive both ignition and spread. Useful metrics for quantifying such risk are extreme quantiles of aggregated burnt area conditioned on predictor variables that describe climate, biosphere and environmental states, as well as the abundance of fuel. Typically, these quantiles lie outside the range of observable data and so, for estimation, require specification of parametric extreme value models within a regression framework. Classical approaches in this context utilize linear or additive relationships between predictor and response variables and suffer in either their predictive capabilities or computational efficiency; moreover, their simplicity is unlikely to capture the truly complex structures that lead to the creation of extreme wildfires. In this paper, we propose a new methodological framework for performing extreme quantile regression using artificial neutral networks, which are able to capture complex non-linear relationships and scale well to high-dimensional data. The "black box" nature of neural networks means that they lack the desirable trait of interpretability often favored by practitioners; thus, we combine aspects of linear, and additive, models with deep learning to create partially interpretable neural networks that can be used for statistical inference but retain high prediction accuracy. To complement this methodology, we further propose a novel point process model for extreme values which overcomes the finite lower-endpoint problem associated with the generalized extreme value class of distributions. Our approach is applied to U.S. wildfire data with a high-dimensional predictor set and we illustrate vast improvements in predictive performance over linear and spline-based regression techniques.

[1]KAUST: King Abdullah University of Science and Technology

CONFERÊNCIAS PLENÁRIAS

C9

Coordenador: Flavio Ziegelmann (UFRGS)

Online Action Learning in High Dimensions: A Conservative Perspective

Marcelo Cunha Medeiros[1]; Claudio Flores[2].

Sequential learning problems are common in several fields of research and practical applications. Examples include dynamic pricing and assortment, design of auctions and incentives and permeate a large number of sequential treatment experiments. In this paper, we extend one of the most popular learning solutions, the ϵ_t-greedy heuristics, to high-dimensional contexts considering a conservative directive. We do this by allocating part of the time the original rule uses to adopt completely new actions to a more focused search in a restrictive set of promising actions. The resulting rule might be useful for practical applications that still values surprises, although at a decreasing rate, while also has restrictions on the adoption of unusual actions. With high probability, we find reasonable bounds for the cumulative regret of a conservative high-dimensional decaying ϵ_t-greedy rule. Also, we provide a lower bound for the cardinality of the set of viable actions that implies in an improved regret bound for the conservative version when compared to its non-conservative counterpart. Additionally, we show that end-users have sufficient flexibility when establishing how much safety they want, since it can be tuned without impacting theoretical properties. We illustrate our proposal both in a simulation exercise and using a real dataset.

[1]PUC/Rio
[2]BCB

C10

Coordenador: Marcia Branco (IME/USP)

Large-Scale Spatial Data Science with ExaGeoStat

Marc Genton [1].

Spatial data science aims at analyzing the spatial distributions, patterns, and relationships of data over a predefined geographical region. For decades, the size of most spatial datasets was modest enough to be handled by exact inference. Nowadays, with the explosive increase of data volumes, High-Performance Computing (HPC) can serve as a tool to handle massive datasets for many spatial applications. Big data processing becomes feasible with the availability of parallel processing hardware systems such as shared and distributed memory, multiprocessors and GPU accelerators. In spatial statistics, parallel and distributed computing can alleviate the computational and memory restrictions in large-scale Gaussian process inference and prediction. In this talk, we will describe cutting-edge HPC techniques and their applications in solving large-scale spatial problems with the new software ExaGeoStat and its R version ExaGeoStatR.

[1]King Abdullah University of Science and Technology (KAUST)

SESSÕES TEMÁTICAS

ST1: Modelagem Estatística da Evolução do Covid ou Estatística na COVID: Desafios e Contribuições

Organizadores: Dani Gamerman-UFRJ e Benilton Carvalho-UNICAMP

Two Years of COVID-19 in Brazil: A Statistician Point of View from Early-Warning of First Cases to Vaccine Efficiency

Leonardo Bastos[1]

In this talk I will show a chronological presentation of COVID-19 events in Brazil together with the analysis made by our group. Starting from adapting an existing early-warning surveillance system for severe acute respiratory infection (SARI) in order to deal with the large amount of data, then analysis of the higher risk groups helping the Brazilian ministry of health to define the COVID-19 vaccine priority order, some results on the impact of vaccination on reducing hospitalized cases, and some results on incidence of SARS-CoV-2 in a favela in Rio de Janeiro

[1]Fiocruz

ST1: Modelagem Estatística da Evolução do Covid ou Estatística na COVID: Desafios e Contribuições

Organizadores: Dani Gamerman-UFRJ e Benilton Carvalho-UNICAMP

Why Geometrical Risks Must be Combined with Statistical Analysis to Improve Emergency Response

Jones Albuquerque[1]

These are some of our [likainstitute.org and irrd.org/covid-19] results on these 2.5 years (and going on) of emergency response to COVID-19. In order to improve the accuracy of analyses and their scenarios, several scientists have applied dynamic approaches based on mathematical epidemiological models. Compartmental models, R(t) metrics, and other previous tools have been used to capture the dynamic behaviour of COVID-19, but several of them have been found to be wrong or misunderstood. Some misconstrued or even wrong uses of these metrics and approached prompt policymakers to formulate misguided public policies, such as early relaxation of physical distancing, etc., and cause the public to lose situational awareness of risks to life. In this webinar, the authors will present a geometric, data-driven parameter-free approach to COVID-19 analysis supported by solid statistical analyses to show how seriously the pandemic is in a place where stochastic and statistical data alone does not always produce reliable insights.

[1]UFRPE

SESSÕES TEMÁTICAS

ST1: Modelagem Estatística da Evolução do Covid ou Estatística na COVID: Desafios e Contribuições

Organizadores: Dani Gamerman-UFRJ e Benilton Carvalho-UNICAMP

Os Desafios Metodológicos, Logísticos e Políticos de se Fazer Pesquisa em Meio a Uma Pandemia

Pedro Hallal[1]

A condução de pesquisas epidemiológicas, nas quais muitas vezes as pessoas são abordadas em suas residências, é naturalmente desafiadora. Em várias ocasiões, o tamanho de amostra é grande, de forma que haja poder estatístico para a análise das questões de pesquisa definidas. Os desafios de conduzir pesquisas epidemiológicas em meio a uma pandemia são ainda mais complexos. Além dos cuidados habituais em termos de segurança e logística, a pandemia trouxe novos desafios, especialmente a dificuldade de as pessoas confiarem na pesquisa em função do risco de transmissão e dos discursos anti-ciência amplamente disseminados no país. Para abordagem esses temas, será utilizado o exemplo do EPICOVID-19, a maior pesquisa epidemiológica sobre Covid-19 no Brasil. Em pouco mais de 12 meses, os pesquisadores conduziram 11 inquéritos no Rio Grande do Sul, com amostra total de 45 mil gaúchos e quatro inquéritos nacionais, com amostra total de mais de 200 mil pessoas.

[1]UFPel

ST2: Aprendizado Estatístico - Núcleo de Data Science da FGV/EESP

Organizador: Marcelo Medeiros - PUC-Rio

Generalized Information Criteria for Structured Sparse Models

Eduardo Mendes[1].

Regularized M-estimators are widely used due to their ability of recovering a low-dimensional model in high-dimensional scenarios. Some recent efforts on this subject focused on creating a unified framework for establishing oracle bounds and deriving conditions for support recovery. Under this same framework, we propose a new Generalized Information Criteria that takes into consideration the sparsity pattern one wishes to recover. We obtain non-asymptotic model selection bounds and sufficient conditions for model selection consistency of the GIC. Furthermore, we show that one may use the GIC for selecting the regularization parameter in a way that the sequence of model subspaces contains the true model with probability converging to one. This allows practical use of the GIC for model selection in high-dimensional scenarios. We illustrate those conditions on examples including group LASSO generalized linear regression and low rank matrix regression.

[1]FGV

ST2: Aprendizado Estatístico - Núcleo de Data Science da FGV/EESP

Organizador: Marcelo Medeiros - PUC-Rio

A score-driven model of short-term demand forecasting for retail distribution centers.

Álvaro Veiga[1]

Forecasting is one of the fundamental input to support planning decisions in retail chains. Frequently, forecasting systems in retail are based on Gaussian models, which may be highly unrealistic when considering daily retail data. In addition, the majority of these systems rely on point forecasts, limiting their practical use in retailing decisions, which often requires the full predictive density for decision making. The main contribution of this paper is the modeling of daily distribution centers (DCs) level aggregate demand forecasting using a recently proposed framework for non-Gaussian time series called score-driven models or Generalized Autoregressive Score (GAS) models. An experimental study was carried out using real data from a large retail chain in Brazil. A log-normal GAS model is compared to usual benchmarks, namely neural networks, linear regression, and exponential smoothing. The results show the GAS model is a competitive alternative to retail demand forecasting in daily frequency, with the advantage of producing a closed form predictive density by construction.

[1]PUC-Rio

ST2: Aprendizado Estatístico - Núcleo de Data Science da FGV/EESP

Organizador: Marcelo Medeiros - PUC-Rio

Forecasting Global Inflation

Marcelo Medeiros[1]

Forecasting inflation is an important and difficult task. Most of the papers usually focus on a single or a small set of countries. In this paper, we consider the problem of simultaneously forecasting inflation from a large panel of countries. Our strategy is to explore potential (nonlinear) links among countries, and we do not rely on any additional variables apart from inflation and deterministic components, such as seasonal dummies.

[1]PUC-Rio

SESSÕES TEMÁTICAS

ST3: Metodologias Ativas para Ensino de Estatística

Organizador: Marcos Magalhães

Irene M. Garzola[1]

[1]UESC

ST3: Metodologias Ativas para Ensino de Estatística

Organizador: Marcos Magalhães

Maria Beatriz Assunção Mendes da Cunha[1]

[1]UNIRIO

SESSÕES TEMÁTICAS

ST3: Metodologias Ativas para Ensino de Estatística

Organizador: Marcos Magalhães

Maurem Porciúncula[1]

[1]FURG

ST4: Ciência de Dados

Organizador: Marcos Prates - UFMG

Estoque seguro: Estatística, Data Science e Programação

Daniel Falbel[1]

Experiências pessoais relacionadas à carreira de cientista de dados pela perspectiva de um profissional que é formado em estatística, atuou como cientista de dados e hoje trabalha como engenheiro de software.

[1] Rstudio

SESSÕES TEMÁTICAS

ST4: Ciência de Dados

Organizador: Marcos Prates - UFMG

Communicating Science Through Shiny Apps

Douglas R. M. Azevedo[1]

A atuação do cientista de dados não se encerra após o ajuste do modelo e análise dos resultados. A apresentação e comunicação dos resultados é também de grande importância para o sucesso de um projeto nessa área. Em particular, é necessário transmitir os resultados obtidos para diferentes públicos o que pode ser um grande desafio. O uso de ferramentas interativas permite que os resultados de um projeto sejam claramente expostos além de dar liberdade aos usuários para explorarem diferentes aspectos de interesse pessoal. Essa interatividade garante o engajamento dos usuários com o projeto. Esse comportamento se torna ainda mais interessante para projetos de cunho ambiental e/ou social, cujo resultado depende de um engajamento em massa. Dada a importância da tecnologia para um melhor entendimento de problemas ambientais e sociais, iniciativas como o Data for Good (Appsilon) são muito bem-vindas. Essas iniciativas tem a capacidade de causar um impacto positivo na sociedade atrvés da comunicação dos resultados de maneira assertiva e interativa promovendo o engajamento e a mudança de atitude dos usuários.

[1]Appsilon

ST4: Ciência de Dados

Organizador: Marcos Prates - UFMG

Ciência de Dados para Performance de Negócios

Roberto C. S. N. P. Souza[1]

A Big Data, fundada em 2012, é pioneira na área de big data analytics no Brasil. Nessa palestra vamos trazer um pouco da nossa experiência na aplicação de soluções de IA e ML em grandes empresas e mostrar como a aplicação dessas tecnologias tem resultados reais. Vamos contar como a cerveja que você bebe quando vai ao bar (ou pede no delivery) e o preço do hambúrguer que você como são determinados pelos nossos algoritmos.

[1]Big Data

SESSÕES TEMÁTICAS

ST5: Estatística para Dados em Alta Dimensão e Alta Frequência

Organizador: Ronado Dias - UNICAMP

Wavelet Spatio-Temporal Change Detection Method for Multi-Temporal SAR Images

Rogério G. Negri; Aluísio Pinheiro[1]; Abdourrahmane Atto

We introduce WECS (Wavelet Energies Correlation Screening), an unsupervised procedure to detect spatio-temporal change points on multi-temporal SAR images. The procedure is based on wavelet approximation for the multi-temporal images, wavelet energy apportionment, and ultra-high dimensional correlation screening for the wavelet coefficients. We show WECS performance on simulated multi-temporal image data. We also evaluate the proposed method on a time series of 84 satellite images in a forest region at the border of Brazil and the French Guiana. The proposed method displays good results in covering change regions, with the additional benefit of having simple and fast computation

[1]UNICAMP

ST5: Estatística para Dados em Alta Dimensão e Alta Frequência

Organizador: Ronado Dias - UNICAMP

From One to n-Dimensional Curve Time Series Objects

Chengqian Xian[1]

[1]University of Western Ontario, Canadá

ST5: Estatística para Dados em Alta Dimensão e Alta Frequência

Organizador: Ronado Dias - UNICAMP

From One to n-Dimensional Curve Time Series Objects

Flavio Ziegelmann[1]

The curve time series framework provides suitable approaches to accommodate some nonstationary features into a stationary setup. In the first part of this talk we review a way to identify the dimensionality of a one-dimensional curve time series object based on the dynamical dependence across different curves. Its practical implementation boils down to an eigenanalysis of a finite-dimensional matrix. The second part of the talk takes us to an initial adventure over n-dimensional curve time series objects, describing one way of tackling vectors of functionals via factor models. It is just a theoretical short description of a "just launched" approach, which can be explored in many directions.

[1]UFRGS

ST5: Estatística para Dados em Alta Dimensão e Alta Frequência

Organizador: Ronado Dias - UNICAMP

Modeling the Evolution of Infectious Diseases with Functional Data Models: The Case of COVID-19 in Brazil

Ronaldo Dias[1]; Julian Collazos[1]; Marcelo C. Medeiros[1]

In this paper, we apply statistical methods for functional data to explain the heterogeneity in the evolution of number of deaths of Covid-19 over different regions. We treat the cumulative daily number of deaths in a specific region as a curve (functional data) such thatthe data comprise of a set of curves over a cross-section of locations. We start by using clustering methods for functional data to identify potential heterogeneity in the curves and their functional derivatives. This first stage is an unconditional descriptive analysis, as we do not use any covariate to estimate the clusters. The estimated clusters are analyzed as "levels of alert" to identify cities in a possible critical situation. In the second and final stage, we propose a functional quantile regression model of the death curves on a number of scalar socioeconomic and demographic indicators in order to investigate their functional effects at different levels of the cumulative number of deaths over time. The proposed model showed a superior predictive capacity by providing better curve fit at different levels of the cumulative number of deaths compared to the functional regression model based on ordinary least squares.

[1]UNICAMP

ST6: Estatística e Sociedade: Inovação e Ferramentas para Combate à Desinformação

Organizador: Francisco Louzada Neto - USP - São Carlos

Estoque Seguro: Previsão de Demanda por Suprimentos Durante a Pandemia de COVID-19

Cibele Maria Russo Novelli[1]

A falta de insumos hospitalares e equipamentos de proteção individual (EPI), devido à demanda explosiva que se desenvolvia em diversas regiões do país, foi um dos grandes desafios enfrentados pelos gestores da área de saúde no Brasil no início da Pandemia do COVID-19. A desigualdade de realidades se intensificou com a aquisição excessiva de materiais por alguns hospitais e o consequente desabastecimento em outros. A falta de insumos hospitalares, somada ao número de leitos disponíveis e o número de profissionais por turno, foram fatores que impuseram limitações nos atendimentos e dificultaram a gestão da pandemia em muitas cidades. Nesse cenário, foi estabelecida uma parceria entre a empresa Bionexo e o CeMEAI/USP para o desenvolvimento de um sistema simples para a previsão de demanda de equipamentos de proteção individual em hospitais, que permitiria, inclusive, o empréstimo de materiais entre hospitais. A Bionexo é uma empresa de tecnologia que oferece soluções digitais para gestão de processos na saúde. Para este sistema, foi proposto uma combinação de modelos estatísticos para a demanda, com base em dados históricos de consumo de EPIs por hospitais, considerando os protocolos vigentes para seus usos e os dados epidemiológicos relacionados à doença, para construir modelos preditivos de demanda por EPIs em hospitais brasileiros durante os primeiros meses da pandemia. A modelagem desenvolvida foi incorporada no sistema gratuito chamado "Estoque Seguro", que fornece informações úteis para os hospitais, principalmente o nível de estoque de segurança e a previsão de consumo/demanda para cada equipamento de proteção individual desejado. Considerando as previsões fornecidas pelo sistema, um hospital poderia estimar seu estoque de segurança para atender a demanda futura, considerando seus níveis históricos de estoque e possíveis compras programadas. A ferramenta permite adotar estratégias para controlar e manter o estoque em níveis de segurança adequados à demanda, mitigando o risco de ruptura do sistema de saúde local. O sistema possibilitou o intercâmbio e a cooperação entre hospitais, visando maximizar a disponibilidade de equipamentos durante a pandemia. Os resultados desta pesquisa foram condensados em O. A. Gonzatto Jr., D. C. Nascimento, C. M. Russo, M. J. Henriques, C. P. Tomazella, M. O. Santos, D. Neves, D. Assad, R. Guerra, E. K. Bertazo, J. A. Cuminato, and F. Louzada (2022). Safety-stock: Predicting the demand for supplies in brazilian hospitals during the covid-19 pandemic. Knowledge-Based Systems 247, 108753.

[1]ICMC-USP

ST6: Estatística e Sociedade: Inovação e Ferramentas para Combate à Desinformação

Organizador: Francisco Louzada Neto - USP - São Carlos

Previsões Covid-19: Curto e Longo Prazo

Marcos Oliveira Prates[1]

O grupo CovidLP é formado por professores, alunas e alunos de pós-graduação em Estatística da UFMG. Ele teve origem como um desafio em uma disciplina de pós-graduação após a suspensão das aulas devido à Covid-19. Seu objetivo é oferecer um aplicativo online que possui dois principais resultados: previsões de curto prazo e de longo prazo da Covid-19. O primeiro se refere a previsões de mortes e número de casos confirmados para o futuro imediato (até 1 a 2 semanas). O segundo tipo de previsões é mais abrangente e visa traçar um panorama mais completo da pandemia: quando o número de casos deixará de crescer e começará a decair? Quantas pessoas essa pandemia irá adoecer? Quando podemos esperar que a pandemia seja encerrada? Nossas previsões são atualizadas diariamente, e podem se alterar com base nos novos dados que chegam todo dia. As previsões são acompanhadas dos respectivos intervalos de credibilidade para que o usuário tenha sempre noção da verdadeira incerteza associada a cada previsão fornecida. Outro ponto importante é que essas previsões são sempre baseadas na manutenção das condições no dia em que a previsão foi feita, incluindo as condições de isolamento. Alterações podem causar mudanças substanciais nas previsões. Nessa palestra irei mostrar a estrutura do time por trás do nosso aplicativo e algumas possíveis análises e resultados que podem ser obtidas pelo aplicativo da nossa modelagem.

[1]UFMG

ST6: Estatística e Sociedade: Inovação e Ferramentas para Combate à Desinformação

Organizador: Francisco Louzada Neto - USP - São Carlos

Fakenewsbr: Identificação Automática de Notícias Falsas

Francisco Louzada[1]

O uso intenso das tecnologias de comunicação tem levado ao crescimento da disseminação de notícias falsas e a necessidade de se desenvolver meios para combatê-las. Neste contexto, a estatística computacional desempenha importante papel no desenvolvimento de novas tecnologias capazes de identificar quais são as características do fenômeno da disseminação da desinformação. Nesta palestra apresentamos a plataforma fakenewsbr e suas ferramentas, as quais estão disponíveis para detecção de notícias falsas em textos em português. Este trabalho é co-autorado por D. C. F. Guzmán e M. J. Henriques do PIPGEs-UFSCar-USP e D. K. Neiva, G. H. Darú, L. G. Giordani, R. G. S. Queiroz, e V. W. Buzinaro do MECAI-USP.

[1]CeMEAI-ICMC-USP

ST7: Inferência Estatística para Processos Complexos

Organizador: Florencia Leonardi - IME/USP

Inferência Bayesiana para Cadeias de Markov de Alcance Variável e Campos Aleatórios Markovianos

Nancy Lopes Garcia [1]

Cadeias de Markov de Alcance Variável e Campos Aleatórios Markovianos são processos estocásticos para os quais, conhecida a vizinhança que afeta o estado atual, ou o sítio a ser atualizado, pode-se facilmente estimar as probabilidades de transição. Entretanto, a pergunta mais difícil de ser respondida é: como determinar esta vizinhança? Neste trabalho apresentamos, sob o paradigma Bayesiano, procedimentos para responder a esta pergunta.

[1]UNICAMP

ST7: Inferência Estatística para Processos Complexos

Organizador: Florencia Leonardi - IME/USP

Estimação com caudas pesadas e erros adversariais

Lucas Resende [1]

Dez anos atrás, Catoni [1] investigou as propriedades não assintóticas da média empírica para distribuições com cauda pesada, mostrando que era possível obter estimadores melhores para a média. Investigamos esse problema sob duas complicações adicionais: estimação uniforme e contaminação adversarial. O problema de estimação uniforme consiste em estimar uniformemente a média de f(X) em uma família de funções f. Já a hipótese de contaminação adversarial assume que parte dos dados é contaminada, mas nenhuma informação além da quantidade de entradas corrompidas é disponibilizada. Nossos resultados generalizam cotas anteriores presentes na literatura, em especial, fornecem uma relação ótima entre momento, contaminação e nível de confiança. A mesma teoria também é aplicada para regressão quadrática em dados com caudas pesadas e contaminação, melhorando resultados recentes de Lecué e Lerasle [2].

[1]

ST7: Inferência Estatística para Processos Complexos

Organizador: Florencia Leonardi - IME/USP

Detecção de pontos de mudança para a identificação de ilhas de homozigose

Lucas de Oliveira Prates [1]

Florencia Leonardi [1]

Neste trabalho propomos estimadores regularizados para pontos de mudança nos parâmetros de uma distribuição multidimensional quando temos várias amostras independentes alinhadas, de tamanho limitado. O estimador pode ser calculado eficientemente por um algoritmo de programação dinâmica ou aproximado por segmentação binária. Mostramos que ambos estimadores convergem quase certamente para o conjunto de pontos de mudança, sem a necessidade de especificar a priori o número de pontos de mudança. Além disso, apresentamos uma nova metodologia para a seleção da constante de regularização que tem a vantagem de ser automática, consistente e menos propensa à análise subjetiva. O trabalho está motivado pelo problema de identificar ilhas de homozigose no genoma dos indivíduos de uma população. Nosso método aborda diretamente a questão da identificação das ilhas de homozigose ao nível populacional, sem a necessidade de analisar indivíduos isolados e depois combinar os resultados, como é feito hoje em dia em abordagens estado-da-arte. Este é um trabalho em colaboração com os pesquisadores Renan B. Lemes, Tábita Hünemeier do Instituto de Biociências da Universidade de São Paulo.

1

1

SESSÕES ESPECIAIS

SE1: Homenagem a Julio Singer

Organizador: Pedro Morettin - USP

Homenagem a Julio Singer

Dalton de Andrade[1]; Mariana Curi[2]; Viviana Giampaoli[3]

Nesta ST pretendemos prestar uma homenagem ao Professor Julio da Motta Singer, que faleceu prematuramente no mês de maio. Julio foi um dos mais destacados estatísticos brasileiros, reconhecido pelas suas contribuições tanto em forma de artigos como de livros publicados no Brasil e no exterior. Recebeu o Prêmio ABE em 2019. Nesta homenagem, colegas e alunos falarão sobre suas vivências com o Julio.

Mini currículo Julio

Julio da Motta Singer é formado em Engenharia pela Escola Politécnica da Universidade de São Paulo, tem mestrado em Estatística pelo Instituto de Matemática e Estatística da mesma universidade e doutorado em Bioestatística pela Universidade da Carolina do Norte em Chapel Hill. É autor de vários artigos e livros sobre metodologia e aplicações da Estatística. É detentor do prêmio "Grizzle Distinguished Alumnus", oferecido pela Universidade da Carolina do Norte em Chapel Hill e do "Prêmio ABE", oferecido pela Associação Brasileira de Estatística. Também foi escolhido como patrono, paraninfo ou professor homenageado ao longo de 45 anos de atuação como professor do Departamento de Estatística da USP. Atualmente é professor titular desse departamento e tem atuado como diretor ou vice-diretor do Centro de Estatística Aplicada da mesma instituição.

[1]UFSC
[2]ICMC-USP
[3]USP

SE2: Homenagem a Heleno Bolfarine

Organizadora: Márcia Branco - USP

Estimation for Partially Linear Models with Autoregressive Skew-Normal Errors

Clécio da Silva Ferreira[1]

This paper proposes a partially linear model with the random error following an autoregressive (AR) process of order p skew-normal.The maximum likelihood estimators are calculated through of the Expectation-Maximization algorithm which have analytic expressions for the M and E-steps. The estimation of the effective degrees of freedom concerning the nonparametric component are obtained through of a linear smoother. The conditional quantile residuals are used to the construction of simulated confidence bands for assessing departures from the error assumptions and to construct graphs of the autocorrelation and partial autocorrelation functions for verifying the adequacy of the AR structure for the errors. A simulation study is also conducted to evaluate the efficiency of the EM algorithm. Finally, the methodology developed through the paper is illustrated with a real data set on cardiovascular mortality.

[1]UFJF, Brasil

SE2: Homenagem a Heleno Bolfarine

Organizadora: Márcia Branco - USP

Heleno's contributions to Latent Variable Models

Jorge Luis Bazán [1].

In this talk we review the contributions of professor Heleno Bolfarine in Latent Variable Models. Applications ilustring the methodology are presented as well. We hope motivate the statistical community in order to increase the contributions in latent variable models as a contribution to the memory from professor Heleno

[1]USP, Brasil

SE2: Homenagem a Heleno Bolfarine

Organizadora: Márcia Branco - USP

Robust Estimation in Functional Comparative Calibration Models via Maximum Lq-likelihood

Manuel Galea[1]; Patricia Giménez[2]; Lucas Guarracino[2]

A parametric estimation procedure is proposed for robust estimation of the structural parameter in a functional comparative calibration model, under normality, based on maximum Lq-likelihood approach. The estimator, called MLqE, depends on a single distortion parameter q, which controls the balance between robustness and efficiency. If q tends to 1 , the maximum likelihood estimator (MLE) is obtained. The estimation procedure can be implemented easily by a simple and fast re-weighting algorithm. For applying the method to practical and real-data scenarios, a data-based choice of an appropriate value of q is proposed. Consistency and asymptotic normality is established and the covariance matrix is given. The influence function is derived, to show the local robustness properties. Theoretical properties, ease of implementability and empirical results on simulated and real data show the satisfactory behavior of the MLqE and its vantages over the MLE in presence of observations discordant with the assumed model.

[1]PUC – Santiago, Chile
[2]Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

MINICURSOS

MC1

Recent Advances in Statistical Modeling of Spatial Extremes

Raphael Huser [1].

The classical modeling of spatial extremes relies on asymptotic models (i.e., max-stable processes or r-Pareto processes) for block maxima or peaks over high thresholds, respectively. However, at finite levels, empirical evidence often suggests that such asymptotic models are rigidly constrained, and that they do not always adequately capture the situation where the most severe events tend to be spatially localized. Another well-known limitation of classical spatial extremes models is that they are either computationally prohibitive to fit in high dimensions, or they need to be fitted using less efficient techniques. In this short course, we will start by describing classical asymptotic models for univariate and spatial extremes defined as block maxima and threshold exceedances. Then, in the second part, we will describe recent progress in the modeling and inference for spatial extremes, focusing on new models that have more flexible tail structures that can bridge asymptotic dependence classes, and that are more easily amenable to likelihood-based inference for large datasets. In particular, we will discuss various types of random scale constructions, as well as the conditional spatial extremes model, which have recently been getting increasing attention within the statistics of extremes community. We will illustrate the practical usefulness of some of these spatial extreme-value models on various environmental applications.

[1]King Abdullah University of Science and Technology (KAUST)

MC2

Modelos geoestatísticos para fenômenos espaço-temporais

Guilherme Vieira Nunes Ludwig [1]

Este minicurso tem como objetivo introduzir conceitos de estatística espacial (ou espaço-temporal) para a modelagem de dados espacialmente estruturados pelo estatístico aplicado, e alguns temas contemporâneos de pesquisa metodológica para alunos interessados na carreira acadêmica. Serão introduzidos modelos de variograma espaciais e espaço-temporais, sua caracterização teórica (e conexão com funções de covariância) e critérios para estimação de variogramas a partir de dados. Com a caracterização de processos estocásticos espaciais (em particular o processo Gaussiano), apresentaremos a técnica de interpolação espacial (ou Kriging), tomando como exemplo em particular dados meteorológicos disponibilizados pelo INMET (Ministério da Agricultura, Pecuária e Abastecimento: Instituto Nacional de Meteorologia, 2011) (Disponíveis em `http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas`). A referência principal será Cressie and Wikle (2011). O curso também irá explorar alguns modelos não-separáveis de variograma espaço-temporal, como os propostos em Cox and Isham (1988), Gneiting (2002) e Ludwig et al (2017); estudaremos a caracterização de modelos de variograma (ou covariância) em geral através de funções condicionalmente negativas-definidas (ou positivas definidas, respectivamente), e algumas das dificuldades teóricas e computacionais na estimação de funções de variograma.

[1]UNICAMP

Design and Analysis of Replication Studies - with an introduction to the R package Replication Success

Leonhard Held [1]

Charlotte Micheloud [1]

Replication studies are increasingly conducted in order to confirm original findings. However, there is currently no consensus on how to design such studies and to define replication success. The purpose of this tutorial is to describe and compare statistical approaches for the design and analysis of replication studies. The standard method based on significance is discussed, as well as alternative approaches based on effect sizes or meta-analysis. Participants will learn how to use the R-package ReplicationSuccess and will apply the methods to real data from large-scale replication projects. Prerequisites include basic knowledge of R and familiarity with standard concepts of statistical inference.

[1]University of Zurich
[1]University of Zurich

MC4

Aprendizado profundo (Deep learning)

Renato Assunção [1]

Daniel Falbel [1]

Algoritmos de Aprendizado Profundo (Deep Learning) aprendem representações de dados em vários níveis, com cada nível explicando os dados de maneira hierárquica. Tais modelos têm sido eficazes na descoberta da estrutura subjacente em dados. Eles foram usados com imenso sucesso em muitos problemas de análise de dados, incluindo a classificação de imagens, o reconhecimento de fala e o processamento de linguagem natural. As redes generativas adversariais (GANs) formam uma proposta revolucionária para geração Monte Carlo de sistemas muito complexos tais como gerar imagens realistas de faces humanas. O minicurso vai misturar aulas expositivas e código em R cobrindo a teoria subjacente e uma variedade de aplicações enfatizando o uso de grandes conjuntos de dados.

[1]UFMG
[1]RStudio Inc.

T1: Um Tutorial sobre GAMLSS no R

Um Tutorial Sobre GAMLSS no R

Fernanda De Bastiani[1]; Cristian Villegas[2];

Os modelos aditivos generalizados para locação, escala e forma, GAMLSS (Generalized Additive Models for Location, Scale and Shape), são modelos de regressão univariada, em que todos os parâmetros da distribuição assumida para a variável resposta podem ser modelados como uma função aditiva das variáveis explicativas. GAMLSS fornece uma estrutura para abordar problemas como a escolha de uma distribuição apropriada para a variável de resposta e explicando como essa distribuição, e seus parâmetros, variam em função das variáveis explicativas. A classe GAMLSS considera diferentes termos aditivos para modelar os parâmetros da distribuição, como linear, suavização não paramétrica e termos de efeitos aleatórios. E contém diferentes técnicas de seleção de modelos e diagnósticos (resíduo quantílico para uma variável aleatória continua ou resíduo quantílico aleatorizado para uma variável aleatória discreta) para verificar a adequação do modelo. Existem mais de dez pacotes disponíveis no R para dar suporte aos GAMLSS, entre eles estão o pacote gamlss que pode ser chamado de pacote principal, o pacote gamlss.dist com um conjunto de distribuições (discretas, contínuas e misturas) que podem ser usadas para modelar as variáveis de resposta no GAMLSS, gamlss.data com conjuntos de dados utilizados em exemplos. O objetivo deste tutorial é apresentar a implementação do GAMLSS na plataforma computacional do R; Apresentaremos a classe GAMLSS; apresentaremos os pacotes do R; Mostraremos como podemos utilizar a função gamlss e detalharemos alguns dos seus componentes; Apresentaremos os diferentes tipos de distribuições implementadas no pacote gamlss.dist; Mostraremos alguns dos termos aditivos disponíveis; Mostraremos a seleção de variáveis explicativas; Explicaremos a parte de diagnóstico; Exemplos práticos serão apresentados e discutidos.

[1]UFPE
[2]ESALQ/USP

T2: Estatística Aplicada com R e Python: Uma Abordagem Integrada

Estatística Aplicada com R e Python: Uma Abordagem Integrada

Wagner Hugo Bonat[1]; Walmes Marques Zeviani[1]

Com a popularização do termo Ciência de Dados uma série de profissionais de todas as áreas têm voltado sua atenção para análise de dados como um suporte para a tomada de decisões nos mais diversos tipos de negócios. A grande disponibilidade de dados oferecida pelo desenvolvimento das tecnologias da informação aliada com o desenvolvimento de ferramentas computacionais eficientes e acessíveis vêm transformando a forma de diversos tipos de negócios. Neste contexto, duas ferramentas computacionais têm ganhado destaque, o R e o Python. Ambas oferecem um ambiente para análise de dados e modelagem estatística. O objetivo deste curso é oferecer uma visão ampla das capacidades das linguagens de programação R e Python para manipulação e visualização de dados, bem como para modelagem estatística. Ambas linguagens trabalham com uma estrutura básica que é estendida por uma série de pacotes ou libraries que aumentam ou melhoram sua capacidade de lidar com tarefas específicas, tais como manipulação e visualização de dados ou implementam metodologias estatísticas mais avançados como modelos lineares generalizados, modelos aditivos generalizados entre outros. Dentre as diversas opções que cada linguagem oferece, nós optamos por focar nos pacotes conhecidos como tidyverse no ambiente R e pandas no ambiente Python. Essa escolha se deve a popularidade destas implementações. Com relação a visualização de dados as escolhas foram pelo pacote ggplot2 e matplotlib (e suas extensões) nos ambientes R e Python, respectivamente. Para oferecer uma visão geral das capacidades de modelagem estatística em R, exploramos as funcionalidades já nativas oferecidas pelas funções lm() e glm() e complementamos com as do pacote mgcv. Em Python optamos pelo pacote statmodels que oferece um conjunto ainda incompleto mas em desenvolvimento de diversos modelos estatísticos. Por fim, deixamos alguns caminhos para futuros desenvolvimentos principalmente em termos de modelagem estatística e indicamos algumas extensões que estamos desenvolvendo em Python em relação a classe de modelos de regressão multivariados intitulado mcglm (Multivariate Covariance Generalized Linear Models).

[1]Universidade Federal do Paraná

T3: Modeling Longitudinal Data using Robust Mixed Models in R

Modeling Longitudinal Data using Robust Mixed Models in R

Fernanda Lang Schumacher[1]; Larissa Avila Matos[2]; Victor Hugo Lachos[3]

In clinical trials, studies often present longitudinal data. They are commonly analyzed using linear mixed models (LMM), which, for mathematical convenience, usually assume that both random effect and error follow normal distributions. These models are frequently used in applications, and a brief review of the standard tools available in R to estimate them will be presented. Nevertheless, real data frequently present non-normal features, such as heavy tails and skewness. In these cases, the normality assumptions may result in a lack of robustness against departures from the normal distribution and in invalid statistical inferences. Aiming to facilitate the use of more complex models in applications, in this tutorial, we will present a flexible extension of the normal LMM using the R package skewlmm, which considers the scale mixture of skew-normal class of distributions, accommodating skewness and heavy-tails, and accounts for a possible within-subject serial dependence, considering some useful dependence structures. The use of the package will be illustrated in real longitudinal data, exploring tools for models selection and evaluation. Palavras-chave: Longitudinal data; Mixed-effects models; Robust models; Scale mixture of skew-normal distributions; Data analysis in R.

[1]Division of Biostatistics, CPH – OSU
[2]Departamento de Estatística, IMECC – UNICAMP
[3]Department of Statistics, UCONN

T4: Regressão Geograficamente Ponderada

Regressão Geograficamente Ponderada

Alan Ricardo da Silva[1]

A tentativa de representar a realidade por meio de modelos, matemáticos ou não, continua sendo um grande desafio para a ciência que, década após década, procura sempre aprimorar tais ferramentas. Uma das técnicas de modelagem matemática mais utilizada é a análise de regressão, que vem sendo atualizada nos últimos anos devido à incorporação de fatores que ajudam a explicar e entender os fenômenos. Dentre essas atualizações destacam-se a regressão espacial tratada de forma global e a regressão espacial tratada de forma local, na qual se destaca a Regressão Geograficamente Ponderada (RGP), ou do inglês Geographically Weighted Regression (GWR). Esta última se diferencia da primeira por analisar as relações entre as variáveis de forma específica para cada unidade de estudo, e não conjuntamente como é feito em um processo global. No caso, tem-se como pressuposto que as regiões j mais próximas da região i possuem maior influência nas estimativas dos coeficientes da regressão do que regiões mais afastadas. Assim, tendo um ajuste específico para cada área, o resultado final é uma melhor representatividade do processo como um todo. O tutorial abordará as características do modelo de regressão geograficamente ponderado, bem como suas vantagens e problemas, além do que vem sendo atualmente desenvolvido sobre o tema. Será utilizado o software SAS On Demand, que pode ser acessado de forma gratuita pelo navegador, sem a necessidade de instalação.

[1]Departamento de Estatística, Universidade de Brasília

COMUNICAÇÕES ORAIS

CO1 - Séries Temporais

Estimação da Dinâmica da Desocupação no Brasil com Modelos de Espaço de Estados para Pesquisas Amostrais

Caio César Soares Gonçalves[1]; Luna Hidalgo[2]; Denise Britz do Nascimento Silva[3]

Este trabalho examina a dinâmica da desocupação no país com base em estimativas mensais, produzidas a partir da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC), do total de desocupados do Brasil e de unidades da federação (UFs) selecionadas, utilizando a abordagem baseada em modelos univariados e bivariado de séries temporais na formulação de espaço de estados. Adicionalmente, os modelos decompõem o efeito que o desenho amostral da PNADC exerce sobre a autocorrelação da série de estimativas e o modelo bivariado incorpora a série de pessoas que recebem seguro-desemprego como informação auxiliar. Além disso, estimativas da primeira diferença foram obtidas para identificar se as variações mensais são estatisticamente significativas durante o período de análise de 2012 a 2020. Os resultados encontrados apontam que a série de total de beneficiários do seguro-desemprego não apresentou tendência comum com a série de total de desocupados, indicando que existem descompassos entre as duas, o que explica a diferença entre o caso brasileiro e as experiências internacionais na utilização do modelo bivariado. As divergências envolvem questões burocráticas, conceituais e de elegibilidade do seguro-desemprego, além de questões associadas ao alto nível de informalidade do mercado de trabalho brasileiro. Adicionalmente, as estimativas de primeira diferença mostraram-se significativas por longos períodos, desde o início da PNADC, evidenciando que as frequentes mudanças no país, sejam econômicas, políticas, ou até mesmo as ações para enfrentamento da pandemia da COVID19, provocaram, e continuam provocando, alterações na dinâmica do mercado de trabalho.

Palavras-chave: Desocupação; Seguro-Desemprego; Modelo de Espaço de Estados; Pesquisa Amostral; Séries Temporais.

[1]Fundação João Pinheiro (FJP), Belo Horizonte-MG e Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro-RJ, Rio de Janeiro-RJ – ccsgonc@gmail.com
[2]Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ – luna.hidalgo@ibge.gov.br
[3]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro-RJ – denise.silva@ibge.gov.br

CO1 - Séries Temporais

Censored Autoregressive Regression Models with Student-t Innovations

Katherine A. L. Valeriano[1]; Fernanda L. Schumacher[2]; Christian E. Galarza[3]; Larissa A. Matos[4]

Data collected over time is common in applications and may contain censored or missing observations, making it impossible to use standard statistical procedures. This paper proposes an algorithm to estimate the parameters of a censored linear regression model with errors serially correlated and innovations following a Student-t distribution. This distribution is widely used in statistical modeling of data containing outliers since its longer-than-normal tails provide a robust approach to handling such data. The maximum likelihood estimates of the proposed model are obtained through a stochastic approximation of the EM algorithm. The methods are applied to an environmental dataset regarding ammonia-nitrogen concentration, which is subjected to a limit of detection (left censoring) and contains missing observations. Additionally, two simulation studies are conducted to examine the asymptotic properties of the estimates and the robustness of the model.

Palavras-chave: Autoregressive Models; Censored Data; Heavy-Tailed Distributions; Missing Data; SAEM Algorithm.

[1]Departamento de Estatística, Universidade Estadual de Campinas - katandreina@gmail.com
[2]Departamento de Estatística, Universidade Estadual de Campinas - fernandalschumacher@gmail.com
[3]Departamento de Matemáticas, Escuela Superior Politécnica del Litoral - chedgala@espol.edu.ec
[4]Departamento de Estatística, Universidade Estadual de Campinas - larissam@unicamp.br

Stochastic Volatility with Missing Data: Assessing the Effects of Holidays

Omar Abbara[1]; Mauricio Zevallos[2]

In empirical finance, it is usual to consider holidays as if they do not exist. The main goal of this paper is to assess the effects of holidays on volatility estimation and prediction. Holidays are taken into account by assuming they are missing values in a time series of returns generated by a Stochastic volatility (SV) model. Estimation is evaluated through Monte Carlo experiments. In addition, we assess the effects of holidays on one-step ahead Value-at-Risk forecasting using several time series returns. The results are slightly better when we take into account the missing values, especially for VaR forecasting.

Palavras-chave: Estimation; Forecasting; Value-at-Risk.

[1]Canvas Capital S.A. – muhieddine@gmail.com
[2]Departamento de Estatística – IMECC/Unicamp – amadeus@unicamp.br

CO1 - Séries Temporais

Robust Semiparametric Nonlinear Autoregressive Models

Marcelo M. Taddeo[1]; Pedro A. Morettin[2]

In this paper we consider a nonlinear autoregression for time series data and estimate the target function using the EM algorithm. In order to get estimators more resitant to ouliers than those based on Normal errors, we also assume them to follow a scale mixture of Gaussian distributions. This class of distributions includes, among others, the Student's t distribution and symmetric stable distributions. The methodology is extended to the case of the partially linear model. To illustrate the methodology and assess its performance, we conduct a simulation study and make an application to a real series.

Palavras-chave: Scale Mixtures; Splines; EM Algorithm; Nonlinear Model; Autoregression.

[1]Departamento de Estatística, Universidade Federal da Bahia – marcelo.taddeo@gmail.com
[2]Departamento de Estatística, Universidade de São Paulo, São Paulo – pam@ime.usp.br

COMUNICAÇÕES ORAIS

CO1 - Séries Temporais

CO2 - Ciência de Dados

Machine Learning Auto-Tuning Processes: a Study on Support Vector Machines Applied to Smile Detection

Anderson Ara[1]; João Gondim[2]; Alexandre Loch[3];

Support vector machines are a set of statistical machine learning models that have great performance in several tasks as well as on image classification and object recognition. However, the proper choice of model's hyperparameters has a great influence on the outcomes and the general capacity performance. In this paper, we explore some different traditional auto-tuning processes to estimate σ hyperparameter for SVM Gaussian kernel. These processes are common and also implemented on standard software of data science languages. The paper considers some different situations on smile detection. The results are composed by simulation studies and two benchmark image applications.

Palavras-chave: Machine Learning; SVM; Gaussian Kernel; Tuning; Smile Detection.

[1]Departament of Statistics, Federal University of Paraná (UFPR), Curitiba – ara@ufpr.br
[2]Institute of Computing, Federal University of Bahia (UFBA), Salvador – joao.gondim@ufba.br
[3]Institute of Psychiatry, University of São Paulo (USP), São Paulo – alexandre.loch@usp.br

COMUNICAÇÕES ORAIS

CO2 - Ciência de Dados

Aprendizado Não-Supervisionado para Textos Curtos

Gustavo Machado Utpott[1]; Márcia Helena Barbian[2,3]

Com a evolução da tecnologia na área da comunicação, quantidades enormes de textos têm sido escritas e compartilhadas em diversas plataformas ao longo da internet, levando a uma demanda crescente de algoritmos de Processamento de Linguagem Natural (NLP). Os objetivos das análises são diversos e buscam desde a identificação de *spams*, tradução ou classificação de textos a análise de sentimentos. Dentre esses temas, descobrir tópicos de documentos de textos que não possuem uma classificação prévia tornam-se cada dia mais comuns, tais métodos, denominados Modelos de Tópicos são definidos como uma classe de algoritmos de Aprendizado não Supervisionado. Especificamente, documentos que possuem uma quantidade limitada de caracteres, os textos curtos, necessitam de metodologias diferentes daquelas comumente aplicadas, como o conhecido algoritmo *Latent Dirichlet Allocation* (LDA). O presente trabalho visa aplicar uma dessas técnicas, o *Biterm Topic Modeling* (BTM), em uma base de dados composta por descrições de diferentes mercadorias para que, após o agrupamento, seja possível selecionar os tópicos com mais semelhança a um dado produto de interessse. Além da aplicação do BTM à base, será proposto um algoritmo para substuição de abreviações contidas nos documentos a serem analisados.

Palavras-chave: Aprendizado Não Supervisionado; Modelagem de Tópicos; Processamento de Linguagem Natural; Textos Curtos; Biterm Topic Model.

[1]Estudante de graduação em Estatística – UFRGS, Porto Alegre – gustavo.utpott@gmail.com
[2]Departamento de Estatística – UFRGS, Porto Alegre
[3]Programa de Pós Graduação em Estatística – UFRGS – Porto Alegre – mhbarbian@gmail.com

CO2 - Ciência de Dados

Rafa 2030 - Classificação de Textos Jurídicos em ODS da Agenda 2030

Pamella Sada Dias Edokawa[1]; Thaís Carvalho Valadares Rodrigues[2];
Ariane Hayana Thomé de Farias[3]; Euler Rodrigues de Alencar[4];
Lucas José Gonçalves Freitas[5]

O Supremo Tribunal Federal (STF), instância máxima do sistema judiciário brasileiro, produz imensa quantidade de dados geralmente organizados em forma de texto, por meio de decisões, petições, liminares, recursos e outros documentos legais. Neste contexto, uma ferramenta de aprendizagem supervisionada para classificação de documentos jurídicos nos 17 objetivos (ODS) da Agenda 2030 da ONU para o Desenvolvimento Sustentável pode ser de grande utilidade para o tribunal, uma vez que essa tarefa é realizada manualmente por um grande grupo de funcionários e a adoção da Agenda 2030 da ONU é um dos principais objetivos da corte atualmente. O objetivo geral deste projeto, denominado RAFA 2030, é gerar valor para o tribunal por meio da construção de sistemas de classificação baseados em Processamento de Linguagem Natural – NLP. Atualmente, as principais entregas deste projeto consistem em ferramentas gráficas para NLP (Coocurrence graphs, nuvem de palavras), algoritmos de aprendizagem de máquina, redes neurais e contagem de palavras-chave, além de outras ferramentas disponíveis em R e Python (Keras, Tensorflow e Pytorch). Os resultados iniciais sugerem imenso potencial para aplicações de NLP e aprendizagem de máquina na classificação de documentos jurídicos em temas da Agenda 2030. RAFA é um acrônimo para Redes Artificiais com Foco na Agenda 2030.

Palavras-chave: Agenda 2030 da ONU; Aprendizagem de Máquina; Aprendizagem Profunda; Classificação de Textos.

[1]Secretaria de Gestão Estratégica, STF – pamella.edokawa@stf.jus.br
[2]Departamento de Estatística, UnB – thaisrodrigues@unb.br
[3]Escritório de Estatística, STF/Departamento de Estatística, UFAM – ariane.hayane@stf.jus.br
[4]Secretaria de Gestão Estratégica, STF – euler.alencar@stf.jus.br
[5]Escritório de Estatística, STF/Departamento de Estatística, UnB – lucas.freitas@stf.jus.br

Seleção de Modelos Discretos Baseado em Delineamentos por Conjuntos Ordenados

Vinicius Ricardo Riffel[1]; Cesar Augusto Taconeli[2]; Paulo Justiniano Ribeiro Júnior[3]

Os delineamentos por conjuntos ordenados são métodos de amostragem que, em geral, fornecem melhores inferências do que a amostragem aleatória simples. Este trabalho teve como objetivo verificar se tais delineamentos são mais eficientes em relação à amostragem aleatória simples na seleção de modelos probabílisticos para dados de contagem, e também a perda de informação na seleção dos modelos selecionados em relação ao modelo teórico. Através de um extenso estudo de simulação, verificamos que os delineamentos por conjuntos ordenados, se operam sob ordenação perfeita ou próximo ao cenário de ordenação perfeita, desempenharam melhor que a amostragem aleatória simples na seleção dos modelos e apresentam menor perda de informação na seleção. Também, foi mostrado que conforme o tamanho amostral aumenta, as taxas de seleção corretas aumentam e a perda de informação diminui. O trabalho se mostrou inovador, tendo em visto que há pouca bibliografia envolvendo dados discretos com a utilização de delineamentos por conjuntos ordenados.

Palavras-chave: Critério de Informação de Akaike; Extreme Ranked Set Sampling; Mixture Ranked Set Sampling; Dados de Contagem; Superdispersão; Inflação de Zeros.

[1]Departamento de Estatística, Universidade Federal do Paraná – viniciusriffel@ufpr.br
[2]Departamento de Estatística, Universidade Federal do Paraná – taconeli@ufpr.br
[3]Departamento de Estatística, Universidade Federal do Paraná – paulojus@ufpr.br

CO2 - Ciência de Dados

Modified Shrinkage Discriminant Algorithms for High-Dimensional Heteroscedastic Data Classification

Olawale Awe[1]; Ronaldo Dias[2]

Discriminant analysis is a standard statistical learning tool for modern data analysis. In many practical applications, there are often a large numbers of pre-processed heteroscedastic features. It is well known that the Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are quite sub-optimal for the analysis of high dimensional heteroscedastic data because of the inherent singularity and instability of the within-class scatter matrix. However, shrinkage discriminant analysis (SDA) and its variants often perform better due to their robustness against multicollinearity and heteroscedasticity. In this work, we propose some newly modified dimension reduction and classification methods based on the SDA. The classification performance of SDA and its modification methods were evaluated by applying them to simulated and real data experiments. We show an estimation consistency property of these supervised classification methods, and compare them with a few other competitors. Our results confirm that LDA and QDA are less helpful when the number of features in a data set is moderate. Methods based on shrinkage maximum uncertainty linear discriminant analysis works better for classification problems with moderate dimensions.

Palavras-chave: Discriminant Analysis; High-Dimensional Data; Heteroscedasticity; Shrinkage Methods; Multicollinearity.

[1]Department of Statistics, Institute of Mathematics, Statistics and Scientific Computing (IMECC), University of Campinas, Brazil – olawaleawe@gmail.com

[2]Department of Statistics, Institute of Mathematics, Statistics and Scientific Computing (IMECC), University of Campinas, Brazil – dias@unicamp.br

Use of Big Data for Official Statistics in Latin America and the Caribbean

Andrea Diniz da Silva[1]; Beatriz Menezes Marques de Oliveira[2]; Ísis Gonçalves Peixoto[3]; Lidiane Braga Sales de Souza[4]

In the last two years, the challenges related to the decline in the financing of statistical production and the cooperation of respondents have been exacerbated by the Covid-19 pandemic. This scenario led the National Institutes of Statistics to consider alternative sources of data to replace or complement traditional research. In this context, the use of big data to produce statistics has become promising. To learn the extent of the use of big data for official statistics in Latin America and the Caribbean, the Regional Hub for Big Data in Brazil conducted a study with the national statistical offices (NSO) of the region. The study includes research on the pages of the NSO (web scraping) and a direct consultation to the NSO representatives. The research showed a rather timid use of big data, but the preliminary results of the consultation already reveal a very positive and promising scenario regarding the use of big data from satellite imagery, web scraping and other sources in applications such as the production of price statistics, coverage and land use and migration.

Palavras-chave: Big Data; Official Statistics; Experimental Statistics.

[1]Professor at the National School of Statistical Sciences, ENCE/IBGE – andrea.silva@ibge.gov.br
[2]Master's student at the National School of Statistical Sciences, ENCE/IBGE – biameny@gmail.com
[3]Master's student at the National School of Statistical Sciences, ENCE/IBGE – isis.peixoto@gmail.com
[4]Master's student at the National School of Statistical Sciences, ENCE/IBGE – lidiane.braga2@yahoo.com.br

CO3 - Políticas Públicas

Cálculo do Indicador ODS 11.7.1: Proporção da Área Construída Cidades que é Espaço Público Aberto para Uso de Todos com Base em Dados de Fontes Abertas

Ísis Gonçalves Peixoto[1]; Andrea Diniz da Silva[2]

Os sistemas estatísticos nacionais vêm testemunhando uma crescente necessidade de se buscar fontes alternativas de dados, para compensar as perdas provenientes das limitações impostas à produção de dados por meio das tradicionais pesquisas. Com o avanço da tecnologia, o acesso gratuito a dados mais estruturados passa a estar na rede de formas mais compreensíveis, como é o caso de mapas e imagens de satélite. No presente trabalho, busca-se mostrar o potencial de dados de fontes abertas, como imagens de satélite e mapas, para produzir o indicador 11.7.1, por meio de uma ilustração considerando o bairro de Ipanema, no município do Rio de Janeiro.

Palavras-chave: Objetivos de Desenvolvimento Sustentável, ODS 11, Agenda 2030, Indicador 11.7.1.

[1]Mestranda na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – isis.peixoto@gmail.com
[2]Professora na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – andrea.silva@ibge.gov.br

Impactos da Não Resposta de Item e Imputação Sobre a Variância do Rendimento de Trabalho da PNAD Contínua

Elizabeth Belo Hypólito[1]; Denise Britz do Nascimento Silva[2]

Para garantir a qualidade das estimativas produzidas, os institutos de estatística realizam um rigoroso processo de crítica dos dados coletados por suas pesquisas, tratando erros de não resposta de itens do questionário e valores inconsistentes por meio de imputação. No entanto, muitas vezes, ao calcular a variância dessas estimativas, consideram o desenho amostral como única fonte de variabilidade. O presente estudo tem como principal objetivo fornecer estimativas de variância da média do rendimento efetivo do trabalho principal produzida pela Pesquisa Nacional por Amostra de Domicílios Contínua, separando as parcelas da variância referentes ao desenho amostral e à não resposta de item tratada por meio de imputação. Os dados utilizados são do terceiro trimestre de 2013 e do primeiro de 2018. O método escolhido para a estimativa da variância foi proposto por Beaumont et al. (2010), com alguns ajustes implementados no *software R*. Os resultados mostram que, em 2013, quando a imputação é realizada pelo método do vizinho mais próximo, a estimativa da variância da não resposta é alta para alguns indicadores. Para o período de 2018, quando as taxas de imputação foram muito baixas, os percentuais da variância por não resposta também são reduzidos Esses resultados indicam que as estimativas de variância obtidas ao considerar o desenho amostral como única fonte de variabilidade podem subestimar a variância, especialmente na presença de altas taxas de não resposta e imputação. Além disso, reforçam a necessidade de estudos sobre erros não amostrais em pesquisas brasileiras, como forma de aprimorar o processo.

Palavras-chave: Pesquisa Amostral; Não Resposta de Item; Imputação; Variância Total.

[1]Escola Nacional de Ciências Estatísticas, ENCE/IBGE – denisebritz@gmail.com
[2]Escola Nacional de Ciências Estatísticas, ENCE/IBGE – elizabeth.hypolito@gmail.com

CO3 - Políticas Públicas

Trabalho Precário e Estatísticas Públicas: Uma Abordagem Possível via PNADC

Beatriz Menezes Marques de Oliveira[1]

O trabalho é um dos elementos centrais para se investigar a organização das sociedades. Através do trabalho e da técnica, o ser humano foi capaz de produzir grandes mudanças sobre o espaço geográfico, sobre os modos de vida e até mesmo sobre o tempo social. Como todo fenômeno social, o trabalho sofre modificações ao longo do curso histórico. Dentre as diversas modalidades de trabalho existentes, o trabalho assalariado se constituiu como uma das mais importantes para o surgimento e consolidação do capitalismo. Apesar de se manter até os dias de hoje, o trabalho assalariado sofreu modificações consideráveis a partir do final do século XX. Com o advento da globalização e da Revolução Tecno-Científica-Informacional, o antigo regime de acumulação fordista foi substituído pelo regime de acumulação flexível. Esse novo regime trouxe consequências significativas para o mundo do trabalho nos últimos anos de modo a aumentar a flexibilização e precarização das relações de trabalho. Nesse sentido, torna-se imprescindível a compreensão dessa nova realidade através do uso das Estatísticas Públicas. Sendo assim, o objetivo deste artigo é indicar possíveis caminhos para essa investigação, bem como realizar um breve diagnóstico para o caso brasileiro.

Palavras-chave: Trabalho Precário; Precarização; Flexibilização; Fordismo; PNADC

[1]Mestranda do Programa de Pós-graduação em População, Território e Estatísticas Públicas, ENCE/IBGE – biameny@gmail.com

COMUNICAÇÕES ORAIS

CO3 - Políticas Públicas

Estudo Sobre o Uso de Big Data em Estatísticas Públicas

Andrea Diniz da Silva[1]; Elizabeth Belo Hypolito[2]; Antonia Xavier[3]
Átila Kopplin Chiquito[4]; Lucas Uchoa Moreira Gomes[5]

A necessidade de estatísticas públicas de alta qualidade é cada vez mais premente em nossa sociedade. Em contrapartida, os recursos destinados à sua produção são cada vez mais escassos. Diante desse cenário, pesquisadores e Institutos de Estatística vêm investindo cada vez mais em fontes de dados alternativas, incluindo de *big data*, como é o caso das redes sociais, registros de transações comerciais ou bancárias, imagens de satélite, dados de sensores de tráfego, câmeras de segurança, GPS, sinal de telefones celulares, entre outras. Embora sejam dados com grande volume, velocidade e variedade, apresentam limitações metodológicas e exigem formas inovadoras e econômicas de coleta, processamento e análise, para que produzam estatísticas confiáveis. O objetivo do presente trabalho é contribuir para o avanço da discussão sobre essa temática, analisando documentos extraídos do Google Scholar, cujos títulos contêm o termo "big data" e um outro tópico considerado de relevância para as estatísticas públicas. Os resultados mostram que de 2004 para 2021 a produção anual saltou de 1 documento para 970. Ademais, os temas saúde e meio-ambiente estão presente em cerca de 50% dos documentos encontrados. Adicionalmente, o trabalho apresenta as etapas futuras da pesquisa, que incluem a análise de uma amostra de artigos livres de custo selecionada de 6012 documentos extraídos e a preparação de um banco de metadados, que poderá ser utilizada como referência para o estudo dos limites e das potencialidades do uso de *big data* na produção de estatísticas públicas.

Palavras-chave: Big Data; Estatísticas Públicas; Pesquisa Bibliográfica.

[1]Professora na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – andrea.silva@ibge.gov.br
[2]Professora na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – elizabeth.hypolito@ibge.gov.br
[3]Graduanda na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – antonia2301@outlook.com
[4]Graduando na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – atilakopplin@gmail.com
[5]Graduando na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – lucasgomes.uchoa@gmail.com

CO3 - Políticas Públicas

PNAD-COVID19: Uma Resposta para Produção de Estatística Pública em Tempos de Pandemia

Natália da Silva Castro[1]

A produção de estatísticas públicas foi fortemente impactada pela pandemia da COVID19, amplificando desafios nos diferentes campos envolvidos na sua produção, em um mundo onde a demanda por informações aumenta sobremaneira. Este artigo apresenta uma análise descritiva da PNAD-COVID19 (Pesquisa Nacional por Amostra de Domicílios COVID19), que surgiu como resposta à preocupação de ocorrência de um "apagão estatístico", e tem como objetivo expor resultados no que se refere aos sintomas de síndrome gripal, com vistas às adaptações utilizadas na produção da estatística classificada pelo IBGE (Instituto Brasileiro de Geografia e Estatística) como experimental. O percentual de relatos de sintomas de COVID19 mostra uma trajetória de declínio, com momentos de estabilidade, mas sem elevações. Ao passo que, o percentual de casos e de óbitos por COVID19 do painel Conass mostram progressão nas primeiras semanas e, posteriormente, uma variação entre estabilidade e elevações.

Palavras-chave: PNAD COVID19; Pandemia; Produção de Estatísticas Públicas.

[1]ENCE, Rio de Janeiro – nscastro.ufrrj@gmail.com.br

COMUNICAÇÕES ORAIS

CO4 - Estatística Educacional e Oficial

A Expansão da Educação Superior no Estado do Rio de janeiro e Seus Efeitos: Análise Multivariada para Construção de Índice Municipal

Lidiane Braga Sales de Souza[1]

A expansão da educação superior no Brasil ganhou força com as políticas públicas adotadas a partir de 2000 em todo o território brasileiro, em especial no Estado do Rio de janeiro a expansão não ocorreu de forma equânime em todos os seus municípios. O presente estudo propõe construir um índice municipal fluminense a partir dos efeitos do avanço da educação superior no Estado do Rio de janeiro, através da elaboração de técnica de análise multivariada, formado por vários indicadores que reflitam os impactos da massificação desse ensino educacional em relação ao desenvolvimento regional, considerando as seguintes dimensões: Educação, Trabalho e Renda, Economia. O resultado do estudo norteará o governo do Estado, no processo de melhoria, aperfeiçoamento e/ou novas políticas públicas de educação superior. Os resultados da análise de componentes principais e análise de agrupamento evidenciaram que 32 dos 92 municípios do Rio de janeiro estão com índices municipais muito baixos, ou seja, com uma maior necessidade de desenvolvimento de políticas públicas nessas localidades.

Palavras-chave: Análise de Componentes Principais; Análise de Agrupamento; Educação Superior; Municípios do Rio de Janeiro.

[1]Mestranda na Escola Nacional de Ciências Estatísticas, ENCE/IBGE – lidiane.braga2@yahoo.com.br

CO4 - Estatística Educacional e Oficial

Desafio em Sala de Aula: Melhorar a Reflexão de Conceitos Estatísticos

Marcos Nascimento Magalhães[1]

A estatística tem ocupado nos últimos anos mais espaço no cotidiano da sociedade. A pandemia de COVID-19 tornou conhecidos do grande público vários termos e gráficos estatísticos que são frequentemente mencionados na mídia. Já há algum tempo conteúdos estatísticos fazem parte do currículo de matemática da Educação Básica e, também, em disciplinas específicas de várias carreiras universitárias. Entretanto, apesar de toda essa exposição, ainda permanecem muitas dificuldades na interpretação correta das ideias básicas de estatística. A formação escolar é um fator multiplicador importante para enfrentar o desafio de melhorar a compreensão dessas ideias. Neste artigo, comentamos algumas iniciativas, utilizadas em disciplinas básicas de estatística para Licenciatura em Matemática, que foram realizadas com o objetivo de reforçar entre os estudantes a reflexão crítica de conceitos estatísticos básicos.

Palavras-chave: Atividades; Estatística Básica; Ensino de Estatística.

[1]Departamento de Estatística, IME-USP – marcos@ime.usp.br

A Expansão da Educação Superior no Brasil: Uma Análise Descritiva de sua Evolução por Modalidade e Rede de Ensino

Mariza Rayanne da Silva Pereira[1]; Lidiane Braga Sales de Souza[2]

A materialização da proteção social no Brasil, através da promulgação da Constituição Federal do Brasil em 1988, representou a garantia de diversos direitos sociais, tal como a universalização da educação. No entanto, a expansão da educação superior no Brasil ganhou força com as políticas públicas adotadas a partir de 2000, quando a onda política neoliberal implantada nos países latinos no século XX ocasionou mudanças estruturais, políticas e ideológicas, caracterizadas pela aquisição de novas formas de articulação do capital privado e de novos espaços para a sua circulação, com a expansão em diversos campos de atuação, inclusive nos serviços essenciais, e consequentemente na educação, contudo, em menor magnitude, observou-se também a expansão da rede pública principalmente no interior do país com a modalidade de ensino à distância. Este estudo apresenta uma análise descritiva sobre a expansão da educação superior no Brasil, considerando além dos marcos históricos, o cenário internacional, os seus instrumentos normativos e a elaboração de uma análise descritiva temporal dos principais indicadores fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) no período de 2000 a 2021, com foco na oferta e demanda da educação superior a nível nacional por rede de ensino e modalidade.

Palavras-chave: Educação Brasileira; Educação Superior; Políticas Públicas.

[1]Mestranda na Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro – mariza_una@hotmail.com
[2]Mestranda na Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro – lidiane.braga2@yahoo.com.br

CO4 - Estatística Educacional e Oficial

Estatísticas à Serviço da Sociedade: InfoVis Bahia - Portal de Visualização de Dados Públicos Usando R/Shiny

Rodrigo Barbosa de Cerqueira[1]; José Roberto Santos da Silva[2];
Jonatas Silva do Espírito Santo[3]

O presente trabalho corresponde a um relato de experiência, cujo objetivo é descrever brevemente e apresentar o processo de construção do InfoVis Bahia, portal de visualização de dados sobre a realidade baiana. Construído utilizando ferramentas de código aberto e baixo custo, como R, shiny, PostgreSQL, entre outras, a plataforma compreende a primeira iniciativa do SEIDataLab, laboratório de dados do Governo do Estado da Bahia. Os resultados positivos e a repercussão alcançados evidenciaram a importância da transparência de dados na esfera pública, promovendo uma integração e maior engajamento da sociedade, aproximando governo, academia e cidadãos. Espera-se que a plataforma sirva de inspiração para outras iniciativas, ampliando o uso de técnicas estatísticas a serviço da sociedade.

Palavras-chave: Visualização de Dados; Dashboard; R Shiny; Indicadores.

[1]Mestre em Economia pela Universidade Federal da Bahia (UFBA). Coordenador de Pesquisas Sociopopulacionais na Superintendência de Estudos Econômicos e Sociais da Bahia (SEI) – rodrigobarbosa@sei.ba.gov.br.

[2]Bacharel em Estatística pela Universidade Federal da Bahia (UFBA). Estatístico na Superintendência de Estudos Econômicos e Sociais da Bahia (SEI) - joserobertosilva@sei.ba.gov.br.

[3]Mestre em Estatística pela Universidade Federal de São Carlos (UFSCar) e Doutorando em Ciência da Computação pela Universidade Federal da Bahia (UFBA). Diretor de Pesquisas da Superintendência de Estudos Econômicos e Sociais da Bahia (SEI) – jonatassanto@sei.ba.gov.br.

Os Desafios e as Controvérsias no Uso de Big Data em Estatísticas Oficiais

Marcus André Alves Zimmermann Vieira[1]; Andréa Diniz da Silva[2]

Nos últimos tempos, os institutos nacionais de estatísticas têm enfrentado grandes cortes de orçamento sem grandes reduções de demandas por estatísticas públicas. Neste cenário de redução de custos, o uso de big data como uma fonte de dados alternativa surge como uma opção mais ágil, eficiente e econômica. Contudo, o uso de big data também apresenta desafios e controvérsias que não podem ser desprezadas. Este artigo pretende contribuir com esse debate a partir de uma revisão bibliográfica sobre o tema, a qual faz parte do trabalho de pesquisa para a elaboração da tese de doutorado do autor principal.

Palavras-chave: Estatísticas Oficiais; Big Data; Desafios; Controvérsias.

[1]Doutorando e Mestre em População, Território e Estatísticas Públicas. Escola Nacional de Ciências Estatísticas (ENCE) – marcusazimmermann@gmail.com

[2]Professora e pesquisadora. Escola Nacional de Ciências Estatísticas (ENCE) – andrea.silva@ibge.gov.br

CO4 - Estatística Educacional e Oficial

Um Estudo Sobre a Desigualdade de Renda no Nordeste a Partir de Dados da PNAD Contínua

Marina Oliveira Cunha[1]; Rita de Cássia de Lima Idalino[2];
Antonio Hermes Marques da Silva Júnior[3]

Resumo

Estudos a respeito da distribuição de renda no Brasil tem sido fortemente estimulados pela Pesquisa Nacional por Amostra de Domicílios Contínua (PNADc) em razão de sua periodicidade curta, multiplicidade de quesitos e constância no questionário de renda. A importância destes estudos está relacionada principalmente à elaboração de políticas públicas com base na identificação de disparidades na distribuição de renda entre os grupos que constituem a população. O notável aumento das disparidades entre os mais pobres e os mais ricos sugere que medidas baseadas na comparação de renda dos mais pobres com a média de toda a população podem não ser adequadas. O índice de Gini clássico e os índices de desigualdade econômica relacionados, no entanto, são baseados exatamente nessas comparações. Neste trabalho discutimos e comparamos o índice de Gini com o índice Zenga, este último baseado em comparações das médias de subpopulações mais pobres e mais ricas, independentemente do limiar que possa ser usado para delinear as duas subpopulações. Como aplicação prática, é apresentada uma análise da desigualdade de renda na Região Nordeste a partir da comparação destes dois índices, utilizando os dados da PNAD Contínua para o 4º trimestre de 2021 no Brasil.

Palavras-chave: Distribuição de Renda; Índice de Gini; Índice de Zenga; Região Nordeste; PNAD Contínua.

[1]Curso de Graduação Bacharelado em Estatística – Universidade Federal do Piauí – UFPI. Teresina-PI – marina_oliveira@ufpi.edu.br
[2]Curso de Graduação Bacharelado em Estatística – Universidade Federal do Piauí – UFPI. Teresina-PI – rita@ufpi.edu.br
[3]Departamento de Estatística – Universidade Federal do Rio Grande do Norte – UFRN. Natal-RN – hermes.marques@ufrn.br

2-D Rayleigh ARMA Model for Anomaly Detection in SAR Imagery

Bruna G. Palm[1]; Fábio M. Bayer[2]; Renato J. Cintra[3,4,5]

Two-dimensional (2-D) autoregressive moving average (ARMA) models are commonly applied to describe real-world image data, usually assuming Gaussian or symmetric noise. However, real-world data often present non-Gaussian signals, with asymmetrical distributions and strictly positive values. In particular, SAR images are known to be well characterized by the Rayleigh distribution. In this context, the ARMA model tailored for 2-D Rayleigh-distributed data is introduced—the 2-D RARMA model. A SAR image anomaly detection experiment was performed, resulting in competitive results with the traditional 2-D ARMA models.

Palavras-chave: Anomaly Detection; ARMA Modeling; Rayleigh Distribution; SAR Images; Two-Dimensional Models.

[1]Department of Mathematics and Natural Sciences, Blekinge Institute of Technology, Sweden – bruna.palm@bth.se
[2]Departamento de Estatística and LACESM, Universidade Federal de Santa Maria, Brazil – bayer@ufsm.br
[3]Programa de Pós-graduação em Estatística, Universidade Federal Pernambuco, Brazil – rjdsc@de.ufpe.br
[4]Department of Electrical and Computer Engineering, Florida International University, USA
[5]School of Science and Mathematics, Howard Payne University, Texas, USA

CO5 - Estatística Aplicada

Envelhecimento Populacional, Produtividade do Trabalho e Razão de Dependência Efetiva

Álvaro de Moraes Frota[1]; Miguel Antonio Pinho Bruno[2]; Ana Carolina Soares Bertho[3]

As avaliações pessimistas sobre o peso dos idosos na economia desconsideram que um aumento da produtividade do trabalho poderia tornar viável à população em idade ativa sustentar as populações idosa e jovem ao longo do processo de envelhecimento populacional. Segundo as projeções da Divisão de População da ONU, tivemos no Brasil em 2019 a razão de dependência demográfica mínima (43,3%), sendo projetada para 2050 uma razão de dependência de 61,8%. Com base na definição da razão de dependência efetiva, que introduz o efeito da produtividade na razão de dependência, uma simulação mostra que um crescimento de 0,72% a.a. no PIB per capita até 2050 seria suficiente para manter a razão de dependência efetiva igual aos de 43,3% de 2019. O exercício realizado permite concluir que uma estratégia de desenvolvimento que resulte neste aumento da produtividade, com distribuição de renda, contrabalançaria o aumento do percentual de idosos no período.

Palavras-chave: Razão de Dependência Demográfica; Produtividade do Trabalho; Razão de Dependência Efetiva; Envelhecimento Populacional.

[1]IBGE, Rio de Janeiro – alvarofrota@gmail.com
[2]ENCE, Rio de Janeiro – miguel.pbruno@gmail.com
[3]ENCE, Rio de Janeiro – carolbertho@gmail.com

Detection of Outliers in Proportional Data. Case Study: Disabled People Data

Paulo Tadeu Meira e Silva de Oliveira[1]

Atypical observations or outliers are almost always present in any dataset due to factors such as storage error or being different from others. Disabled people are considered to be those who have physical, hearing, intellectual or sensory disability, factors that, in interactions with different barriers, can obstruct their full and effective participation in society like other people. According to the IBGE in the 2010 Demographic Census, there were 45.6 million people with disabilities in Brazil distributed in different municipalities, justified by the fact that the level of service varies according to infrastructure and availability of resources in different locations. Data sets from the 2010 Demographic Census Sample carried out by the IBGE were considered, aggregated by municipality considering variables related to disability, identification, education, family, work, income, housing conditions, occupation, basic improvements, other goods and quality of life and UNDP data from all Brazilian municipalities. Compositional data are those that establish the relative information, they are parts of a whole, the sum of these data in each line is a constant, in such a way that it represents 100%. In this work, a comparative study of univariate, bivariate and multivariate outliers was applied to compositional data.

Palavras-chave: Outliers; Compositional Data; Disabled People; 2010 Demographic Census; UNDP.

[1]IEA-USP, São Paulo, SP – poliver@usp.br

CO5 - Estatística Aplicada

A Spatial Autocorrelation Index for Symbolic Interval Data

Wanessa W.L. Freitas[1]; Renata M.C.R. de Souza[2];
Getúlio J.A. Amaral[3]; Fernanda De Bastiani[4]

This work proposes to extent the Moran's spatial autocorrelation index to symbolic interval data. Symbolic data analysis is a domain of research and application related to the areas of machine learning and statistics that provide tools to describe objects through intervals, histograms or lists of categories. Spatially correlated data are geospatial data with spatial autocorrelation, and the variability that comes from each region and neighborhood may be better expressed by intervals. In the context, our index considers the variability present in the interval variable and the variability present in geographical information. An application on Covid-19 interval data illustrates the usefulness of the proposed approach.

Palavras-chave: Interval Data; Moran's Index; Spatial Analysis; Symbolic Data Analysis; Spatial Variability.

[1]Universidade Federal de Pernambuco, Centro de Informática, Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, 50.740-560 Recife, PE, Brazil – wwlf@cin.ufpe.br
[2]Universidade Federal de Pernambuco, Centro de Informática, Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, 50.740-560 Recife, PE, Brazil – rmcrs@cin.ufpe.br
[3]Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Departamento de Estatística, Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, 50740-560, Recife, PE, Brazil – gjaa@de.ufpe.br
[4]Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Departamento de Estatística, Av. Jornalista Aníbal Fernandes, s/n – Cidade Universitária, 50740-560, Recife, PE, Brazil – debastiani@de.ufpe.br

Estimativas Trimestrais da Taxa de Pobreza com a PNAD Contínua Usando Imputação Múltipla

Guilherme A. P. Jacob[1]; Pedro Luis do Nascimento Silva[2]

A PNADC usa um desenho de painel rotativo, uma estratégia que permite análises longitudinais curtas, estimação de fluxos entre trimestres e incorporar módulos específicos nos questionários em determinada visita ou trimestre do ano. Este é o caso das informações sobre rendas de outras fontes, coletadas nas 1ª e 5ª visitas ao domicílio. A renda domiciliar é a soma das rendas de outras fontes e das rendas de todos os trabalhos dos moradores de um domicílio. Portanto, a renda total de um domicílio só é conhecida para 40% da amostra planejada de cada trimestre. Raghunathan e Grizzle (1995) descrevem esta situação como um problema de dados ausentes e aplicam a técnica de imputação múltipla. Assim, este artigo apresenta uma aplicação deste método para aumentar a precisão das estimativas trimestrais de pobreza para Brasil e UFs. Em média, o ganho de eficiência foi de aproximadamente 25% ao longo dos trimestres de 2012 a 2019, apresentando ganhos mais discretos nos trimestres de 2020. Além do ganho de precisão, há um efeito de suavização das séries temporais, especialmente em UFs menos populosas.

Palavras-chave: Pobreza; Pesquisa Domiciliar; Painéis Rotativos; Amostragem Complexa; Imputação Múltipla.

[1]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro – guilhermejacob91@gmail.com
[2]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro – pedronsilva@gmail.com

CO6 - Estatística Longitudinal e Sobrevivência

Metodologia para Pesquisas com Egressos: o Caso do Senac

Antônio Etevaldo Teixeira Júnior[1]; Hyago Souza Nascimento[2]

Este trabalho tem como principal objetivo contribuir com o debate acerca de metodologias utilizadas em pesquisas, principalmente aquelas que têm egressos como unidade de pesquisa. Como cenário é apresentado o caso da Pesquisa Nacional de Avaliação dos Egressos do Senac (PNAES). O instrumento de coleta dos dados da pesquisa deixou de ser aplicado por telefone para ter sua aplicação realizada por meio de um questionário estruturado implementado em ambiente web, o que permitiu que a operação de coleta passasse a ser totalmente realizada no âmbito da Instituição. Além disso, a pesquisa deixou de ser realizada por amostragem para ser feita de forma censitária. Como em qualquer pesquisa, na PNAES verifica-se a ocorrência de não resposta, fenômeno que tem a capacidade de introduzir vieses nas estimativas. Para reduzir essa possibilidade, o tratamento da não resposta em pesquisas por amostragem tem se tornado uma prática cada vez mais comum. Entretanto, essa prática não tem sido aplicada em pesquisas censitárias. A principal inovação deste trabalho é tratar a não resposta em uma pesquisa censitária, de forma a reduzir possíveis vieses em suas estimativas. O tratamento da não resposta foi realizado em duas etapas, na primeira foram utilizados os modelos de propensão de resposta, e na segunda foi aplicado um método de calibração denominado *raking*. Apesar das baixas taxas de resposta verificadas na pesquisa, o tratamento da não resposta pode ser considerado exitoso, em razão da baixa variabilidade verificada na maioria das estimativas em âmbito nacional e as variáveis consideradas nos procedimentos de modelagem.

Palavras-chave: Não Resposta; Propensão de Resposta; Egressos.

[1]Senac – Departamento Nacional, Rio de Janeiro – antonio.junior@senac.br
[2]Senac – Departamento Nacional, Rio de Janeiro – hyago.nascimento@senac.br

Identifying Implausible Values in Longitudinal Big Data: An Example Applied to Child Anthropometric Data From the Brazilian Food and Nutrition Surveillance System

Juliana Freitas de Mello e Silva[1]; Natanael de Jesus Silva[1,2]; Thaís Rangel Bousquet Carrilho[3]; Elizabete de Jesus Pinto[1,4]; Gustavo Velasquez-Melendez[5]; Rosemeire Leovigildo Fiaccone[1,6]; Enny Santos da Paixão[8]; Gilberto Kac[3]; Rita de Cássia Ribeiro-Silva[1,7]; Maurício Lima Barreto [1,9]

Surveillance systems are important to keep track of nutritional characteristics of children growth. In Brazil, the Food and Nutrition Surveillance System (SISVAN) monitors information such as length/height and weight of children who attended public health services. Such measures may be computed with errors that can become biologically implausible values (BIVs), possibly due to the seemingly trivial nature of these data. As a result, the database will need to undergo a cleaning process to remove the mentioned values, otherwise, a statistical analysis can be compromised. Furthermore, the methodological process of identifying BIVs in longitudinal anthropometric data has been recently studied but works with large datasets are scarce. The structure of longitudinal databases like SISVAN allows one to study not only population outliers (POs) but also the so-called longitudinal outliers (LOs) as well. In the latter, the individual's trajectory helps identify values that are too far from the overall and the individual specific longitudinal information. Our goal with this work is to shed a light on the data cleaning process of huge data bases focusing on the identification of POs and LOs.

Palavras-chave: Longitudinal Anthropometric Data, Longitudinal Outliers, Mixed Effects Model with Splines, Growth Curves, Weight Data, Height Data.

[1]Centre for Data and Knowledge Integration for Health, Gonçalo Moniz Institute, Oswaldo Cruz Foundation, Salvador, BA, Brazil
[2]ISGlobal, Hospital Clinic. Universitat de Barcelona, Barcelona, Spain
[3]Nutritional Epidemiology Observatory, Josué de Castro Nutrition Institute, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
[4]Federaral University of Recôncavo da Bahia, Santo Antônio de Jesus, BA, Brazil
[5]Department of Maternal and Child Nursing and Public Health, Nursing School, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil
[6]Institute of Mathematics and Statistics, Federal University of Bahia, Salvador, BA, Brazil
[8]London School of Hygiene & Tropical Medicine, London, Reino Unido
[7]School of Nutrition, Federal University of Bahia, Salvador, BA, Brazil
[9]Institute of Collective Health, Federal University of Bahia, Salvador, BA, Brazil

CO6 - Estatística Longitudinal e Sobrevivência

Survival Models Induced by Zero-Modified Power Series Discrete Frailty: Application with a Melanoma Dataset

Vera Tomazella[1]; Katy C. Molina[2]

Survival models with a frailty term are presented as an extension of Cox's proportional hazard model, in which a random effect is introduced in the hazard function in a multiplicative form with the aim of modeling the unobserved heterogeneity in the population. Candidates for the frailty distribution are assumed to be continuous and non-negative. However, this assumption may not be true in some situations. In this paper, we consider a discretely-distributed frailty model that allows units with zero frailty, that is, it can be interpreted as having long-term survivors. We propose a new discrete frailty-induced survival model with a Zero-Modified Power Series family, which can be zero-inflated or zero-deflated depending on the parameter value. Parameter estimation was obtained using the maximum likelihood method, and the performance of the proposed models was performed by Monte Carlo simulation studies. Finally, the applicability of the proposed models was illustrated with a real melanoma cancer dataset.

Palavras-chave: Discrete Frailty Models; Long-term Model; Zero-Modified Power Series Distributions; Zero Frailty; Melanoma.

[1]DEs-UFSCar-São Carlos-SP – veratomazella@gmail.com
[2]PIPGEs/UFSCar-USP, São Carlos-SP – rocio.cm@usp.br

Causal Mediation for Survival Data: A Unifying Approach via GLM

Marcelo M. Taddeo[1]; Leila D. Amorim[2]

Mediation analysis has been receiving much attention from the scientific community in the last years, mainly due to its ability to disentangle causal pathways from exposures to outcomes. Particularly, causal mediation analysis for time-to-event outcomes has been widely discussed using accelerated failures times, Cox and Aalen models, with continuous or binary mediator. We derive general expressions for the Natural Direct Effect and Natural Indirect Effect for the time-to-event outcome when the mediator is modeled using generalized linear models, which includes existing procedures as particular cases. We also define a responsiveness measure to assess the variations in continuous exposures in the presence of mediation. We consider a community-based prospective cohort study that investigates the mediation of hepatitis B in the relationship between hepatitis C and liver cancer. We fit different models as well as distinct distributions and link functions associated to the mediator. We also notice that estimation of NDE and NIE using different models leads to non-contradictory conclusions despite their effect scales. The survival models provide a compelling framework that is appropriate to answer many research questions involving causal mediation analysis. The extensions through GLMs for the mediator may encompass a broad field of medical research, allowing the often necessary control for confounding.

Palavras-chave: Mediation; Causal Inference; Survival Analysis; Generalized Linear Models.

[1]Departamento de Estatística, IME/UFBA – marcelo.magalhaes@ufba.br
[2]Departamento de Estatística, IME/UFBA – leiladen@ufba.br

CO6 - Estatística Longitudinal e Sobrevivência

Unobserved Heterogeneity for Multiple Repairable Systems Under ARA and ARI Classes of Imperfect Repair

Éder Silva de Brito[1]; Vera Lúcia Damasceno Tomazella[2]; Paulo Henrique Ferreira[3]; Francisco Louzada Neto[4]; Oilson Alberto Gonzatto Junior[5]

In repairable systems, different types of repair can be performed after the occurrence of each failure, which can impact the system reliability over time. Furthermore, due to the nature of the recurrence of events inherent to this type of systems, it is not reasonable to ignore the possibility of the existence of dependence between the failure times of each system and/or the unobserved heterogeneity (or frailty) between multiple systems analyzed together. Thus, the main purpose of this work is to present two models of parametric frailty shared between multiple repairable systems under an imperfect repair. We consider the both classes ARA (arithmetic reduction of age) and ARI (arithmetic reduction of intensity) of the imperfect repair family model and any number of failure memory. It is assumed that the failure intensity function of the model follows a Power Law Process and that the parametric frailty terms of all systems follow the same Gamma distribution. The frequentist approach is used to construct the likelihood function of the model and numerical methods are suggested to obtain the maximum likelihood estimators and their respective confidence intervals. Finally, the presented procedures are applied to a real data set known in the literature.

Palavras-chave: Repairable Systems; Imperfect Repair; Frailty Model; Unobserved Heterogeneity; Power Law Process.

[1]Programa de Pós-Graduação em Estatística (USP/UFSCar), São Carlos/SP – eder.brito@usp.br
[2]Departamento de Estatística (UFSCar), São Carlos/SP – vera@ufscar.br
[3]Departamento de Estatística (UFBA), Salvador/BA – paulohenri@ufba.br
[4]Instituto de Ciências Matemáticas e de Computação (USP), São Carlos/SP – louzada@icmc.usp.br
[5]Departamento de Ingeniería en Informática y Ciencias de la Computación (UDA), Atacama/Chile – oilson.agjr@gmail.com

Custo mínimo de Energia no Transporte de Sensores em uma Rede Bicolor Via Diferença Absoluta Esperada Entre Processos de Poisson Distintos

Cira Etheowalda Guevara Otiniano[1]; Adolfo Manoel Dias da Silva[2]

Neste trabalho, uma fórmula analítica fechada para a diferença absoluta esperada entre dois processos de Poisson independentes com taxas de chegada $\lambda_1 > 0$ e $\lambda_2 > 0$ e respectivos tempos de chegada $X_1, X_2, \cdots$ e $Y_1, Y_2, \cdots$ é determinada ao usarmos a função H de Fox, Fox (1962). Com isto, generalizamos os trabalhos de Kapelko (2020) e e Kranakis (2014). Ao considerar que um par de sensores de duas cores $\{X_k, Y_j\}$ são inicialmente colocados de acordo com os processos descritos, o custo de transporte que minimiza o consumo de energia é dado pela soma das diferenças absolutas esperadas entre os dois processos. Aqui, um intervalo exato para o custo de transporte é obtido. Além disso, mostramos que o custo amostral é um estimador fortemente consistente e imparcial do custo teórico de transporte. A consistência do custo da amostra é ilustrada com experimentos de Monte Carlo e com algumas ilustrações gráficas.

Palavras-chave: Processos de Poisson; Funções Gamma; Transporte de Sensores.

[1]Departamento de Estatística, Universidade de Brasília, Brasil – ciragotiniano@gmail.com
[2]Departamento de Estatística, Universidade de Brasília, Brasil – adolfomanoel@hotmail.com

CO7 - GLM e Processos Estocásticos

Recent theoretical results about Hilbert space embeddings of probabilities.

Jean Carlo Guella[1];

We review some recent theoretical results about Hilbert space embeddings of probabilities, like the fact that Gaussian kernels on Hilbert spaces define an inner product in the space of measures with bounded variation and that the standard metric in Hilbert and real/complex hyperbolic spaces are of the strong negative type. We also present the concept of positive definite independent kernels, which generalizes the concepts of Hilbert Schmidt Independence Criterion and Distance Covariance, provides a metric in the space of couplings, and are related to Bernstein functions with 2 variables.

Palavras-chave: Kernel Methods; Distance Covariance; Independence Tests; Hilbert Schmidt Independence Criterion.

[1]Unicamp-IMECC, Campinas - jcguella@unicamp.br

Count Data and Underdispersion: Models, Software and Applications

Eduardo E. Ribeiro Jr.[1]; Clarice Garcia Borges Demétrio[2]; John Hinde[3]

The baseline distribution for the analysis of count data is the Poisson distribution. However, the assumption of equal mean and variance is often unreasonable and there has been a large body of work on the occurrence and modelling of overdispersion, where the variance is greater than the mean. The topic of underdispersion, where the variability is smaller than the mean, is less widely studied, although of growing interest in recent years. Initially, we will look at possible mechanisms than can lead to underdispersed counts. These include specific forms of dependence in the data, such as a count data process with particular types of non-exponential inter-event times, aggregated dependent data, other features of the data collection process such as recording maxima, or minima, of independently observed counts, and situations where competition is present. We will then survey count data models that can explicitly accommodate underdispersion, including Poisson-Tweedie, COM-Poisson and some of the other specific and modified Poisson distributions that have recently been proposed. We will discuss R packages for the various models for underdispersed count data, illustrated with one, or more, real data examples. We will include a consideration of inferential aspects, such as estimation and testing, and computational issues associated with various approaches. The idea that count data may be over- or under-dispersed is rather simplistic and in more complex situations we may have both aspects in different subsets of the data. Here the most useful families of count models are those that can accommodate both under- and over-dispersion, particularly when used with joint models for the mean and dispersion. We will illustrate this using an aquatic toxicity study. We will conclude with a discussion of potential diagnostics, both for model checking and model comparison and highlight issues for further development.

Palavras-chave: Underdispersion; Count Data.

[1]Departamento de Ciências Exatas, USP/ESALQ, Piracicaba, SP
[2]Departamento de Ciências Exatas, USP/ESALQ, Piracicaba, SP – clarice.demetrio@usp.br
[3]3School of Mathematics, Statistics and Applied Mathematics, NUI Galway, Galway, Ireland – john.hinde@nuigalway.ie

CO7 - GLM e Processos Estocásticos

A Modelling Approach for Forecast Data with Reporting Delay

Izabel Nolau[1]; Dani Gamerman[2]; Leonardo S Bastos[3]

Establishing patterns of progression of illnesses and predicting them is challenging because of reporting delay. The lack of timeliness may be due to laboratory confirmation, logistical problems, and infrastructure difficulties. Correcting the available information is crucial to decision-making as issuing warnings to the public and local authorities. We propose a Bayesian hierarchical modeling approach to correct the reporting delays, quantify the associated uncertainty, and make short-term and long-term predictions. We want to apply this methodology to important illness data, such as dengue fever incidence and severe acute respiratory infection data.

Palavras-chave: Bayesian Hierarchical Model; Deporting Delay; Dengue.

[1]Department of Statistical Methods, Federal University of Rio de Janeiro, Brazil – nolau@dme.ufrj.br
[2]Department of Statistical Methods, Federal University of Rio de Janeiro, Brazil – dani@im.ufrj.br
[3]Scientific Computing Program, Oswaldo Cruz Foundation, Brazil – leonardo.bastos@fiocruz.br

Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling

Luis Alejandro Masmela-Caita[1]; Thaís Paiva-Galletti[2]; Marcos Oliveira-Prates[3]

Missing data theory deals with the statistical methods in the occurrence of missing data. Missing data occurs when some values are not stored or observed for variables of interest. However, most of the statistical theory assumes that data is fully observed. An alternative to deal with incomplete databases is to fill in the spaces corresponding to the missing information based on some criteria, this technique is called imputation. We introduce a new imputation methodology for databases with univariate missing patterns based on additional information from fully-observed auxiliary variables. We assume that the non-observed variable is continuous, and that auxiliary variables assist to improve the imputation capacity of the model. In a fully Bayesian framework, our method uses a flexible mixture of multivariate normal distributions to model the response and the auxiliary variables jointly. Under this framework, we use the properties of Gaussian Cluster-Weighted modeling to construct a predictive model to impute the missing values using the information from the covariates. Simulations studies and a real data illustration are presented to show the method imputation capacity under a variety of scenarios and in comparison to other literature methods.

Palavras-chave: Cluster-Weighted Modeling; Gaussian Mixture Models; Imputation Method; Missing Data.

[1]Universidad Distrital F.J.C. Bogotá D.C. – Colombia. – lmasmela@udistrital.edu.co
[2]Universidade Federal de Minas Gerais. Belo Horizonte – MG – Brasil. – thaispaiva@est.ufmg
[3]Universidade Federal de Minas Gerais. Belo Horizonte – MG – Brasil. – marcosop@est.ufmg.br

CO7 - GLM e Processos Estocásticos

Geração de Coordenadas Geográficas Sintéticas para Banco de Dados Confidenciais com Aplicação a Dados de COVID-19 em Montes Claros, MG

Thaís Paiva Galletti[1]; Fernanda Buzza Alves Barros[2]

Com a crescente produção de dados das últimas décadas, um dos principais problemas é a violação da privacidade de indivíduos. O desafio é desenvolver mecanismos que preservem o sigilo dos dados e, ao mesmo tempo, permitam que os dados sejam divulgados e utilizados para análises estatísticas. Os métodos de imputação múltipla para simulação de dados sintéticos têm se mostrado uma alternativa interessante para resolver esse tipo de problema, podendo ser aplicado inclusive para localizações espaciais. O objetivo deste trabalho é propor uma extensão para a metodologia de geração de coordenadas geográficas sintéticas com covariáveis discretas e contínuas, além de aplicar o método para imputação de localizações sintéticas de indivíduos com suspeita de COVID-19 na cidade de Montes Claros, MG.

Palavras-chave: Confidencialidade; Imputação Múltipla; Dados Sintéticos; Dados Espaciais.

[1]Departamento de Estatística, UFMG – thaispaiva@est.ufmg.br
[2]Departamento de Estatística, UFMG – fernandabarros@ufmg.br

Inducing High Spatial Correlation with Randomly Edge-Weighted Neighborhood Graphs

Danna L. Cruz[1]; Rosangela H. Loschi[2]; Renato M. Assunção[3]

Traditional models for areal data assume a hierarchical structure where one of the components is the random effects that spatially correlate the areas. The conditional autoregressive (CAR) model is the most popular distribution to jointly model the prior uncertainty about these spatial random effects. One limitation of the CAR distribution is the inability of producing high correlations between neighboring areas. We propose a robust model for areal data that alleviates this problem. We represent the map by an undirected graph where the nodes are the areas and randomly-weighted edges connect nodes that are neighbors. The model is based on a multivariate Student-t distribution, spatially structured, in which the precision matrix is indirectly built assuming a multivariate distribution for the random edges effects. The edges effects' joint distribution is a spatial multivariate Student-t that induces another t distribution for the areas' spatial effects which inherit its capacity to accommodate outliers and heavy-tail behavior. Most important, it can produce a higher marginal correlation between the spatial effects than the CAR model overcoming one of the main limitations to this model. We fit the proposed model to analyze real cancer maps and compared its performance with several state-of-art competitors. Our proposed model provides better fitting in almost all cases.

Palavras-chave: Bayesian Inference; Graph of Edges; Spatial Autoregression; Student-t Distribution.

[1]Grupo de Investigación Clínica, Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia. Departamento de Estatística, Universidade Federal de Minas Gerais – danna.cruz@urosario.edu.co

[2]Departamento de Estatística, Universidade Federal de Minas Gerais – loschi@est.ufmg.br

[3]Esri Inc., USA, and Departamento de Ciência da Computação, Universidade Federal de Minas Gerais – assuncao@dcc.ufmg.br

CO8 - Estatística Bayesiana

Inference on Missing Locations in Geostatistics Under Preferential Sampling

Gustavo da Silva Ferreira[1]; Dani Gamerman[2]

This paper deals with the inverse problem in Geostatistics in situations where the researcher performs inference about missing locations under a specific type of informative sampling design. More specifically, this paper considers the case where a observation y^* is taken in a unknown location $x^*, x \in D, D \subset R^d$, assuming that a vector of measures $\tilde{\mathbf{y}} = (y_1, \cdots, y_n)'$ was previously observed in a set of locations $\tilde{\mathbf{x}} = (x_1, \cdots, x_n)'$ planned preferentially. Since the sampling design is preferential, the missing location x^* is not independent of its associated measure y^*. In order to perform inference about x^* we will present a methodology based on gaussian approximation for the distribution of the underlying stochastic process given the set of locations $x = (\tilde{\mathbf{y}}, x^*)$.

Palavras-chave: Missing Data; Geostatistics; Preferential Sampling; Point Process.

[1]National School of Statistical Sciences, Brazilian Institute of Geography and Statistics – gustavo.ferreira@ibge.gov.br
[2]Department of Statistical Methods, Mathematical Institute, Federal University of Rio de Janeiro – dani@im.ufrj.br

Handling Categorical Features with Many Levels Using a Product Partition Model

Tulio L. Criscuolo[1]; Renato M. Assunção[2]; Rosangela H. Loschi[3];
Wagner Meira Jr. [4]; Danna Cruz-Reyes [5]

A common difficulty in data analysis is how to handle categorical predictors with a large number of levels or categories. Few proposals have been developed to tackle this important and frequent problem. We introduce a generative model that simultaneously carries out the model fitting and the aggregation of the categorical levels into larger groups. We represent the categorical predictor by a graph where the nodes are the categories and establish a probability distribution over meaningful partitions of this graph. Conditionally on the observed data, we obtain a posterior distribution for the levels aggregation, allowing the inference about the most probable clustering for the categories. Simultaneously, we extract inference about all the other regression model parameters. We compare our and state-of-art methods showing that it has equally good predictive performance and more interpretable results. Our approach balances out accuracy versus interpretability, a current important concern in statistics and machine learning.

Palavras-chave: Categorical Predictors; Clustering Effects; Random Partition; Dimension Reduction.

[1]Departamento de Ciência da Computação, UFMG – tcriscuolo@gmail.com
[2]Departamento de Ciência da Computação, UFMG – assuncao@dcc.ufmg.br
[3]Departamento de Estatística-UFMG – loschi@est.ufmg.br
[4]Departamento de Ciência da Computação, UFMG – meira@dcc.ufmg.br
[5]Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Colombia – dcruzreyes@gmail.com

CO8 - Estatística Bayesiana

Impacto do Desbalanceamento Amostral em Reconstruções Filogeográficas Discretas

Felipe Lazzari Vargas[1]; Gabriela Betella Cybis[2]

Métodos estatísticos para filogeografia ajudam substancialmente a aumentar nossa compreensão do ritmo e dos meios pelos quais os diferentes organismos, particularmente vírus como HIV, dengue, gripe e SARS-COV2 se dispersam geograficamente. Tais modelos assumem que a variável geográfica evolui de acordo com um processo estocástico sobre a árvore filogenética. Contudo a inferência desses modelos geralmente depende de suposições implícitas relacionadas ao design de amostragem espacial. Embora os locais amostrados transmitam sinais sobre os processos que moldam a biodiversidade espacial, quando a amostragem espacial é viesada isso potencialmente leva a estimativas tendenciosas de parâmetros. Nesse trabalho será realizada uma revisão bibliográfica das abordagens mais recentes para a correção dessas deficiências, tanto para o caso em que a variável espacial é tratada como contínua, quanto quando ela é tratada como discreta. Além disso, apresentamos estudos de simulação, no software R, para a caracterização do impacto do desbalanceamento amostral na inferência filogeográfica.

Palavras-chave: Filogrografia; Método de Monte Carlo; Processo de Coalescência.

[1]Aluno de mestrado no Programa de Pós-graduação em Estatística na Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre – felipelvargas@hotmail.com

[2]Professora no Programa de Pós-graduação em Estatística na Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre – gcybis@yahoo.com.br

Bayesian Wavelet Estimation of Dynamic Weights in Gaussian Mixture Models

Flávia Castro Motta[1]; Michel Helcias Montoril[2]

In this work we propose a new Bayesian method to estimate the weight of a Gaussian mixture model in a dynamic context. The estimation of the dynamic weight is performed using wavelet bases. The full conditional posteriors of the component parameters are generated along with the dynamic weights through an efficient MCMC algorithm. The performance of the proposed method is evaluated through simulated and real data sets, where promising results were observed.

Palavras-chave: Mixture Problem; Nonparametric Regression; Wavelets; Wavelet Mixture Priors; Wavelet Empirical Bayes.

[1]Departamento de Estatística – UFSCar, São Carlos – flavia.motta@estudante.ufscar.br
[2]Departamento de Estatística – UFSCar, São Carlos – michel@ufscar.br

CO9 - Modelos de Regressão

Estimation of Nonparametric Regression Models by Wavelets

Pedro A. Morettin[1]; Rogério F. Porto[2]

In this work we survey the estimation of nonparametric regression models using wavelets, under different conditions on the innovations, on the predictor variables, on the function spaces involved and on the regularity conditions imposed. We begin with the seminal works of Donoho and co-authors, in the regular fixed design and independent and identically Gaussian noise and move towards non-regular designs, random designs and correlated errors.

Palavras-chave: Regression Models; Nonparametric; Wavelets.

[1]University of São Paulo, Brazil – pam@ime.usp.br
[2]Bank of Brazil, Brazil – rdporto1@gmail.com

COMUNICAÇÕES ORAIS

CO9 - Modelos de Regressão

A Class of Categorization Methods for Credit Scoring Models

Diego M. B. Silva[1]; Gustavo H. A. Pereira[2]; Tiago M. Magalhães[3]

Credit scoring models are usually developed using logistic regression. For several reasons, professionals of this area frequently categorize the quantitative covariates before using them in the model. In this work, we introduce a class of methods for covariate categorization in regression models for binary response variables. Applications to real data and a Monte Carlo simulation study suggest that one of the methods of this class has a better predictive performance and a smaller computational cost than other methods available in the literature.

Keywords: Risk Analysis; Covariate Categorization; Credit Scoring models; Discretization; Logistic Regression.

[1]Department of Statistics, Federal University of São Carlos, São Carlos-SP-Brazil – diegow00t@gmail.com
[2]Department of Statistics, Federal University of São Carlos, São Carlos-SP-Brazil – gpereira@ufscar.br
[3]Department of Statistics, Federal University of Juiz de Fora, Juiz de Fora-MG-Brazil – tiago.magalhaes@ufjf.edu.br

CO9 - Modelos de Regressão

K Vizinhos Mais Próximos Circular

Maicon Facco[1]; Fábio M. Bayer[2]

Dados circulares estão presentes em várias áreas da ciência e carecem de métodos estatísticos específicos para seu tratamento. No âmbito de modelos de regressão, a literatura apresenta modelos de regressão paramétricos para dados circulares, os quais fazem suposições de determinadas distribuições de probabilidade circulares para seus ajustes. Por outro lado, na área de aprendizado de máquina, uma abordagem de predição supervisionada para dados contínuos envolve modelos de regressão não paramétricos, os quais podem não ser adequados para situações em que a variável de interesse é circular. Neste contexto, o presente trabalho propõem um modelo de aprendizado de máquina para predição de dados circulares, o qual é denominado k vizinhos mais próximos circular.

Palavras-chave: Aprendizado de Máquina; Dados Angulares; Estatística Circular; Modelos Preditivos; Regressão Não Paramétrica.

[1]Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Maria – maicon_facco@yahoo.com.br
[2]Departamento de Estatística, Universidade Federal de Santa Maria – bayer@ufsm.br

Hierarchical and Multivariate Regression Models to Fit Correlated Asymmetric Positive Continuous Outcomes

Lizandra C. Fabio[1]; Francisco J. A. Cysneiros[2]; Gilberto A. Paula[3]; Jalmar M. F. Carrasco[4]

In the extant literature, hierarchical models typically assume a flexible distribution for the random-effects. The random-effects approach has been used in the inferential procedure of the generalized linear mixed models . In this paper, we propose a random intercept gamma mixed model to fit correlated asymmetric positive continuous outcomes. The generalized log-gamma (GLG) distribution is assumed as an alternative to the normality assumption for the random intercept. Numerical results demonstrate the impact on the maximum likelihood (ML) estimator when the random-effect distribution is misspecified. The extended inverted Dirichlet (EID) distribution is derived from the random intercept gamma-GLG model that leads to the EID regression model by supposing a particular parameter setting of the hierarchical model. Monte Carlo simulation studies are performed to evaluate the asymptotic behavior of the ML estimators from the proposed models. Analysis of diagnostic methods based on quantile residual and COVARATIO statistic are used to assess departures from the EID regression model and identify atypical subjects. Two applications with real data are presented to illustrate the proposed methodology.

Palavras-chave: Generalized Linear Mixed Model; Generalized Log-Gamma Distribution; Misspecation of the Random-Effects; Extended Inverted Dirichlet Model; Diagnostic Analysis.

[1]Departamento de Estatística, Universidade Federal da Bahia (UFBA) – lizandra.fabio@ufba.br
[2]Departamento de Estatística, Universidade Federal de Pernanbuco (UFPE) – cysneiros@de.ufpe.br
[3]Departamento de Estatística, Universidade de São Paulo (USP) – giapaula@ime.usp.br
[4]Departamento de Estatística, Universidade Federal da Bahia (UFBA) – carrasco.jalmar@ufba.br

CO9 - Modelos de Regressão

Unsupervise Bayesian Classification for Mixed Regression Models with Scalar and Functional Covariates

Nancy L. Garcia[1]; Mariana Rodrigues-Motta[2]; Helio S. Migon[3];
Eva Petkova[4]; Thaddeus Tarpey[4]; R. Todd Ogden[5]

We perform unsupervised classification of a scalar response into one of L components of a mixture modeling a scalar response according to a mixture of parametric distributions and the second level modeling the mixture probabilities by means of a generalised linear model with functional and scalar covariates. We use b-spline expansion to reduce the dimensionality and a Bayesian approach for estimating the parameters and providing predictions of the latent classification vector. The Bayesian approach considers the classical Markov Chain Monte Carlo method as well as the Variational Bayes method. We apply our methodology to a data set considering a normal mixture model for the scalar response and functional EEG (Eletroencefalogram) curves from patients with depression in the probability regression.

Palavras-chave: Latent Vector, Functional Covariates, Mixture model, Unsupervised Clustering, Variational Inference.

[1]Department of Statistics, State University of Campinas, São Paulo, Brazil – nancyg@unicamp.br
[2]State University of Campinas, Brazil
[3]Federal University of Rio de Janeiro, Brazil
[4]New York University, USA
[5]Columbia University, USA

COMUNICAÇÕES ORAIS

CO10 - Inferência

Variational Inference for Bayesian Bridge Regression

Carlos Tadeu Pagani Zanini[1]; Helio dos Santos Migon[2]; Ronaldo Dias[3]

We study the implementation of Automatic Differentiation Variational inference (ADVI) for Bayesian inference on regression models with bridge penalization. The bridge approach uses ℓ_α norm, with $\alpha \in (0, +\infty)$ to define a penalization on large values of the regression coefficients, which includes the Lasso ($\alpha = 1$) and ridge ($\alpha = 2$) penalizations as special cases. Full Bayesian inference seamlessly provides joint uncertainty estimates for all model parameters. Although MCMC aproaches are available for bridge regression, it can be slow for large dataset, specially in high dimensions. The ADVI implementation allows the use of small batches of data at each iteration (due to stochastic gradient based algorithms), therefore speeding up computational time in comparison with MCMC. We illustrate the approach on non-parametric regression models with B-splines, although the method works seamlessly for other choices of basis functions. A simulation study shows the main properties of the proposed method.

Palavras-chave: Variational Inference, Bayesian Inference, Bridge Regression, Penalization.

[1]Departamento de Métodos Estatísticos, Instituto de Matemática – UFRJ – carloszanini@dme.ufrj.br
[2]Departamento de Métodos Estatísticos, Instituto de Matemática – UFRJ – migon@dme.ufrj.br
[3]Departamento de Estatística e Computação Científica, Instituto de Matemática Unicamp – dias@unicamp.br

CO10 - Inferência

A Bayesian Approach for Estimating the Parameters of an α-Stable Distribution

Maicon J. Karling[1]; Sílvia R. C. Lopes[2]; Roberto M. de Souza[3]

The lack of closed representations for the density functions of the α-stable distributions, when considering Bayesian inference using Markov Chain Monte Carlo methods, has historically lead to the use of bivariate probability density functions (Buckle, D. J., 1995. "Bayesian inference for stable distributions". *Journal of the American Statistical Association 90 (430)*, 605-613) and Fast Fourier Transforms of their characteristic functions (Lombardi, M. J., 2007. "Bayesian inference for α-stable distributions: a random walk MCMC approach". *Computational Statistics & Data Analysis 51 (5)*, 2688-2700). In the present work, a novel approach resorting on a full power series representation for the probability density functions is considered. The Bayesian analysis for the estimation of all parameters, based on the posterior distributions, is provided for two different parameterization systems for one-dimensional stable distributions. We provide an algorithm that makes use only and exclusively of the power series representation for all the distribution's support. Three goodness-of-fit tests, based on the empirical distribution functions are included, as well as two types of loss functions with their respective decision rules to minimize the Bayesian risk. A simulation study and two empirical applications demonstrate the advantages of our methodology presented here.

Palavras-chave: α-Stable Distribution, Parameterization Systems, Bayesian Techniques, Power Series Representations, Goodness-of-fit Tests, Loss Functions, Simulations.

[1]CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia, 23955-6900 – maicon.karling@kaust.edu.sa

[2]Mathematics and Statistics Institute, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil, 91509-900 – silviarc.lopes@gmail.com

[3]Dean's office of Research and Graduate Studies, Federal Technology University of Paraná (UTFPR), Curitiba, PR, Brazil, 80230-000 – rmolinasouza@utfpr.edu.br

Birnbaum-Saunders Semi-Parametric Additive Modelling: Estimation, Smoothing, Diagnostics, and Application

Esteban Cárcamo[1]; Carolina Marchant[2]; Ibacache-Pulgar[1]; Víctor Leiva[3]

Inclusion of non-parametric functions enhances the modelling when accommodating non-linear effects of covariates. Semi-parametric models have been successfully used to describe non-linear structures by means of parametric and non-parametric components. In this work, we formulate a semi-parametric additive regression model based on a Birnbaum-Saunders distribution and carry out influence diagnostics for this model. This semi-parametric structure permits us to model the mean and variance simultaneously. We use a back-fitting algorithm to obtain the penalized maximum likelihood estimates by using natural cubic smoothing splines. We derived methods of local influence by calculating the normal curvatures under different perturbation schemes. The obtained results are computationally implemented in the R software so that diverse practitioners will have available this model. Finally, an application of the proposed model with real data from one of the most polluted cities in the world is presented.

Palavras-chave: Local Influence; Penalized Maximum Likelihood Estimators; R Software; Splines; Weighted Back-Fitting Algorithm.

[1]Institute of Statistics, Universidad de Valparaíso, Chile
[2]Faculty of Basic Sciences, Universidad Católica del Maule, Talca, Chile – carolina.marchant.fuentes@gmail.com
[3]School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Chile

CO10 - Inferência

Análise do desempenho acadêmico de alunos de Ciências Exatas e da Terra da Unicamp através de um Modelo de Regressão Linear Misto.

Danielle Ap. F. Carvalho[1]; Rafael P. Maia[2]

A partir de dados socioeconômicos e acadêmicos de 2.408 alunos dos cursos de Ciências Exatas e da Terra, com exceção das Engenharias, que ingressaram na Universidade Estadual de Campinas (Unicamp) entre 2014 e 2018, foi realizada uma análise do desempenho acadêmico considerando o Coeficiente de Rendimento (CR) dos mesmos obtido no último semestre de 2018. O CR dos estudantes foi padronizado por ano e curso de ingresso, possibilitando a comparação entre diferentes turmas. Assim, utilizando como variável resposta o CR padronizado, foi ajustado um modelo de regressão linear misto, onde o curso do estudante foi considerado como efeito aleatório. Como a variável resposta apresentou assimetria à esquerda, foram testadas várias distribuições e concluiu-se que a que melhor representa os dados é a distribuição t assimétrica do tipo 2. Uma vez definidos o modelo e a distribuição da variável resposta, foi realizada a análise da qualidade do ajuste do modelo, através de gráficos quantil-quantil e da aplicação do Teste de Shapiro-Wilk nos resíduos obtidos no ajuste, e a seleção de variáveis e interações, através do processo de *forward* e da aplicação do Teste de Razão de Verossimilhança. O modelo final obtido apontou que alunos que já concluíram outra graduação, não trabalhavam, não cursaram Ensino Médio na rede pública, não se autodeclaram pretos, pardos ou indígenas, cursaram Ensino Médio técnico na rede pública e/ou possuíam menos de 20 anos quando ingressaram em seu curso de graduação possuem maiores coeficientes de rendimento padronizado em relação aos demais, consequentemente obtendo um melhor desempenho acadêmico.

Palavras-chave: Desempenho Acadêmico; Ciências Exatas e da Terra; Dados Socioeconômicos; Modelo de Regressão Linear Misto; Distribuição T Assimétrica do Tipo 2.

[1]Departamento de Estatística, IMECC – Unicamp – d169616@dac.unicamp.br
[2]Departamento de Estatística, IMECC – Unicamp – rpmaia@unicamp.br

Minimum Distance Estimation of Long-Memory Stochastic Duration Models

Mauricio Zevallos[1]

This paper proposes a minimum distance estimator for long-memory stochastic duration models which satisfies a central limit theorem. Distinctive features of the proposed method are: it is easy to calculate and implement, allows fast estimation even for huge data sets, and provides asymptotic standard errors for the estimators. Monte Carlo experiments indicate that the proposed estimator performs very well. The proposed method is illustrated with the estimation of a real-life time series of nearly a million observations.

Palavras-chave: Autocovariance Differences; High Frequency; Intertrade Durations.

[1]Department of Statistics, University of Campinas, SP, Brazil – amadeus@unicamp.br

P1

O Software R como Ferramenta para Auxiliar no Processo de Ensino-Aprendizagem da Análise Combinatória

Adriane Caroline Teixeira Portela[1]; Hugo Henrique Gonsalves dos Santos Oliveira[2]; Denise Nunes Viola[3].

O Software R é utilizado como ferramenta para análises estatísticas, manipulação e visualização de dados. Suas funcionalidades variam desde sua utilização para construção de tabelas ou gráficos até para abordagens mais elaboradas, como modelagem estatística. O R se destaca por ser um software livre e de código aberto, além disto, possui uma crescente comunidade de usuários que contribui com pacotes e bibliotecas, expandindo as suas funcionalidades. Diante deste potencial, foi proposta uma nova ferramenta para auxiliar no entendimento da análise combinatória, um dos conteúdos em que alunos da educação básica e superior tendem apresentar dificuldade no aprendizado, que geralmente está associada a interpretação do enunciado ou na utilização das fórmulas, que são facilmente confundidas. Por outro lado, para os professores fica o desafio de transmitir o conteúdo sem que o mesmo fique massivo ou com uso demasiado de fórmulas, o que acontece na maioria das vezes, em que os conceitos e propriedades são transmitidos automaticamente seguidos de exercícios-padrão. O processo de ensino-aprendizagem pode ser simplificado com o auxílio do R. O objetivo deste trabalho é utilizar o software R como ferramenta para auxiliar no direcionamento e compreensão do uso na análise combinatória, por meio da construção de um pacote com diferentes funções, como por exemplo, uma árvore de decisão que após o usuário responder três perguntas dicotômicas, ela indica se estamos em um caso de permutação, arranjo ou combinação, podendo ser com ou sem repetição. Esta ferramenta tem se apresentado eficiente e satisfatória para melhorar o ensino-aprendizagem da análise combinatória.

Palavras-chave: Software R; Ferramenta; Ensino-Aprendizagem; Análise Combinatória.

[1]Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo/Universidade Federal de São Carlos – adrianeportela@usp.br
[2]Departamento de Ciência da Computação, Universidade Federal da Bahia – hugoholiveira45@gmail.com
[3]Departamento de Estatística, Universidade Federal da Bahia – viola@ufba.com.br

Resposta Binária Longitudinal Usando Ligações Alternativas para Dados Médicos

Alex de la Cruz Huayanay[1]; Jorge L. Bazán[2]; Carlos A. Ribeiro Diniz[3].

Motivados por um conjunto de dados médico sobre sintoma de esquizofrenia, onde se observa a resposta binária desbalanceada, apresentamos uma ampla classe de funções de ligação, denominadas potência e reversa de potência, como alternativa para analisar dados binários longitudinais, principalmente quando estão desbalanceadas como é comum em dados médicos. A estimação Bayesiana usando um procedimento MCMC através do algoritmo No-U-Turn Sampler é proposta. verificações preditivas posteriores, resíduos quantílicos aleatorizados bayesianos e uma medida de influência bayesiana são consideradas para o diagnóstico do modelo. Diferentes modelos são comparados usando critérios de seleção de modelo. Um estudo de simulação é desenvolvido para analisar a sensibilidade a priori da variância do efeito aleatório e avaliar o desempenho do modelo proposto na presença de dados desbalanceados. Por fim, considera-se uma aplicação da metodologia estudada em um conjunto de dados médicos sobre a presença do sintoma de esquizofrenia "transtorno do pensamento", neste conjunto de dados, a presença de sintomas é muito menor do que a ausência, assim mostramos, na prática, a utilidade do uso de funções de ligação alternativas em dados desbalanceados.

Palavras-chave: Ligação assimétrica; Diagnóstico Bayesiano; Resposta Binária; Dados Desbalanceados; Modelo de Efeitos Mistos.

[1]Programa Interinstitucional de Pós-Graduação em Estatística, USP/UFSCar, Brasil – aldehu@usp.br
[2]Departamento de Matemática Aplicada e Estatística, USP, Brazil, Local – jlbazan@icmc.usp.br
[3]Departamento de Estatística, UFSCAR, Brasil – dcad@ufscar.br

P3

Modelo de Cox na Estimação da Sobrevida Global de Mulheres com Câncer de Mama

Tainá Cabalheiro[1]; Alexandra Bautista[1]; Ivaine T. S. Sartor[2]; Fernando M. Obst[2]; Rita Costamilan[3]; Roberta Pozza[2]; Sheila Schuch Ferreira[4]; Silvana Schneider[1]; Juliana Giacomazzi[1,2,5]; Patrícia K. Ziegelmann[1,5].

O câncer de mama (CM) é o tumor mais frequentemente diagnosticado entre as mulheres. No Brasil, especialmente no Rio Grande do Sul, há poucos estudos estimando sobrevida global de mulheres diagnosticadas com CM. Objetivo: utilizar o modelo de riscos proporcionais de Cox na estimação da sobrevida global em 5 anos de mulheres diagnosticadas com CM. Método: estudo de coorte retrospectiva de base secundária utilizando dados do Sistema de Registro Hospitalar de Câncer (SisRHC) das Unidades de Alta Complexidade em Oncologia do Hospital Tacchini (Bento Gonçalves) e do Hospital Geral (Caxias do Sul) linkados com dados do Sistema de Informações sobre Mortalidade (SIM) do RS. Foram avaliadas mulheres acima de 15 anos com diagnóstico de CM entre os anos de 2005 e 2018. Sobrevida foi definida como tempo entre o diagnóstico e a morte ou censura. Modelo de Cox com interação entre faixa etária e estadiamento ao diagnóstico (ED) foi utilizado para estimação da sobrevida. Estudo aprovado no CEP-UFRGS sob parecer no 4.330.186. Resultados: Foram analisadas 3.568 mulheres com idade mediana de 56 anos (IQR 47,4-65,1), sendo 59,2% classificadas no ED-I, 31,1% no ED-II-III, e 9,7% no ED-IV. As sobrevidas estimadas para ED-I foram: 95.0%, 96.0% e 90.0%, nas faixas etárias, < 35, 35-64 e $\neq 65$ anos, respectivamente. Para ED-II-III foram 77%, 83% e 67%. E para ED-IV foram 65%, 49% e 39%. Conclusões: Com o modelo de Cox e os dados do SisRHC vinculados aos do SIM foi possível estimar a sobrevida de mulheres diagnosticadas com CM na serra gaúcha.

Palavras-chave: Modelo de Cox; Sobrevida; Câncer de Mama; Análise de Sobrevivência.

[1]Departamento de Estatística da UFRGS, Porto Alegre, Brasil
[2]Instituto Tacchini de Pesquisa em Saúde e Hospital Tacchini, Tacchini Sistema de Saúde, Bento Gonçalves, Brasil
[3]Hospital Geral - Fundação Universidade de Caxias do Sul, Caxias do Sul, Brasil
[4]Secretaria Estadual de Saúde do RS, Rio Grande do Sul, Brasil
[5]Programa de Pós-Graduação em Epidemiologia da UFRGS, Porto Alegre, Brasil

COMUNICAÇÕES PÔSTERES 1

P4

Uma Comparação das Teorias Abstrata e Topológica para Medidas Aleatórias e Processos Pontuais

Alexandre Reggiolli Teixeira[1].

A teoria das medidas aleatórias e processos pontuais, definidos enquanto processos estocásticos com valores de medida, pode ser desenvolvida com o uso de diferentes formalismos. Historicamente, estes podem ser classificados com base na seguinte dicotomia: métodos abstratos e métodos topológicos. No primeiro caso, as propriedades analíticas e de mensurabilidade dos espaços de medidas, como s-finitude, são o foco, com as características distribucionais, de existência e regularidade dos processos sendo derivadas destas utilizando propriedades finas/intrínsecas destes espaços. No caso topológico, as propriedades dos processos, no mesmo contexto acima, são obtidas a partir de princípios gerais que dependem de particularidades topológicas dos espaços de medidas, como separabilidade, metrizabilidade e outras. Técnicas comuns da abordagem topológica incluem, por exemplo, limites projetivos de medidas. O objetivo primário deste trabalho é, através da utilização de técnicas da teoria da medida e da topologia, explorar as conexões entre estas abordagens, obter resultados gerais de existência, integrabilidade e regularidade de medidas aleatórias que possam ser aplicados em ambos os contextos, e exibir particularidades que dependem de atributos dos espaços trabalhados.

Palavras-chave: Medidas Aleatórias; Processos Pontuais; Teoria da Medida; Limites Projetivos; Integrabilidade de Processos.

[1]Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica (UNICAMP) – a163407@dac.unicamp.br

P5

Análise Bayesiana de Modelos Beta Autoregressivos de Médias Móveis

Aline Foerster Grande[1]; Guilherme Pumi[1]; Gabriela Bettella Cybis[1].

A modelagem de séries temporais tomando valores no intervalo (0, 1), como razões e proporções, é uma área que vem crescendo rapidamente nos últimos anos. Tais modelos são não-Gaussianos por natureza e, embora várias abordagens foram propostas na literatura, uma abordagem que vem se destacando é a abordagem GARMA (do inglês, modelos Au- toregressivos de Médias Móveis generalizados), sistematizada por (Benjamin et al., 2003). A ideia da abordagem GARMA é combinar a simplicidade dos modelos ARMA com a flexibilidade dos modelos lineares generalizados, dentro de uma estrutura de modelos *observation driven* (Cox et al., 1981). Um dos modelos mais utilizados na prática, e que serão o objeto de interesse neste estudo, são os modelos βARMA, introduzidos por Rocha and Cribari-Neto (2009). Nele, a componente aleatória é especificada através de uma distribuição beta enquanto a componente sistemática do modelo segue uma especificação ARMA. Em Rocha and Cribari-Neto (2009), uma abordagem frequentista baseada na verossimilhança é utilizada para a estimação dos parâmetros do modelo. O objetivo deste estudo é descrever em detalhes uma abordagem Bayesiana baseada em métodos MCMC para a estimação dos parâmetros de modelo βARMA. Uma simulação de Monte Carlo considerando modelos βARMA de diversas ordens, sob diversos cenários é conduzida, bem como uma análise de sensibilidade com relação à escolha dos hiperparâmetros das prioris utilizadas.

Palavras-chave: Séries Temporais; Análise Bayesiana; Modelos Não-Gaussianos; Modelos GARMA.

[1]Programa de Pós-Graduação em Estatística, UFRGS

Um Aplicativo em Shiny para Ajuste do Modelo de Regressão Quantílica Chen

Alisson Rosa Pereira[1]; Laís Helen Loose[2].

Os modelos de regressão mais conhecidos geralmente utilizados em ajustes a dados reais são baseados no pressuposto de normalidade. No entanto, esta suposição nem sempre é satisfeita na prática. Como consequência, nos últimos anos, tem crescido o interesse no desenvolvimento e análise de modelos não gaussianos. Nesse sentido, o presente trabalho tem como objetivo desenvolver o modelo de regressão quantílico Chen e apresentar um aplicativo em Shiny para ajuste do modelo proposto. Para isso, reparametrizamos a distribuição Chen em termos do quantil e inserimos uma estrutura de regressão para sua modelagem. Na estimação dos parâmetros utilizamos os estimadores de máxima verossimilhança (EMV) e as inferências de testes de hipóteses são realizadas utilizando propriedades assintóticas dos EMV. Propomos ainda o uso do resíduo quantílico e gráficos para avaliar a adequação do modelo. Para a utilização do modelo disponibilizamos um aplicativo *web* desenvolvido em Shiny, em que o usuário pode fazer upload de um arquivo em formato csv e escolher as covariáveis, a variável resposta, o valor para o quantil e a função de ligação que deseja. Ao especificar as quantidades necessárias e solicitar o ajuste do modelo um quadro resumo é apresentado com as estimativas pontuais, o erro padrão estimado, o valor da estatística de teste, o p-valor e outras medidas de interesse. Ainda, fornecemos os principais gráficos para verificar a adequação do modelo. Em trabalhos futuros buscaremos disponibilizar um pacote em linguagem R para ajuste do modelo proposto.

Palavras-chave: Regressão; Quantil; Estimador de Máxima Verossimilhança; Shiny; Linguagem R.

[1]Acadêmico do curso de Estatística, UFSM – alirpereira887@gmail.com
[2]Departamento de Estatística, UFSM – lais.loose@ufsm.br

P7

Preenchimento de Valores Faltantes em Séries Temporais Utilizando Árvores de Decisão

Alisson Silva Neimaier[1]; Taiane Schaedler Prass[2].

Na literatura existem diversas técnicas para o tratamento de observações faltantes para dados que não são séries temporais. No contexto de séries temporais encontra-se alguns trabalhos focados em modelos lineares da família ARIMA. Entretanto, em geral, os artigos não discutem a validade das metodologias propostas para o caso de um grande volume de dados faltantes. A identificação da ordem do modelo apropriado para utilização das metodologias baseadas em modelos é outro ponto desafiante nesse contexto. Tendo em vista esses fatos, este trabalho aborda uma metodologia para recomposição de séries temporais que não assume um modelo paramétrico para os dados. A abordagem proposta utiliza árvores de decisão, um método de aprendizado de máquina que pode ser utilizado tanto para regressão quanto para classificação. Nesta abordagem os valores conhecidos da série temporal fazem o papel de variável resposta, enquanto que defasagens correspondentes a tais valores são utilizadas como preditores. A árvore selecionada pelo algoritmo de treinamento é então utilizada para prever os valores faltantes na resposta. Para investigar a metodologia proposta, consideramos simulações de Monte Carlo variando o tamanho das séries temporais, os parâmetros dos modelos, a proporção de valores faltantes e os preditores. Para avaliar a qualidade das previsões utilizando esse método, o comparamos com alguns métodos de inputação conhecidos e implementados no R. Os resultados encontrados até o momento são promissores.

Palavras-chave: Valores Faltantes; Árvores de Decisão; Séries Temporais.

[1]PPGEst – Instituto de Matemática e Estatística UFRGS – alissonneimaier@hotmail.com
[2]PPGEst – Instituto de Matemática e Estatística UFRGS – taianeprass@gmail.com

P8

Análise Espaço-Temporal do Uso e Cobertura da Terra em Feira de Santana-BA

Taíze da Silva Sousa[1]; Nilton de Souza Ribas Júnior[2]; Aloisio Machado da Silva Filho[3].

O aumento global dos espaços urbanos e a forma como o homem explora o uso e cobertura da terra tem impulsionado a diminuição das áreas de vegetação, desencadeando uma série de consequências ambientais como a alteração do microclima, principalmente nas áreas urbanas. Nesse sentido, o desenvolvimento de pesquisas que visem compreender como estes fenômenos interagem, podem ajudar na gestão da inter-relação homem/espaço natural. A presente pesquisa tem o objetivo de analisar a tendência temporal e espacial das classes de uso e cobertura da terra no município de Feira de Santana, Bahia, Brasil, entre os anos de 2000 e 2020 e identificar se as mudanças na cobertura vegetal exercem influência na estrutura térmica do município. Para isso, utilizou-se dados de *Land Surface Temperature* (LST) do sensor MODIS e para a análise espacial, mapas de uso e cobertura da terra adquiridos a partir da coleção 6 do projeto MapBiomas. Para estimar a tendência da série temporal aplicou-se o modelo de regressão linear simples com correção de Prais e Winsten (PRAIS; WINSTEN, 1954). Como resultados temos que houve um aumento dos valores de LST de aproximadamente 1°C no valor máximo da temperatura. Os maiores valores de LST foram encontrados na classe de área urbana, já os menores valores em áreas com presença de água e vegetação. Constatou-se que as áreas de vegetação diminuíram, entretanto, o modelo de regressão linear simples com correção de Prais e Winsten não identificou tendência decrescente estatisticamente significativa. Os resultados via modelo de regressão linear apontam que a área urbana e silvicultura apresentaram tendência crescente, por outro lado, as classes de rio/lago e mosaico de agricultura e pastagem possuem tendência decrescente.

Palavras-chave: Vegetação; LST; Modelo de Regressão; Uso e Cobertura da Terra; MODIS.

[1]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Modelagem em Ciências da Terra e do Ambiente, Feira de Santana-BA – taize.sousa04@gmail.com
[2]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Ciências Ambientais, Feira de Santana-BA – niltonribasjr@gmail.com
[3]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Ciências Ambientais, Feira de Santana-BA – aloisioestatistico@uefs.br

P9

Standardized Average Weighted Biallelic Statistic (SAWB): a New Method for Identifying Genetic Correlation Networks

Janaína Pacheco Jaeger[1]; Felipe G. Pinheiro[2]; Silvana Schneider[2]; Eduardo Horta[2]; Gabriela B. Cybis[3]

Genome Wide Association Studies (GWAS) test thousands of genome variants searching for genetic markers associated with characteristics of interest. However, the interest lies not only in testing these variants independently, but also in the interactions between them. Climer et al. (2014) proposed a method that, through the Custom Correlation Coefficient (CCC), computes correlations between pairs of SNPs to build allelic networks, which are subsequently tested between case and control individuals in association studies. Nevertheless, the probability distribution and statistical properties of this coefficient have not been studied. The present study derives statistical properties of the CCC under the null hypothesis of independence between variants of different biallelic loci. In particular, its expected value suggested strong frequency-dependent selection. In order to eliminate this bias, we proposed a new correlation statistic, the Standardized Average Weighted Biallelic Statistic (SAWB), which we denoted by Sij, calculated from the same weight matrix used in the CCC. For Sij, asymptotic normality was demonstrated and a corresponding statistical test was defined. The statistical properties of the CCC and Sij, as well as of their related statistics, were compared by simulation studies and through an application on a database for Attention Deficit Hyperactivity Disorder (ADHD). These studies demonstrated the frequency-dependent selection effects of CCC and corroborated that Sij corrects this bias. Furthermore, the Sij statistic was able to identify pairs of correlated SNPs through a statistical test with controlled Type I Error and more power than the test based on the CCC.

Palavras-chave: SNP networks; CCC statistic; SAWB statistic; GWAS.

[1]Federal Institute of Education, Science and Technology Sul-rio-grandense, RS, Brazil – janainajaeger@ifsul.edu.br.
[2]Department of Statistics, Federal University of Rio Grande do Sul, RS, Brazil
[3]Department of Statistics, Federal University of Rio Grande do Sul, RS, Brazil - gcybis@gmail.com

COMUNICAÇÕES PÔSTERES 1

P10

Algumas Simulações em Modelos de Regressão Quantílica

Andrey Bezerra Sarmento[1]; Maria Regina Madruga[2].

Este trabalho teve o intuito de estudar diferentes tipos de modelos de regressão quantílica e suas propriedades por meio de simulações computacionais, especificamente com o apoio do pacote "quantreg" implementado no software estatístico R. Verificou-se que as estimativas dos parâmetros do modelo quantílico possui diferentes características quando considerados erros homocedásticos ou heterocedásticos. Além disso, as propriedades da regressão quantílica no que diz respeito a sua robustez na presença de outilers, e da equivariância em tranformações monótonas, foram comprovadas nas simulações. Os modelos estudados foram: modelo linear simples com erros independentes e identicamente distribuídos e modelo com duas covariáveis considerando erros independentes e não identicamente distribuídos. Para os modelos considerando erros independentes e identicamente distribuídos, certificou-se que as covariáveis influenciam apenas na localização da distribuição condicional da variável dependente, significando que as retas para cada quantil são diferenciadas apenas pelo intercepto do modelo. Considerando os modelos com os erros independentes e não identicamente distribuídos, as estimativas dos parâmetros são diferentes para cada quantil de ordem τ, $0 < \tau < 1$, ou seja, para cada quantil da distribuição condicional da variável dependente Y, os coeficiente estimados do modelo caracterizam diferentes efeitos das covariáveis na distribuição condicional desta variável.

Palavras-chave: Regressão Quantílica; Simulação; Quantis; Modelos Lineares.

[1]Faculdade de Estatística, UFPA, Brasil – Aluno PIVIC/UFPA – andrey.sarmento@icen.ufpa.br
[2]Faculdade de Estatística, UFPA, Brasil – madruga@ufpa.br

P11

Optimal Plot Size in a Field Experiment with Yellow Passion Fruit

Beatriz Garcia Lopes[1]; Taciana Villela Savian[2]; Glaucia Amorim Faria[3].

Brazil is the largest center of origin of passion fruit, and the country with the greatest predominance of the genus *Passiflora*; therefore, Brazil becomes important in socio-economic issues, since jobs are generated in various sectors around the species. The yellow passion fruit has a strong commercial expression, being the most important among the species of the genus *Passiflora* in Brazil, with a high acidity content. An adequate design of experiments is extremely important and, for that, it is essential to choose the appropriate plot size. Hence, the purpose of this work was to recommend the optimal plot size using the segmented linear model with plateau in an experiment with yellow passion fruit in the field. 100 plants were used and each plant was considered as a basic unit, in which 17 different sizes were simulated with 31 different shapes; the variables analyzed were the length and diameter of the fruits and the thickness of the peel. According to the method used, it is suggested to use 10 plants per plot in field experiments with yellow passion fruit.

Palavras-chave: *Passiflora Edulis* Sims; Blank Test; Coefficient of Variation; Experimental Precision.

[1]Doctoral student in Statistics and Agronomic Experimentation (USP/ESALQ), Piracicaba/SP – beatrizgl@usp.br
[2]Doctor, Exact Sciences department (USP/ESALQ), Piracicaba/SP – tvsavian@usp.br
[3]Doctor, Mathematics department (UNESP/FEIS), Ilha Solteira/SP – glaucia.a.faria@unesp.br

P12

Aplicativo em Shiny para Monitoramento de Anomalias Congênitas no Rio Grande do Sul

Bruno Alano da Silva[1]; Guilherme Rodrigues Boff[2]; Márcia Helena Barbian[3]; Luiza Monteavaro Mariath[4]; Thayne Woycinck Kowalski[4,5]; Fernanda Sales Luiz Vianna[5]; Lavínia Schüler-Faccini[5].

Anomalias congênitas (ACs) são anormalidades estruturais ou funcionais que têm origem antes do nascimento, sendo uma das principais causas de mortalidade infantil no Brasil. Sistemas de vigilância epidemiológica em ACs são importantes para estabelecer políticas de atenção e cuidado à saúde. Nesse sentido, o objetivo deste trabalho é apresentar um aplicativo de acesso livre na web que pode auxiliar pesquisadores e administradores públicos no monitoramento de ACs no estado do Rio Grande do Sul (RS). O aplicativo foi desenvolvido em linguagem de programação R, fazendo-se uso do pacote shiny, a partir do qual é possível criar aplicações web interativas funcionalmente acessíveis. A base de dados utilizada para geração dos resultados requeridos pelo usuário foi obtida através do Sistema de Informações sobre Nascidos Vivos (SINASC) e refere-se a nascimentos no RS entre os anos de 2010 e 2019. Os casos são registrados pelo município de residência da mãe e de acordo com a Classificação Internacional de Doenças (CID-10). Nove grupos de ACs foram considerados: Cardiopatias congênitas, Defeitos de parede abdominal, Defeitos de redução de membros/pé torto/artrogripose/polidactilia, Defeitos de tubo neural, Fendas orofaciais, Hipospadia, Microcefalia, Sexo indefinido e Síndrome de Down. A ferramenta oferece diversas funcionalidades e integra importantes métodos de vigilância epidemiológica: estatísticas descritivas, como número de nascidos vivos, número de nascidos vivos com ACs e prevalência ao nascimento de ACs; análises gráficas; mapas que permitem entender a variação espacial de casos de ACs ao longo do tempo nos municípios ou macrorregiões de saúde do RS; análise da associação espacial entre os municípios no que diz respeito à prevalência de ACs; e detecção de conglomerados espaçotemporais ativos no estado. Assim, espera-se que o aplicativo possa contribuir para as estratégias de vigilância em saúde de ACs no estado do RS, indicando como os números de casos são distribuídos entre os municípios e diferentes regiões de saúde. Essas informações podem colaborar nas políticas de distribuição de recursos para cuidado e atenção à saúde no estado. Este estudo faz parte de um projeto piloto aprovado pelo CEP-HCPA 30886520.9.1001.5327 e financiado pelo convênio OPAS/Ministério da Saúde/Fundação Médica do RS (Projeto 2178-4 SCON2020-00173 - Vigilância e Atenção em Anomalias Congênitas no RS).

Palavras-chave: R; Shiny; Visualização de Dados; Dashboard; Anomalias Congênitas.

[1]Estudante de graduação em Estatística - UFRGS - Porto Alegre – alano.bruno31@gmail.com
[2]Bacharel em Estatística - UFRGS - Porto Alegre – guilherme.rboff@hotmail.com
[3]Programa de Pós Graduação em Estatística - UFRGS - Porto Alegre – mhbarbian@gmail.com
[4]Programa de Pós-Graduação em Genética e Biologia Molecular -UFRGS - Porto Alegre
[5]Centro Universitário CESUCA - Cachoeirinha - Brasil

P13

Aplicação de Análise de Sobrevivência com Dados de Pacientes com Doença Renal Crônica

Carla Patrícia de Carvalho Oliveira[1]; Daisy Santana Ferreira[2]; Liciana Vaz de Arruda Silveira[3].

Em análise de sobrevivência é comumente estudado dados de sobrevivência, observando-se apenas o tempo de ocorrência do evento e suas covariáveis, considerando-se que estas auxiliam e avaliam o desempenho do modelo proposto. Aqui, discutimos a relevância prática dos modelos paramétricos, das técnicas não paramétricas e o modelo de regressão de Cox a um conjunto de dados sobre ocorrência ou não de doença renal crônica em adultos e idosos. Este trabalho consiste em realizar um estudo sobre análise de sobrevivência utilizando métodos paramétricos, técnicas não paramétricas, e o modelo semi-paramétrico de riscos proporcionais de Cox, com a finalidade de selecionar as covariáveis que melhor descrevem o banco de dados obtido no Hospital das Clínicas da Faculdade de Medicina de Botucatu da Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) e o evento de interesse é o tempo até a morte do indivíduo.

Palavras-chave: Modelos Paramétricos; Técnicas Não Paramétricas; Modelo de Regressão de Cox; Análise de Sobrevivência; Doença Renal Crônica.

[1]Doutoranda em Biometria - IBB/UNESP, Brasil – carla.patricia@unesp.br
[2]Mestranda em Biometria - IBB/UNESP, Brasil – daisy.santana@unesp.br
[3]Docente - IBB/UNESP, Brasil – liciana.silveira@unesp.br

P14

Tell Me Who You Follow and I'll Tell You Who You Are: Bayesian Estimate of Feminist Ideology on Twitter

Camila Lainetti de Morais[1]; Márcia D'Elia Branco[2].

A common challenge in the social sciences is to understand the political positioning of a population and its representatives. Measuring and comparing these locations can be a difficult task, requiring a wealth of always-up-to-date data from the people being surveyed, as well as a highly complex understanding of how these data are related. To address this challenge, Barberá (2015) proposes a Bayesian ideal point statistical model that measures ideology across the conservative-liberal political spectrum using latent variables. Barberá's technique uses connections made by users of the social network Twitter as the main source to understand political positions and opinions. The model proposed by Barberá is a Bayesian latent model, called Model of Ideal Points (MIP). Using the MIP, this paper analyzes a set of Brazilian influencers and citizens active on Twitter, measuring their ideological positions on feminism, with the aim of understanding more about feminist and anti-feminist groups, the relationship between them and their possible divisions. The estimates of the influencer's ideal points, indicate that there are two groups, one feminist and one anti-feminist, quite separate and with few common followings. In relation to the citizens, preliminars results indicate more moderate positions than those of public figures.

The second author received financial support from FAPESP´s project 2012/21788-2.

Palavras-chave: Latent Model; Bayesian Inference; Political Science; Twitter's Data; Big-Data.

[1]Universidade de São Paulo, São Paulo, SP – cami.lainetti@gmail.com
[2]Departamento de Estatística, Universidade de São Paulo, São Paulo, SP – mbranco@ime.usp.br

P15

Crime Modeling in São Carlos Using Machine Learning Techniques

Thales de Lima Kobosighawa[1]; Cibele Maria Russo[2]; Luis Gustavo Nonato[3].

Brazil has been dealing with crime problems for a long time. Public security government agencies, along with the military police, are often interested in new ways to prevent the occurrence of any type of crime. Nowadays, much information about crime is available through police reports, as well as urban and infrastructure information from military police research. The aim of this study is to investigate associations between urban and crime information and, using machine learning techniques, to obtain a classification model able to predict the occurrence of a crime in a certain São Carlos city corner from infrastructure information. For this, feature selection, normalization techniques and handling imbalanced data techniques are applied to fit the data to the classification model and then, a Bayesian neural network model was proposed. Visualization of the importance of each urban information in crime prediction is derived graphically. In further studies, alternatives as other regression models can be considered for comparison of performance and results, as well as using time series techniques to identify patterns of crime on São Carlos city corners. In addition, the association between video surveillance cameras and the behavior of the amount of crime over the years are investigated.

Palavras-chave: Feature Selection; Normalization; Imbalanced Data; Classification Model; Bayesian Neural Network.

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil – thales.kobosighawa@usp.br

[2]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil – cibele@icmc.usp.br

[3]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brasil – gnonato@icmc.usp.br

Modelo de Regressão Quantílico com Base na Distribuição Lomax Exponencializada

Cleber Bisognin[1]; Guilherme da Silva Machado[2]; Vanessa Siqueira Peres da Silva[3]

A distribuição Lomax Exponencializada (LE) foi proposta por Abdul-Moniem e Abdel-Hameed em 2012. Por se tratar de uma distribuição bastante flexível e eficaz na análise de dados com suporte nos reais positivos é muito útil para tratar de perdas financeiras devido a catástrofes causadas por vento, tensão de ruptura de fibras de carbono, tempo de falha do pára-brisa de aeronaves, entre outras áreas. O objetivo deste trabalho é propor um modelo de regressão para modelagem dos quantis da distribuição LE. Seja $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^\top$ variáveis aleatórias independentes, onde cada Y_t, para $t = 1, \cdots, n$ possui função densidade de probabilidade reparametrizada LE com quantil μ e parâmetros $(\alpha, \delta, \lambda)$ não negativos, onde α e δ são parâmetros de forma e λ o parâmetro de escala. Considerando essa reparametrização temos a possibilidade de incluir uma estrutura de regressão para modelagem do quantil por meio da relação $g(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ é o vetor de parâmetros da regressão ($\boldsymbol{\beta} \in \mathbb{R}^{k+1}$) e $\boldsymbol{x}_t = (x_{t0}, x_{t1}, \cdots, x_{tk})^\top$ são observações de $k+1$ covariáveis ($k+1 < n$), as quais são supostamente fixas e conhecidas e $g : \mathbb{R}^+ \to \mathbb{R}$ é uma função de ligação duas vezes diferenciável e estritamente monótona. A estimação do vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \delta, \boldsymbol{\beta}^\top)^\top$ do modelo proposto será realizada utilizando os estimadores de máxima verossimilhança (EMV), onde a função de log-verossimilhança de $\boldsymbol{y}$ é dada por $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{t=1}^{n} \ell_t(\alpha, \delta, \mu_t)$, onde $\ell_t(\alpha, \delta, \mu_t) = \log\left\{\frac{\alpha\delta}{\mu_t}\left[\left(1-\tau^{\frac{1}{\alpha}}\right)^{-\frac{1}{\delta}} - 1\right]\right\} + (\alpha - 1)\log\left\{1 - \left[1 + \left(\frac{1}{\mu_t}\left[\left(1-\tau^{\frac{1}{\alpha}}\right)^{-\frac{1}{\delta}} - 1\right]\right) y_t\right]^{-\delta}\right\} - (\delta+1)\log\left\{1 + \frac{y_t}{\mu_t}\left[\left(1-\tau^{\frac{1}{\alpha}}\right)^{-\frac{1}{\delta}} - 1\right]\right\}$ com a reparametrização $\lambda = \frac{1}{\mu_t}\left[\left(1-\tau^{\frac{1}{\alpha}}\right)^{-\frac{1}{\delta}} - 1\right]$ para $\tau \in (0,1)$. Foram realizadas simulações de Monte Carlo para avaliar as propriedades dos EMV. Tais simulações indicaram que os estimadores são assintoticamente consistentes e normalmente distribuídos.

Palavras-chave: Distribuição Lomax Exponencializada; Método Máxima Verossimilhança; Regressão Quantílica; Simulação de Monte Carlo.

[1]Departamento de Estatística, UFSM – cleber.bisognin@ufsm.br
[2]Departamento de Estatística, UFSM – guilhermesv2015@gmail.com
[3]Departamento de Estatística, UFSM – vanessa@ufsm.br

P17

Regressão Linear-Circular para Modelagem de Dados Meteorológicos da Cidade de São João Del Rei – MG

Clodoaldo Teodosio Santana da Silva[1]; Carla Regina Guimarães Brighenti[2]

Dados meteorológicos são importantes pois auxiliam na tomada de decisão para gestores e são um campo fértil para pesquisa. Para análise de dados de umidade relativa do ar, velocidade dos ventos, pressão e temperatura, estatística descritiva, séries temporais e modelos de regressão são os mais utilizados. Além disso, dados circulares como direção dos ventos ou observações que podem ser transformadas em circulares, como dias e meses do ano, têm sido frequentemente pesquisados, pois a análise destes não deve ser realizada da mesma maneira que os dados na reta real. Neste trabalho dados provenientes do INMET referentes a estação meteorológica situada na cidade de São João del Rei-MG, no período de 01/01/2020 a 31/12/2021, foram avaliados a partir da correlação entre variáveis circulares e estimação dos parâmetros do modelo linear-circular para radiação solar, tendo como variáveis explicativas os dias e meses do ano. Para ajustar o modelo foi utilizado o software R. O ajuste foi signicativo apenas para os meses do ano e a covariável umidade foi inserida no modelo. Todos os parâmetros foram signicativos ao nível de signicância de 1%. Para validar o modelo foi realizada a análise dos resíduos, que indicou ajuste satisfatório. Através da estatística circular foi possível vericar os períodos de picos de radiação, e em que meses do ano ocorreram altas radiações.

Palavras-chave: Ventos; Radiação; Correlação; Direção.

[1]Departamento de Ciências Exatas da Universidade Federal dos Vales do Jequitinhonha e Mucuri-UFVJM, Teólo Otoni-MG e Discente do Programa de Pós Graduação em Estatística e Experimentação Agropecuária(PPGEE) da Universidade Federal de Lavras-UFLA, Lavras-MG – teoelania@gmail.com

[2]Departamento de Zootecnia da Universidade Federal de São João del Rei-UFSJ, São João del Rei-MG e Docente do Programa de Pós Graduação em Estatística e Experimentação Agropecuária(PPGEE) da Universidade Federal de Lavras-UFLA, Lavras-MG – carlabrighenti@ufsj.edu.br

P18

The Defective Beta-Gompertz Distribution for Cure Rate Regression Models

Cynthia Arantes Vieira Tojeiro[1]; Vera Lucia Damasceno Tomazella[2]; Heberth Duarte dos Santos[3]

An alternative to the standard mixture model is proposed for modeling data containing cured elements or a cure fraction. This approach is based on the use of defective distributions to estimate the cure fraction as a function of the estimated parameters. Defective distributions model cure rates by changing the usual domain of its parameters in a way that their survival functions converge to a value $p \in (0, 1)$. A new way to generate defective distributions to model cure fractions is proposed. The new way relies on a property derived from the Beta-G family of distributions. We take a special attention when G comes from an defective Gompertz distribution, that is, when we have the defective Beta-Gompertz distribution. We use some simulation studies to show the nite sample convergence of the parameters in the distribution, as well as to compare the proposed model with the standard mixture approach. We use a real cancer related data set to show that the new family can outperform the standard mixture model. A regression approach for these models is also proposed.

Palavras-chave: Cure Fraction; Defective Distributions; Gompertz Distribution; Beta-G Family; Survival Analysis, Regression Models.

[1]Instituto de Matemática e Estatística, UFG, Goiânia-GO – cynthiatojeiro@ufg.br
[2]Departamento de Estatística, UFSCar, São Carlos-SP – vera@ufscar.br
[3]Instituto de Matemática e Estatística, UFG, Goiãnia-GO – hbtsantos@msn.com

P19

Estudo Comparativo de Metodologias para Mapeamento de QTLs em Dados Familiares

Lara Midena João[1]; Daiane Aparecida Zuanetti[1].

O mapeamento de regiões no genoma associadas a traços quantitativos (QTLs) através de marcadores genéticos do tipo SNP tem sido um dos problemas centrais em Genética e Biologia Molecular e vários métodos de detecção e identificação de QTLs tem sido propostos na literatura. Metodologias que analisam amostras em família são geralmente mais complexas do que metodologias que assumem independência entre os indivíduos, pois elas precisam considerar e descrever a associação genética que existe entre indivíduos da mesma família, além de identificar e selecionar os SNPs mais relevantes para explicar a variabilidade do traço quantitativo em estudo. Neste trabalho, apresentamos um estudo comparativo da performance dos principais métodos que tem sido utilizados para esse fim em dados familiares. O desempenho das diferentes metodologias estudadas é comparado via análise dos dados familiares GAW17. O primeiro método já analisado e aplicado, chamado FAMSKAT-RC, consiste em um modelo misto que verifica a relevância dos efeitos aleatórios dos SNPs através de um teste de significância sob a variância associada a eles. Para esse método utilizamos tanto o nível de significância de 5% corrigido por Bonferroni quanto o nível de significância de 5% sem a correção. Até a apresentação do trabalho, duas outras metodologias, fastGWA e Mendel, serão estudadas e comparadas.

Projeto financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) no processo 2021/06131-6.

Palavras-chave: Dados familiares GAW17; Modelos mistos; Seleção de variáveis; Teste de significância.

[1]Departamento de Estatística, Universidade Federal de São Carlos - lara.midena@estudante.ufscar.br

P21

Inferência para a Distribuição Very Flexible Weibull Baseada em Censura Tipo II Progressiva

Daniele S. Baratela Martins Neto[1]; E.S. Brito[2]; P.H. Ferreira[3]; V.L.D. Tomazella[4]; R.S. Ehlers[5]

Neste trabalho, apresentamos métodos inferenciais clássicos e bayesianos baseados em amostras na presença de censura tipo-II progressiva sob a distribuição *Very Flexible Weibull*. Obtemos os estimadores de máxima verossimilhança dos parâmetros do modelo, bem como suas medidas de variação assintótica. Propomos o uso de métodos Monte Carlo em cadeia de Markov para o cálculo das estimativas de Bayes. Um estudo de simulação é realizado para avaliar o desempenho dos estimadores propostos sob diferentes tamanhos de amostras e esquemas de censura tipo-II progressiva. A metodologia é ilustrada por meio de um conjunto de dados reais.

Palavras-chave: Estimação de Bayes; Estimadores de Máxima Verossimilhança; Censura Tipo-II Progressiva; Distribuição *Very Flexible Weibull*.

[1]Departamento de Matemática – UnB, Brasília-DF – danielebaratela@unb.br
[2]PIPGEs – UFSCar/USP, São Carlos-SP – eder.brito@ifg.edu.br
[3]Departamento de Estatística – UFBA, Salvador-BA – paulohenri@ufba.br
[4]Departamento de Estatística – UFSCar, São Carlos-SP – vera@ufscar.br
[5]Instituto de Matemática e Ciência da Computação – USP, São Carlos-SP – ehlers@icmc.usp.br

P22

Métodos de Diagnóstico em Modelos de Regressão com Resposta Inversa Gaussiana Ampliada em Zero

Danilo V. Silva[1]; Gilberto A. Paula[2]

Em muitas situações práticas em que a variável resposta é estritamente positiva pode haver interesse em ampliar o domínio para incluir também o valor zero, resultando em uma mistura de distribuições. Esse é caso, por exemplo, do estudo do valor total pago pela seguradora (incluindo zero) ao segurado em apólices de seguros de automóveis. Há artigos e bibliotecas, por exemplo `gamlss` em `R`, que discutem e ajustam esse tipo de modelo, contudo as análises de diagnóstico têm sido restritas a análises de resíduos. Neste texto desenvolvemos vários procedimentos de diagnóstico em modelos de regressão com resposta inversa Gaussiana ampliada em zero (ZAIG) quando todos os três componentes (localização, dispersão e probabilidade de zero) são modelados. Derivamos um próprio processo iterativo para o ajuste e ilustramos com a análise de uma grande base de dados da área de seguros de automóveis na qual poucas observações amostrais afetam de maneira desproporcional as estimativas e alteram decisões inferências sobre os parâmetros do modelo.

Palavras-chave: Modelos ZAIG; Inversa Gaussina; MLGs Duplos; Métodos de Diagnóstico; GAMLSS.

[1]Departamento de Estatística, Universidade de São Paulo – danilo.silva@ime.usp.br
[2]Departamento de Estatística, Universidade de São Paulo – giapaula@ime.usp.br

P23

Inducing High Spatial Correlation with Randomly Edge-Weighted Neighborhood Graphs.

Danna L. Cruz[1]; Rosangela H. Loschi[2]; Renato M. Assunção[3].

Traditional models for areal data assume a hierarchical structure where one of the components is the random effects that spatially correlate the areas. The conditional autoregressive (CAR) model is the most popular distribution to jointly model the prior uncertainty about these spatial random effects. One limitation of the CAR distribution is the inability of producing high correlations between neighboring areas. We propose a robust model for areal data that alleviates this problem. We represent the map by an undirected graph where the nodes are the areas and randomly-weighted edges connect nodes that are neighbors. The model is based on a multivariate Student-t distribution, spatially structured, in which the precision matrix is indirectly built assuming a multivariate distribution for the random edges effects. The edges effects' joint distribution is a spatial multivariate Student-t that induces another t distribution for the areas' spatial effects which inherit its capacity to accommodate outliers and heavy-tail behavior. Most important, it can produce a higher marginal correlation between the spatial effects than the CAR model overcoming one of the main limitations to this model. We fit the proposed model to analyze real cancer maps and compared its performance with several state-of-art competitors. Our proposed model provides better fitting in almost all cases.

Palavras-chave: Bayesian Inference; Graph of Edges; Spatial Autoregression; Student-t Distribution

[1]Grupo de Investigación Clínica, Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia. Departamento de Estatística, Universidade Federal de Minas Gerais – danna.cruz@urosario.edu.co
[2]Departamento de Estatística, Universidade Federal de Minas Gerais – loschi@est.ufmg.br
[3]Esri Inc., USA, and Departamento de Ciência da Computação, Universidade Federal de Minas Gerais – assuncao@dcc.ufmg.br

P24

Estimation in the Multivariate Regression Model with General Parameterization via a Generalized Entropy Measure

Diego Ramos Canterle[1]; Alexandre Galvão Patriota[2]

This work considers a parameter estimation based on a generalized entropy measure, called nonextensive q-entropy, in the multivariate regression model with general parameterization (MRMGP). The MRMGR is a very large class of regression models that encompasses mixed models, error-in-variables models, nonlinear models, among others. All results obtained from the general model can be readily applied to all its submodels. The estimator based on the nonextensive q-entropy is called the maximum Lq-likelihood estimator (MLqE) and depends on a tuning parameter q, which is chosen conveniently to yield robust inferential results against outliers. We derived all required quantities to attain the parameter estimation and the variance matrix of the MLqE in the MRMGP; some of these quantities are detailed for three particular cases. A scheme to choose the tuning parameter q is considered. In order to evaluate the performance of the MLqE in the MRMGP with and without outliers, we conduct some Monte Carlo simulations. Finally, we consider an application to show the practical utility of our proposal under actual data.

Palavras-chave: General Parameterization; Nonextensive Entropy; Outliers; Regression Model; Robust Estimation.

[1]Programa de Pós-Graduação em Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo/SP – diegocanterle@gmail.com

[2]Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo/SP – patriota.alexandre@gmail.com

P25

Teoria da Resposta ao Item e Testes Adaptativos Computadorizados em Ambientes Coletivos

Dionisio Alves da Silva Neto[1]; Héliton Ribeiro Tavares[2]

A Teoria da Resposta ao Item (TRI) compõe um conjunto de metodologias inferenciais voltadas à construção de indicadores para o desempenho em avaliações educacionais. Nos últimos anos, destacaram-se as abordagens para Testes Adaptativos Computadorizados (TAC ou CAT, sigla em inglês), os quais visam oferecer uma avaliação mais individualizada e/ou eficiente, utilizando um número menor de itens, dado algum critério de parada. Objetivo: Este trabalho visa comparar a abordagem tradicional da TRI com a agilidade de um teste adaptativo, estimando a distribuição do número de itens administrados. Métodos: Para realizar o estudo, foi utilizado um banco de itens simulado contendo 1.000 itens calibrados pelo Modelo Logístico de 3 parâmetros e, então, estimou-se as habilidades de 10.000 respondentes pelo método *Expected A Posteriori* (EAP). Por meio de um código próprio, simulou-se um TAC para cada respondente através do software R, estabelecendo-se o erro padrão (EP) como critério de parada e o Critério da Máxima Informação para a seleção do próximo item. Resultados: Com o TAC foi possível estimar a proficiência estabelecida com apenas 18 itens aplicados, considerando o EP=0,35; enquanto que EP=0,30, passou-se a 32 itens aplicados. Conclusão: O TAC proporciona uma ótima inovação para centros de aplicação e estudantes, ao mensurar de forma equivalente o real traço latente em um processo menos oneroso e, consequentemente, diminuindo o desgaste de uma longa avaliação. Em adição, comprova-se que à medida que mais itens são aplicados, menor é o erro padrão da habilidade, o que torna a predição mais precisa.

Palavras-chave: TRI; Avaliação; Proficiência; Item; Erro Padrão.

[1]Faculdade de Estatística da Universidade Federal do Pará, Belém – PA – dionisio.neto@icen.ufpa.br
[2]Faculdade de Estatística da Universidade Federal do Pará, Belém – PA – heliton@ufpa.br

P26

Estimador de Cores de Olhos para a População do Sul do Brasil: uma Ferramenta Forense de Fenotipagem

Eduardo Ávila Carlos[1]; Eduardo Ibaldo Gonçalves[2]; Alessandro Kahmann[3]; Márcio Dorn[4]; Clarice Sampaio Alho[5]

Variantes genéticas relacionadas com características externas visíveis são de grande relevância para a predição da cor dos olhos de seres humanos. Nesse estudo foram testadas a habilidade de sete SNPs (Single *nucleotide polimorphism*) preverem a cor de olhos em uma população miscigenada do sul do Brasil. Para tanto, foram genotipados 478 observações com diferentes cores de olhos. A partir destas observações, foram estimados modelos de regressão logística para diferentes categorizações de cores de olhos, com o intuito de predizer esta informação através do genótipo de indivíduos. A melhor combinação de SNPs para cada categorização de cor de olhos foi determinada pelo AIC (Critério de Informação de Akaike) e a acurácia média das amostras de teste no 10-fold *cross-validation* foi de: 0,75 quando a variável resposta foi dividida em cinco cores; 0,82 quando a variável resposta foi dividida em três classes e; 0,93 quando a variável resposta foi dividida em duas classes. Este é o primeiro estudo aplicado na população brasileira e produz resultados melhores, quando comparado ao padrão ouro da área.

Palavras-chave: Genótipo; Fenotipagem; AIC; Regressão Logística; Forense.

[1]Instituto Nacional de Ciência e Tecnologia Forense, Porto Alegre – efavilas@yahoo.com
[2]Programa de Pós Graduação em Genética – UFPR, Curitiba – carlosibaldo@gmail.com
[3]Departamento Interdisciplinar – UFRGS, Tramandaí – alessandro.kahmann@ufrgs.br
[4]Instituto de Informática – UFRGS, Porto Alegre – mdorn@inf.ufrgs.br
[5]Escola de Ciências da Saúde e da Vida – PUCRS, Porto Alegre – csalho@pucrs.br

P27

Variograma Espaço-temporal para Modelagem da Quantidade Anual de Dias Sem Chuvas

Elias Silva de Medeiros[1]; Carolina Cristina Bicalho[2]

O objetivo desta pesquisa consistiu em apresentar uma estrutura de variograma espaço-temporal para modelar a quantidade anual de dias sem chuvas (DWR) no Estado da Paraíba, Brasil. O estado é dividido em quatro mesorregiões (Sertão, Borborema, Agreste e Mata Paraibana) com condições climáticas e ambientais distintas, sendo esta região caracterizada por apresentar uma alta variabilidade espacial e temporal da precipitação. Assim, faz-se necessário empregar uma metodologia que seja capaz de modelar a dependência espaço-temporal da distribuição das chuvas na região. O banco de dados utilizado neste estudo foi constituído de 238 estações pluviométricas, em que foi avaliada a DWR durante 27 anos (1994 a 2020). Para modelagem da DWR foi empregada a metodologia geoestatística espaço-temporal, seguido do ajuste do variograma teórico produto-soma generalizado. Os resultados do ajuste da regressão indicam que os efeitos linear e quadrático da longitude foram estatisticamente significativos. Notou-se que 55,64% (R^2) da variabilidade da DWR foi explicada para ajuste da tendência. Removida a tendência, ajustou-se o variograma espaço-temporal produto-soma generalizado. A variabilidade da componente espacial (sill = 272,274) é superior à da componente temporal (sill = 3,176), indicando que a variabilidade da DWR está mais relacionada ao espaço do que ao tempo. Os resultados mostram que existe uma dependência espacial da DWR entre os pontos vizinhos separados por um raio de até 39 km. Por outro lado, a DWR de um local apresenta uma dependência temporal, o que ocorre hoje na região influência nos resultados nos próximos 2,8 anos.

Palavras-chave: Geoestatística Espaço-Temporal; Chuvas; Nordeste; Krigagem.

[1]Faculdade de Ciências Exatas e Tecnologia, UFGD – eliasmedeiros@ufgd.edu.br
[2]Departamento de Matemática, UEMS – carolinabicalho@gmail.com

P28

Construção de um Indicador para Aferir a Percepção Discente Acerca das Condições Didático-Pedagógicas dos Docentes dos Cursos de Graduação da UFSCar

Samantha Navarro Janine[1]; Estela Maris Pereira Bereta[2]; Márcio Luis Lanfredi Viola[3]

O Sistema Nacional de Avaliação da Educação Superior (SINAES) promove a avaliação de instituições, de cursos e de desempenho dos estudantes. Cada instituição promove processos avaliativos internos, os quais contribuem para o aperfeiçoamento dos cursos de graduação e das condições de funcionamento das universidades, permitindo que a Universidade conheça suas "virtudes" e suas "fragilidades", com foco em pontos que podem ser melhorados. A partir do interesse em estudar as "Condições didático-pedagógicas dos docentes", abordadas em um questionário aplicado aos discentes dos cursos de graduação presenciais da UFSCar, foram usadas questões relacionadas à tal dimensão, cujas respostas são dadas em escala Likert. *Este trabalho tem como objetivo construir um indicador usando o método de componentes principais considerando a natureza da escala de resposta das questões.* O procedimento adequado para reduzir a dimensionalidade de variáveis medidas em escala do tipo Likert, é a análise de componentes principais usando o procedimento Prinqual. Este método atribui valores numéricos às categorias de cada variável qualitativa, utilizando o escalonamento ótimo, fazendo com que seja possível executar a ACP nas variáveis transformadas. O valor numérico atribuído a cada variável qualitativa é obtido pelo "método dos mínimos quadrados alternados". Este procedimento iterativo faz com que as quantificações numéricas em cada variável possuam propriedades métricas. O indicador é construído a partir dos autovalores e dos escore de das dimensões associadas as componentes principais. A partir da distribuição da pontuação atribuída por cada indivíduo, pode-se concluir que as condições didático-pedagógicas dos docentes foi bem avaliada pelos discentes.

Palavras-chave: Sistema Nacional de Avaliação da Educação Superior; Análise de Componentes Principais; Procedimento PRINQUAL; Escala Likert.

[1]Departamento de Estatística, UFSCar – samantha.janine@estudante.ufscar.br
[2]Departamento de Estatística, UFSCar – estela@ufscar.br
[3]Departamento de Estatística, UFSCar – lanfredi@ufscar.br

P29

Redes Neurais Estocásticas: Deep Learning Aplicado na Solução de Equações Diferenciais Estocásticas

Felipe Cavalcante do Rosário[1]; Valcir João da Cunha Farias[2]

Os métodos clássicos para resolver equações diferenciais são bastantes efetivos, porém enfrentam graves restrições em situações de alta dimensionalidade e, sobretudo, da aleatoriedade existente em vários problemas, principalmente, devido à problemas com dependência de espaço-temporal. Em contrapartida, algumas técnicas modernas tentam contornar o problema da dimensionalidade, como, por exemplo, aproximando a solução de equações diferenciais estocásticas por meio de uma rede neural. Podemos escrever a solução de uma equação diferencial estocástica como uma função determinística no tempo e do processo de estado. Nesse viés, Raissi (Raissi (2018)) desenvolveu uma técnica, no qual aproxima a função determinística de tempo e espaço por meio de uma rede neural com aprendizagem profunda (Deep Learning Neural Network). Assim, este trabalho têm como objetivo utilizar o procedimento desenvolvido por Raissi (Raissi (2018)) para resolver equações difenrenciais estocástica, sendo feito algumas adaptações, de acordo com os problemas trabalhados. A motivação para este trabalho vem pelo fato de que, as redes neurais estocásticas possuem um papel significativo em estudos nas áreas da inteligência artificial, logo se faz necessário a utilização das mesmas no contexto estatístico. Para ilustrar os resultados, uma análise gráfica após a realização das simulações é proposta com intuito de verificar adequabilidade da solução fornecida pela rede em relação a solução conhecida da equação diferencial estocástica.

Palavras-chave: Equações Diferenciais Estocásticas; Rede Neural; Aprendizagem Profunda; Inteligência Artificial.

[1]Faculdade de Estatística, UFPA – felipe08.cavalcante@gmail.com
[2]Faculdade de Estatística, UFPA – valcir@ufpa.br

P30

A Flexible Hierarchical Quantile Spatial Autoregressive Model

Rafael Cabral Fernandez[1]; Kelly Cristina Mota Gonçalves[2];
João Batista de Morais Pereira[3]

The paper introduces a new class of nested models, named hierarchical extended quantile spatial autoregressive models, which extends the literature standard combination of spatial autoregressive model for areal data with parametric quantile regression by changing the usual Asymmetric Laplace distribution for the random errors to a Generalized Asymmetric Laplace distribution. Besides, the new proposed model can incorporate a hierarchical structure, allowing it to deal with clustered data. Such approach produces a robust and flexible statistical method for modelling the quantiles of areal data distributed in a hierarchically geographical setting. The proposed model is evaluated using a well-known house pricing data and through a simulated study. The hierarchical version is also applied to a real dataset concerning math scores related to the public high schools within the Metropolitan area of Rio de Janeiro, Brazil.

Palavras-chave: Quantile Regression Model; Spatial Autoregressive Model; Generalized Asymmetric Laplace Distribution; Hierarchical Models; Bayesian Inference.

[1]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro – rafaelc@dme.ufrj.br
[2]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro – kelly@dme.ufrj.br
[3]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro – joao@dme.ufrj.br

P31

Sistema de Recomendação Baseado em Modelo de Diagnóstico Cognitivo

Lohan Rodrigues Narcizo Ferreira[1]; Fernanda Tostes Marana[2]; Mariana Curi[3].

Com a introdução da economia online, fortalecida pelo rápido desenvolvimento das redes sociais e plataformas de e-commerce, um novo fenômeno surgiu para ser enfrentado por vendedores e compradores online: a sobrecarga de informação. Como toda plataforma possui mais informações do que seus usuários podem consumir, o excesso de escolhas irrelevantes diminui as taxas de experiência do usuário e dificulta a obtenção de produtos, informações e serviços desejados. Portanto, estratégias como Sistemas de Recomendação visam entender como os itens se relacionam entre si, analisando seus padrões e recomendando aqueles que devem interessar a cada usuário em particular. Este projeto propõe o desenvolvimento de um novo sistema de recomendação utilizando ferramentas que já possuem grande repercussão no meio educacional tais como os Modelos de Diagnóstico Cognitivo (CDMs) e Testes Adaptativos Computadorizados (CAT). Os primeiros testes realizados no CDM DINO indicam o potencial desta migração do modelo para uma nova interpretação quando levado ao contexto de recomendação de mídias de entretenimento, mais especificamente os filmes. Na sequência foram feitas comparações entre um sistema baseado no DINO e um sistema comum da literatura baseado em agrupamento. Os resultados indicaram eficiência similar entre os dois com potencial de junção para resultados melhores.

Palavras-chave: Sistemas de Recomendação, Modelo de Diagnóstico Cognitivo, Teste Adaptativo Computadorizado.

[1]lohanext@gmail.com
[2]Departamento de Matemática Aplicada e Estatística, ICMC USP – fernanda.marana@usp.br
[3]Departamento de Matemática Aplicada e Estatística, ICMC USP – mcuri@icmc.usp.br

P32

Aggregated Functional Data Model Applied on Clustering and Disaggregation of UK Electrical Load Profiles

Gabriel Franco[1]; Camila P. E. Souza[2]; Nancy Lopes Garcia[3]

Understanding electrical energy demand at the consumer level plays an important role in planning the distribution of electrical networks and offering of off-peak tariffs, but observing individual consumption patterns is still expensive. On the other hand, aggregated load curves are normally available at the substation level. The proposed methodology separates substation aggregated loads into estimated mean consumption curves, called typical curves, including information given by explanatory variables. In addition, a model-based clustering approach for substations is proposed based on the similarity of their consumers' typical curves and covariance structures. The methodology is applied to a real substation load monitoring dataset from the United Kingdom and tested in eight simulated scenarios.

Palavras-chave: Blind Source Separation; Functional Aggregated Model; Gaussian Process; Basis Function Expansion.

[1]Department of Statistics, University of Campinas, Brazil – gabrielfranco89@gmail.com
[2]The University of Western Ontario, London, Canada
[3]Department of Statistics, University of Campinas, Brazil

P33

A New Extended Weibull model: Simulations, Regression and Application

Gabriela M. Rodrigues[1]; Roberto Vila[2]; Edwin M. M. Ortega[3]; Gauss M. Cordeiro[4]; Victor Serra[5]

The area of statistics that studies the time until the occurrence of an event of interest is called survival analysis, whose main characteristic of the data is the presence of censored observations. In practical applications, it is common for time to be affected by one or more explanatory variables, and regression models must be used to analyze such effects. In the survival data analysis literature, it is convenient to consider more flexible distributions to capture a wide variety of symmetric, asymmetric and bimodal behaviors with non-monotonic failure rate function. Thus, this work proposes a new distribution called log exponentiated odd log-logistic Weibull (LEOLLW), which has greater flexibility, such as bimodal density, which allows modeling data whose failure rate function takes different forms. Mathematical properties of the new distribution are presented and a new location-scale regression models are defined based on it. Estimates are obtained by the maximum likelihood method and a simulation study verifies their accuracy. We propose the use of modified deviance residuals to verify the adequacy of the model, and some simulations show that its empirical distribution approaches the standard normal for small percentages of censoring. To illustrate the usefulness of the models presented, we carried out an application considering censored data from Japanese-Brazilian emigration. The proposed regression model proved to be adequate for this dataset.

Palavras-chave: Weibull Distribution; Regression Model; Censored Data; Residual Analysis; Simulation Studies.

[1]Department of Exact Sciences, University of Sao Paulo, Piracicaba, SP, Brazil – gabrielar@usp.br
[2]Department of Statistics, University of Brasilia, DF, Brazil – rovig161@gmail.com
[3]Department of Exact Sciences, University of Sao Paulo, Piracicaba, SP, Brazil – edwin@usp.br
[4]Department of Statistics, Federal University of Pernambuco, Recife, PB, Brazil – gausscordeiro@gmail.com
[5]Department of Statistics, University of Brasilia, DF, Brazil – victorserra92@gmail.com

P34

Modelagem Bayesiana da Precipitação Máxima de Petrópolis-RJ e Poços de Caldas-MG

Gilberto Rodrigues Liska[1]; Thales Rangel Ferreira[2]; Sandra Valéria Coelho da Silva[3]; Fabricio Goecking Avelar[4]; Luiz Alberto Beijo[5]

Cidades localizadas em regiões serranas como Petrópolis-RJ e Poços de Caldas-MG frequentemente sofrem danos decorridos de precipitações extremas, sobretudo enchentes e deslisamentos de terra. Visando amenizar ou mesmo solucionar estes problemas, pode-se para estas localidades realizar planejamentos de atividades vulneráveis aos efeitos deste fenômeno por meio da análise e previsão da ocorrência de precipitações extremas. A modelagem desta variável pode ser feita por meio da distribuição generalizada de valores extremos (GEV), sendo que a metodologia Bayesiana tem apresentado bons resultados na estimação de seus parâmetros. Portanto, este trabalho teve como objetivo ajustar a distribuição GEV às séries históricas de precipitação máxima de Petrópolis e Poços de Caldas, e avaliar diferentes estruturas de distribuições a priori, informativas e não informativas, na predição da precipitação máxima esperada para diferentes tempos de retorno. Foram analisadas a acurácia e a precisão a fim de avaliar as previsões obtidas com as informações advindas das precipitações máximas de diferentes localidades para elicitação da distribuição a priori. A obtenção das distribuições marginais a posteriori foi realizada usando-se o método Monte Carlo via cadeias de Markov. A utilização da distribuição a priori informativa fundamentada nos dados de Poços de Caldas foi mais precisa e acurada para predizer as precipitações máximas para Petrópolis, enquanto que para Poços de Caldas foi a priori informativa fundamentada nas informações de São João da Boa Vista-SP. Para ambas as localidades, espera-se que, em um tempo médio de cinco anos, ocorra pelo menos um dia com precipitação igual ou superior 100mm.

Palavras-chave: Chuva Extrema; Distribuição GEV; Priori Informativa; Tempos de Retorno.

[1]Departamento de Tecnologia Agroindustrial e Sócio economia Rural, Araras – gilbertoliska@ufscar.br
[2]Departamento de Estatística da Universidade Federal de Alfenas, Alfenas – thales.rangel8@gmail.com
[3]Departamento de Estatística da Universidade Federal de Alfenas, Alfenas – sandravcds78@gmail.com
[4]Departamento de Estatística da Universidade Federal de Alfenas, Alfenas – fabricio.avelar@unifal-mg.edu.br
[5]Departamento de Estatística da Universidade Federal de Alfenas, Alfenas – luiz.beijo@unifal-mg.edu.br

P35

Aplicações de Modelos da Classe Box-Cox Simétrica a Dados de Consumo Alimentar

Giovana Fumes-Ghantous[1]; José Eduardo Corrente[2];
Ana Flávia Giacondino Soligo Lezcano Tatis[3]

Para estimar o consumo habitual de um grupo populacional, o recordatório 24 horas (R24h) é o instrumento de coleta de dados mais comum, e para a análise estatística, a forma de avaliação mais frequente, utiliza uma transformação Box-Cox e analisa os dados de consumo de nutrientes transformados sob a normalidade. Esta metodologia apresenta problemas quando a distribuição dos dados apresenta uma alta assimetria e presença de pontos discrepantes. Como alternativa, um modelo Box-Cox t com efeitos aleatórios foi proposto para modelar a distribuição usual de consumo alimentar, apresentando resultados satisfatórios. A distribuição Box-Cox t pertence a uma classe intitulada Box-Cox Simétrica (BCS). Este trabalho apresenta aplicações de modelos da classe Box-Cox simétrica a um conjunto de dados de consumo alimentar, com a finalidade de verificar se outras distribuições da referida classe poderiam ser utilizadas para modelagem de dados com tais características. O conjunto de dados avaliado é proveniente de um estudo epidemiológico, que teve a participação de idosos, realizado em 2011, no município de Botucatu, São Paulo, Brasil. O instrumento utilizado para a coleta de dados foram três recordatórios 24 horas, e o consumo bruto de 33 micro e macronutrientes foram utilizados. Modelos da classe Box-Cox simétrica foram ajustados, a estimação foi feita por máxima verossimilhança e o critério de Akaike foi usado para seleção dos modelos. Em suma, os modelos da classe Box-Cox simétrica foram adequados para modelagem dos dados, sendo os modelos Box-Cox t e Box-Cox Slash os que apresentaram melhores ajustes dentre as distribuições estudadas.

Palavras-chave: Box-Cox t; Box-Cox Slash; Distribuição de Consumo; Modelos Assimétricos; Recordatório 24 Horas.

[1]Departamento de Ciências Básicas, Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo, giovana.fumes@usp.br

[2]Escritório de Apoio à Pesquisa – Faculdade de Medicina de Botucatu – Universidade Estadual Paulista, jecorren@unesp.br

[3]Graduanda em Engenharia de Alimentos, Faculdade de Zootecnia e Engenharia de Alimentos – Universidade de São Paulo, anatatis@usp.br

P36

Um Modelo de Regressão Quantílica Espaço-Temporal Longitudinal para Taxa de Ocupação de Leitos por COVID-19

Giovanni Pastori Piccirilli[1]; Marcia D'Elia Branco[2]; Jorge Luis Bazán Guzmán[3]

Neste trabalho propomos um novo modelo espaço-temporal para variáveis respostas limitadas e longitudinais sob a abordagem Bayesiana. O algoritmo No-U-Turn-Sampler (NUTS) é empregado para simular valores da distribuição a posteriori. A análise de resíduos é realizada considerando os resíduos quantílicos aleatorizados. Um conjunto de dados reais no contexto da pandemia de COVID-19 nos motivou o uso desse modelo. Na aplicação, estudamos a evolução da proporção de ocupação de leitos de UTI exclusivos para COVID-19 nos Departamentos Regionais de Saúde (DRS) do estado de São Paulo semanalmente no período de Outubro de 2020 a Março de 2021. Além disso, apresentamos as previsões à curto prazo de ocupação de leitos de UTI a nível regional do modelo proposto. Como conclusão podemos dizer que o modelo proposto aqui é uma alternativa para a analise de variáveis respostas limitadas e longitudinais.

Palavras-chave: Resposta Limitada; Análise Longitudinal; COVID-19; Inferência Bayesiana.

[1]Instituo de Matemática e Estatística, USP – giovannipcl@usp.br
[2]Instituo de Matemática e Estatística, USP – mbranco@ime.usp.br
[3]Instituto de Ciências Matemáticas e de Computação, USP – jlbazan@icmc.usp.br

P37

Modelos Observation-Driven para Dados Assimétricos

Gisele de Oliveira Maia[1]; Glaura da Conceição Franco[2].

Neste trabalho estendemos o Modelo Autorregressivo Média Móvel Linear Generalizado (GLARMA) para dados assimétricos não-negativos e limitados, mais especificamente, quando a série temporal condicionada nas observações passadas e covariáveis segue as distribuições Gama, Normal Inversa e Beta. Apresentamos propriedades do modelo, incluindo estudo de estacionariedade, e estimamos os parâmetros utilizando o método de máxima verossimilhança. Um estudo de simulação, realizado com o intuito de verificar a metodologia proposta, mostrou que os resultados são válidos. Por fim,uma análise da série temporal do número diário de casos de Covid19 em Minas Gerais (Brasil) e apresentada assim como previsões para valores futuros.

Palavras-chave: GLARMA; Série Temporal Limitada; Covid-19.

[1]Departamento de Estatística, Universidade Federal de Minas Gerais – glm2018@ufmg.br
[2]Departamento de Estatística, Universidade Federal de Minas Gerais – glaura@est.ufmg.br

P38

Efeito do Cálcio na Interação Estenose Aórtica Induzida, Treinamento Físico e Diltiazem na Tensão Desenvolvida

Gislaine Cristina Batistela[1]; Vitor Loureiro da Silva[2]; Lívia Paschoalino de Campos[3]; Carlos Roberto Padovani[4]

Estenose Aórtica caracteriza-se pelo estreitamento do diâmetro da passagem de sangue do ventrículo esquerdo para a aorta. Em ratos simula-se a patologia com a colocação de um clipe de prata acima da válvula aórtica. Objetivou-se analisar a influência do trânsito de cálcio sobre o parâmetro mecânico tensão desenvolvida (TD) em ratos *Wistar* no esquema de três fatores: indução à estenose aórtica supravalvar (Sham ou EAo); Treinamento Físico (ausência=NTF e presença=TF) e Inibidor específico dos canais tipo L do coração denominado Diltiazem (ausência e presença). Os grupos compostos por 58 animais Sham; 16 NTF Diltiazem, 16 NTF placebo, 13 TF Diltiazem e 13 TF placebo, e 40 EAo; 10 NTF Diltiazem, 10 NTF placebo, 10 TF Diltiazem e 10 TF placebo. Avaliou-se a TD (g/mm^2) no músculo papilar submetido à elevação sequencial equiespaçada ($0,5\ mM$) de cálcio, iniciando-se em $0,5\ mM$ até $3,5\ mM$. Os perfis médios da TD nos grupos considerando simultaneamente os sete níveis de cálcio foram avaliados por meio da MANOVA e pelos intervalos de confiança simultâneos de Bonferroni. Destacam-se: nos NTF diferenças entre o grupo Sham e EAo independentemente do Diltiazem, contudo quando tem-se o TF esta situação é equilibrada. Além disso, o Diltiazem tem um efeito significativo na TD quando o animal é exposto ou não ao TF, independentemente de ser do Sham ou EAo. O fato do animal ser submetido ou não ao TF, não se fez mudança na resposta da TD. O Diltiazem mostrou-se como protetor independentemente do TF e da submissão à EAo.

Palavras-chave: MANOVA; Tensão Desenvolvida; Manobra de Cálcio.

[1]Departamento de Engenharia de Produção, Instituto de Ciências e Engenharia, Unesp, Câmpus de Itapeva – gislaine.batistela@unesp.br
[2]Departamento de Clínica Médica, Faculdade de Medicina, Unesp, Câmpus de Botucatu – vitor-loureiro_ed.fisica@hotmail.com
[3]UniBR – Faculdade de Botucatu – lipaschoalino@gmail.com
[4]Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia, Instituto de Biociências, Unesp, Câmpus de Botucatu – cr.padovani@unesp.br

P39

Um Estudo a Respeito de Aspectos não Lineares e Dinâmicos em Modelos para Séries Temporais

Guilherme dos Santos[1]; Daniel Takata Gomes[2]; Larissa de Carvalho Alves[3]

Dentro da comunidade estatística, a previsão de séries temporais é uma tarefa de grande interesse, isto é, tendo as observações decorrentes de um processo até o presente momento, deseja-se estimar os valores que este irá assumir no futuro. Para tal, é necessário aproximar o processo gerador dos dados utilizando modelos estatísticos. Este processo pode apresentar relações complexas de dependência temporal, envolvendo não-linearidades ou ainda evoluindo ao longo do tempo. Contudo, grande parte dos modelos para séries temporais vistos em uma graduação de estatística dependem de suposições de linearidade, e de que os parâmetros são estáticos, isto é, não variam no tempo. Neste trabalho são estudadas duas classes de modelos que vêm desempenhando um papel importante na previsão de séries temporais nos últimos anos, os modelos dinâmicos bayesianos, que ganham flexibilidade permitindo que os parâmetros variem no tempo, e as redes neurais recorrentes, que possibilitam que sejam capturadas relações de dependência temporal não lineares.

Palavras-chave: Modelos Dinâmicos; Redes Neurais; Séries Temporais; Previsão.

[1]Departamento de Métodos Estatísticos, IM/UFRJ – guidossantos@dme.ufrj.br
[2]ENCE/IBGE – daniel.gomes@ibge.gov.br
[3]Coordenação de Graduação, ENCE/IBGE – larissa.alves@ibge.gov.br

P40

Selecting Genetic Variants and Interactions Associated with Amyotrophic Lateral Sclerosis: a Group LASSO Approach

Hellen Geremias dos Santos[1]; Maria Luiza Matos Silva[2]; Sofia Galvão Feronato[3]; Rafael Izbicki[4]

Amyotrophic Lateral Sclerosis (ALS) is a complex disease, resulting from relationships between genetic, environmental and lifestyle characteristics. Genome-wide Association Studies (GWAS) is the most common approach to detect relationships between single nucleotide polymorphisms (SNPs) and such disease. As SNP data are high-dimensional, the association of each SNP with the disease is typically individually tested, resulting in a large number of simultaneous hypothesis tests. Furthermore, this approach does not support the detection of SNPs dependent on genetic interactions. We aim to overcome this by applying a two-step Group LASSO procedure to select SNPs and pairwise-interactions associated with the ALS phenotype. We analyzed SNP data from 276 ALS patients and 268 controls from the National Institute of Neurological Disorders and Stroke Repository available at the database of Genotypes and Phenotypes. A two-step analysis was applied in 2,000 iterations. For each iteration, we fitted a Group LASSO model to a bootstrap sample and a random subset of the variables (25%) from the original data set aiming to screen for important SNPs, thus restricting the pairwise-interaction search space (first step). Subsequently, we fitted a Hierarchical Group LASSO model in order to consider pairwise-interactions between all the selected SNPs (second step). A set of variables was prioritized according to their bootstrap selection frequency to assess biological implications. We identified 7 SNPs (*rs16984239*, *rs10459680*, *rs1436918*, *rs1037666*, *rs4552942*, *rs10773543* and *rs2241493*) and two pairwise-interactions (*rs16984239:rs2118657* and *rs16984239:rs3172469*), mostly related to conservation and function of the nervous system, including proteins associated with resting potential and survival of neurons.

Palavras-chave: Amyotrophic Lateral Sclerosis; Genome-wide Association Studies; Group LASSO Regularization; Single Nucleotide Polymorphisms; Pairwise-interaction.

[1]Instituto Carlos Chagas, Fundação Oswaldo Cruz, Curitiba, Paraná – hellen.santos@fiocruz.br
[2]Departamento de Estatística, Universidade federal de São Carlos, São Carlos, São Paulo – marialuizamatossilva@gmail.com
[3]Instituto Carlos Chagas, Fundação Oswaldo Cruz, Curitiba, Paraná – sofiagf1803@gmail.com
[4]Departamento de Estatística, Universidade federal de São Carlos, São Carlos, São Paulo – rizbicki@ufscar.br

P41

The Use of Quasi U-statistics to Test Homogeneity of Variants of Coronavirus

Hildete Prisco Pinheiro[1]; Bruno Martinez Farias[2]

One of the goals in genetic studies is the comparison of groups of genetic sequences. The main purpose of this work is to test homogeneity among groups of RNA sequences of different variants of coronavirus. These comparisons are based on a class of weighted U-statistic, the so-called quasi U-statistics. We use the test statistic proposed by Pinheiro et al. (2005, 2009, 2011) with diversity measures between sequences built based on Hamming distances. Asymptotic properties of the test statistic are discussed when either the number of sequences (n) and/or the length of the sequences (K) are large. For small sample sizes resampling techniques, such as bootstrap or jackknife can be used to generate the empirical distribution and p-values.

Palavras-chave: Coronavirus; Bootstrap; Hamming Distance; RNA Sequence; U-statistics.

[1]Department of Statisics, IMECC, Unicamp – hildete@unicamp.br
[2]Department of Statistics, IMECC, Unicamp – b181980@dac.unicamp.br

P42

Transition Models for Grouped Data Applied to Psyllid Movement

Idemauro Antonio Rodrigues de Lara[1]; Rafael A. Moral[2]; Cesar Augusto Taconeli[3]; Carolina Reigada[4]; John Hinde[5]

This work is a part of recently paper (Lara et al., 2020), where we present a methodology to analyze the changes of response categories over time, using transition models, for grouped data. In Entomology, it is common categorized data to be recorded as groups, i.e. different categories with a number of individuals in each. We have used likelihood ratio tests to assess non-stationarity. As motivation, we present an application to understand the movement patterns of female adults of *Diaphorina citri*, a pest of citrus plantations. Also, simulation studies about the methodology were presented, showing that procedure is an alternative to analyse categorized data over discrete time with grouped structure.

Palavras-chave: Transition Probabilities; Stationarity; Longitudinal Data; Stochastic Process.

[1]Departamento de Ciências Exatas, ESALQ/USP – idemaurol@usp.br
[2]Maths and Statistics Department, Maynooth University, Ireland – rafael.deandrademoral@mu.ie
[3]Departamento de Estatística, UFPR – cetaconeli@gmail.com
[4]Departamento de Ecologia e Biologia Evolutiva, UFSCar – ca.reigada@gmail.com
[5]School of Mathematics and Statistics, National University of Ireland – john.hinde@nuigalway.ie

P43

Estimação do Número Ótimo de Grupos k em Análises de Cluster de Séries Temporais: Estudo de Caso Utilizando o Índice Silhouette e Algoritmo K-Means

Inácio Puntel dos Passos[1]; Raquel da Fontoura Nicolette[2]

O Índice Silhouette de validação interna de análises de cluster vem sendo usado desde a sua proposição como um estimador do número ótimo de grupos k, visto que a maximização da largura média das silhuetas dos grupos é tida como uma medida da qualidade de uma solução. Visto isso, realizou-se um experimento a fim de averiguar a acurácia do índice na estimação do número de grupos k em análises de cluster de séries temporais por meio do algoritmo K-Means. Foram utilizados 3 bancos de dados publicamente disponíveis. No primeiro banco, encontrou-se que o índice falhou em estimar k real ($k = 4$), porém, a solução sugerida pelo índice ($k = 7$) alcançou concordância quase perfeita (escore ARI = 0.96) com os dados reais, apesar de sugerir um número maior de grupos. Em relação ao segundo banco, o índice acertou o k real ($k = 3$), porém, a solução sugerida obteve escore ARI (0.52) inferior a outras soluções (ARI = 0.67 com $k = 2$). Finalmente, em relação ao terceiro banco, não somente o índice falhou em estimar o k real ($k = 7$), como também a solução sugerida por ele ($k = 2$) obteve o pior escore ARI (0.04). Os dados obtidos sugerem que o Índice Silhouette não deve ser usado como uma evidência genérica da qualidade de uma solução ou como estimador genérico do número real de classes, mas os contextos em que ele serve a essas funções devem ser sistematizados a fim de orientar o seu uso.

Palavras-chave: Análise de Cluster; K-Means; Séries Temporais; Índice Silhouette.

[1]Pós-graduando em Ambientometria, IMEF/FURG – rbkpuntel@gmail.com
[2]Professora PPG Ambientometria, IMEF/FURG – raquelnicolette@furg.br

P44

Generalizações da Função de Ligação Cloglog para Modelos de Regressão Binomial

Jessica Suzana Barragan Alves[1]; Jorge Luis Bazán Guzmán[2]; Reinaldo B. Arellano-Valle[3]

Várias funções de ligação assimétricas com um parâmetro extra têm sido propostas na literatura ao longo dos últimos anos para lidar com dados desbalanceados em regressão binária (quando uma das classes é menor que a outra), mas estas não possuem a função de ligação *cloglog* como caso particular, com exceção da função de ligação baseada em distribuição generalizada. Neste trabalho, propomos funções de ligação generalizadas baseadas na função de ligação *cloglog* como uma alternativa para modelar dados de resposta binomial na presença de dados médicos desbalanceados. Um procedimento de estimação Bayesiano baseado nas cadeias MCMC foi desenvolvido para esses modelos, utilizando para isso o algoritmo NUTS, no pacote `rstan` em R. Foi realizado um estudo de simulação para avaliar o desempenho na recuperação dos parâmetros e bons resultados foram obtidos com esse estudo. Uma aplicação foi realizada com dados das meninas de Varsóvia já utilizadas na literatura para modelar dados desbalanceados, e como resultados obtivemos melhores estimativas do que as funções de ligação mais comumente utilizadas. Com esse trabalho podemos verificar a importância do uso das funções de ligação generalizadas quanto à assimetria de dados, controlada por meio de um parâmetro extra de forma, λ.

Palavras chaves: Estimação Bayesiana; Regressão Binomial; *Cloglog*; Função de Ligação Generalizada; Algoritmo NUTS.

[1]Departamento de Matemática Aplicada e Estatística Universidade de São Paulo, São Carlos, Brasil – jessicasbarragan@usp.br

[2]Departamento de Matemática Aplicada e Estatística Universidade de São Paulo, São Carlos, Brasil – jlbazan@icmc.usp.br

[3]Departamento de Estatística Pontifícia Universidade Católica do Chile, Santiago, Chile – reivalle@mat.uc.cl

P45

Sigma Júnior Consultoria Estatística: Vivência Empresarial na Área de Dados Durante a Graduação

Thiago Tavares Lopes[1]; Rafael Adamy Monteiro[2]; João Inácio Scrimini[3]; Laís Helen Loose[4]; Renata Rojas Guerra[5]

O processo de qualificação de um futuro profissional da estatística inicia-se ainda dentro do ambiente acadêmico. Entretanto, com os novos desafios do mercado de trabalho, é cada vez mais importante que o domínio do conhecimento teórico seja complementado por experiências baseadas em projetos que permitem aprimorar habilidades dos profissionais em formação. Pensando nisto, no ano 2017, foi fundada a Sigma Júnior Consultoria Estatística, primeira empresa júnior da área de estatística no interior do Rio Grande do Sul, sediada na Universidade Federal de Santa Maria (UFSM). A Sigma Júnior é composta por discentes da UFSM e conta com a orientação de uma professora e colaboração de mais quatro professores, todos do departamento de Estatística da UFSM, sendo atribuição dos discentes a administração da empresa. Os acadêmicos têm a possibilidade de ainda durante a graduação, atuarem em consultorias estatísticas, aplicando, nos mais diversos tipos de projetos, as técnicas de análise de dados aprendidas ao longo da graduação. Além de proporcionar aos alunos a vivência profissional como estatísticos, a Sigma Jr também proporciona experiências de empreendedores e conhecimentos de organização empresarial, bem como a ampliação e o desenvolvimento de diferentes habilidades. Cabe a Sigma Júnior contribuir para a cultura de dados na sociedade local e regional, viabilizar o acesso a serviços de qualidade e em condições acessíveis para empresas e toda comunidade acadêmica. Nos últimos cinco anos a Sigma Jr possibilitou que mais de 85 acadêmicos da UFSM adquirissem experiências na realização de consultorias estatísticas e conhecimentos em termos de organização empresarial por meio do desenvolvimento de mais de 80 projetos.

Palavras-chave: Consultoria Estatística; Cultura de Dados; Empreendedorismo; Empresa Júnior; Vivência Empresarial.

[1]Curso de Bacharelado em Estatística, UFSM – thiago.tlopess@gmail.com
[2]Curso de Bacharelado em Estatística, UFSM – raphael_adamy@hotmail.com
[3]Curso de Bacharelado em Estatística, UFSM – joao.inacio.scrimini@gmail.com
[4]Departamento de Estatística, UFSM – lais.loose@ufsm.br
[5]Departamento de Estatística, UFSM – renata.r.guerra@ufsm.br

P46

Semi-parametric Models Based on the Scale Mixtures of Centred Skew Normal Distribution for Independent and Longitudinal Data

João Victor Bastos de Freitas[1]; Caio Lucidius Naberezny Azevedo[2]; Juvêncio Santos Nobre[3]

In this work, two classes of regression models for continuous, skewed and/or heavy tailed data were developed. One for independent data and another for dependent data. We considered a semi-parametric approach using Generalized Additive Partially Linear Models (GAPLM), for independent data, and GAPLM with Generalized Estimation Equations (GEE), for dependent data. In both cases, semi-parametric predictors for response means and scale mixtures of centered skew-normal (SMCSN) distributions for the (marginal) errors were considered. For dependent data, the dependence structures were modelled through GEE. Concerning the SMCSN distributions we considered usual mixing measures (gamma, beta and binary distributions) as well as never used ones (generalized gamma, Birnbaum-Saunders and beta prime distributions). Estimation methods, goodness of model fit and diagnostic tools for these models, under the frequentist paradigm, were developed. Computational routines were created, to allow for the use of the developed methodologies, as well as simulation studies were performed to study their performance. Also, the modelling of real problems, through such methodologies, were considered, illustrating the potential of the obtained results.

Palavras-chave: Longitudinal Data; Scale Mixtures of Centered Skew-Normal; Generalized Estimating Equations; Generalized Additive Partially Linear Models; Frequentist Inference.

[1]Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Brasil – jvbfreitas@ime.unicamp.br
[2]Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Brasil
[3]Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Brazil

P47

Análise de Sobrevivência com Censura Dependente Baseada em Cópulas Arquimedianas

José Ailton Nunes de Lima[1]; Paulo Cerqueira dos Santos Junior[2]

Em Análise de Sobrevivência, quando algum evento concorrente está representando a censura, a suposição tradicional de censura independente pode não ser satisfeita e, assim, a aplicação usual com o estimador Kaplan-Meier produz uma estimativa enviesada para a função de sobrevivência do tempo de vida. Considere um tempo de sobrevivência T que está sujeito a censura aleatória à direita e suponha que T seja dependente do tempo de censura U. Essa situação é frequentemente encontrada na prática. Por exemplo, na área clínica, se o motivo de um indivíduo com uma determinada doença sair do estudo estiver relacionado ao seu estado de saúde, influenciando o tempo até a morte, portanto é possível que T e U sejam dependentes. Diante disso, o interesse geral está no comportamento da distribuição marginal de T dado o tempo de censura dependente U. Neste estudo, é utilizado um modelo com uma estrutura de dependência baseada em funções de cópulas arquimedianas para modelar a relação entre T e U, usando o modelo exponencial por partes potência para as distribuições marginais, sob a estrutura de riscos proporcionais. O método de estimação por máxima verossimilhança é utilizado para a obtenção das estimativas dos parâmetros dos modelos marginais considerando o parâmetro de dependência da cópula ϕ conhecido e especificado, assim como também é realizada a estimação simultaneamente com ϕ, seguindo algumas condições de identificabilidade. É realizada uma análise de sensibilidade e um estudo de simulação, com a finalidade de avaliar o desempenho do modelo em questão e o processo de estimação, especificando as cópulas paramétricas Ali-Mikhail-Haq (AMH), Clayton e de Frank. Finaliza-se com uma aplicação a dados reais comparando com os modelos usuais.

Palavras-chave: Idetificabilidade; Máxima Verossimilhança; Modelo Exponencial por Partes Potência; Parâmetro de Dependência.

[1]Instituto de Ciências Exatas e Naturais, UFPA – ailton.g1721@gmail.com
[2]Faculdade de Estatística, UFPA – pauloest16@gmail.com

P48

Análise do Preço do Gás de Cozinha na Bahia entre 2002 e 2021 Usando Métodos de Análises de Séries Temporais

Rodrigo Barbosa de Cerqueira[1]; José Roberto Santos da Silva[2]; Denilson Lima Santos[3]; Hildete Karla Borba Andrade[4]; Jonatas Silva do Espírito Santo[5]

O gás de cozinha é um produto fundamental na vida da família brasileira. Mais de 90% dos domicílios usam esse combustível diariamente. Este trabalho tem como objetivos identificar quais os elementos que compõem o preço do botijão de 13 kg, entender como os preços em cada uma destas etapas são gerados e como se comportaram ao longo dos últimos anos, numa tentativa de explicar a dinâmica recente dos preços do GLP. Foram utilizadas análises de séries temporais, correlações cruzadas, cointegração e quebra de estrutura para analisar do comportamento do seu preço e a influência dos fatores que compõem o seu valor final. Os dados foram retirados das publicações mensais da ANP entre janeiro de 2002 e dezembro de 2021. Como resultados, notou-se que a série do preço do gás de cozinha ao consumidor final apresentou uma tendência crescente no período, sendo a Margem Bruta de Revenda e o Preço ao Produtor os fatores que mais contribuíram para esse crescimento como demonstrado na análise de correlação cruzada. Na análise de dissimilaridade entre o preço final e cada uma das suas componente, o Preço ao Produtor e Margem Bruta de Revenda apresentaram as menores distâncias, na série da Bahia, considerando o *Dynamic time warping* (DTW) e a distância euclidiana, respectivamente para a série completa. Quando analisada a série após a quebra estrutural, em 2017, ambas as técnicas apontaram para o Preço ao Produtor como menor distância, tanto para a Bahia como para o Brasil.

Palavras-chave: GPL; Preço; Séries Temporais; Renda; Variação.

[1]Superintendência de Estudos Econômicos e Sociais da Bahia – rodrigobarbosa@sei.ba.gov.br
[2]Superintendência de Estudos Econômicos e Sociais da Bahia – joserobertosilva@sei.ba.gov.br
[3]Superintendência de Estudos Econômicos e Sociais da Bahia – denilsonlima@sei.ba.gov.br
[4]Superintendência de Estudos Econômicos e Sociais da Bahia – hildeteandrade@sei.ba.gov.br
[5]Superintendência de Estudos Econômicos e Sociais da Bahia – jonatassanto@sei.ba.gov.br

P49

Performance of Confirmatory Factor Analysis with Binary Indicators: A Simulation Study

Josemir R. Almeida[1]; Ythalo Hugo[2]; Michelle Passos[3]; Maria del P.F. Quispe[4]; Valentina Martufi[5]; Rosana Aquino[6]; Elzo P. Pinto Junior[7]; Roberta Freitas[8]; Leila D.A.F. Amorim[9]

Confirmatory Factor Analysis (CFA) aims to reduce the dimensionality of data from a larger number of observed and highly correlated variables to a smaller number of latent dimensions. The CFA is a well-known methodology used to assess the psychometric properties of scales in pre-determined theoretical models, especially when all variables are continuous and normally distributed. However, the literature is scarcer for analysis of binary data. This work aims to investigate the performance of estimation methods for CFA when all indicators for each factor/construct are binary in nature. We reviewed several estimation procedures, including DWLS (diagonally weighted least squares), WLS (weighted least squares), WLSM (weighted average adjusted least squares), and evaluated their performance using Monte Carlo simulation studies. We simulated data varying elements such as: the number of factors, correlation between factors, correlation between factor indicators, factor loadings, sample size, number of indicators per factor and the frequency distribution of the indicators. In the simulation, it was possible to verify that the lower the correlation between the factors and the smaller the sample size, the greater the bias in the estimates of factor loadings. As for the percentage of convergence, the smaller sample size (n=200) and the reduction in the correlation between the factors, provide a decrease in the percentage of convergence. In the simulations, 2000 samples were used. Further simulations studies should be encouraged and are welcome to advance the understanding of situations in which CFA with binary indicators is appropriate.

Palavras-chave: Confirmatory Factor Analysis; Binary Indicators; Monte Carlo Simulation.

[1]Fiocruz-Cidacs – josemir.almeida@fiocruz.br
[2]Fiocruz-Cidacs – ythalo.santos@fiocruz.br
[3]Fiocruz-Cidacs/PGMAT-UFBA – michelle.passos@fiocruz.br
[4]Fiocruz-Cidacs – mariadelpilarfloresq@hotmail.com
[5]Fiocruz-Cidacs – valentina.martufi@gmail.com
[6]ISC/UFBA – aquino@ufba.br
[7]Fiocruz-Cidacs – elzo.junior@fiocruz.br
[8]Fiocruz-Cidacs – roberta@fiocruz.br
[9]IME/UFBA – leiladen@ufba.br

P50

Sorte e Habilidade: a Influência Destes Dois Aspectos ao Longo de 56 Temporadas das Cinco Principais Ligas Europeias

Juliana Sena de Souza[1]; Márcia Helena Barbian[2]

Um dos esportes mais populares do mundo é o futebol, que possui alta quantidade de torcedores, admiradores e simpatizantes. A concorrência acirrada no futebol cria uma imprevisibilidade mais acentuada nos resultados das partidas, o que gera um interesse maior do público. Outros fatores que influenciam esse interesse é a mistura entre sorte e habilidade. Este trabalho tem como objetivo utilizar o coeficiente de sorte e habilidade (φ) em dados de futebol das ligas nacionais da primeira divisão dos cinco principais campeonatos europeus para medir a sorte e a habilidade ao longo de 56 temporadas. O coeficiente permite avaliar se há diferentes habilidades entre as equipes de uma temporada ou se o campeonato é determinado aleatoriamente. Além disso, para verificar se o coeficiente é estatisticamente significativo, foi utilizado simulações de Monte Carlo. Por fim, para valores de φ significativos, foi utilizado um algoritmo para verificar quais clubes que, se retiradas do campeonato, deixam a temporada aleatória. Para a definição dos principais campeonatos europeus, foi utilizado o ranking baseado no "coeficiente de clubes por país" desenvolvido pela UEFA. Esta aplicação foi aplicada nas últimas 56 edições dos campeonatos. Os resultados encontrados neste trabalho são, de forma geral, evidências que corroboram com o senso comum. Os times mais ricos do mundo dominam suas ligas, além de deixar o campeonato menos aleatório.

Palavras-chave: Estatística Esportiva; Futebol Europeu; Coeficiente de Sorte e Habilidade.

[1]Programa de Pós-Graduação em Estatística, UFRGS – Porto Alegre, RS – julianass.estatistica@gmail.com
[2]Programa de Pós-Graduação em Estatística, Departamento de Estatística, UFRGS – Porto Alegre, RS – mhbarbian@gmail.com

P51

Visualização e Modelagem dos Dados do Enem e Análise de Impacto da Pandemia de COVID-19

Julie Camolesi da Silva[1]; Eduardo Próspero Santana[2]; Cibele Maria Russo[3]

Neste trabalho analisamos dados educacionais do Exame Nacional do Ensino Médio (Enem) nas provas aplicadas nos anos de 2019 e 2020, utilizando técnicas de visualização e exploração de dados e modelos de Teoria de Resposta ao Item - TRI (de Andrade et al., 2000). O estudo inclui uma análise comparativa entre os resultados da prova nos dois anos de observação, identificando possíveis consequências do primeiro ano de quarentena e ensino à distância impostos pela pandemia, dentre eles a diminuição do número de inscrições provenientes das regiões menos favorecidas do país. Essa comparação foi feita também para a prova digital, aplicada pela primeira vez em 2020. Pôde-se observar um perfil bastante diferente para os estudantes que realizaram a prova digital em 2020 dos estudantes que realizaram a prova física, tanto em 2019 quanto em 2020, sobretudo por ter sido disponibilizada somente em noventa e nove municípios brasileiros, não tendo sido acessível a grande parte das regiões norte e nordeste do país. Com os modelos de TRI estudados, foi possível identificar as questões mais fáceis e mais difíceis em cada área do conhecimento. Foi desenvolvida uma comparação dos modelos de Rasch, 2 parâmetros e 3 parâmetros, considerando a estimação das habilidades pela esperança a posteriori (EAP) e moda a posteriori (MAP), e os melhores resultados foram observados para o modelo de 3 parâmetros utilizando a MAP. As análises foram implementadas utilizando pacotes girth em Python com o Jupyter notebook e google colab e MIRT na linguagem R.

Palavras-chave: Teoria de Resposta ao Item; COVID-19; Enem; Impactos da Pandemia.

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – juliecamolesi@usp.br

[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – eduardosantana2001@usp.br

[3]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – cibele@icmc.usp.br

P52

The Zero-Modified Negative Binomial Distribution

Katiane S. Conceição[1]; Marinho G. Andrade[2]; Nalini Ravishanker[3]

In this paper, we give detailed descriptions of the Zero-Modified Negative Binomial distribution for analyzing count data. In particular, we study the characterizations and properties of this distribution, whose main advantage is its flexibility which makes it suitable for modeling a wide range of overdispersed and underdispersed count data (which may or may not be caused by zero-modification, i.e., the inflation or deflation of zeroes), without requiring previous knowledge about any of these inherent data characteristics. We derive maximum likelihood estimation of the model parameters based on positive observations, and evaluate the loss of efficiency by considering this procedure. We illustrate the suitability of this distribution on real data sets with different types of zero-modification.

Palavras-chave: Negative Binomial Distribution; Underdispersion; Zero-Deflated Count Data; Relative Efficiency.

[1]Department of Applied Mathematics and Statistics, Institute of Mathematical and Computer Science, University of São Paulo, São Carlos/SP – katiane@icmc.usp.br

[2]Department of Applied Mathematics and Statistics, Institute of Mathematical and Computer Science, University of São Paulo, São Carlos/SP – marinho@icmc.usp.br

[3]Department of Statistics, University of Connecticut, Storrs/CT, United States of America – nalini.ravishanker@uconn.edu

P53

Avaliação Numérica de Testes de Hipóteses Utilizados no Modelo de Regressão Quantílica Chen

Gabriel Hagemann Behling Alves[1]; Laís Helen Loose[2]

O modelo de regressão quantílico Chen é uma proposta que vem sendo desenvolvida e é adequada para modelagem de quantis positivos. Neste estudo avaliamos, por meio de simu- lações de Monte Carlo, o desempenho de estatísticas de teste usuais para testar hipóteses neste modelo. Utilizamos as estatísticas razão de verossimilhanças (LR), gradiente (G), escore (S) e Wald (W), 5.000 réplicas de Monte Carlo e consideramos três dife- rentes tamanhos amostrais ($n \in \{30, 50, 100\}$). Avaliamos a taxa de rejeição das estatísticas de teste em dois cenários, considerando os parâmetros do modelo, três valores para o quantil (0,25; 0,5; 0,9) e a função de ligação logarítmica. Consideramos ainda a avaliação dos testes com um e dois parâmetros sendo testados. Quando apenas um parâmetro foi testado, as estatísticas LR e G apresentam um comportamento mais conservador, com valores mais próximos aos níveis de significância fixados, os resultados melhoram à medida que o tamanho amostral aumenta. Também foi possível observar que a estatística escore é mais liberal que as demais em amostras menores. Ao testarmos dois parâmetros, observamos que ao acrescentar um parâmetro ao teste de hipótese, as taxas de rejeição da hipótese nula convergem com menor rapidez para os níveis de significância assumidos. Por fim, em ambos os cenários, os resultados são similares e portanto não recomendamos a utilização da estatística escore em pequenas amostras. Ainda, em geral a estatística gradiente é a estatística de teste que apresenta o melhor desempenho.

Palavras-chave: Distribuição Chen; Estatísticas de Teste; Taxas de Rejeição; Regressão.

[1]Acadêmico de Estatística, Universidade Federal de Santa Maria, Santa Maria, Brasil – gabrielmsr9@gmail.com
[2]Departamento de Estatística, Universidade Federal de Santa Maria, Santa Maria, Brasil – lais.loose@ufsm.br

P54

Modelos de Classes Latentes com Desfechos Distais: Estratégias de Análise para a Relação entre a Síndrome Metabólica e o Diabetes Mellitus Tipo 2.

Nila Mara Smith Galvão[1]; Leila Denise Alves Ferreira Amorim[2].

Diferentes técnicas de modelagem de classes latentes com desfechos distais foram exploradas na investigação da relação entre a síndrome metabólica (SM), definida como um construto categórico, e o diabetes mellitus tipo 2 (DM2), em mulheres. Foram analisados dados do baseline e da Onda 2 do Estudo Longitudinal de Saúde do Adulto (ELSA-Brasil) para 4.794 mulheres sem DM2 no início do estudo. Padrões de SM foram identificados via Análise de Classes Latentes (LCA). As associações entre estes padrões e o DM2 foram mensuradas sob diferentes abordagens: convencional – modelo logístico e modelo de Cox combinado com procedimentos classify-analyze - e via modelagem com variáveis latentes - método LTB (Lanza-Tay-Bray, 2013) e modelo de mistura de sobrevivência, com curvas de sobrevivência estimadas pelo método de Kaplan-Meier. Três padrões latentes de SM foram identificados: 'SM ausente', 'SM parcial' e 'SM total'. As probabilidades de ausência do diabetes após 4 anos foram 0,98, 0,91 e 0,88 para os padrões de menor a maior risco metabólico. O risco de desenvolver DM2 foi pelo menos três vezes maior em mulheres pertencentes aos padrões 'SM parcial' e 'SM total', em comparação ao padrão 'SM ausente'. Em contraste com modelos de regressão convencionais, a modelagem com variáveis latentes leva em conta o erro de mensuração inerente ao uso de fatores observados para representar um construto de interesse, resultando em inferências estatísticas mais adequadas. Modelos LCA com desfechos distais mostraram-se úteis para identificar padrões de SM e explorar a sua relação com o DM2, considerando a expressão heterogênea da síndrome. **Palavras-chave:** Análise de classes latentes; Desfechos distais; Modelos de mistura; Análise de Sobrevivência.

[1]Departamento de Ciências Exatas e da Terra – UNEB, Salvador - BA – ngalvao@uneb.br
[2]Departamento de Estatística – IME/UFBA, Salvador - BA – leiladen@ufba.br

P55

Previsão em Tempo Real de Séries Temporais para Operação de Reservatórios e Distribuição de Água

Leonardo Fonseca Larrubia[1]; Chang Chiann[2]; Olga Satomi Yoshida[3]

Aplicamos técnicas de previsão de séries temporais visando auxiliar os Centros de Controle Operacionais de distribuição de água em suas operações e tomadas de decisões diárias. Para tanto, foram utilizadas séries temporais geradas por equipamentos de medição de nível, vazões de entrada e de saídas do reservatório e de pressões a montante e a jusante de válvulas que controlam o uxo de água ao longo da rede de abastecimento. Os dados, referentes ao sistema de distribuição de água da cidade de Peruíbe, foram fornecidos pela Sabesp da Baixada Santista e sua amostragem temporal é a cada hora, indo das 1:00 do dia 1^o de janeiro de 2017 até às 23:00 do dia 31 de dezembro de 2018. Após um tratamento adequados das séries, isto é, correção de valores atípicos e preenchimento de valores omis- sos, usamos os modelos SARIMA, modelos de regressão com erros auto correlacionados e modelos BATS e TBATS para gerar as previsões de cada série. Para a vericação da performance dos métodos preditivos foram utilizadas técnicas de *rolling analysis* modicada, na qual os valores previstos foram avaliados através de medidas de acurácias, como o RMSE, e comparados com métodos *benchmarks* de previsão. Os resultados demonstraram que os métodos propostos possuem bons desempenhos para a maioria das séries.

Palavras-chave: Séries Temporais, Previsão, Indústria da Água, Tempo Real, *Big Data*.

[1]Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo – leonardo.larrubia@usp.br
[2]Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo – chang@ime.usp.br
[3]Instituto de Pesquisas Tecnológicas do Estado de São Paulo, São Paulo – olga@ipt.br

P56

Redução Amostral Bivariada na Presença de Correlação Espacial

Letícia Ellen Dal' Canton[1]; Tamara Cantú Maltauro[2]; Luciana Pagliosa Carvalho Guedes[3]; Miguel Angel Uribe-Opazo[4]

A partir de informações sobre uma coleta amostral prévia e tendo em vista o menor investimento financeiro possível com análises de solo, é fundamental conhecer as correlações espaciais entre as observações amostrais. Isto porque, pontos amostrais autocorrelacionados espacialmente geram informações redundantes. A proposta do tamanho amostral efetivo é utilizar a autocorrelação espacial para calcular qual é o número de pontos amostrais independentes, tomando este como o novo tamanho amostral. Dentro deste contexto, existem estudos que envolvem duas variáveis georreferenciadas, que além da autocorrelação espacial, também apresentam correlação espacial entre si. O intuito desse trabalho foi obter uma redução amostral, por meio de uma proposta para o cálculo do tamanho amostral efetivo bivariado (ESS_{bi}), considerando simultaneamente duas variáveis correlacionadas espacialmente. Para simular, modelar e descrever o padrão espacial destas variáveis foi considerado o modelo espacial gaussiano bivariado com componente de correlação parcialmente comum (BGCCM), que considera tanto a correlação espacial entre as duas variáveis, quanto a associação espacial individual. Foi construído um cenário com cinco ensaios utilizando diferentes estruturas de dependência espacial, com parâmetros de alcance simulados que variaram de 75 a 275 metros. Os parâmetros de dispersão e alcance estimados pelo BGCCM, foram utilizados para calcular a estimativa do ESS_{bi}. Após 102 simulações para cada ensaio, foi verificado que o ESS_{bi} diminui à medida que se aumenta simultaneamente a correlação espacial, tanto entre as variáveis quanto a autocorrelação.

Palavras-chave: Agricultura de Precisão; BGCCM; Geoestatística; Simulação; Tamanho Amostral Efetivo Bivariado.

[1]Centro de Ciências Exatas e da Terra, Programa de Pós-Graduação em Engenharia Agrícola, Universidade Estadual do Oeste do Paraná, Cascavel-PR, Brasil – leticiacanton@hotmail.com
[2]Centro de Ciências Exatas e da Terra, Programa de Pós-Graduação em Engenharia Agrícola, Universidade Estadual do Oeste do Paraná, Cascavel-PR, Brasil – tamara_ma02@hotmail.com
[3]Centro de Ciências Exatas e da Terra, Programa de Pós-Graduação em Engenharia Agrícola, Universidade Estadual do Oeste do Paraná, Cascavel- PR, Brasil – luciana_pagliosa@hotmail.br
[4]Centro de Ciências Exatas e da Terra, Programa de Pós-Graduação em Engenharia Agrícola, Universidade Estadual do Oeste do Paraná, Cascavel- PR, Brasil – mopazo@uol.com.br

P57

Classificação de Áreas de Cana-de-açúcar Utilizando Imagens LANDSAT e Algoritmos de Aprendizado de Máquina

Ana Clara A. V. B. de Barros[1]; Marcelo A. da Silva[2]; Ana Claudia S. Luciano[3]

A cana-de-açúcar possui grande importância no agronegócio e na economia nacional, já que o Brasil é o maior produtor de açúcar e um dos grandes mercados de biocombustível internacional. Nos últimos anos, pesquisas sobre o uso de dados de sensoriamento remoto têm sido amplamente aplicadas no monitoramento de cana-de-açúcar, gerando grandes avanços e inovações tecnológicas na automação da identificação de cana-de-açúcar. O objetivo deste trabalho é classificar áreas com cana-de-açúcar a partir de imagens LANDSAT explorando três importantes métodos de classificação em aprendizado de máquina: a Regressão Logística, o modelo de Árvores de Classificação e o algoritmo Random Forests. Para isso, foram selecionados índices de vegetação e bandas espectrais obtidos a partir das imagens de satélite LANDSAT de uma das regiões com maior concentração de área plantada com cana-de-açúcar do estado de São Paulo. A qualidade dos resultados obtidos pelos diferentes algoritmos foi mensurada e comparada com a literatura através de critérios computacionais e estatísticos. Especificamente, os algoritmos utilizados forneceram acurácia acima de 75%, estando dentro do esperado.

Palavras-chave: Aprendizado de Máquina; Imagens por Satélite; Classificação de Imagens; Regressão Logística; Sensoriamento Remoto.

[1]Graduanda em Engenharia Agronômica, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brasil – anaarantes@usp.br
[2]Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brasil – silva.marcelo@usp.br
[3]Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brasil – analuciano@usp.br

P58

Um Estudo Sobre os Impactos de Algumas Medidas Sanitárias na Incidência e Mortalidade por COVID-19

Maria Sílvia de Assis Moura[1]; Rhayani Aparecida Paiuta[2]

Desde março de 2020 o Brasil, assim como o mundo todo, vem enfretando uma grande pandemia causada pela COVID-19 (SARS-CoV-2). A COVID-19 é uma infecção respiratória com elevada transmissibilidade e por esse motivo governantes do mundo todo tomaram medidas sanitárias para controlar os casos da doença e assim frear a pandemia. O *Lockdown* foi uma medida sanitária utilizada em algumas cidades para evitar a circulação de pessoas nas ruas e assim reduzir a contaminação de pessoas com a doença. Além disso, depois de quase um ano do início da pandemia alguns países começaram a aplicar vacinas também a fim de diminuir os impactos da pandemia. Diante da incerteza sobre o que de fato seria eficaz para o controlar o números de casos e de mortes por COVID-19, pesquisadores e pessoas do senso comum começaram a ter interesse sobre essas medidas sanitárias e começaram se questionar se o *Lockdown* ou se o início da vacinação ajudou no controle da pandemia. Nesse contexto, este trabalho estudou os efeitos do *Lockdown* e da vacinação no número de casos novos e no número de mortes por COVID-19. Para incorporar o efeito das medidas sanitárias citadas acima na série do número de casos e na série do número de mortes, utilizamos análise de intervenção com base em modelos da classe ARIMA e modelos de suavização exponencial. Verificamos que o *Lockdown* reduziu o número de casos de COVID-19 e que o início da vacinação colaborou para a redução de casos e óbitos.

Palavras-chave: Análise Intervenção; ARIMA; COVID-19; *Lockdown*; Suavização Exponencial.

[1]Departamento de Estatística – UFSCar, São Carlos – msilvia@ufscar.br
[2]Departamento de Estatística – UFSCar, São Carlos – rpaiuta@estudante.ufscar.br

P59

Comparação de Diferentes Métodos para Simulação de Amostras com Erros de Ranqueamento em Amostragem por Conjuntos Ordenados

Vinicius Ricardo Riffel[1]; Cesar Augusto Taconeli[2]

Devido à sua eficiência, a amostragem por conjunto ordenados (ACO) é particularmente útil quando não é possível coletar grandes amostras. Esse método de amostragem se baseia na ordenação dos indivíduos dispostos em pequenos conjuntos usando algum critério de baixo custo. Esse critério pode ser baseado em uma inspeção visual, no julgamento de um especialista ou em outra variável que que esteja relacionada com a variável de interesse. Essa outra variável usada para fins de ordenamento leva o nome de variável concomitante. Quando não há erros no ranqueamento, dizemos que a ordenação é perfeita, caso contrário, dizemos que ela é imperfeita. É comum que propriedades de estimadores em amostras obtidas por ACO sejam verificadas por meio de simulação, e portanto deve-se definir um método para induzir os erros de ranqueamento na simulação das amostras. É incomum encontrar na literatura de ACO uma escolha criteriosa desses métodos. No presente trabalho, através de um estudo de simulação verificamos os impactos dos diferentes métodos para simulação de amostras com ordenação imperfeita na estimação da média populacional. Para tal, utilizamos diversas distribuições de probabilidade, tamanhos amostrais e diferentes delineamentos amostrais baseados em conjuntos ordenados. Os resultados mostram que a escolha do método para indução de erros de ordenação pode ter um impacto significativo nas conclusões do estudo. Também, propomos um novo método para a indução dos erros de ranqueamento e uma metodologia empírica para a verificação da qualidade de cada um dos métodos aplicado a dezenas de conjuntos de dados reais.

Palavras-chave: Extreme Ranked Set Sampling; Percentile Ranked Set Sampling; Erros de Ranquamento; Estimação.

[1]1Departamento de Estatística, Universidade Federal do Paraná – viniciusriffel@ufpr.br
[2]1Departamento de Estatística, Universidade Federal do Paraná – taconeli@ufpr.br

P60

Modelo de Regressão Quantílico com Base na Distribuição Dagum Exponencial Generalizada Exponencializada

Vitor Bernardo Silveira Pereira[1]; Cleber Bisognin[2]; Laís Helen Loose[3]

A distribuição EGED foi proposta por Suleman Nasiru, Peter N. Mwita e Oscar Ngesa em 2019 e é utilizada para modelagem de dados contínuos e positivos, com aplicações nas áreas de análise de confiabilidade, finanças, ciências atuariais entre outros. A menos de nosso conhecimento, modelos de regressão para variáveis com distribuição EGED não foram amplamente explorados. Nesse sentido, o objetivo deste trabalho é propor um modelo de regressão para modelagem dos quantis da distribuição EGED. Seja $\boldsymbol{Y} = (Y_1, \cdots Y_n)^\top$ variáveis aleatórias independentes, onde cada Y_t, para $t = 1, \cdots, n$ possui função densidade de probabilidade reparametrizada EGED com quantil μ_t e parâmetros $(\alpha, \delta, \sigma, \lambda, \eta, \gamma)$ não negativos, onde $\sigma, \delta, \lambda, \eta$ e γ são parâmetros de forma e α o parâmetro de escala. Considerando essa reparametrização temos a possibilidade de incluir uma estrutura de regressão para modelagem do quantil por meio da relação $g(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ é o vetor de parâmetros da regressão ($\boldsymbol{\beta} \in \mathbb{R}^{k+1}$) e $\boldsymbol{x}_t = (x_{t0}, x_{t1}, \cdots, x_{tk})^\top$ são observações de $k+1$ covariáveis ($k+1 < n$), as quais são supostamente fixas e conhecidas e $g : \mathbb{R}^+ \to \mathbb{R}$ é uma função de ligação duas vezes diferenciável e estritamente monótona. A estimação do vetor de parâmetros $\boldsymbol{\theta} = (\delta, \sigma, \lambda, \eta, \gamma, \boldsymbol{\beta}^\top)^\top$ do modelo proposto será realizada utilizando os estimadores de máxima verossimilhança (EMV), onde a função de log-verossimilhança de $\boldsymbol{y}$ é dada por $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{t=1}^{n} \ell_t(\alpha, \delta, \sigma, \lambda, \eta, \gamma, \mu_t)$, μ_t é o quantil da observação t e $\ell_t(\alpha, \delta, \sigma, \lambda, \eta, \gamma, \mu_t) = \log(\alpha\lambda\sigma\delta\eta\gamma) - (\delta+1)\log(y_t) - (\sigma+1)\log(z_t) + (\gamma+1)\log\left(1 - z_t^{-\sigma}\right) + (\eta-1)\log\left[1 - \left(1 - z_t^{-\sigma}\right)^{\gamma}\right] + (\lambda-1)\log\left\{1 - \left[1 - \left(1 - z_t^{-\sigma}\right)^{\gamma}\right]^{\eta}\right\}$, com $z_t = (1 + \alpha y_t^{-\delta})$ e a reparametrização $\alpha = \mu_t^{\delta}\{[(1 - (1 - (1 - \tau^{\frac{1}{\lambda}})^{\frac{1}{\eta}})^{\frac{1}{\gamma}})^{-\frac{1}{\sigma}} - 1]\}$, com $\tau \in (0, 1)$. Foram realizadas simulações de Monte Carlo para avaliar as propriedades dos EMV. Os resultados indicaram que os estimadores são assintoticamente não viesados, consistentes e normalmente distribuídos.

Palavras-chave: Regressão Quantílica, Estimação de Máxima Verossimilhança, Simulações de Monte Carlo

[1]Departamento de Estatística, UFSM – vitorpereira3115@gmail.com
[2]Departamento de Estatística, UFSM – cleber.bisognin@ufsm.br
[3]Departamento de Estatística, UFSM – lais.loose@ufsm.br

P61

Análise das Séries Temporais do Nível do Mar de Estações Maregráficas Localizadas Nordeste Brasileiro

Nilton de Souza Ribas Júnior[1]; Taíze da Silva Sousa[2]; Everaldo Freitas Guedes[3]; Aloisio Machado da Silva Filho[4]

As mudanças climáticas e o consequente avanço do nível do mar vem preocupando há algum tempo pesquisadores e governantes na criação de medidas que ajudem a minimizar os impactos causados por esses fenômenos. Além disso, o nível do mar tem fundamental importância na engenharia quanto a determinação do datum altimétrico de um país. Segundo o quinto relatório apresentado pelo *Intergovernmental Panel on Climate Change*-IPCC a expansão térmica dos oceanos e o derretimento de geleiras têm sido os contribuintes dominantes para o aumento do nível médio do mar global do século 20 (IPCC, 2013). Portanto, visando contribuir com pesquisas relacionadas ao tema pretende-se neste estudo estimar a tendência e autocorrelação das séries temporais diárias do nível do mar de estações maregráficas localizadas no nordeste brasileiro (Salvador e Fortaleza) no período de 2004 a 2020. Para tal, será utilizado principalmente o método conhecido na literatura como *Detrended Fluctuation Analysis*-DFA (Peng, et al, 1994) capaz de avaliar autocorrelação em séries não estacionárias. Por opção metodológica, analisamos as séries temporais considerando as seguintes medidas durante o dia: nível mínimo, máximo, mediano e médio. Nossos achados constataram um padrão de comportamento entre as séries temporais do nível do mar (mínima, máxima, mediana e média) em termos de autocorrelação nas estações maregráficas. Tendo como alicerce o método DFA, as séries temporais do nível mínimo e máximo foram classificadas como antipersistente e a mediana e média como persistente, independentemente da estação maregráfica avaliada. Foi possível também, via DFA, identificar componentes sazonais em diferentes escalas de tempo.

Palavras-chave: Nível Médio do Mar; Dados Meteorológicos; Séries Temporais; Autocorrelação.

[1]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Ciências Ambientais, Feira de Santana-BA – niltonribasjr@gmail.com
[2]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Modelagem em Ciências da Terra e do Ambiente, Feira de Santana-BA – taize.sousa04@gmail.com
[3]Universidade Federal do Recôncavo da Bahia – efgestatistico@gmail.com
[4]Universidade Estadual de Feira de Santana, Departamento de Ciências Exatas, Programa de Pós-graduação em Ciências Ambientais, Feira de Santana-BA – aloisioestatistico@uefs.br

P62

Bootsrapping em Regressão Logística

Carla Patrícia de Carvalho Oliveira[1]; Albaro Ramon Paiva Sanz[2]; Liciana Vaz de Arruda Silveira[3]

As aplicações atuais das distribuições da estatística de teste da amostra aleatória, aplicadas nas suposições das distribuições utilizadas nas amostras por reamostragem, requerem um elevado desempenho computacional para a validação da suposta distribuição escolhida. O Bootstrap é um método computacional que serve para estimar estimativas, estimar a variabilidade de estimadores para uma estatística de interesse, estimar coeficientes e erro padrão do modelo de regressão logística, fazendo uma reamostragem dos dados observados. Neste trabalho, apresentamos um ajuste no modelo clássico de regressão logística e utilizamos o bootstrap paramétrico e não paramétrico, para estimar o intervalo de confiança dos parâmetros para o modelo logístico e odds ratio. Comparando os métodos bootstrap, percebemos que foram semelhantes às da regressão logística clássica. O modelo proposto foi aplicado em um conjunto de dados reais obtido em 2018 do SINAN da FMS - Prefeitura Municipal de Teresina (PI).

Palavras-chave: Modelos Lineares Generalizados; Bootstrap Paramétrico; Bootstrap Não Paramétrico; Regressão Logística; Intervalos de Confiança.

[1]1Doutoranda em Biometria – IBB/UNESP, Brasil – carla.patricia@unesp.br.
[2]Instituto de Estudios Científicos y Tecnológicos – IDECY/Venezuela – albaropaiva@gmail.com.
[3]Docente - IBB/UNESP, Brasil – liciana.silveira@unesp.br.

P63

Previsão de Consumo de Medicamentos Durante a Pandemia de COVID-19 Utilizando Técnicas de Bootstrap

Adenilso da Silva Simão[1]; Flaviane Louzeiro da Silva[2]; Maristela Oliveira dos Santos[3]; Fabio Ricardo Carrasco[4]; João Soares De Campos Junior[4]; Kleber José Maximiano Soares[4]; Thiago Luiz de Russo[4,5]; Renata Pedrolongo Basso Vanelli[4]; Cibele Maria Russo[6];

Neste trabalho propomos um sistema de controle de estoque de medicamentos utilizados durante a pandemia de COVID-19 em um hospital público, com base em informações históricas e situação epidemiológica na cidade de São Carlos, SP. Utilizamos métodos não-paramétricos baseados em bootstrap (Efron, 1982) para prever o consumo de cada medicação nos próximos doze meses, com base no consumo registrado nos meses anteriores (dados históricos). São consideradas restrições impostas pela validade dos produtos e a possibilidade do medicamento substituir ou ser substituído por outros de mesmo princípio ativo ou efeitos similares. Como resultado, o sistema pode auxiliar na tomada de decisão para a compra de medicamentos com previsão de esgotar e evitar a aquisição de medicamentos que tem uma previsão de sobra. O sistema permite a seleção de diversos perfis de probabilidade para os meses anteriores, podendo assim refletir diferentes características de uso. Além disso, para auxiliar na tomada de decisão, diversos filtros estão disponíveis para que o especialista possa considerar diferentes cenários, tais como, medicamentos com alta probabilidade de se esgotar e sem ata de compra válida. Deve-se investigar outros métodos de previsão baseados em reamostragem, os quais serão comparados com possibilidade de inclusão no sistema desenvolvido.

Palavras-chave: Previsão de Demanda; Bootstrap; Simulação.

[1]Departamento de Sistemas de Computação, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – adenilso@icmc.usp.br

[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – flaviane.silva@usp.br

[3]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – mari@icmc.usp.br

[4]Hospital Universitário da UFSCar - Empresa brasileira de serviços hospitalares (Ebserh), São Carlos, SP – Fabio.Carrasco@ebserh.gov.br, joao.campos@ebserh.gov.br, kleber.soares@ebserh.gov.br, thiago.russo@ebserh.gov.br, renata.vanelli@ebserh.gov.br

[5]Departamento de Fisioterapia, Universidade Federal de São Carlos, São Carlos, SP – russo@ufscar.br

[6]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – cibele@icmc.usp.br

P64

Métodos para Mapeamento de QTL Através de Marcadores Tipo SNP: Uma Revisão

Lara Midena João[1]; Daiane Aparecida Zuanetti[1]

O mapeamento de regiões no genoma associadas a traços quantitativos (QTLs) através de marcadores genéticos do tipo SNP tem sido um dos problemas centrais em Genética e Biologia Molecular e vários métodos de detecção e identificação de QTLs tem sido propostos na literatura. Neste trabalho, três diferentes metodologias foram aplicadas e comparadas nos dados GAW17 (sem estrutura familiar) quanto ao seus desempenhos em identificar corretamente SNPs relevantes e reguladores de um traço quantitativo. São elas: teste de significância do coeficiente de regressão associado a cada SNP em um modelo de regressão linear simples com e sem a correção de Bonferroni no nível de significância, LASSO e SPLS, do inglês *sparse partial least squares regression*, que faz a colapsagem das informações genéticas dos SNPs através de alguns poucos fatores latentes. A fim de comparar o desempenho dessas metodologias, utilizamos a sensibilidade e a especificidade como métricas e notamos que o LASSO e o teste com nível de significância de 5% no modelo de regressão linear simples apresentam os melhores resultados, uma vez que equilibram valores relativamente altos de sensibilidade e especificidade. O LASSO, por sua vez, também identifica SNPs influentes mais raros. Os dados GAW17 se tratam de dados simulados para 697 indivíduos e mais de 24000 marcadores genéticos tipo SNP, sendo a maior parte deles raro, com frequência muito baixa do alelo menor. Dessa maneira, o problema em pauta se trata de uma questão de seleção de variáveis em dados de alta dimensão e na presença de muitas covariáveis com baixa variabilidade.

Palavras-chave: Dados Independentes GAW17; Seleção de Variáveis; Teste de Significância; LASSO; SPLS.

[1]Departamento de Estatística, Universidade Federal de São Carlos - lara.midena@estudante.ufscar.br

P65

Aplicação da Teoria de Valores Extremos no Estudo de Dados da Covid-19: Análise Probabilística do Número de Novos Óbitos Diários

Ana Carolina Matiussi[1]; Gilberto Rodrigues Liska[2]; Carolini Stefani Ramos[3]; Daniel Wagner dos Santos Araújo[4]

A pandemia de COVID-19 causada pelo novo Coronavírus (Sars-CoV-2) é preocupante em todo mundo e tem desafiado pesquisadores e gestores a encontrar medidas que evitem o colapso do sistema de saúde pública. O desenvolvimento de métodos matemáticos no estudo da doença é fundamental para impedir uma possível superlotação do sistema de saúde, bem como a escassez de insumos para a população. Nesse sentido, o presente projeto tem como objetivo, analisar o número de óbitos decorrentes da COVID-19 no Brasil e, com base nesses dados, utilizar a Teoria de Valores Extremos (TVE) para avaliar a probabilidade de eventos extremos ocorrerem futuramente, investigando o seu comportamento durante o tempo de disseminação. Os dados partiram de publicações oficiais e diárias do *Our World in Data*, ligada à Universidade de Oxford - Reino Unido. Será adotada uma abordagem estocástica no estudo de comportamentos extremos, usando a modelagem TVE para calcular estimativas posteriores dos parâmetros de interesse associados a um conjunto de variáveis. A partir dos resultados parciais, para a metodologia de Blocos Máximos e para a distribuição Gumbel, a uma probabilidade de 66% dos números de óbitos serem maiores que 1.000 e 33% de probabilidade de ser maior que 5.000. Já para a metodologia dos Picos Acima do Limiar e para a distribuição Exponencial, a uma probabilidade de 18% dos números de óbitos serem maiores que 1.000 e 2% de chances de ser maior que 5.000. Logo, foi possível descrever probabilidade de ocorrência mensal de óbitos de Covid-19 e seu valor máximo esperado em duas metodologias da TVE.

Os resultados apresentados serão publicados com menção do projeto FAPESP de auxílio à participação.

[1]Discente do bacharelado em Biotecnologia, UFSCar-CCA – anamatiussi@estudante.ufscar.br

[2]Professor do Departamento de Tecnologia Agroindustrial e Sócio economia Rural, UFSCar-CCA — gilbertoliska@ufscar.br. Os resultados apresentados serão publicados com menção do projeto FAPESP de auxílio à participação.

[3]Aluno da Escola Estadual Profa. Maria Rosa Nucci Pacífico Homem, Rua Marcos Freire, s/nº, Parque das Árvores, Araras, SP, contemplado com bolsa PIBIC-EM – ramoscarol755@gmail.com

[4]Aluno da Escola Estadual Profa. Maria Rosa Nucci Pacífico Homem, Rua Marcos Freire, s/nº, Parque das Árvores, Araras, SP, contemplado com bolsa PIBIC-EM – daniel14wagner@gmail.com

Palavras-chave: Sars-CoV-2; Eventos Extremos; Saúde Pública; Probabilidade; Distribuição Gumbel; Distribuição Exponencial.

P66

Modelos de Séries Temporais Semiparamétricos com Fator Latente

Gisele de Oliveira Maia[1]; Wagner Barreto de Souza[1]; Fernando de Souza Bastos[1]; Hernando Ombao[1].

Introduzimos uma classe de modelos de séries temporais semiparamétricos assumindo uma abordagem de quase-verossimilhança conduzida por um processo latente. Mais especificamente, dado o processo latente, apenas especificamos a média e variância condicionais das séries temporais e utilizamos uma abordagem de quase-verossimilhança para estimar os parâmetros relacionados à média. Essa metodologia proposta possui três características marcantes: (i) nenhuma forma paramétrica é assumida para a distribuição condicional das séries temporais, dado o processo latente; (ii) capaz de modelar séries temporais não-negativas, contagens, limitadas/binárias e com valores reais; (iii) não se assume que o parâmetro de dispersão seja conhecido. Além disso, obtemos expressões explícitas para os momentos marginais e para a função de autocorrelação das séries temporais, para que o método de momentos possa ser empregado para estimar o parâmetro de dispersão e também os parâmetros relacionados ao processo latente. Resultados simulados com o objetivo de verificar o procedimento de estimação proposto são apresentados. Procedimentos de previsão são propostos e avaliados em dados simulados e reais. A análise de dados reais sobre séries temporais do número de internações em um hospital devido à asma e insolação total ilustram o desempenho de nossa metodologia em situações práticas.

Palavras-chave: Série Temporal Limitada; Processo Gaussiano; Análise de Regressão; Processo Gama Deslocada; Estimação por Quase-verossimilhança.

[1]Universidade Federal de Minas Gerais – UFMG

P67

Estimação das Habilidades das Equipes do Brasileirão de 2019 por Meio da Versão Dinâmica do Modelo de Bradley-Terry

Juliana Sena de Souza[1]; Márcia Helena Barbian[1,2]

Um dos esportes mais populares do mundo é o futebol, que possui alta quantidade de torcedores, admiradores e simpatizantes. Mesmo estando atrás de esportes como o beisebol e o basquete em termos de análises sofisticadas e acesso a dados mais complexos, tem feito com que aos poucos o futebol venha criando espaços dentro dos clubes para analistas utilizarem os dados disponíveis para tomadas de decisões. Entre os principais objetivos dessas análises esportivas está o interesse na predições acerca dos resultados das partidas, de modo que as técnicas disponíveis na literatura para predição do desfecho do jogo são amplas e abordam diferentes métodos. Apesar de alguns autores considerarem que a incerteza a respeito desse resultado tenham perdido sua relevância à audiência do público, que evoluiu para preferências que também dizem respeito ao talento dos jogadores em campo, o aumento na competitividade nos mercados de apostas, faz com que esportes como o futebol, que possuem muita incerteza quanto ao resultado do jogo, sejam atrativos. Logo, há uma maior demanda por modelos preditivos de classificações para previsão dos resultados das partidas. Nesse trabalho, a estimação da habilidade das equipes do Brasileirão de 2019 é feita através de um ajuste com base nos resultados de todas as partidas de uma temporada, utilizando uma versão dinâmica do modelo de Bradley Terry, o modelo busca captar a evolução temporal das habilidades das equipes em termos de jogos dentro e fora de casa utilizando médias móveis exponencialmente ponderadas.

Palavras-chave: Futebol; Habilidades; Modelo de Bradley Terry Dinâmico; Predição.

[1]Programa de Pós Graduação em Estatística – UFRGS – Porto Alegre - julianass.estatistica@gmail.com
[2]Departamento de Estatística – UFRGS, Porto Alegre – mhbarbian@gmail.com

P68

Modelagem de Rating de Crédito por Meio da RLMO

Kévin Allan Sales Rodrigues[1]; Dafne Martins do Prado Sousa[2]; Silvia Nagib Elian[3]

Neste trabalho utilizamos a regressão logística multivariada ordinal para modelar rating de crédito de empresas brasileiras. O objetivo é fornecer um modelo relativamente simples que permita ao investidor avaliar o rating de crédito de empresas de modo simples e rápido sem ter pleno domínio de conhecimentos contábeis ou econômicos. O modelo obtido no trabalho só depende de 3 variáveis explicativas: margem líquida em 12 meses, alavancagem financeira em 12 meses e CAPEX vs ativo total.

Palavras-chave: Risco de Crédito; Regressão Logística Ordinal; Finanças; Ciência de Dados Aplicada a Finanças.

[1]IME-USP, São Paulo/SP – kevin@usp.br
[2]IME-USP, São Paulo/SP – dafne.martins@usp.br
[3]IME-USP, São Paulo/SP – selian@ime.usp.br

P69

A Misspecification Test for Beta Prime Regression Models

Kleber Henrique dos Santos[1]; Tarciana Liberal Pereira[2]; Tatiene Correia Souza[3]; Marcelo Bourguignon[4]

The beta prime regression model is an alternative to the generalized linear models and useful to model positive asymmetric data. In this paper, we propose two general misspecification tests based on the RESET test for beta prime regression models with varying precision. In the first test, we add the testing variable in the mean submodel, whereas the second test focuses on adding the testing variables in all submodels. We conduct an extensive Monte Carlo simulation study to evaluate the performance of the proposed tests in finite sample size in terms of their sizes and powers, thus obtaining information about the best combination of test statistics and testing variables to perform the proposed tests. We also present and discuss two empirical applications to show the applicability and importance of the proposed tests.

Palavras-chave: Misspecification Test; Beta Prime Regression Model; Incorrect Link Function; Monte Carlo Simulation; Omitted Variables.

[1]Departamento de Estatística, Universidade Federal da Paraíba – kleber.statistic@gmail.com
[2]Departamento de Estatística, Universidade Federal da Paraíba – tarcianalp@gmail.com
[3]Departamento de Estatística, Universidade Federal da Paraíba – tatiene@de.ufpb.br
[4]Departamento de Estatística, Universidade Federal do Rio Grande do Norte – m.p.bourguignon@gmail.com

COMUNICAÇÕES PÔSTERES 2

P70

A Distribuição Gompertz Unitária Inflacionada em Zero ou Um

Bruna Freitas dos Santos[1]; Laís Helen Loose[2]

Nos últimos anos novas distribuições de probabilidade foram propostas, em especial no contexto de dados unitários, como taxas, índices e proporções. Uma recente proposta é a distribuição Gompertz Unitária, desenvolvida por Mazucheli, Menezes e Dey (2019). A distribuição proposta é adequada para modelagem de dados contínuos e restritos ao intervalo unitário, além de apresentar ajustes mais adequados, em dados assimétricos, que as distribuições mais comumente utilizadas e conhecidas, como as distribuições Beta e Kumaraswamy. Em dados do tipo taxas e proporções valores zeros ou uns podem ser frequentemente observados, nesse sentido, o objetivo do presente trabalho é propor a distribuição Gompertz unitária inflacionada em zero ou um, que possibilitará modelar taxas e proporções na presença de zeros ou uns. Inicialmente reparametrizamos a densidade Gompertz unitária em termos do quantil, a partir desta reparametrização obtivemos a função densidade de probabilidade da distribuição Gompertz unitária inflacionada. Para isso, foi necessário combinar duas distribuições, uma distribuição discreta para a inclusão da inflação em zero ou um e a outra contínua, adequada para os dados restritos ao intervalo unitário $(0, 1)$. Na estimação dos parâmetros do modelo utilizamos os estimadores de máxima verossimilhança (EMV) e a fim de avaliar as propriedades dos estimadores do modelo proposto utilizamos simulações de Monte Carlo. A implementação computacional foi desenvolvida utilizando o *software R* e para obtenção das estimativas utilizamos o método BFGS, por meio da função `optim`. Os resultados das simulações indicam que os EMV para os parâmetros da distribuição proposta apresentam boas propriedades, são assintoticamente não viesados, consistentes e normalmente distribuídos.

Palavras-chave: Distribuição Gompertz Unitária; Reparametrização; Quantil; Inflacionada; Dados Unitários.

[1]Departamento de Estatística, UFSM – bruna.freitas@acad.ufsm.br
[2]Departamento de Estatística, UFSM – lais.loose@ufsm.br

P72

Modelagem Espacial das Chuvas Intensas no Estado da Paraíba

Larissa da Silva Souza[1]; Elias Silva de Medeiros[2]; Carolina Cristina Bicalho[3]

O Nordeste Brasileiro (NEB) é mundialmente conhecido pela sua escassez de recursos hídricos, mas esse fator não ocorre em toda sua extensão territorial, a região apresenta como característica irregularidade em acumulados de precipitações, o que é decorrência de vários fatores, tais como as características ambientais e de sistemas atmosféricos. Assim, faz-se necessária uma análise estatística, com o intuito de se ter conhecimento dos níveis de retorno da precipitação em uma região, para a prevenção de desastres naturais causados por chuvas intensas. O objetivo dessa pesquisa foi modelar a precipitação máxima diária no estado da Paraíba, o qual faz parte da NEB. Os dados foram obtidos através do portal da Agência Executiva de Gestão das Águas (AESA), sendo composto por 238 estações pluviométricas, abrangendo o período de 27 anos (1994 a 2020). Para a modelagem estatística foi ajustada a distribuição Gumbel, por meio do método da máxima verossimilhança. Os resultados apontaram que essa distribuição foi adequada para modelagem dos dados do primeiro semestre ano. As estimativas para os níveis de retorno apontam uma grande heterogeneidade de precipitação máxima, com períodos chuvosos distintos em cada mesorregião da Paraíba, sendo a Mata Paraibana de abril a junho e Sertão de janeiro a março, sendo estas regiões, tipicamente, mais chuvosas. As informações obtidas fornecem informações dos níveis de retorno da precipitação para possíveis empregos de políticas públicas visando a diminuição dos impactos causados pelas chuvas intensas.

Palavras-chave: Software R; Gumbel; Nordeste.

[1]Faculdade de Ciências Exatas e Tecnologia, UFGD – larissa.souza040@academico.ufgd.edu.br
[2]Faculdade de Ciências Exatas e Tecnologia, UFGD – eliasmedeiros@ufgd.edu.br
[3]Departamento de Matemática, UEMS – carolinabicalho@gmail.com

COMUNICAÇÕES PÔSTERES 2

P73

Detecção de Anomalias, Interpolação e Previsão em Tempo Real de Séries Temporais para Operação de Reservatórios e Distribuição de Água

Palavras-chave:

P74

Detecção Offline de Pontos de Mudança para Dados Binários Via Métodos de Regularização

Lucas de Oliveira Prates[1]; Florencia Graciela Leonardi[2].

Em análise de séries temporais, o problema de detecção de pontos de mudança consiste em estimar os tempos nos quais a distribuição de probabilidade sofre alguma alteração. Sob à hipótese de que os dados têm distribuição Bernoulli, o problema pode ser visto como estimar os tempos nos quais o parâmetro de probabilidade se altera. Neste trabalho, apresentaremos métodos estatísticos para estimar o número e a localização dos pontos de mudança quando os dados têm distribuição Bernoulli. Os métodos escolhidos foram verossimilhança penalizada, Fused LASSO e métodos baseados em validação cruzada. Provamos a consistência de alguns dos métodos propostos, e fornecemos um estudo de simulação para comparação de modelos. Por fim, aplicamos os modelos no problema de identificação de regiões de homozigose em arrays de SNPs.

Palavras-chave: Detecção de Pontos de Mudança, Regularização, SNPs.

[1]Departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo – lucasdelprates@gmail.com

[2]Departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo – florencia@usp.br

P75

Análise de Risco e Modelos de Classificação para Óbitos por COVID-19 no Estado de São Paulo

Francisco Rosa Dias de Miranda[1]; Lucas Roberto de Oliveira Lopes[2]; Thaís Parron Alves[3]; Cibele Maria Russo[4]

Neste trabalho investigamos a associação entre a ocorrência do óbito por COVID-19 e a incidência de doenças preexistentes ou comorbidades em pacientes hospitalizados com o vírus, no Estado de São Paulo. Por meio de ferramentas estatísticas e de *machine learning*, os riscos relativos relacionados ao sexo e idade dos pacientes irem a óbito foram estudados, e um modelo logístico de resposta binária foi ajustado, com o propósito de auxiliar no dimensionamento e prevenção dos impactos da pandemia de COVID-19. Utilizamos ferramentas como a análise de dados categorizados, de risco relativo, razões de chances, assim como modelos para classificação, obtidos a partir do banco de dados de hospitalizações por COVID-19 da Fundação Seade (SEADE, 2022), para medir tais associações. Foi encontrado que, com exceção da Síndrome de Down, indivíduos portadores das comorbidades registradas no banco de dados apresentam risco aumentado para ocorrência de óbito como consequência da infecção por SARS-CoV-2. O mesmo acontece para indivíduos do sexo masculino, que têm maior risco de vir a óbito do que os do sexo feminino. Dentre as comorbidades as que possuíam os maiores riscos relativos foram as puérperas ou pessoas com obesidade. As análises foram implementadas utilizando pacotes scipy e PyCaret do Python. Este trabalho é parte do Grupo de Extensão PREDICT, do ICMC USP.

Palavras-chave: Modelo Linear Generalizado; Modelos de classificação, COVID-19; Comorbidades; Risco Relativo.

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – francisco.miranda@usp.br
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – lucas1308@usp.br
[3]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP – thaisparron@usp.br
[4]Departamento de Matemática Aplicada e Estatística, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, cibele@icmc.usp.br

P76

Amostragem Estratificada Otimizada para a Redução Amostral de Atributos Químicos do Solo

Tamara Cantú Maltauro[1]; Letícia Ellen Dal' Canton[2]; Luciana Pagliosa Carvalho Guedes[3]; Miguel Angel Uribe-Opazo[4]

A amostragem estratificada em áreas agrícolas consiste em delimitar a área em estratos com características semelhantes, permitindo assim, a aplicação de fertilizantes a taxa variável. Estes estratos podem ser definidos por métodos de agrupamentos e uma de suas vantagens é direcionar a determinação de uma futura amostragem do solo, com uma possível redução amostral. Desta forma, o objetivo deste trabalho foi obter configurações amostrais reduzidas usando amostragem estratificada e processo de otimização. Para isso, utilizaram-se atributos químicos do solo de uma área agrícola comercial, referentes ao ano-safra da soja (2013-2014). Primeiramente, foram avaliados três métodos de agrupamento para a obtenção dos estratos. Os estratos foram obtidos por meio de uma matriz de dissimilaridade que agrega a dependência espacial dos atributos. Dentro de cada estrato selecionou-se 50% dos pontos amostrais da configuração inicial, gerando uma configuração amostral reduzida. Essa escolha foi realizada por um processo de otimização (Algoritmo Genético), cuja eficiência foi avaliada com base na predição espacial, utilizando o somatório do índice de acurácia Exatidão Global, de todos os atributos do solo. Os resultados indicaram a divisão da área agrícola em dois estratos, considerando o método K-means. Para a maioria dos atributos do solo, quando se compararam as configurações amostrais original e a reduzida, observou-se uma similaridade nas estatísticas descritivas e uma baixa acurácia nas estimativas dos índices de acurácia, mostrando assim uma influência da redução amostral na caracterização da variabilidade espacial, a qual pode ser um indicativo de que houve uma redução drástica do tamanho amostral.

Palavras-chave: Agricultura de Precisão; Agrupamento; Matriz de Dissimilaridade Espacial Multivariada; Redução Amostral.

[1]Centro de Ciências Exatas e da Terra (CCET), Programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI), Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel, PR – tamara_ma02@hotmail.com
[2]CCET, PGEAGRI, UNIOESTE, Cascavel, PR – leticiacanton@hotmail.com
[3]CCET, PGEAGRI, UNIOESTE,, Cascavel, PR – luciana_pagliosa@hotmail.com
[4]CCET, PGEAGRI, UNIOESTE,, Cascavel, PR – mopazo@uol.com.br

P77

Finite Mixture of Birnbaum–Saunders Distributions Using The k-bumps Algorithm

Luis Benites[1]; Rocío Maehara[2]; Filidor Vilca[3]; Fernando Marmolejo-Ramos[4]

Mixture models have received a great deal of attention in statistics due to the wide range of applications found in recent years. This paper discusses a finite mixture model of Birnbaum–Saunders distributions with G components, which is an important supplement to that developed by Balakrishnan et al. (2011) who considered a model with two components. Our proposal enables the modeling of proper multimodal scenarios with greater flexibility for a model with two or more components, where a partitional clustering method, named k-bumps, is used as an initialization strategy in the proposed EM algorithm to the maximum likelihood estimates of the mixture parameters. Moreover, the empirical information matrix is derived analytically to account for standard error, and bootstrap procedures for testing hypotheses about the number of components in the mixture are implemented. Finally, we perform simulation studies to evaluate the results and analyze two real dataset to illustrate the usefulness of the proposed method.

Palavras-chave: Birnbaum–Saunders Distribution; EM Algorithm; k-bumps Algorithm; Maximum Likelihood Estimation; Finite Mixture.

[1]Departamento de Ciencias, Pontificia Universidad Católica del Perú, Lima, Perú – lben – itess@pucp.edu.pe
[2]Departamento de Ingeniería, Universidad del Pacífico, Lima, Perú – rp.maeharaa@up.edu.pe
[3]Departamento de Estatística, Universidade Estadual de Campinas, Brazil – fily@unicamp.br
[4]Centre for Change and Complexity in Learning, University of South Australia, Adelaide, Australia – Fernando.Marmolejo-Ramos@unisa.edu.au

P78

A Robust Approach to the Continuous-State Dynamic Bayesian Networks

Luiz E. S. Gomes[1]; Thais C. O. Fonseca[1]; Kelly C. M. Gonçalves[1]; Guilherme L. Oliveira[2]

This work proposes a robust approach to the continuous-state dynamic Bayesian networks. The usual inference method for this class is based on dynamic normal linear models. In contrast, our proposal is based on dynamic quantile linear models. Quantile regression quantifies the association of explanatory variables with a conditional quantile of a continuous outcome without assuming any specific conditional distribution. Thus, it models the quantiles, instead of the mean as done in the standard regression approach. In cases where either the assumptions for standard linear regression are violated or interest lies in the outer regions of the conditional distribution, quantile regression can explain dependencies more accurately than usual methods. Our method will be applicable in cases where there is interest in events at the 'bounds of probability' (low and high quantiles), the conditional distribution does not follow a known distribution, there are a lot of outliers in the conditional distribution, and in the case of heteroscedasticity. This method combined with dynamic Bayesian networks will be able to capture both non-linearities and complex cause-effect relationships. The outcome of this project is a robust modeling tool that will apply in many applications of Machine Learning and Artificial Intelligence.

Keywords: Bayesian Networks; Dynamic Models; Quantile Regression; Robustness.

[1]Federal University of Rio de Janeiro, Brazil
[2]Federal Center for Technological Education of Minas Gerais, Brazil

P79

Desempenho de Gráficos de Controle para Monitorar Variáveis Contínuas Duplamente Limitadas com Parâmetro Estimado

Letícia Garcez Corrêa da Costa[1]; Luiz Medeiros Araujo Lima–Filho[2]; Marcelo Bourguignon[3]

Em diversas situações práticas o interesse é modelar variáveis contínuas duplamente limitadas, tais como as taxas e proporções. Há circunstâncias em que essas taxas e proporções não resultam de um experimento de Bernoulli. Assim, quando a característica da qualidade é proveniente de números contínuos, a utilização do gráfico de controle p é inapropriada. Sabe-se que o desempenho dos gráficos de controle se deteriora significativamente quando os parâmetros são desconhecidos. No entanto, em situações reais, os parâmetros precisam ser estimados a partir de uma amostra finita (Fase I). Com isso, este trabalho tem como objetivo avaliar gráficos de controle do tipo Shewhart para monitorar variáveis contínuas duplamente limitadas quando os parâmetros são desconhecidos. Foram consideradas três distribuições contínuas definidas no intervalo (0,1), sendo elas: beta, gama unitária e simplex. Os gráficos de controle foram avaliados e comparados através de um extenso estudo de simulação de Monte Carlo, por meio da medida de desempenho que representa a média de amostras até que uma causa especial seja detectada (average run length - ARL). Os resultados sugerem que para obter um menor que 10%, faz-se necessário um tamanho de amostra mínimo de 200, para $\alpha = 0,0027$. Observa-se também que a performance dos gráficos de controle melhora com o aumento do tamanho da amostra na Fase I. O estudo traz contribuições importantes a área de Controle Estatístico de Processos quando o objetivo é modelar dados de processos contínuos no intervalo unitário padrão.

Palavras-chave: Gráfico de Controle; Parâmetros Desconhecidos; Variáveis Duplamente Limitadas; Taxas e Proporções.

[1]Departamento de Estatística – Universidade Federal da Paraíba – UFPB – leticiagaccosta@gmail.com
[2]Departamento de Estatística – Universidade Federal da Paraíba – UFPB – luizmalf@gmail.com
[3]Departamento de Estatística – Universidade Federal do Rio Grande do Norte – UFRN

P80

Seleção de ordem em modelos autorregressivos

Palavras-chave:

P81

Um Modelo Hierárquico da Teoria da Resposta ao Item Multidimensional com Aplicação na Percepção da Sustentabilidade Ambiental

Marcelo A. da Silva[1]; Jorge L. Bazán[2]; Ren Liu[3]; Edna Possan[4]; Silvana Vincenzi[5]

Os modelos hierárquicos da teoria de resposta ao item multidimensionais incorporam um traço geral e mais de uma dimensão de traços específicos por meio de diferentes estruturas latentes. Neste estudo, propomos inserir duas diferentes estruturas hierárquicas no conhecido modelo de Samejima. Essas estruturas hierárquicas consideram diferentes conceitos sobre a relação entre o traço latente geral e os específicos. Para estimar os parâmetros do modelo proposto, utilizamos uma abordagem Bayesiana através do algoritmo No-U-Turn Sampler, um método de Monte Carlo em cadeia de Markov. Um estudo de simulação foi conduzido para avaliar a recuperação de parâmetros do modelo em diferentes cenários. Os resultados indicaram que o algoritmo NUTS recupera adequadamente todos os parâmetros do modelo. Consideramos dados reais sobre a percepção de sustentabilidade de moradores da Bacia do Paraná III, no Brasil, avaliando através de uma característica geral e três dimensões de características específicas: econômica, ambiental e social. Os resultados obtidos sugerem que o modelo é adequado aos dados e, comparado com o modelo de Samejima unidimensional, o modelo proposto descreve melhor os dados, fornecendo informações úteis sobre as dimensões de traços específicos e a relação entre eles e o traço geral.

Palavras-chave: Bayesian Estimation; Sustainability Perception; Graded Response Model; Hierarchical MIRT Model; Multidimensional Item Response Theory

[1]Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, Brasil – silva.marcelo@usp.br
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil – jlbazan@icmc.usp.br
[3]Universidade da Califórnia, Merced, CA, USA – rliu45@ucmerced.edu
[4]Universidade Federal da Integração Latino-Americana, Foz do Iguaçu, Brasil – edna.possan@unila.edu.br
[5]Universidade Federal de Santa Catarina, Florianópolis, Brasil – sligie@globo.com

P82

Geração Aleatória e Correção Automática de Questões Através do R/exams.

Markus Chagas Stein[1]; Márcia Helena Barbian[1,2]; Rodrigo Citton Padilha dos Reis[1,3]; Lisiane Priscila Roldão Selau[1]

Planejar e executar avaliações é um dos grandes desafios no processo de ensino e aprendizagem para qualquer curso, na condução de turmas EAD esse aspecto também necessita de maior atenção, visto que é importante criar mecanismos avaliativos que evitem a repetição de questões entre os alunos e inibam potenciais cópias/plágios. Por fim, se destaca também a criação de questões e a sua correção, bem como a análise dos resultados e o retorno da avaliação ao estudante. Esse trabalho visa abordar o uso do pacote *exams* do software R na criação de questões para disciplinas de estatística. As questões são construídas de forma que os valores de cada avaliação sejam aleatórios, além disso, é possível facilitar o processo de elaboração da prova, por meio do sorteio de questões, da correção da avaliação da atividade de ensino e da disponibilização do gabarito. As questões criadas podem facilmente ser adaptadas ao ambiente virtual de aprendizagem adotado pelo docente. Quando as avaliações são impressas, é possível fazer a leitura óptica das provas, economizando muito tempo na correção das avaliações. O uso do *exams* torna escalável a aplicação e geração de avaliações, dessa forma o docente terá uma ferramenta para facilitar na condução de disciplinas.

Palavras-chave: R; Pacote Exams; Avaliação Educacional; Ensino.

[1]Departamento de Estatística - UFRGS, Porto Alegre – markus.stein@ufrgs.br
[2]Programa de Pós Graduação em Estatística – UFRGS – Porto Alegre – mhbarbian@gmail.com
[3]Programa de Pós-Graduação em Epidemiologia – UFRGS, Porto Alegre

P83

Desenvolvimento de uma Aplicação Web para Monitoramento e Predição de Taxas de Detecção de Câncer de Mama em Mulheres Rastreadas na Unidade de Prevenção do Hospital de Amor Barretos

Marco Antonio de Oliveira[1]; Maria Cecília Evangelista[2]; Thiago Buosi Silva[3]; Adriane Feijó Evangelista[4]

A Unidade de Prevenção do Hospital de Amor Barretos é uma porta de entrada para os casos novos de câncer na instituição, através do programa de rastreamento. A aplicação web (dashboard) tem sido utilizada para realizar a inspeção visual de indicadores em saúde ao longo do tempo (série temporal) e acredita-se ser plausível que um gestor tenha acesso a série histórica das taxas de detecção de maneira dinâmica através de uma aplicação web. Objetivo: Desenvolver uma aplicação web de monitoramento e predição da taxa de detecção de câncer de mama na Unidade de Prevenção. Materiais e métodos: A série histórica foi composta pelas taxas mensais de detecção no período de 2011 a 2019. As análises estatísticas executados no ambiente R v.4.0.0 com o ajuste do Modelo de Suavização Exponencial e a desenvolvimento da aplicação web através do pacote R-Shiny. Resultados: A aplicação web apresenta no Painel Principal algumas medidas descritivas, as estimativas e os gráficos de linhas da série histórica. A taxa de detecção foi 1,81%, a série apresentou uma tendência crescente, sem sazonalidade e a taxa média estimada para o ano de 2020 foi 1,94%. A aplicação web está disponível através do link: https://marcooliveiraha.shinyapps.io/HA_PrevencaoIV/. Conclusão: O produto pode se tornar uma ferramenta de gestão para inspeção visual da série histórica da taxa de detecção do câncer de mama nas Unidades de Prevenção do Hospital de Amor e permitir acesso às estimativas futuras.

Palavras-chave: Série Temporal, Dashboard, Câncer de Mama, Shiny

[1]Fundação PIO XII – Hospital de Câncer de Barretos – Núcleo de Epidemiologia e Bioestatística – Barretos/S – marco.oliveira@hcancerbarretos.com.br
[2]Fundação PIO XII – Hospital de Câncer de Barretos – Departamento de Prevenção – Barretos/SP – maria.evangelista@hcancerbarretos.com.br
[3]Fundação PIO XII – Hospital de Câncer de Barretos – Departamento de Prevenção – Barretos/SP – tbuosi@gmail.com
[4]Fundação PIO XII – Hospital de Câncer de Barretos – Centro de Pesquisa em Oncologia Molecular – Barretos/SP – adriane.feijo@gmail.com

P84

Comparison of Record Linkage Methods

Marcus André Alves Zimmermann Vieira[1]; Karoline Louise e Silva[2]

Record linkage is an important tool to enhance database integration. This is even more valuable in a scenario with more hefty budget cuts and a growing drop in response rate in traditional surveys. This strategy makes it possible to expand the crossing alternatives with variables not present in the original base. However, there are many different data pairing methods exposed in the literature. In this sense, the objective of this paper is to compare well-known methods of record linkage. The comparison was made in synthetic dataset. To compare the methods, it was adopted a quantitative approach based on the Precision, Recall, and F-Statistics metrics, using two comparison functions: Levenshtein and Jaro-Winkler. Among the six types of classifiers analyzed, the supervised methods had the best results.

Palavras-chave: Record Linkage; Data Cleaning; Comparison; Classification; Quality.

[1]PhD candidate and MsC in Population, Territory and Public Statistics at Nation School of Statiscal Sciences – marcusazimmermann@gmail.com

[2]MsC in Population, Territory and Public Statistics at Nation School of Statiscal Sciences – louise.ksm@gmail.com

P85

Análise do Padrão de Despesas Individuais de Homens e Mulheres Utilizando a POF 2017-2018

Maria Eduarda Campello Gallo[1]; Júlia Valles Marques[2]; Alinne de Carvalho Veiga[3]; Maria Salet Ferreira Novellino[4]

Estudar o padrão de consumo dos indivíduos é fundamental pois pode contribuir para o direcionamento de políticas públicas. Em particular, comparar esse padrão entre homens e mulheres, que foi o foco deste trabalho. Sendo assim, o objetivo dessa monografia foi verificar a existência de um padrão de despesas por sexo, incorporando as diferenças socioeconômicas de cada indivíduo e outras possíveis variáveis indicadas pela literatura. Para isso foi considerada como população de análise as pessoas de referência do domicílio de 18 anos ou mais e que declararam possuir rendimento. A análise se utilizou de dados da Pesquisa de Orçamentos Familiares (POF) de 2017-2018, realizada pelo IBGE, que é classificada como uma pesquisa amostral complexa. Sendo assim as informações do plano amostral foram incorporadas em todas as análises, por meio da utilização do pacote *survey*, do *software R*. Para possibilitar a análise das despesas individuais, estas foram divididas em 10 grupos de gastos, indicados pelo IBGE. Para cada grupo, ajustou-se as despesas positivas por meio de um modelo de regressão linear do logaritmo do gasto naquele grupo. Os resultados encontrados indicaram a existência de um padrão de gastos distinto para homens e mulheres, em que se identificou a influência da variável sexo para todos os grupos de gastos. Tal influência foi encontrada no efeito principal para todos os grupos de gastos, com exceção de *Recreação e Cultura* e *Vestuário*, nos quais foi encontrado uma relação distinta por sexo entre a despesa e a Renda do Trabalho.

Palavras-chaves: Pesquisa de Orçamentos Familiares; Plano Amostral Complexo; Análise de Gastos.

[1]ENCE/IBGE – dudacgallo@gmail.com
[2]ENCE/IBGE – juliavallesm@gmail.com
[3]ENCE/IBGE – alinne.veiga@ibge.gov.br
[4]ENCE/IBGE – salet.novellino@ibge.gov.br

P86

Analysis of Multinomial Data with Overdispersion: Diagnostics and Application

Maria Letícia Salvador[1]; Eduardo Eliás Ribeiro Junior[2]; César Augusto Taconeli[3]; Idemauro Antonio Rodrigues de Lara[4]

In agronomic experiments, the presence of polytomous variables is common, and the generalized logit model can be used to analyze these data. One of the characteristics of the generalized logit model is the assumption that the variance is a known function of the mean, and the observed variance is expected to be close to that assumed by the model. However, it is not uncommon for extra-multinomial variation to occur, due to the systematic observation of data that are more heterogeneous than the variance specified by the model, a phenomenon known as overdispersion. In this context, the present work discusses a diagnostic of overdispersion in multinomial data, with the proposal of a descriptive measure for this problem, as well as presenting a methodological alternative through the Dirichlet-multinomial model. The descriptive measure is evaluated through simulation, based on two particular scenarios. As a motivational study, we report an experiment applied to fruit growing, whose objective was to compare the flowering of adult plants of an orange tree, grafted on "Rangpur" lime or "Swingle" citrumelo, with as response variable the classification of branches into three categories: lateral flower, no flower or aborted flower, terminal flower. Through the proposed descriptive measure, evidence of overdispersion was verified, indicating that the generalized logit model may not be the most appropriate. Thus, as a methodological alternative, the Dirichlet-multinomial model was used, which proved to be more suitable to fit the data with overdispersion, by allowing the inclusion of an additional parameter to accommodate the excessive extra-multinomial dispersion.

Palavras-chaves: Model Selection; Dirichlet-Multinomial; Maximum Likelihood; Dispersion Index.

[1]University of São Paulo, Luiz de Queiroz College Agriculture, Department Exact Sciences, Piracicaba, São Paulo, Brazil – mariale_salvador@usp.br
[2]University of São Paulo, São Paulo, Department of Statistics, São Paulo, São Paulo, Brazil – jreduardo@ime.usp.br
[3]Federal University of Paraná, Exact Sciences Sector, Department of Statistics, Curitiba, Paraná, Brazil – taconeli@ufpr.br
[4]University of São Paulo, Luiz de Queiroz College Agriculture, Department Exact Sciences, Piracicaba, São Paulo, Brazil – idemauro@usp.br

P87

Avaliando Mudanças na Relação Entre o Nível Socioeconômico e o Desempenho dos Candidatos nos Processos Seletivos 2020 e 2021 da UFPA

Carla Eloiany Mata do Nascimento[1]; Maria Regina Madruga[2]; Heliton Ribeiro Tavares[3]

Devido à pandemia da Covid-19 as atividades escolares presenciais foram suspensas no início do ano de 2020, trazendo impactos na preparação dos candidatos que concorriam a vagas em instituições superiores para o ano de 2021. Esse impacto foi maior em candidatos de baixa renda, que tiveram mais dificuldade com o acesso à internet e computadores, necessários para as aulas remotas. Neste trabalho foi estimado um escore socioeconômico para os candidatos ao Processo Seletivo da Universidade Federal do Pará do ano de 2020 e 2021, a fim de identificar se houve mudanças na relação entre esse escore e o desempenho dos candidatos. Com base no questionário socioeconômico aplicado pela UFPA no ato da inscrição foi ajustado um modelo híbrido da Teoria da Resposta ao Item para estimar o escore, pois havia itens nominais e graduais no questionário. Os itens referentes à renda familiar líquida mensal do candidato, o tipo de estabelecimento onde estudou, acesso à internet e seu meio de transporte foram os que mais forneceram informação para a estimação do escore. Com base no Critério Brasil foram construídos grupos sociais para analisar o desempenho dos candidatos por grupo e ano de realização do PS-UFPA, e observou-se em ambos os anos, exceto em Ciências da Natureza no ano de 2020, um melhor desempenho dos candidatos de maior grupo social. Nas demais áreas notou-se uma queda no desempenho dos candidatos do ano de 2021, quando comparados aos de 2020 nos grupos sociais de menor nível socioeconômico.

Palavras-chave: Pandemia; Modelo Híbrido; Escore Socioeconômico; Desempenho.

[1]Faculdade de Estatística, UFPA, Brasil – carlaeloiany@gmail.com
[2]Faculdade de Estatística, UFPA, Brasil – madruga@ufpa.br
[3]Faculdade de Estatística, UFPA, Brasil – heliton@ufpa.br

P88

Efeito do Lockdown na Cidade de Araraquara, SP

Maria Sílvia de Assis Moura[1]

O enfrentamento à pandemia de COVID-19 levou governantes a tomarem diferentes medidas. Na cidade de Araraquara, SP, o prefeito declarou confinamento de maneira mais estrita que em outras cidades, por duas vezes. Este trabalho apresenta análise do número de casos registrados da doença por meio de séries temporais interrompidas para avaliar o impacto da adoção das medidas extermas. Os dados foram coletados na plataforma CORONAVIRUS BRASIL referentes à cidade de Araraquara e foram agrupados em semanas. As duas vezes que o lockdown foi imposto, as intervenções extremas tiveram resultado positivo, ou seja, após a declaração do confinamento, o número de casos na cidade teve diminuição significativa, semana após semanas.

Palavras-chave: COVID-19; *Lockdown*; Séries Temporais Interrompidas.

[1]Departamento de Estatística – UFSCar, São Carlos – msilvia@ufscar.br

COMUNICAÇÕES PÔSTERES 2

P89

Redes Neurais

Mariana Curi

Palavras-chave:

P90

Caracterização das Hospitalizações Infantis no Estado do Pará: Um Estudo com Técnicas Estatísticas e Computacionais

Júlio Henri Maciel Bezerra da Silva[1]; Marinalva Cardoso Maciel[2];
Vânia Cristina Campelo Barroso Carneiro[3]; Maria Regina Madruga Tavares[4]

A internação de crianças no ambiente hospitalar é um tema importante da saúde pública considerando que, em muitos casos tais doenças poderiam ser evitadas, pois são decorrentes de doenças sensíveis à atenção primária. Desse modo, identificar os fatores relacionados com as hospitalizações é importante por possibilitar o planejamento de ações e políticas públicas específicas que promovam a redução das internações evitáveis. O presente estudo tem por objetivo analisar dados do Estado do Pará, nos anos de 2010 e 2019 oriundos das bases do DATASUS a partir do Sistema de Informação Hospitalar do SUS (SIH), com informações sobre internação hospitalar em crianças menores de 6 anos de idade, objetivando caracterizar as internações por condições sensíveis à atenção primária (ICSAP) e identificar possíveis alterações ocorridas nos períodos. Para tanto foram utilizados testes de hipóteses, regressão logística e modelos de aprendizado de máquina (*Naives Bayes* e *Random Forest*). Os resultados evidenciaram uma redução significativa nas internações hospitalares na faixa etária avaliada, de 2010 para 2019, principalmente influenciada pela queda nas ICSAPs. Na modelagem verificou-se que as hospitalizações ICSAP estão mais relacionadas com crianças do sexo feminino, com idade superior a 1 ano de idade, que não utilizaram UTI e não evoluíram a óbito. Espacialmente, houve redução em Icsap em todas as mesorregiões do Estado. Em 2010, a mesorregião Sudoeste do Pará apresentava menor chance de ocorrência de ICSAP, enquanto em 2019 a mesorregião Sudeste se equiparou a essa última com bons resultados.

Palavras-chave: Estatística; ICSAP; Epidemiologia; Hospitalização infantil.

[1]Faculdade de Estatística – Universidade Federal do Pará – henrijulio2@gmail.com.br
[2]Faculdade de Estatística – Universidade Federal do Pará – marinalvamaciel@gmail.com
[3]Universidade Federal do Pará – vania_barroso@yahoo.com.br
[4]Faculdade de Estatística – Universidade Federal do Pará – madruga@ufpa.br.

P91

A Bayesian Approach for ZMPS-GARMA Model Applied to Influenza Count Data Time Series

Marinho G. Andrade[1]; Katiane S. Conceição[2]; Naline Ravishanker[3]

The paper aims to develop Zero-Modified (e.g., Inflated or Zero Deflated) models for time series with discrete data. The models introduced in this paper is an extension of Zero-Modified models of the Power Series family. The main advantage of the proposed model is the suitability of these models to fit time series data with both characteristics (zero-inflation and zero-deflation) present in the same time series. Inference methods based on the Bayesian approach, together with Hamiltonian Monte Carlo (HMC) techniques were considered. The paper also provides a relevant application for a real problem by modeling and forecast the series of notifications of the number of deaths from Influenza in the city of São Paulo, Brazil.

Palavras-chave: Hamiltonian Monte Carlo; Influenza notification; Power Series distribution; Zero-Modified model; ZMPS-GARMA models.

[1]Department of Applied Mathematics and Statistics, Institute of Mathematics and Computer Science, University of São Paulo, São Carlos/SP, Brazil – marinho@icmc.usp.br

[2]Department of Applied Mathematics and Statistics, Institute of Mathematics and Computer Science, University of São Paulo, São Carlos/SP, Brazil – katiane@icmc.usp.br

[3]Department of Statistics, University of Connecticut, Storrs/CT, United States of America – nalini.ravishanker@uconn.edu

P92

Detecção de Fraudes na Utilização de Cartões com o Uso da Técnica de Regressão Logística: Uma Aplicação em Situações de Desbalanceamento Severo

Mário Hissamitsu Tarumoto[1]; Vitória de Oliveira Silva[2]; Olga Lyda Anglas Rosales Tarumoto[1]

Com o passar dos anos, o número de fraudes em cartões de crédito e débito vem crescendo e as maneiras como os fraudadores atuam são inovadas diariamente. Isso se dá por conta da ascensão do uso de cartões como forma de pagamento, que acompanha o avanço da tecnologia. Para identificar as fraudes e conhecer os seus mecanismos, é necessário recorrer às estratégias, estudos e técnicas estatísticas que ajudarão a prever e detectar as ocorrências de fraudes. Uma situação comum nestes casos, é que nas bases de dados desta natureza, a proporção de fraudes é muito pequena comparada a não fraude, consequentemente, os dados se tornam desbalanceados e necessitam ser tratados. No presente trabalho foram utilizados os métodos de *Oversampling* e *Undersampling* para balancear os dados utilizados e a técnica de Regressão Logística para detectar transações realizadas em cartões de crédito e débito que possuem cunho fraudulento. Para a aplicação, foram utilizados dados sintéticos gerados por um simulador, o qual se baseia em uma amostra de dados reais. Observou-se um severo desbalanceamento dos dados, tendo em vista que apenas 1,3% da base, após os devidos tratamentos, eram transações fraudulentas. Assim, foram feitas três aplicações do modelo, sendo uma com os dados desbalanceados e as outras duas usando os métodos de balanceamento, e notou-se que o *Undersampling* foi o método que apresentou melhores resultados.

Palavras-chave: Detecção de Fraude; Dados desbalanceados; Regressão Logística; oversampling; undersampling.

[1]Departamento de Estatística – FCT/Unesp – mario.tarumoto@unesp.br
[2]Curso de Estatística – FCT/Unesp - vitoriaoliveira131997@gmail.com

P93

O Potencial de Dados de Telefonia Móvel para Análise de Movimento Pendular: Um Estudo de Caso a Partir de Niterói e de São Gonçalo

Mariza Rayanne da Silva Pereira[1]

A pesquisa que está em execução, como parte do desenvolvimento da dissertação de mestrado da autora, destaca–se por apresentar técnicas para o uso de big data, precisamente de telefonia celular. O estudo contribuirá para a criação de um protocolo para futuro uso de tais dados em pesquisas oficiais e busca mostrar que é possível realizar estudos sobre movimento pendular com dados de telefonia móvel, utilizando como ilustração os municípios de Niterói e São Gonçalo, ambos localizadas na Região Metropolitana (MP) no estado do Rio de Janeiro. Serão utilizadas técnicas de tratamento e análise da mobilidade na área de estudo, com o monitoramento do ponto de origem (residência) aos diferentes destinos ao longo do dia. Sendo possível verificar a demanda em relação ao deslocamento intraurbano, serão apresentadas as técnicas usadas para compor o quadro metodológico da pesquisa, como passo a passo da execução do tratamento e análise do banco de dados, com sua comparação com os indicadores sobre mobilidade urbana e movimento pendular calculados com dados do Censo 2010.

Palavras-chave: Mobilidade; Movimento Pendular; Big Data; Dados de Telefonia Móvel.

[1]Mestranda na Escola Nacional de Ciências Estatísticas (ENCE/IBGE), Rio de Janeiro - mariza_una@hotmail.com

P94

Long Memory in High Frequency Time Series Using Wavelets and Conditional Volatility Models

Mateus Gonzalez de Freitas Pinto[1]; Guilherme de Oliveira Lima Cagliari Marques[2]; Chang Chiann[3]

The presence of spikes or cusps in high-frequency return series might generate problems in terms of inference and estimation of the parameters in volatility models. For example, the presence of jumps in a time series can influence the sample autocorrelations, which can cause misidentification or generate spurious ARCH effects. On the other hand, these jumps might also hide the proper heteroskedastic behavior of the dependence structure of a series, leading to identification issues and a poorer fit of a model. We propose a method to separate jumps with wavelet shrinkage in high-frequency financial series, fitting a suitable model that accounts for its stylized facts. We also perform simulation studies to assess the effectiveness of the proposed method, whereas also to exemplify the effect of the jumps in time series. Finally, we use the methodology to model real high-frequency time series of stocks traded in the Brazilian Exchange and OTC and a series of cryptocurrencies.

Palavras-chave: Volatility; Wavelets; High-Frequency Data; Jumps; Long Memory.

[1]Instituto de Matemática e Estatística, Universidade de São Paulo – mateusgf@ime.usp.br
[2]Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas, Universidade Federal do ABC – guilherme.lima@ufabc.edu.br
[3]Instituto de Matemática e Estatística, Universidade de São Paulo – chang@ime.usp.br

P95

Implementação de Modelos de Mensuração com Indicadores Binários em Softwares Estatísticos

Michelle Passos[1]; Josemir Almeida[2]; Ythalo Santos[3]; Elzo P. Pinto Junior[4]; Maria del Pilar Flores-Quispe[5]; Valentina Martufi[6]; Rosana Aquino[7]; Leila Amorim[8]

A Análise Fatorial Confirmatória (CFA) é uma metodologia classicamente utilizada para relacionar indicadores observados a variáveis latentes através de um modelo de mensuração. A função de ligação probito ou logito é utilizada quando os indicadores são binários. Alternativamente, têm-se a Análise Fatorial Confirmatória Bayesiana (BCFA), que inclui distribuições a priori para os parâmetros do modelo. Com o objetivo de identificar as vantagens e desvantagens da utilização dos softwares estatísticos na implementação do modelo de mensuração com indicadores binários, utilizando a CFA e BCFA, comparou-se as estimativas das cargas fatoriais, escores fatoriais, medidas de confiabilidade (alfa de Cronbach ordinal e variância média extraída) e medidas de bondade de ajuste (raiz do erro quadrático médio de aproximação, índice de ajuste comparativo (CFI) e de Tucker-Lewis (TLI)). Vários métodos de estimação para CFA com indicadores binários estão disponíveis tanto no R versão 4.0.2 quanto no MPlus versão 8.7. Verificou-se que os escores gerados com CFA no R e no Mplus são aproximadamente iguais e que as estimativas (padronizadas e não padronizadas) e medidas de bondade de ajuste são idênticas. No entanto, o Mplus não calcula as medidas de confiabilidade dos construtos. Ainda, os escores gerados pela CFA e pela BCFA são muito próximos, em ambos os softwares. Para ilustrar a implementação dessas metodologias utilizou-se dados disponibilizados pelo Departamento de Atenção Básica do Ministério da Saúde (PMAQ 2011) para mensurar a qualidade do pré-natal considerando-se 10 indicadores binários, classificados como adequado ou inadequado.

Palavras-chave: Análise Fatorial Confirmatória; Análise Fatorial Confirmatória Bayesiana; R; MPlus; Indicadores Binários.

[1]Fiocruz-Cidacs – michelle.passos@fiocruz.br
[2]Fiocruz-Cidacs – josemir.almeida@fiocruz.br
[3]Fiocruz-Cidacs – ythalo.santos@fiocruz.br
[4]Fiocruz-Cidacs – elzo.junior@fiocruz.br
[5]Fiocruz-Cidacs – maria.quispe@fiocruz.br
[6]Fiocruz-Cidacs – valentina.martu@gmail.com
[7]ISC/UFBA – aquino@ufba.br
[8]IME/UFBA – leiladen@ufba.br

P96

Dependência Espacial Utilizando os Modelos Espaciais Lineares t-Student Reparametrizados

Rosangela Carlin Schemmer[1]; Miguel Angel Uribe-Opazo[2]; Manuel Galea[3]; Fernanda De Bastiani[4]; Rosangela Aparecida Botinha Assumção[5]

Este trabalho apresenta um estudo de dependência espacial e de diagnósticos de influência local para dados espacialmente georreferenciados, usando a distribuição t-student reparametrizada. A distribuição t-student reparametrizada tem a mesma forma da matriz de covariância da distribuição normal, que permitem uma comparação direta entre elas. Além disso, o parâmetro de forma é limitado, e supondo a existência do segundo momento finito, permite a estimação dos parâmetros por máxima verossimilhança. As técnicas de diagnósticos de influência local são usadas para identificar as presenças de observações influentes que podem interferir na seleção do modelo de variabilidade espacial, na estimação dos parâmetros e na construção de mapas de contorno. A metodologia foi aplicada a um conjunto de dados reais na área agrícola. Os resultados mostraram que a presença de observações consideradas influentes nos dados tem uma forte influência na escolha, estimação da estrutura de covariância e na construção de mapas de contorno pela técnica da krigagem.

Palavras-chave: Agricultura de Precisão; Geoestatística; Krigagem; Variabilidade Espacial.

[1]Programa de Pós-Graduação em Engenharia Agrícola-PGEAGRI, UNIOESTE, Cascavel, PR, Brasil- roschemmer@hotmail.com
[2]Centro de Ciências Exatas e da Terra (CCET), Programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI), Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel, PR, Brasil – miguel.opazo@unioeste.br
[3]Pontificia Universidad Catolica de Chile, Santiago, Chile – mgalea@mat.uc.cl
[4]Universidade Federal de Pernambuco, Recife PE, Brasil – fernandadebastiani@ufpe.br
[5]Universidade Tecnológica Federal do Paraná, Toledo PR, Brasil – rosangelaa@utfpr.edu.br

P97

Um Novo Gráfico de Controle para o Monitoramento de Séries Temporais de Contagem

Moizés da Silva Melo[1]; Airlane Pereira Alencar[2]

Vários gráficos de controle têm sido propostos para auxiliar no processo de vigilância em saúde pública. Os gráficos de controle com memória são comumente usados para monitorar esse tipo de dados, pois detectam com eficiência pequenas mudanças. Nesse contexto, propomos um novo gráfico de controle com memória, no qual a média progressiva é utilizada como estatística de plotagem para o monitoramento de dados de contagem autocorrelacionados, tipicamente encontrados nos sistemas de vigilância em saúde pública. O desenvolvimento do novo gráfico de controle é baseado nos resíduos quantílicos aleatorizados do modelo Conway-Maxwell-Poisson autorregressivo de médias móveis (CMP-ARMA) ajustado. Realizamos um estudo de simulação de Monte Carlo para avaliar e comparar o desempenho do gráfico de controle proposto com duas abordagens tradicionais; gráficos de controle do tipo Shewhart e da média móvel exponencialmente ponderada (EWMA). Os resultados mostram que a proposta supera as duas cartas de controle convencionais consideradas. Por fim, ilustramos a aplicabilidade do novo gráfico por meio do monitoramento do número semanal de internações de pessoas com mais de 60 anos por doenças respiratórias na cidade de São Paulo - SP.

Palavras-chave: Dados de Contagem, Gráfico de Controle, Média Progressiva, Séries Temporais, Modelo CMP-ARMA.

[1]Instituto de Matemática, Estatística e Física – FURG, Rio Grande, RS – mozas.silva@yahoo.com.br
[2]Instituto de Matemática e Estatística – USP, São Paulo, SP – lanealencar@usp.br

P98

Modelo de Regressão Bell Misto para Dados de Contagem

Naiara Caroline Aparecido dos Santos[1]; Jorge Luis Bazán[2]

Embora o modelo Poisson sirva como uma ferramenta padrão para modelar dados de contagem, está bem consolidado que essa abordagem é limitada pela suposição de equidispersão. A distribuição Bell de um parâmetro proposta recentemente na literatura estatística, demonstrou-se como uma alternativa viável para dados de contagem. Diante disso, com base na distribuição Bell, propomos um novo modelo de regressão misto, o qual pode ser uma alternativa interessante aos modelos de regressão misto usuais para dados de contagem. Consideramos uma abordagem Bayesiana para realizar inferências, em que o modelo proposto é implementado utilizando o Stan em R, que faz uso do algoritmo No-U-Turn-Sampler (NUTS) para obter os valores simulados da distribuição a posteriori. Simulações de Monte Carlo indicam que o método considerado é bastante eficaz para estimar os parâmetros do modelo Bell. Também propomos os resíduos quantílicos aleatorizados para checar adequabilidade do ajuste do modelo. Por fim, uma aplicação empírica é considerada para mostrar a utilidade do modelo de regressão Bell misto na prática.

Palavras-chave: Distribuição Bell; Modelo de Regressão Misto; Abordagem Bayesiana; Dados de Contagem.

[1]Departamento de Estatística, Universidade Federal de São Carlos – UFSCar – naicaroline2@usp.br
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo – ICMC/USP – jlbazan@icmc.usp.br

P99

Estimação de Máxima Verossimilhança de Modelos de Volatilidade Estocástica Assimétrica

Omar Abbara[1]; Mauricio Zevallos[2]

O objetivo do trabalho é propor um estimador para os modelos de volatilidade estocástica (SV - *stochastic volatility*) quando a volatilidade de um dado dia tem comportamento assimétrico em relação ao retorno do dia anterior. Esta propriedade é chamada de "efeito alavanca" (*leverage effect*), e pode ser incluída nos modelos SV quando há correlação entre os dois termos de erro. A proposta consiste em aproximar a distribuição conjunta dos termos de erro por uma mistura de normais, combinando as propostas de Shumway and Stoffer (2006, chap.6) e Omori et al. (2007). As propriedades do estimador são estudadas por um experimento de Monte Carlo, e o estimador é aplicado para séries de retornos reais.

Palavras-chave: Volatilidade; Mistura; Assimetria.

[1]Canvas Capital S.A. – muhieddine@gmail.com
[2]Departamento de Estatística – IMECC/Unicamp – amadeus@unicamp.br

P100

Modelo Hierárquico Beta-Bernoulli para Tempo de Resposta e Precisão

Patrícia Stülp[1]; Jorge Luis Bazán[2]

Com o avanço da tecnologia, muitos testes passaram a ser realizados em computadores e, com isso, um examinado deixa registrado suas respostas dos itens do teste e também o tempo que ele gastou para responder aos itens do teste. O tempo para a realização do teste é limitado e não infinito, como supõem alguns modelos na literatura que trabalham com distribuições no suporte $(0, \infty)$. Além da modelagem considerando o tempo de resposta limitado, é interessante modelar a proporção de tempo de resposta, uma vez que o examinado pode distribuir o tempo total do teste entre todos os itens. Neste trabalho, modelamos o tempo de resposta (TR) seguindo uma distribuição limitada e obtivemos um modelo conjunto em combinação com a precisão da resposta, considerando a distribuição Bernoulli. Para a precisão nós consideramos o modelo TRI Normal Ogive de dois parâmetros e para modelar a proporção TR utilizamos a distribuição Beta. Realizamos um estudo de simulação para verificar quanto à recuperação dos parâmetros do modelo e realizamos uma aplicação do modelo a um conjunto de dados reais de leitura computadorizada do PISA 2015. As estimações dos parâmetros foram realizadas no contexto bayesiano, utilizando o pacote `R2jags` no software R, e como resultado de todo o desenvolvimento do trabalho, concluimos que o modelo é uma boa alternativa para ajustar conjunto de dados de proporção de TR e precisão de resposta, uma vez que os resultados foram muito satisfatórios.

Palavras-chave: Distribuição Beta; Proporção do Tempo de Resposta; Modelo Hierárquico; Estimação Bayesiana.

[1]Universidade Federal de São Carlos – UFSCar, Departamento de Estatística, Rod. Washington Luiz, km 235, CEP: 13565-905, São Carlos, SP, Brasil – patriciastulp@usp.br

[2]Universidade de São Paulo – USP, Departamento de Matemática Aplicada e Estatística, Avenida Trabalhador São-carlense, 400, CEP: 13566-590, São Carlos, SP, Brasil – jlbazan@icmc.usp.br

P101

Impacto da Pandemia de COVID-19 nas Infecções de Corrente Sanguínea nos Hospitais da cidade de São Paulo: Uma Abordagem Usando Modelos de Regressão Segmentados para Dados de Contagem

Paulo Henrique Dourado da Silva[1]; Antônio Carlos Pedroso de Lima[2]

A pandemia pela COVID-19 afetou significativamente os serviços hospitalares, muitas instituições observaram aumento de infecções relacionadas à saúde (IRAS) mesmo após o aumento da adesão aos protocolos de isolamento e da higienização das mãos. Segundo a OMS, as IRAS estão entre as maiores causas de morte e aumento de morbidade entre os pacientes hospitalizados. O objetivo do estudo é estudar o impacto da COVID -19 no número de infecções da corrente sanguínea (ICS) relacionada à cateter venoso central (CVC) levando em consideração a natureza, o tamanho e a densidade do uso do CVC de cada hospital. Para tanto foram utilizados modelos de regressão segmentado Poisson e Binomial Negativo com ponto de mudança fixo e aleatório. As funções de segmentação utilizadas foram a linear e a quadrática. Para o caso aleatório, foram propostos intervalos de confiança (IC) para o ponto de mudança baseado nas estatísticas *Score test* e Gradiente dado que no pacote `segmented` do R apenas IC baseado no método delta está presente para MLGs segmentados. Os resultados mostram aumento da taxa de ICS após o início da pandemia.

Palavras-chave: COVID-19; ICS; MLG Segmentado; Ponto de Mudança; Intervalo de Confiança.

[1]Departamento de Estatística, Universidade de São Paulo – phdsilva@ime.usp.br
[2]Departamento de Estatística, Universidade de São Paulo – acarlos@ime.usp.br

P102

Regression Analysis for Compositional Data by Municipality and Disabled People in Brazil

Paulo Tadeu Meira e Silva de Oliveira[1]

According to 2010 IBGE Demographic Census, there were 45.6 million disabilities people in Brazil distributed in different municipalities. Data were considered by municipality justified by the fact that the level of service provided varies according to the infrastructure and availability of existing resources in the most diverse locations. Data sets of the 2010 Census aggregated by municipality and the UNDP data available in the Human Development Atlas for each of the 5565 Brazilian municipalities, were considered. Compositional data are those that establish the relative information, they are parts of a whole. Regression analysis for compositional data with transformation of the variables by the reason log following the suggestion proposed by Aitchism (1986) as an alternative to simplify the data analysis and considering the different deficiencies as independent variables and the others as independent, generating an adjustment for each of these deficiencies and then, for each of these adjustments a variable selection procedure is applied so that for each adjustment all independent variables considered significant. In this work, for compositional data, regression analysis techniques with logarithmic transformation were applied in conjunction with the variable selection procedure in which the variables for each model were considered significant. It was possible to conclude that for the dependent variables constituted by the different disabilities, adjustment considered a greater number of independent variables as significant was the person with a disability DEF and the one with the lowest number was ND4 and independent variables considered significant in all adjustments were NIA1, NMD3, NTT7 and TE2.

Palavras-chave: Disability People; Compositional Data; Regression Analysis; Variables Selections; IBGE Census 2010; UNDP.

[1]IEA - USP, São Paulo, SP – poliver@usp.br

P103

Construção de Portfólios de Longo Prazo

Rafael Bernardoni Chave[1]; Cleiton Guollo Taufemback[2]

Portfólios de longo prazo são geralmente construídos considerando ativos de baixo risco e com um bom fluxo de dividendos. A metodologia proposta neste trabalho é uma alteração do modelo de Markowitz, considerando a remoção da variabilidade de curto prazo de retorno dos ativos, com o objetivo de construir um portfólio que seja menos sensibilizado por variações rápidas e temporárias na série de retornos. Utilizando dados de 20 anos dos ativos presentes na composição do índice Dow Jones Industrial Average, comparamos os resultados do método proposto com o modelo tradicional de Markowitz e o modelo Naive, para horizontes de três, seis e doze meses. Também implementamos quatro frequências máximas para o "low-pass filter"utilizado na série. Os resultados indicam que portfólios construídos com o modelo proposto, geralmente resultam em retornos superiores em relação aos modelos de benchmark. Análises de drawdown e de dominância estocástica também indicam a superioridade do nosso método.

Palavras-chave: Portfólio; Séries Temporais; Investimentos; Filtragem; Markowitz; Dow Jones; Domínio da Frequência.

[1]Estudante de graduação pela Universidade Federal do Rio Grande do Sul, UFRGS
[2]Doutor pela Universidade Federal do Santa Catarina, UFSC

P104

Modelo de Regressão Logística Bivariado para Estimar as Chances de Um Candidato ao Curso de Medicina Ingressar em Primeira Chamada na Unicamp

Rafael Pimentel Maia[1]; Mariângela Lima Rodrigues[2]

Objetivo deste trabalho é apresentar um modelo de regressão logística vivariado com um efeito aleatório compartilhado para estimar a probabilidade de um candidato ao curso de Medicina da Unicamp pelo vestibular ser convocado em primeira chamada. O Vestibular Unicamp é composto por duas fases, a primeira formada por uma prova de múltipla escolha com 90 questões e a segunda fase por uma redação e provas dissertativas. O curso de medicina é o mais concorrido, ultrapassando nos últimos anos 300 candidatos por vaga; são convocados para a segunda fase cerca de 10 candidatos por vaga (são oferecidas 110 vagas). Seja Y_{1i} a indicadora do i-ésimo candidato ser convocado para a segunda fase e Y_{2i} a indicadora do i-ésimo candidato ser convocado em primeira chamada, $i = 1, \ldots n$. Foi ajustado um modelo de regressão loogistíca marginal para $P(Y_{1i} = 1)$ e outro para a $P(Y_{2i} = 1 \mid Y_{1i} = 1)$. Os modelos foram ajustrado conjutamento com um efeito aleatório de individuo compartilhado entre os modelos marginais. A probabilidade do candidato i ser convocado em primeira chamada foi então calculada por $P(Y_{2i} = 1) = P(Y_{1i} = 1)P(Y_{2i} = 1 \mid Y_{1i} = 1)$. Para ajuste do modelo foram utilizados dados dos candiados ao vestibular nos anos de 2012 a 2018; como co-variáveis foram utilizadas as respostas ao questionário socioeconômico preenchido pelos candidatos no momento da inscrição. Uma das questões foi avaliar os impactos da implementação e mudanças nas políticas de inclusão social ao longo do período estudado nas chances de candidatos oriundos de escolas públicas do ensino médio e candidatos autodelaclarados pretos e partos ingressarem na universidade em um curso de alta demanda como a medicina.

Palavras-chave: Regressão Logística; Modelo Misto; Efeito Aleatório Compartilhado; Políticas de Inclusão; Vestibular.

[1]Departamento de Estatística, Unicamp – rpmaia@unicamp.br
[2]Departamento de Estatística, Unicamp – mariangela.rodrigues@comvest.unicamp.br

P105

Estudo do Desempenho em Disciplinas de Primeiro Ano e o Impacto na Evasão dos Cursos de Exatas da UNICAMP

Rafael Ribeiro Santos[1]; Rafael Pimentel Maia[2]

Este projeto visa estudar e mensurar o impacto do desempenho nas disciplinas cursadas no primeiro ano de graduação e desempenho nas provas do vestibular na probabilidade de evasão de alunos ingressantes pelo Vestibular Unicamp, entre 2012 a 2018, nos cursos da área de Ciências Exatas, Tecnológicas e da Terra; bem como identificar características socioeconômicas e demográficas que possam estar associadas à evasão, que foi caracterizada como o encerramento do vínculo de um aluno com o curso de ingresso, exceto óbitos. O conjunto de dados analisado contém informações provenientes da Comissão Permanente para o Vestibulares da Unicamp, a Comvest, e da Diretoria Acadêmica da UNICAMP, a DAC. Entre os dados estão as respostas do questionário socioeconômico aplicado no momento da inscrição do vestibular e o histórico acadêmico do aluno na UNICAMP. Ao todo temos o registro de 10592 ingressantes. Inicialmente foi realizada uma análise exploratória para avaliar a associação entre as variáveis socioeconômicas e as possíveis situações nas disciplinas ao final do semestre (aprovação, reprovação por nota, reprovação por frequência e desistência) utilizando a técnica de Análise de Correspondência. Para estimar a probabilidade de evasão na presença de covariáveis foi ajustado um modelo de regressão logística. Dado que o modelo logístico não suporta censuras, que são aqueles que não se formaram, nem evadiram, foi modelada a evasão nos três primeiros anos. A seleção de variáveis foi feita utilizando o critério BIC e a qualidade de ajuste do modelo foi feita com base na análise gráfica do resíduos deviance.

Palavras-chave: Regressão Logística; Evasão; Desempenho Acadêmico.

[1]Departamento de Estatística, Unicamp – r243464@dac.unicamp.br
[2]Departamento de Estatística, Unicamp – rpmaia@unicamp.br

P106

A Robust Lasso Regression for Linear Mixed-Effects Models with Diagnostic Analysis

Rafael Rocha de Oliveira Garcia[1]; Cibele Maria Russo Novelli[2]

Variable selection has been a topic of great interest for statisticians and researchers alike. The choice of the best subset of predictors may be carried out with the objective of improving prediction or for easier interpretation of results. However, such methods are not always straightforward, mainly in the context of linear mixed-effects models. Variable selection for such models must be carried out for both fixed and random effects, the first being related to the global mean of data and the second to subject-level variance. There are two possible approaches when selecting variables for mixed-effects models: joint or two-stage procedures. In existing literature on the topic of variable selection for linear mixed-effects model, there is a method of joint selection via lasso for linear mixed-effects models under a normal distribution. Another topic of remarkable importance is diagnostics and residual analysis. While residual analyses are carried out to assess issues with the fitted model and identification of atypical observations, diagnostic analyses are carried out assuming the model as correct and, assessing its conclusions robustness to small disturbances in the data and/or the model. There are many possible ways to deal with such observations. One is using robust models, which are said to be robust to disturbances in the data. That is, models that are better fit to data sets that possess observations considered to be as outliers and/or leverage. This work aims to use the robust method for variable selection in linear mixed-effects model and compare it with the normal method using diagnostic analysis.

Palavras-chave: Mixed Models; Lasso; Robust Models; Diagnostics; Regression Analysis.

[1]Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, São Carlos, Brazil – rafael.rogarcia@gmail.com

[2]Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, São Carlos, Brazil – cibele@icmc.usp.br

P107

Modelo Chen Autoregressivo e de Médias Móveis para Modelagem de Quantis Positivos

Renata F. Stone[1]; Laís H. Loose[2]

A distribuição Chen (Chen, 2000) é bastante flexível e pouco explorada na literatura, esta distribuição tem dois parâmetros de forma e é adequada para modelagem de dados contínuos e positivos. Modelos de séries temporais utilizando a distribuição Chen, ainda não foram amplamente explorados. Nesse sentido, no presente trabalho propomos uma classe de modelos quantílicos, com estrutura autorregressiva e de médias móveis para modelagem do quantil utilizando a distribuição Chen, o modelo CHARMA(p, q), em que p e q correspondem a quantidade de termos autorregressivos e de médias móveis, respectivamente. Para a estimação dos parâmetros do modelo utilizamos os estimadores de máxima verossimilhança condicional (EMVs). A avaliação numérica dos EMVs do modelo foi feita por meio de simulações de Monte Carlo, considerando três cenários: CHARMA(1,0), CHARMA(0,1), CHARMA(1,1). Utilizamos 5.000 réplicas, diferentes tamanhos amostrais $n \in \{100, 250, 500\}$ e diferentes quantis $(0.25, 0.5, 0.75, 0.9)$. A avaliação dos EMVs baseou-se no cálculo da média, viés, viés relativo, erro padrão e erro quadrático médio (EQM). Os resultados das simulações indicaram que para todos os cenários os EMVs possuem boas propriedades. A média dos estimadores é próxima do valor fixado, indicando um baixo viés que diminui à medida que o tamanho amostral aumenta. A consistência do estimador é observada numericamente através da diminuição do EQM. Para trabalhos futuros pretendemos desenvolver aplicações em dados reais a fim de ilustrar a aplicabilidade do modelo proposto.

Palavras-chave: Distribuição Chen; Séries Temporais; Estimadores; Avaliação Numérica.

[1]Departamento de Estatística, Universidade Federal de Santa Maria – renastan@gmail.com
[2]Departamento de Estatística, Universidade Federal de Santa Maria – lais.loose@ufsm.br

P108

Estimação de Efeitos Causais Heterogêneos: Comparação Entre os Métodos Inverse Probability Weighting e Causal Forest

Eduardo de Oliveira Horta[1]; Renato Pedroso Lauris[2]; Rodrigo Citton Padilha dos Reis[3]

Em experimentos aleatorizados, a alocação aleatória do grupo de tratamento garante que o status do tratamento não será confundido com covariáveis pré-tramento. Desta forma, o efeito de tratamento nos desfechos (variáveis respostas) pode ser estimado comparando-se os desfechos entre tratados e não-tratados. Devido a restrições éticas e de ordem prática, nem sempre é possível conduzir um experimento aleatorizado. Assim, a maioria dos efeitos causais é estimada a partir de dados observacionais. Neste contexto, o tratamento é geralmente influenciado pelas características do indivíduo, tal que as diferenças sistemáticas entre os grupos devem ser consideradas na estimação. Métodos baseados no escore de propensão, diferenças em diferenças e controle sintético são alguns exemplos de como estimar efeitos causais em estudos observacionais. Esses métodos, no arcabouço do modelo de desfechos potenciais, estimam o efeito médio de tratamento (*average treatment effect*, ATE). Além do ATE, muitas vezes se está também interessado em eventuais efeitos heterogêneos conforme subgrupos formados pela combinação de níveis de certas características (covariáveis), estimando o efeito médio condicional de tratamento (*condicional average treatment effect*, CATE). O CATE contribui para o aprendizado de um tratamento, pois permite identificar e priorizar os subgrupos com impactos mais significativos. Diante da utilidade dos efeitos causais heterogêneos, esse trabalho tem como objetivo apresentar dois métodos de estimação do CATE: o Inverse probability weighting e o Causal Forest. As diferenças e limitações de ambos são apresentadas e através de um estudo de simulação comparou-se a eficiência dos estimadores e a taxa de cobertura dos intervalos de confiança.

Palavras-chave: Aprendizado de Máquina; Desfechos Potenciais; Inferência Causal.

[1]Professor do Departamento de Estatística do Instituto de Matemática e Estatística e do Programa de Pós-graduação em Estatística da Universidade Federal do Rio Grande do Sul (UFRGS) – eduardo.horta@ufrgs.br

[2]Mestrando do Programa de Pós-graduação em Estatística da Universidade Federal do Rio Grande do Sul (UFRGS) – renato.lauris@gmail.com

[3]Professor do Departamento de Estatística do Instituto de Matemática e Estatística e do Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul (UFRGS) – citton.padilha@ufrgs.br

P109

Influencia Local em Modelos Parcialmente Lineares Aditivos Generalizados

Rodolpho Jordan Domingos Quintela[1]; Roberto F. Manghi[2]; Francisco José A. Cysneiros[3]

Este trabalho tem como objetivo propor resíduos e técnicas para a análise de diagnóstico nos Modelos Parcialmente Lineares Aditivos Generalizados (MPLAGs), tais como: alavancagem generalizada, análise de resíduos, dos quais propomos utilizar os resíduos de Pearson e resíduos aleatorizados, bem como medidas para análise de influência local sob os seguintes esquemas de perturbação: perturbação de casos, perturbação na variável resposta e perturbação em uma das variáveis explicativas. Para isto, derivamos tais medidas fundamentados em uma vasta pesquisa bibliográfica e conceitual sobre estes métodos no contexto dos MPLAGs. Essas técnicas foram utilizadas em exemplos de aplicação a dados reais e os resultados foram discutidos a fim de avaliar o nosso estudo teórico. Para tanto, apresentamos as equações de estimação para os parâmetros do modelo através da função de verossimilhança penalizada, considerando como estrutura não paramétrica o uso de P-*splines*. Assim, definimos tal modelo, buscando apresentar algumas propriedades e vantagens que motivam o uso de P-*splines* no contexto de regressão não paramétrica. Por fim, o método iterativo *backfitting* (Gauss-Seidel) é utilizado para a obtenção das estimativas.

Palavras-chave: Diagnóstico; Influência Local; MPLAGs; P-splines.

[1]Instituto de Matematica e Estatística (IME) da Universidade de Sao Paulo (USP), R. do Matão, 1010 – Butantã, São Paulo – SP, 05508-090, Brasil – rodolpho.jordan@usp.br

[2]Departamento de Estatística, Universidade Federal de Pernambuco, Av. Jornalista Aníbal Fernandes, Recife, 50740-540, Pernambuco – roberto.manghi@de.ufpe.br

[3]Departamento de Estatística, Universidade Federal de Pernambuco, Av. Jornalista Aníbal Fernandes, Recife, 50740-540, Pernambuco – cysneiros@de.ufpe.br

P110

Normal Scale Mixture Copula Marginal Regression with Box-Cox Symmetric Distributions

Rodrigo Matheus Rocha de Medeiros[1]; Silvia Lopes de Paula Ferrari[2]

The class of the Box-Cox symmetric distributions was recently introduced in the statistical literature. The class provides a flexible modeling framework for univariate independent positive continuous data with different levels of skewness and tail-heaviness. Additionally, the relatively easy parameter interpretation makes it attractive for regression purposes. However, more general applications may involve correlated data, such as when observations have a temporal or spatial dependence. Based on Sklar's Theorem, the copula theory provides an approach to modeling dependence through a function (named copula) which describes how the elements of a random vector are associated. Particularly, copulas generated by scale mixtures of normal distributions allow the bivariate associations to determine the dependence structure of the random vector entirely. Moreover, they also achieve positive and negative associations without restrictions on the data dimension. This work introduces a broad class of marginal regression models to analyze correlated positive continuous data with Box-Cox symmetric marginal distributions, where a normal scale mixture copula describes the dependence. Our approach resembles the joint modeling of univariate observations of the classical generalized estimating equations model. It is possible to select one of several association structures specified in terms of nonlinear response transformations, which provides flexibility in modeling independent observations, time series, longitudinal, clustered, or spatially correlated data.

Palavras-chave: Clustered Data; Dependence Structures; Log-Symmetric Distributions; Time Series; Working Correlation Matrix.

[1]Department of Statistics, University of São Paulo, São Paulo, Brazil – rodrigo.matheus@ime.usp.br
[2]Department of Statistics, University of São Paulo, São Paulo, Brazil – silviaferrari@usp.br

P111

Associação Entre Valor Bruto de Produção Agropecuária e Uso da Área Rural: Uma Aplicação do Coeficiente de Correlação de Medidas Repetidas

Sergio Augusto Rodrigues[1]; Caroline Pires Cremasco[2]; Valter Cesar de Souza[3]; Paulo Andre de Oliveira[4]; Carlos Roberto Padovani[5]

O valor da produção agropecuária (VP) é utilizado para estudar desempenho econômico de regiões, sendo importante entender sua associação com uso de áreas rurais. Nestes estudos, medidas repetidas na mesma unidade são comuns, contudo, poucos utilizam alternativas que revelem essa particularidade para medir tais associações. O coeficiente de correlação de Pearson é utilizado para associações entre variáveis quantitativas, assumindo independência entre observações. Quando cada unidade fornece medidas de variáveis em dois ou mais momentos, essa suposição é violada mas, frequentemente, observa-se seu uso, podendo produzir resultados enviesados. Uma opção consiste em utilizar a média das medidas repetidas, porém, isto pode ser razoavelmente entendido quando a variabilidade dentro das unidades for muito pequena, caso contrário, a média será imprecisa para representar repetibilidade das medições, tornando esse procedimento frágil. Uma alternativa que considera a variação intraindividual é o coeficiente de correlação de medidas repetidas. Assim, objetivou-se apresentar uma aplicação em R para medir a associação do VP com áreas rurais mensuradas em 40 escritórios de desenvolvimento rural do estado de São Paulo (EDRs). Cada EDR forneceu dados nos anos de 2008 e 2017 e as correlações com os valores médios dos anos (r_m) e de medidas repetidas (r_{mr}) foram obtidas. Observou-se magnitudes semelhantes das associações, indicando baixa variação dentro das EDRs (entre anos). Concluiu-se que área temporária associou-se com VP ($r_m = 0,695$ e $r_{mr} = 0,689$, ambos com $p < 0,05$) e área de culturas perenes não apresentou correlação significativa, com as médias ($r_m = 0,241, p = 0,135$), mas, apesar de baixa, mostrou-se significativa ($r_{mr} = 0,247, p = 0,033$) em medidas repetidas.

Palavras-chave: Variação Intraclasse; Correlação; Medidas Repetidas; Áreas Produtivas; Indicadores de Produtividade.

[1]Dep. Bioprocessos e Biotecnologia, FCA Unesp, Botucatu-SP, Brasil – sergio.rodrigues@email.com.br
[2]Dep. Bioprocessos e Biotecnologia, FCA Unesp, Botucatu-SP, Brasil – caroline.cremasco@unesp.br
[3]Dep. Bioprocessos e Biotecnologia, FCA Unesp, Botucatu-SP, Brasil – valter.souza@unesp.br
[4]Faculdade de Tecnologia (Fatec), Botucatu-SP, Brasil – paulo.oliveira108@fatec.sp.gov.br
[5]Dep. Bioestatística, IB Unesp, Botucatu-SP, Brasil – cr.padovani@unesp.br

P112

Percepções sobre Tatuagem e sua Relação com Vivências no Mercado de Trabalho no Brasil

Camilla C. Alves[1]; Júlia S. de Araújo[2]; Sofia S. H. de Brito[3]; Denise B. N. Silva[4]

A tatuagem está inserida em vários contextos históricos. Sociedades antigas usavam esse tipo de arte em ritos de passagens e eventos especiais. Marinheiros e aventureiros possuíam o costume de marcar seus corpos para simbolizar pertencimento nacional e identidade religiosa. Mesmo disseminada entre pessoas de todas as crenças, gêneros e idades, ainda há incerteza sobre como a sociedade julga pessoas tatuadas em determinados espaços sociais. No vital ambiente do trabalho podem ser encontrados relatos sobre a possível existência de preconceito. Inspiradas nos desafios que nos aguardam no início da carreira profissional, e tendo como base o debate atual sobre o tema, este estudo teve como objetivo realizar uma pesquisa quantitativa para investigar a percepção e o posicionamento quanto à posse de tatuagem em questões associadas ao mercado de trabalho de brasileiros(as) com idade a partir de 16 anos. A coleta de dados ocorreu entre os dias 31 de outubro e 13 de novembro de 2021 através de plataformas digitais obtendo 938 respostas válidas. O questionário incorporou perguntas relacionadas às características socioeconômicas dos respondentes, à posse de tatuagem, a percepções sociais e ao mercado de trabalho, bem como investigou situações vivenciadas por pessoas tatuadas que apontassem indícios de reprovação num período de referência de 5 anos. A pesquisa também fomentou debates sobre o tema nas redes sociais, ratificando a relevância do estudo e o interesse de profissionais nas informações coletadas. Seus resultados indicam que ainda há preconceito sobre a posse de tatuagem no mercado de trabalho e revelou indícios de preconceito mesmo entre pessoas tatuadas.

Palavras-chave: Pesquisa Quantitativa; Estatística Social; Tattoo; Preconceito; Questionário Eletrônico.

[1]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), RJ – camillaalvesestudos@gmail.com
[2]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), RJ – juliaxluccas@gmail.com
[3]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), RJ – sshb.sofia@gmail.com
[4]Escola Nacional de Ciências Estatísticas (ENCE/IBGE), RJ – denisebritz@gmail.com

P113

Mixture Models for Longitudinal Data of Cognitive Function in Older Adults

Tatiana Benaglia[1]; Eric Barbosa[2]; Hildete P. Pinheiro[1]; Graciela Muniz-Terrera[1]

A logistic Binomial regression model that accounts for change points in the predictor is presented to model cognitive ability in older adults up to their death. A mixture specification rises to describe two prevalent behaviors in the data: one group of older adults declines at constant rate; whilst the other experiences a spontaneous accelerated decline before death. The latter aspect is dealt with random change points nonlinear predictors. The study's goal is to quantify associations amidst cognitive decline and dementia diagnosis after accounting for sociodemographic information. The proposed model is illustrated on data collected in a longitudinal cohort of older adults.

Palavras-chave: Mixture Models; Longitudinal Models; Mixed Effects Models; Random Change Points; Binomial Distribution.

[1]Departamento de Estatística, Unicamp – Campinas/SP
[2]Heritage College of Osteopathic Medicine, Ohio University

P114

Essays on the Unit Burr XII Distribution: Regression and Time Series Models

Tatiane F. Ribeiro[1]; Gauss M. Cordeiro[2]; Fernando A. Peña-Ramírez[3]; Renata R. Guerra[4]; Airlane P. Alencar[5]

There is an interest in modeling bounded random variables to the standard unit interval in many practical situations, such as rates, proportions, and indexes. We propose two new probability distributions to deal with the uncertainty involved by variables of this type and develop its associated regression models. Both distributions are based on a transformation of the Burr XII random variable. We also introduce a new dynamic model for time series data with support in the interval $(0, 1)$. This study is composed of three main and independent sections. In the first, we define the unit Burr XII (UBXII) distribution and its quantile regression model. Some of its mathematical and statistical properties are investigated. In the second, the reflexive UBXII distribution is obtained, and the regression model is proposed. The maximum likelihood (ML) method is considered for parameter estimation of both regression models. In the third, we propose the dynamic class of models: UBXII autoregressive moving average (UBXII-ARMA) for time series taking values in the unit interval. The conditional ML method is used to estimate and construct asymptotic confidence intervals of the parameters that index the UBXII-ARMA model. Closed-form expressions for the conditional score vector are derived. Furthermore, Monte Carlo simulation studies, diagnostic analysis tools, model selection criteria, and applications to the real data are presented and discussed for the three proposed models.

Palavras-chave: BetaRregression; Quantile Regression; Statistical Learning; Time Series; Unit Probability Distributions.

[1]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil – tatianefr@ime.usp.br
[2]Departamento de Estatística, Universidade Federal de Pernambuco, Recife, Brazil – gauss@de.ufpe.br
[3]Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia – fapenara@gmail.com
[4]Departamento de Estatística, Universidade Federal de Santa Maria, Santa Maria, Brazil – renata.r.guerra@ufsm.br
[5]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil – lanealencar@usp.br

P115

Método de Clusterização Aplicado à Classificação da Pobreza Familiar: Um Estudo de Performance Computacional para Dados Mistos

Vinicius Ricardo Riffel[1]; Hellen Paz[2]; Rosemeire Leovigildo Fiaccone[2]; Dandara Ramos[3]; Anderson Ara[4]

Os métodos de clusterização são amplamente utilizados na investigação da relação multivariada entre diversas variáveis. O foco desses métodos é encontrar similaridades entre as observações e classificá-las em *clusters*. Geralmente, as distâncias entre as observações são utilizadas como base do critério de classificação. A definição da medida de distância no cenário de dados mistos, em que existem variáveis qualitativas e quantitativas, pode não ser trivial. Bem como, esses métodos requerem um grande esforço computacional, uma vez que tomamos as medidas de distâncias entre os pares de observações. Neste trabalho, apresentamos um método eficiente de clusterização para dados mistos tendo como base o método k-médias e que envolve a combinação de medidas de distância e/ou similaridade, bem como formas matriciais para o cálculo das distâncias. O método foi validado via simulação e aplicado para dados reais referentes às famílias do Cadastro Único que são beneficiárias do programa Bolsa Família do Ministério da Cidadania, com cerca de 2 milhões de observações. A utilização de tais técnicas na classificação da pobreza representa um avanço metodológico na literatura, tendo em vista que essa classificação geralmente é binária e realizada com base apenas em critérios monetários. O método proposto mostrou-se com melhor performance computacional quando comparado com os métodos tradicionais. Essa metodologia pode ser estendida para outros tipos de dados mistos e/ou clusterização multinível.

Palavras-chave: Clusterização; Distância de Jaccard; Distância Euclidiana; Performance Computacional; Multinível.

[1]Departamento de Estatística, Univer- sidade Federal do Paraná (DEST-UFPR) – viniciusriffel@ufpr.br
[2]Departamento de Estatística, Universidade Federal da Bahia (DEST-UFBA)
[3]Instituto de Saúde Coletiva, Universidade Federal da Bahia (ISC-UFBA) – dandara.ramos@ufba.br
[4]Departamento de Estatística, Universidade Federal do Paraná (DEST-UFPR) – ara@ufpr.br

P116

Aplicação de Um Modelo Multidimensional de Teoria de Resposta ao Item Dicotômico para Uma Escala Psicométrica de Avaliação de Depressão Usando Python

Virginia Pereira[1]; Mariana Curi[2]; Jorge Bazan[3]

O objetivo deste trabalho é o estudo da Teoria de Resposta ao Item Multidimensional com aplicabilidade na mensuração da depressão. Tendo em vista o crescente número de casos de depressão tanto no Brasil como no mundo, principalmente como parte da herança da pandemia de Covid-19 que causou profundas e significativas mudanças na vida das pessoas, compreende-se a importância deste estudo para a comunidade científica. Dentre os diversos testes psicológicos existentes, utilizamos a Escala de Depressão de Beck, ou BDI (Inventário de Depressão de Beck), que visa avaliar a gravidade dos sintomas depressivos. Ela é composta por um grupo de itens que pretende medir o traço latente de intensidade de sintomas depressivos, possibilitando a identificação do estágio da depressão e realizar predições de seu desfecho. Para auxiliar na medição desta escala, aplicamos a Teoria de Resposta ao Item (TRI), já bastante utilizada na área psiquiátrica, que tem como alvo o encontro de um indivíduo com um item, onde o padrão de respostas do indivíduo a um particular grupo de itens (amostra de comportamento) fornece a base para a estimativa do traço latente. Nós consideramos os dados de Fragoso e Curi (2013) e desenvolvemos um processo de estimação considerando além de demais literaturas sobre o modelo multidimensional de resposta ao item considerando Python. A divulgação desta metodologia amplia as possibilidades de sua utilização em diversos estudos com variados grupos de indivíduos e situações.

Palavras-chave: Depressão; TRI; Python; BDI.

[1]ICMC/CeMEAI (pós-graduanda), USP São Carlos – virginia.lani@ufrgs.br
[2]ICMC - Instituto de Ciências Matemáticas e de Computação, USP São Carlos – mcuri@icmc.usp.br
[3]ICMC - Instituto de Ciências Matemáticas e de Computação, USP São Carlos – jlbazan@icmc.usp.br

COMUNICAÇÕES PÔSTERES 2

P117

Modelo DARMA Quantílico

Vitor Bernardo Silveira Pereira[1]; Cleber Bisognin[2]; Laís Helen Loose[3]

O objetivo deste trabalho é propor um novo modelo de séries temporais baseado na distribuição Dagum Exponencial Generalizada Exponencializada (EGED) proposta por Suleman Nasiru, Peter N. Mwita e Oscar Ngesa em 2019, com parâmetros $(\alpha, \delta, \sigma, \lambda, \eta, \gamma) \in \mathbb{R}_+^6$. Seja $\boldsymbol{Y} = (Y_1, \cdots Y_n)^\top$ variáveis aleatórias onde cada Y_t, para $t = 1, \cdots, n$, dada $\mathcal{F}_{t-1} = \mathcal{F}_{t-1}(Y_{t-1}, Y_{t-2}, \cdots)$ a σ-álgebra gerada pelas informações observadas até o instante $t-1$, possui densidade condicional EGED com quantil μ_t. Assim, considere $g(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \kappa_t$, onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ é vetor de parâmetros $(\boldsymbol{\beta} \in \mathbb{R}^{k+1})$ e $\boldsymbol{x}_t = (x_{t0}, x_{t1}, \cdots, x_{tk})^\top$ são observações de $k+1$ covariáveis $(k+1 < n)$, as quais são supostamente fixas e conhecidas, $g : \mathbb{R}_+ \to \mathbb{R}$ é uma função de ligação duas vezes diferenciável e estritamente monótona, e assumimos que κ_t é um processo ARMA(p, q). Ou seja, $\phi(L)(g(y_t) - \boldsymbol{x_t}^\top \boldsymbol{\beta}) = \theta(L) r_t$, onde $\phi(L) = -\sum_{j=0}^{p} \phi_j z^j$ e $\theta(L) = -\sum_{j=0}^{q} \theta_j z^j$, para todo $z \in \mathbb{C}$, onde L é o operador *backward*, $\phi_0 = -1 = \theta_0$ e $\phi(\cdot)$ e $\theta(\cdot)$ não possuem raízes em comum e r_t é um erro aleatório. Logo a estrutura a ser considerada é $g(\mu_t) = \nu + \boldsymbol{x}_t^\top \boldsymbol{\beta} + \sum_{j=1}^{p} \phi_j \left[g(y_{t-j}) - \boldsymbol{x}_{t-j}^\top \boldsymbol{\beta} \right] + \sum_{j=1}^{q} \theta_j r_{t-j}$, onde $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top$ e $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ são os vetores de parâmetros autorregressivos e de médias móveis, respectivamente, e r_t é o termo de erro aleatório, o qual será dado por $r_t = g(y_t) - g(\mu_t)$ e $\nu \in \mathbb{R}$ é uma constante. A estimação do vetor de parâmetros $\boldsymbol{\theta} = (\delta, \sigma, \lambda, \eta, \gamma, \boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top, \boldsymbol{\theta}^\top)^\top$ do modelo proposto será realizada utilizando os estimadores de máxima verossimilhança condicional (EMV), onde a função de log-verossimilhança condicional de $\boldsymbol{y}$ é dada por $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{t=m+1}^{n} \ell_t(\alpha, \sigma, \lambda, \eta, \gamma, \mu_t)$, com $\ell_t(\alpha, \sigma, \lambda, \eta, \gamma, \mu_t) = \log(\alpha\lambda\sigma\delta\eta\gamma) - (\delta+1)\log(y_t) - (\sigma+1)\log(z_t) + (\gamma+1)\log\left(1 - z_t^{-\sigma}\right) + (\eta-1)\log\left[1 - \left(1 - z_t^{-\sigma}\right)^{\gamma}\right] + (\lambda-1)\log\left\{1 - \left[1 - \left(1 - z_t^{-\sigma}\right)^{\gamma}\right]^{\eta}\right\}$, com $m = \max\{p, q\}$, $z_t = (1 + \alpha y_t^{-\delta})$ e reparametrização $\alpha = \mu_t^{\delta}\{[(1 - (1 - (1 - \tau^{\frac{1}{\lambda}})^{\frac{1}{\eta}})^{\frac{1}{\gamma}})^{-\frac{1}{\sigma}} - 1]\}$, com $\tau \in (0, 1)$. Foram realizadas simulações de Monte Carlo para avaliar as propriedades dos EMV, os resultados indicam que os estimadores são consistentes, assintoticamenente não viesados e normalmente distribuídos.

Palavras-chave: Modelo de Séries Temporais DARMA; Estimação de Máxima Verossimilhança; Simulações de Monte Carlo.

[1]Departamento de Estatística, UFSM – vitorpereira3115@gmail.com
[2]Departamento de Estatística, UFSM – cleber.bisognin@ufsm.br
[3]Departamento de Estatística, UFSM – lais.loose@ufsm.br

P118

Modelos Estatísticos para o Índice de Herfindahl

Viviana Giampaoli[1].

O intuito da pesquisa foi avaliar o desempenho de modelos Beta inflacionados aplicados ao índice de Herfindahl, por meio de técnicas de validação cruzada. A presença de zeros requer cuidados adicionais na aplicação destes métodos. Assim, um processo de modelagem foi formulado, levando em conta incluso a possível presença de multicolinearidade. Este índice tem sido amplamente estudado no caso de aglomeração de indústrias. Entre tanto, neste trabalho, este índice foi aplicado ao fenômeno de aglomeração de redes de comida rápida no Brasil. Por meio do processo de modelagem analisou-se que fatores referentes aos munícipios considerados no estudo, favoreceram no processo de agrupamento de lojas no país. Os resultados principais desta análise apontaram que entre outras variáveis a quantidade de homicídios e a população do município têm impactos significantes na aglomeração de lojas estrangeiras e nacionais.

Este projeto foi parcialmente financiado pelo Projeto FAPESP "Instituto Nacional de Ciência e Tecnologia de Fluidos Complexos - INCT-FCx", processo no 2014/50983-3 financiado pelo CNPq/FAPESP.

Palavras-chave: Índice de Herfindahl; Modelos Beta Inflacionados; Validação Cruzada

[1]Departamento de Estatística-Universidade de São Paulo – vivig@ime.usp.br

P119

Indicadores de Desempenho em Testes Adaptativos Informatizados

Viviane do Nascimento Figueiredo[1]; Héliton Ribeiro Tavares[2]

Os KPI (Key Performance Indicator), ou Indicador-Chave de Desempenho, são indicadores largamente utilizados em avaliações de pessoas, empresas, processos, dentre outros. Sua apresentação adequada facilita muito a visualização da quantificação inerentes ao construto avaliado. Os Testes Adaptativos Informatizados (TAI ou CAT - Computerized Adaptive Testing) têm recebido grande atenção nos últimos anos e terá crescimento ainda maior nos próximos anos, com muitas instituições já migrando suas rotinas de forma otimizada com o TAI. Normalmente são apresentados indicadores de domínio por Habilidades ou Competências de Área, individual ou conjunto. Neste trabalho os KPI estarão associados ao contexto de TAI com resultados de simulação e dados da Prova São Paulo (PSP), instrumentalizando as instituições educacionais com informações úteis para definições de políticas. Além dos KPI associados a proficiências individuais e grupos, também serão apresentados KPIs associados aos itens dinamicamente. Ainda, serão explorados diferentes métodos de estimação de domínios associados a critérios de parada e seleção de próximo item, inerentes aos sistemas TAI. Todo o processo foi construído com a utilização de pacotes R e códigos próprios.

Palavras-chave: KPI; Teoria da Resposta ao Item; Proficiência; Critério de Parada; Máxima Informação Global.

[1]Faculdade de Estatística - UFPA, Belém-PA – viviane.figueiredo@icen.ufpa.br
[2]Faculdade de Estatística - UFPA, Belém-PA – heliton@ufpa.br

P120

Spatial Autoregressive (SAR) Modelling of Crimes in the State of São Paulo

Wellington Yuanhe Zhao[1]; Luis Gustavo Nonato[2]; Cibele M. Russo [2]

The issue of public security is a considerable challenge for the Brazilian society and criminality is a great concern in the most populous state in the country, São Paulo. It is often desirable for public management to model and predict crime patterns considering historical data available and the georeference of each municipality (i.e., latitude and longitude). In this context, the use of geospatial models to explain the relationship between predictors and crimes, considering geolocation, can be of great importance. A possible model is the SAR (spatial autoregressive) model, which takes into account the covariates, as well as the underlying spatial dependence (Kazar and Celik, 2012). In this work, SAR model is used to describe and model the number of crimes in the cities of the state of São Paulo in Brazil, including also the monthly seasonality observed in the data. To create the precise model, we make use of packages in Python and R to organize and visualize the data and develop the modelling using the spatial neighborhood matrix. The lasso method is used to select the most significant covariates, for instance inhabitants per household, public elementary school failure rate, public early years, elementary school dropout rate, and then the SAR model is applied to include the spatial information and enrich the modelling of crimes.

Palavras-chave: Crimes Modelling; Geo-Spatial Modelling; SAR Model; State of São Paulo; Security Data.

[1]Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs) de Universidade de São Paulo e Universidade Federal de São Carlos, São Carlos – wellington.zhao@usp.br

[2]Departamento de Matemática Aplicada e Estatística, Universidade de São Paulo, São Carlos – gnonato@icmc.usp.br, cibele@icmc.usp.br

P121

Classificação Supervisionada e RNA para Atividades Físicas Realizadas por Gestantes Registradas via Dados do Acelerômetro

Gleici da Silva Castro Perdoná[1]; Christoph Michael Mitschkao[2]; Rafael B Fazio[3]; Carla Micheli da Silva[4]

A atividade física, mesmo que de intensidade leve, é importante para a saúde da mãe e do feto durante a gravidez e após o parto. O risco de pré-eclâmpsia é reduzido, assim como os problemas após o parto e a prevalência de depressão pós-parto. É fundamental analisar a quantidade de atividade física entre gestantes de diversos níveis socioeconômicos e estilos de vida, a fim de compreender melhor os elementos que influenciam seus hábitos de atividade física, bem como ajudar a desenvolver novas diretrizes e regulamentações públicas. Esta pesquisa faz parte do projeto PPSUS-FAPESP N.2019/03984-8, que tem como objetivo definir e classificar as atividades físicas realizadas por 150 gestantes na cidade de Ribeirão Preto no Brasil no Sistema Único de Saúde (SUS) , com base em dados gerados por meio do uso de acelerômetros e de um aplicativo para registro das atividades físicas realizadas. Para avaliar a necessidade de atividade física, está sendo desenvolvido um assistente virtual no projeto https://eva.fmrp.usp.br/EVA, que visa acompanhar gestantes nesse período e analisar sua necessidade de atividade física e, com base nos resultados, fazer recomendações. Consequentemente, cuidar da saúde da gestante. Para tanto, foram aplicadas técnicas de limpeza e processamento de dados e, então, para classificação supervisionada, foram considerados *LightGBM* (tree-based gradient *boosting*) e redes neurais artificiais do tipo *Long short - memória de termo* (LSTM). Dentre as conclusões, o treinamento utilizando um período de 30 segundos é apontado como a abordagem com as melhores métricas de precisão. Alguns pontos fracos desta abordagem também foram identificados e possíveis melhorias derivadas.

Palavras-chave: Atividades Físicas; Assistente Virtual; Rede LSTM; Reconhecimento de Padrões; LightGBM.

[1]Departamento de Medicina Social,FMRP-USP, Ribeirão Preto – pgleici@fmrp.usp.br
[2]MBA em Ciencias de dados, ICMC-USP – ch.mitschka@gmail.com
[3]Graduado em IBM,FMRP-USP, Ribeirão Preto – rafaelbdefazio@usp.br
[4]4Pós Graduação em Saúde Coletiva-Departamento de Medicina Social,FMRP-USP, Ribeirão Preto – carlasilva@usp.br


Programa de
Pós-Graduação
EM ESTATÍSTICA
IME-UFRGS

Appsilon

R Studio

FAPESP

ABE

CAPES

Big
Data

CNPq

FAPERGS

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL