PROGRAMA E RESUMOS

DOS

ENCONTROS DE BIOMETRIA

2013

I Encontro Português de Biometria
I Encontro Luso-Galaico de Biometria

14-16 Julho 2013

Braga, Portugal

Editado por Giovani Silva, Luzia Gonçalves, Pedro Oliveira

ISBN 978-972-8890-29-2
Número de exemplares: 120
Julho, 2013

Bem-vindos aos EB2013

Caros Colegas,

É com enorme satisfação que organizamos o I Encontro Português de Biometria e o I Encontro Luso-Galaico de Biometria (EB2013), integrados nas comemorações do Ano Internacional da Estatística e que, esperamos, sejam os primeiros de uma longa série.

Importa salientar que esta é uma iniciativa conjunta da Sociedade Portuguesa de Estatística (SPE) e da Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO), o que bem demonstra que as fronteiras são artefactos humanos que não dividem um espaço linguístico e cultural comum de séculos. Prova disso é o facto de as Comissões Científica e Organizadora, constituídas com membros de ambas as comunidades, desenvolverem um trabalho em cooperação constante.

A comunidade científica de raiz ibero-americana, mas não só, respondeu à chamada e, assim, foram submetidas 85 comunicações, contando o programa final com 38 comunicações orais e 34 sob a forma de poster, com um total de 86 participantes.

Os Encontros de Biometria incluem ainda duas mesas-redondas, uma sobre "Registos Clínicos" e outra sobre "Os desafios atuais da Epidemiologia e a sua ligação à Biometria", que contam com o contributo de especialistas a quem agradecemos a disponibilidade para participar nestes encontros. O programa científico compreende também um minicurso sobre "Análise de Dados Categorizados Incompletos", lecionado por Julio da Motta Singer.

Um agradecimento especial merecem os oradores convidados, Alan Agresti (University of Florida), Lucília Carvalho (Universidade de Lisboa), Charmaine Dean (Western University of Ontario), Guadalupe Gómez Melis (Universitat Politècnica de Catalunya) e Julio da Motta Singer (Universidade de São Paulo), que aceitaram o nosso convite sem qualquer hesitação. O seu contributo constitui uma marca relevante nestes encontros.

Várias entidades, públicas e privadas, da Galiza e de Portugal subsidiaram a realização dos encontros. Neste momento de crise, estes apoios são ainda mais relevantes e, por isso, merecem o nosso especial reconhecimento.

Por último, é intenção destes encontros refletir e discutir formas de agrupamento de investigadores em Biometria, fortalecendo e projetando as duas comunidades. Que estes encontros possam ser recordados como um primeiro passo nesse sentido.

Braga, Julho 2013

Pela Comissão Organizadora

Pedro Oliveira

Dois poemas da "mesma" língua

E eu tenho de partir para saber
Quem sou, para saber qual é o nome
Do profundo existir que me consome
Neste país de névoa e de não ser.

Sophia de Mello Breyner Andresen

Este vaise i aquel vaise,
e todos, todos se van.
Galicia, sin homes quedas
que te poidan traballar.
Tés, en cambio, orfos e orfas
e campos de soledad,
e nais que non teñen fillos
e fillos que non ten pais.
E tés corazóns que sufren
longas ausencias mortás,
viudas de vivos e mortos
que ninguén consolará

Rosalía de Castro

PATROCINADORES:

Portugal

- Universidade do Minho: Escola de Ciências - Departamento de Matemática e Aplicações - Centro de Matemática
- Centro de Estatística e Aplicações da Universidade de Lisboa
- FCT - Fundação para a Ciência e a Tecnologia
- ICBAS - Instituto de Ciências Biomédicas Abel Salazar
- Empresa ST+I, Unipessoal, Lda.
- Instituto dos Vinhos do Douro e Porto, I.P.
- PSE - Produtos e Serviços de Estatística
- Escolar Editora

Galiza

- BIOSTATNET
- Comunidade de Traballo - Galiza e Norte de Portugal
- Cooperación Transfronteriza España - Portugal
- Unión Europea - FEDER
- Xunta de Galicia

ORGANIZADORES:

- Sociedade Portuguesa de Estatística (SPE)
- Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO)

COMISSÃO ORGANIZADORA:

Portugal

Pedro Oliveira - ISPUP e ICBAS, Universidade do Porto (Presidente)
Cecília Azevedo - CMat, Universidade do Minho
Luzia Gonçalves - IHMT, Universidade Nova de Lisboa
Denisa Mendonça - ISPUP e ICBAS, Universidade do Porto
Giovani Silva - IST, Universidade Técnica de Lisboa

Galiza

Balbina Casas Mendez - DEIO, Universidade de Santiago de Compostela (Presidente)
Esther López Vizcaino - Instituto Galego de Estadística
María del Carmen Iglesias Pérez - DEIO, Universidade de Vigo
Javier Roca Pardiñas - DEIO, Universidade de Vigo
Marta Sestelo Pérez - DEIO, Universidade de Vigo

COMISSÃO CIENTÍFICA:

Portugal

Carlos Daniel Paulino - IST, Universidade Técnica de Lisboa (Presidente)
Dinis Pestana - DEIO, Universidade de Lisboa
Luís Machado - CMat, Universidade do Minho
Alcindo Maciel Barbosa - Unidade Local de Saúde do Alto Minho, EPE
Henrique Barros - Instituto de Saúde Pública da Universidade do Porto
José Pereira Miguel - Faculdade de Medicina da Universidade de Lisboa e INSA-Ricardo Jorge

Galiza

Antonio Vaamonde Liste - DEIO, Universidade de Vigo (Presidente)
Wenceslao González Manteiga - DEIO, Universidade de Santiago de Compostela
Carmen Cadarso Suárez - DEIO, Universidade de Santiago de Compostela
Jacobo de Uña Álvarez - DEIO, Universidade de Vigo
Ricardo Cao Abad - DM, Universidade da Coruña
Francisco Gude Sampedro - Hospital Clínico Universitario de Santiago

QUADRO DO PROGRAMA

14 Julho	10:30–12:30	Minicurso: Análise de Dados Categorizados Incompletos. Julio da Motta Singer
	12:30–14:00	Almoço
	14:00–16:00	Minicurso: Análise de Dados Categorizados Incompletos. Julio da Motta Singer
	16:00–16:20	Intervalo café
	16:20–16:45	Sessão de Abertura
	16:45–17:30	Sessão de Posters
	17:30–18:30	Sessão Plenária I: Good Confidence Intervals for Categorical Data Analyses. Alan Agresti
	19:00–21:00	Passeio + Receção
15 Julho	09:00–10:00	Comunicações Orais: Sessão I. Análise Longitudinal e de Sobrevivência Sessão II. Epidemiologia Veterinária
	10:00–10:20	Intervalo café
	10:20–11:20	Sessão Plenária II: Epidemias de Gripe: Três Problemas, Três Abordagens. Lucília Carvalho
	11:20–12:40	Mesa-redonda I: Registos clínicos
	12:40–14:00	Almoço
	14:00–15:40	Comunicações Orais: Sessão III. Análise de Sobrevivência Sessão IV. Avaliação e Diagnóstico
	15:40–16:00	Intervalo café
	16:00–17:00	Sessão Plenária III: Joint Analysis of Multivariate Spatial Count and Zero-Heavy Count Outcomes using Common Spatial Factor Models. Charmaine Dean
	17:00–18:00	Assembleia
	19:00–20:00	Concerto
	20:00–22:00	Jantar
16 Julho	09:00–11:00	Comunicações Orais: Sessão V. Análise de Regressão e Problema de Triagem Sessão VI. Aplicações em Genética, Biologia e Ecologia
	11:00–11:20	Intervalo café
	11:20–12:40	Mesa-redonda II: Os desafios actuais da Epidemiologia e a sua ligação à Biometria
	12:40–14:00	Almoço
	14:00–15:40	Comunicações Orais: Sessão VII. Dados Omissos, Imputação e Valores-P Sessão VIII. Epidemiologia, Demografia e Qualidade de Vida
	15:40–16:00	Intervalo café
	16:00–17:00	Sessão Plenária IV: Método para Decidir entre un Evento Compuesto o una de sus Componentes como Variable Principal en en Ensayo Clínico. Plataforma Web para Facilitarlo. Guadalupe Gómez Melis
	17:00–17:15	Sessão de Encerramento

Local das atividades (Escolas de Ciências da Universidade do Minho):

- Minicurso, Sessões Plenárias, Abertura e Encerramento: **Anfiteatro ECUM**
- Comunicações Orais: **CP1-103** (sala A) e **CP1-104** (sala B)
- Comunicações em Poster: **Hall do CP1**
- Intervalos do café: **CP1-201** (sala C).

Programa

14 Julho 2013

10:30-12:30 & 14:00-16:00 Minicurso

Coordenador: Pedro Oliveira

1 ANÁLISE DE DADOS CATEGORIZADOS INCOMPLETOS
Julio da Motta Singer

16:45-17:30 Sessão de Posters

Coordenadoras: Denisa Mendonça, Esther López Vizcaino

2 Análise de dados de NGS para o estudo da expressão diferencial de mutações no gene SETD2 das células do carcinoma renal
Catarina Almeida, Lisete Sousa, Ana Rita Grosso

4 Análise de sobrevivência com informação incompleta nas covariáveis - Estudo de simulação
Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça

5 Desigualdades socioeconómicas na sobrevivência de doentes diagnosticados com tumores do estômago e bexiga na Região Norte de Portugal
Luís Antunes, Bernard Rachet, Maria José Bento, Denisa Mendonça

7 Valor prognóstico de biomarcadores no carcinoma da mama masculina
Mariana Aparício, Giovani L. Silva, António E. Pinto

8 Mapa polínico atmosférico da região de Évora - 2001 a 2007
Ricardo Assunção, Manuela Oliveira, Fernando Afonso, Elsa Caeiro, Rui Brandão, Maria Luísa Lopes

10 Predicción del origen tuberculoso del derrame pleural mediante los modelos aditivos generalizados robustos
Laura Calaza Díaz, Mónica López-Ratón, Carmen Cadarso-Suárez, Francisco Gude-Sampedro, Luis Valdés-Cuadrado

12 Estudo do perfil sensorial de vinhos do Porto
Elisete Correia, Bebiana Monteiro, Alice Vilela

13 Control de Calidade en Aerobioloxía: Comparación de diferentes métodos de lectura de láminas
Tomás R. Cotos-Yáñez, Francisco J. Rodríguez-Rajo, Ana Pérez-González, Victoria Jato

15 História familiar de eventos cardiovasculares, rigidez arterial e pressão arterial central: Estudo de determinação do risco cardiovascular da população de Guimarães e Vizela, incluindo a prevalência de rigidez arterial e envelhecimento arterial precoce
Pedro G. Cunha, Pedro Oliveira, Jorge Cotter, Isabel Vila, Nuno Sousa

16 Estudo da mortalidade por suicidio en Galicia
María José Ginzo-Villamayor, Rosa María Crujeiras-Casais, María Esther López-Vizcaíno, María Isolina Santiago-Pérez

18 Definição operacional de doença periodontal na prática clínica e em epidemiologia
José A. Lobo Pereira, Teresa Oliveira, Antonio Costa, Luzia Mendes

19 Income from influenza in the pandemic period in Hospital Universitario A Coruña
Beatriz López-Calviño, María José Pardo-Landrove, Sonia Pértega-Díaz, Teresa Seoane-Pillado, Salvador Pita-Fernández, José Manuel Suárez-Lorenzo, María Fernández-Albalat

21 Avaliação dos cuidados de saúde: implementação de valores target
Bruno Marques

23 Distribuição espacial de casos de dengue nos estados brasileiros
Natália da Silva Martins, Paulo Justiniano Ribeiro Junior

25 Inferência no modelo doença-morte
Luís Meira Machado, Artur Araújo, Jacobo de Uña-Álvarez

27 Estimação da função condicional de tempos de sobrevivência sucessivos
Luís Meira Machado, Ana Moreira

29 A influência da idade da mãe no desempenho da escala CRIB: curva ROC condicionada
Filipa Mourão, Ana Cristina Braga, Pedro Oliveira

30 Exame à medula óssea e reações adversas
Sara Narciso, Fernanda Diamantino, Ana Batalha Reis

32 Modelos de idade-período-coorte para a projeção da incidência de cancro: novas abordagens
Joana Oliveira, Clara Castro, Luís Meira Machado

33 VIH/SIDA Estimação de probabilidades de atrasos e sub-notificação
Alexandra Oliveira, Joaquim Pinto da Costa, Ana Rita Gaio

34 Regressão quantílica para dados longitudinais: uma aplicação na área da Medicina
Ana Luísa Papoila, Marta Alves, Daniel Virella, Andreia Mascarenhas, Teresa Neto

36 Análise semiparamétrica bayesiana da diversidade do repertório de recetores de células T
Carlos Daniel Paulino, Michele Guindani, Peter Müller

37 Ajuste da função logística a dados de crescimento
Glauber Márcio Silveira Pereira, Lídia Raquel de Carvalho, Martha Maria Mischan

39 Biomonitorização ambiental em Portugal continental - uma análise espacial e temporal
Helena Piairo, Raquel Menezes, Inês Sousa

41 O modelo de Crámer-Lundberg e a probabilidade de ruína
Celine Queirós, Patrícia Gonçalves, Irene Brito

43 Modelagem do tempo de vida de pacientes com leucemia utilizando a distribuição geométrica beta generalizada semi-normal
Thiago Gentil Ramires, Edwin Moises Marcos Ortega

45 Influence of human mitochondrial DNA haplogroups on risk of osteoarthritis progression: an interval-censored approach
Ignacio Rego-Perez, Angel Soto-Hermida, Sonia Pértega-Díaz, Mercedes Fernández-Moreno, Juan Fernández-Tajes, Eugenia Vazquez-Mosquera, Estefania Cortes-Pereira, Sara Relaño-Fernández, Natividad Oreiro-Villar, Carlos Fernández-López, Francisco Blanco-Garcia, Beatriz López-Calviño, Teresa Seoane-Pillado

47 Variabilidade espacial de cálcio no solo da região de Unaí-MG, Brasil
Ana Julia Righetto, Luiz Ricardo Nakamura, Diogo Néia Eberhardt, Paulo Justiniano Ribeiro Junior, Roseli Aparecida Leandro

49 Bivariate kernel smoothers. Applications in thoracic aorta pathology
Javier Roca-Pardiñas, Francisco de Asis López-Alvarez, Pablo G. Tahoces, Juan A. Martínez-Mera, Jose M. Carreira

51 Ajuste de modelos platô de resposta em dados de crescimento de bovinos da raça nelore do estado de Minas Gerais
Tânia Jussara Silva Santana, João Domingos Scalon, Isabel Cristina Costa Leite, Azly Santos Amorim de Santana, Ângela Cristina da Fonseca Mirante

52 O papel do peso das caudas na utilização de testes quantitativos compostos
Rui Santos, Miguel Felgueiras, João Paulo Martins

54 ANOVA não-paramétrica de dados de contagem com medidas repetidas e com excesso de zeros
Soane Mota dos Santos, Julio da Motta Singer

55 Validacão das medidas de curvatura, de um modelo não linear para crescimento em altura de eucalyptus, através da reparametrização
Romulo Barbosa Veloso, Natalino Calegário

57 Avaliação de mecanismos de imputação múltipla na análise de sobrevivência com dados omissos
Ana Margarida Vinhas, Luís Antunes, Inês Sousa

17:30-18:30 Sessão Plenária I

Coordenador: Carlos Daniel Paulino

59 GOOD CONFIDENCE INTERVALS FOR CATEGORICAL DATA ANALYSES
Alan Agresti

15 Julho 2013

9:00-10:00 Sessão I - Comunicação Oral - Sala A

Tema: Análise Longitudinal e de Sobrevivência ***Coordenador: Javier Roca-Pardiñas***

60 Survival and longitudinal analysis of breast cancer at Braga's hospital
Inês Sousa, Ana Borges, Luís Castro

62 Análise conjunta bayesiana de dados longitudinais e de sobrevivência com fracção de cura espacial e sua aplicação ao estudo do VIH
Rui Martins, Giovani L. Silva, Valeska Andreozzi

64 Modelação conjunta de dados longitudinais e eventos competitivos em doentes renais crónicos
Laetitia Teixeira, Anabela Rodrigues, Inês Sousa, Denisa Mendonça

9:00-10:00 Sessão II - Comunicação Oral - Sala B

Tema: Epidemiologia Veterinária ***Coordenador: Tomás R. Cotos Yáñez***

66 Using independent information in distance sampling surveys to account for animal density gradients with respect to transects
Regina Bispo, Tiago A. Marques, Stephen T. Buckland, Brett Howland

67 Critério para descarte de animais em remodelação cardíaca
Renan Mercuri Pinto, Antônio Carlos Cicogna, Carlos Roberto Padovani

69 Simulation model of salmonella dynamics on a farrow-to-finish herd
Carla Correia Gomes, Theodoros Economou, Trevor Bailey, João Niza Ribeiro

10:20-11:20 Sessão Plenária II

Coordenadora: Carmen Cadarso Suárez

71 EPIDEMIAS DE GRIPE: TRÊS PROBLEMAS, TRÊS ABORDAGENS
Maria Lucília Carvalho

11:20-12:40 Mesa-redonda I

73 Registos clínicos
Oradores: Maria de Fátima de Pina, Pilar Gayoso-Diz, Javier Muñiz-García, Paulo Costa; Organizadores: María del Carmen Iglesias-Pérez, Cecília Azevedo

14:00-15:40 Sessão III - Comunicação Oral - Sala A

Tema: Análise de Sobrevivência ***Coordenador: Luís Meira Machado***

75 Survival of branching processes through mutations
Maria Conceição Serra, Serik Sagitov

76 Modelos mecanísticos de sobrevivência
Francisco Louzada Neto

77 A distribuição de Birnbaum-Saunders gerada pela lei Logística
Emilia Athayde, Assis Azevedo

79 Survival of colorectal cancer patients with unknown censorship
Beatriz López-Calviño, Ricardo Cao-Abad, Ewa Strzalkowska-Kominiak, Sonia Pértega-Díaz, Salvador Pita-Fernández, Teresa Seoane-Pillado

81 Modelos de Sobrevivência Weibull modificado aplicados a dados de câncer de mama
Gleici Perdona, Francisco Louzada Neto, Cleyton Zanardo, Hayala Cavenague

14:00-15:40 Sessão IV - Comunicação Oral - Sala B

Tema: Avaliação e Diagnóstico ***Coordenadora: Maria Antónia Amaral Turkman***

83 Comparação de teste de rastreio de glaucoma primário de ângulo aberto na diabetes tipo 2, utilizando curvas ROC
Ana Cristina Braga, Hugo Frade, Lígia Figueiredo, Dália Meira

85 Adesão ao tratamento em doenças crônicas
Fernando Gomes, Cecília Azevedo

87 Avaliação de imagens radiográficas de sementes usando ICA
Isabel Cristina Costa Leite, Thelma Sáfadi, Maria Laene Moreira de Carvalho

89 Adequação de modelos de classes latentes a planos experimentais relevantes no contexto biomédico
Ana Subtil, Luzia Gonçalves, Patrícia de Zea Bermudez

91 Investigating the performances of classification techniques: an application to medical diagnosis data
Derya Ersel, Suleyman Gunay

16:00-17:00 Sessão Plenária III

Coordenador: Kamil Feridun Turkman

93 JOINT ANALYSIS OF MULTIVARIATE SPATIAL COUNT AND ZERO-HEAVY COUNT OUTCOMES USING COMMON SPATIAL FACTOR MODELS
Charmaine Dean

16 Julho 2013

9:00-11:00 Sessão V - Comunicação Oral - Sala A

Tema: Análise de Regressão e Problema de Triagem ***Coordenadora: Ana Pérez González***

94 Bayesian joint analysis of zero-inflated counting and severity data
Giovani L. Silva, Elizabeth Juarez-Colunga, Charmaine Dean

95 Modelos de regressão binária: que ligação escolher?
Isabel Natário, Sílvia Shrubsall

97 Regressão quantílica bayesiana para proporções
Bruno Santos, Heleno Bolfarine

99 Análisis cluster de curvas de regresión no parametrica en recursos marinos
Nora M. Villanueva, Marta Sestelo, Javier Roca-Pardiñas

101 Modelação flexível do problema de triagem via processos de Dirichlet dependentes
Sandra Ramos, Maria Antónia Amaral Turkman, Marília Antunes

102 Identificación de factores de riesgo de lesión en el fútbol profesional
María del Carmen Iglesias-Pérez, Miguel Martínez-González, Luis Casáis-Martínez, Marta Sestelo, Javier Roca-Pardiñas

9:00-11:00 Sessão VI - Comunicação Oral - Sala B

Tema: Aplicações em Genética, Biologia e Ecologia Coordenadora: Patrícia de Zea Bermudez

104 Arrow plot: um novo gráfico para a seleção de genes em dados de microarrays
Carina Silva-Fortes, Maria Antónia Amaral Turkman, Lisete Sousa

105 Efeito do preprocessamento de dados na deteção de genes diferencialmente expressos
Adelaide Freitas, Sara Roque

106 Dinâmica da população do tubarão Centroscymnus coelolepis nas águas continentais portuguesas
Ivone Figueiredo, Isabel Natário, Teresa Moura, Maria Lucília Carvalho

108 Ecologia das comunidades vegetais: utilização da distribuição multinomial numa perspetiva bayesiana
Luís Silva

110 Comparação dos desempenhos de semeadoras manuais por meio da distribuição triangular discreta generalizada
Silvio Sandoval Zocchi, Célestin C. Kokonendji

112 Um modelo de Markov espaço-temporal não homogéneo para a ocorrência de Dengue
Marília Antunes, Maria Antónia Amaral Turkman, Kamil Feridun Turkman, Marco A. Horta, Cristina Catita

11:20-12:40 Mesa-redonda II

113 Os desafios actuais da Epidemiologia e a sua ligação à Biometria
Oradores: Paulo Ferrinho, Xurxo Hervada-Vidal, Helena Sofia Rodrigues; Moderador: Vitor Rodrigues; Organizadores: Luzia Gonçalves, María Esther López-Vizcaíno

14:00-15:40 Sessão VII - Comunicação Oral - Sala A

Tema: Dados Omissos, Imputação e Valores-P ***Coordenador: Julio da Motta Singer***

114 Assessing the limits of multiple imputation in tackling missing genotypes in a scenario of limited genetic information
Nuno Sepúlveda, Alphaxard Manjurano, Taane Clark, Eleanor Riley, Chris Drakeley

115 Análise bayesiana semiparamétrica de respostas binárias com uma covariável contínua sujeita a omissão informativa
Frederico Zanqueta Poleto, Carlos Daniel Paulino, Julio da Motta Singer, Geert Molenberghs

116 Impacto de valores omissos em estudos epidemiológicos - Uma aplicação na modelação do Índice de Massa Corporal
Beatriz Preto Goulão, Valeska Andreozzi, Patrícia de Zea Bermudez

118 Combinação de valores de prova e valores de prova generalizados
Maria de Fátima Brilhante, Dinis Pestana, Fernando Sequeira

120 Adjusted p-values for SGoF multitesting procedure: Definition and properties
Irene Castro-Conde, Jacobo de Uña-Álvarez

14:00-15:40 Sessão VIII - Comunicação Oral - Sala B

Tema: Epidemiologia, Demografia e Qualidade de Vida ***Coordenadora: Valeska Andreozzi***

122 Influência do nível socioeconómico da região no risco de fratura do fémur proximal
Carla Oliveira, Denisa Mendonça, Maria de Fátima de Pina

124 Aplicação de modelos de equações estruturais na avaliação da qualidade de vida em pessoas com doenças metabólicas
Estela Vilhena, José Luís Pais Ribeiro, Luísa Pedro, Isabel Silva, Rute F. Meneses, Helena Cardoso, António Martins da Silva, Denisa Mendonça

126 Las tablas de mortalidad de Galicia y la Región Norte de Portugal empleando el software R
María Martín-Vila, María Esther Calvo-Ocampo, Solmary Silveira-Calviño, Gael Naveira-Barbeito, María Isolina Santiago-Pérez, Carlos Iglesias-Patiño, María Esther López-Vizcaíno

128 Modelação e projeção da incidência de cancro colo-rectal e do estômago no Sul de Portugal
Ricardo São João, Ana Luísa Papoila, Bruno de Sousa, Ana Miranda

130 Técnicas de meta-análise na estimação de uma taxa de prevalência
João Paulo Martins, Miguel Felgueiras, Rui Santos

16:00-17:00 Sessão Plenária IV

Coordenador: Jacobo de Uña Álvarez

132 MÉTODO PARA DECIDIR ENTRE UN EVENTO COMPUESTO O UNA DE SUS COMPONENTES COMO VARIABLE PRINCIPAL EN UN ENSAYO CLÍNICO. PLATAFORMA WEB PARA FACILITARLO
Guadalupe Gómez-Melis

135 **Índice de Autores**

RESUMOS

I Encontro Português de Biometria & I Encontro Luso-Galaico de Biometria

Minicurso

Análise de dados categorizados incompletos

Julio da Motta Singer
Departamento de Estatística, Universidade de São Paulo, Brasil, jmsinger@ime.usp.br
Carlos Daniel Paulino
Instituto Superior Técnico, Universidade Técnica de Lisboa, dpaulino@math.ist.utl.pt
Frederico Zanqueta Poleto
Moody's Analytics, frederico@poleto.com

Palavras–chave: Dados incompletos, Dados faltantes/omissos, MAR, MCAR, MNAR, Análise de dados categorizados.

Resumo: Nosso objetivo é descrever a análise de dados categorizados com omissão modelados por distribuições multinomiais. Com essa finalidade, desenvolvemos os resultados em formulação matricial adequada para a implementação computacional, concretizada por meio das rotinas ACD (*Analysis of Categorical Data*) programadas no ambiente estatístico R. Motivamos o curso com a apresentação de conjuntos de dados originados de problemas de natureza biológica para os quais descrevemos as questões de interesse. Indicamos os modelos probabilísticos adotados e especificamos modelos lineares, log-lineares e mais geralmente, funcionais lineares, cuja finalidade é impor estruturas ditadas pelos propósitos inferenciais aos parâmetros do modelo probabilístico (daí a designação de modelos estruturais). Mostramos como a análise pode ser conduzida por metodologia de máxima verossimilhança (MV), de mínimos quadrados generalizados (MQG) ou híbrida (MV/MQG), em que estimativas dos parâmetros do modelo probabilístico e da correspondente matriz de covariâncias são obtidas por MV num primeiro estágio e utilizadas para modelagem estrutural por meio de MQG, no segundo. Para efeito de familiarização com os procedimentos propostos, consideramos inicialmente a análise de dados completos, detalhando tanto a expressão analítica dos modelos estruturais quanto sua especificação para a análise por intermédio das rotinas ACD. Em seguida, indicamos como os modelos precisam ser adaptados para levar em conta diferentes mecanismos de omissão, nomeadamente, omissão não informativa, seja ela aleatória (MAR, *missing at random*) ou completamente aleatória (MCAR, *missing completely at random*) ou omissão informativa ou não aleatória (MNAR, *missing not at random*). Em particular, consideramos o ajuste de modelos saturados e estruturais lineares e log-lineares por MV para casos em que o mecanismo de omissão é MAR ou MCAR e também utilizamos a metodologia MQG tanto nesse contexto como também naquele em que se considera o ajuste de modelos funcionais lineares, incluindo casos em que o mecanismo de omissão é MNAR. Assim como para dados completos, ilustramos a metodologia por meio da análise de exemplos práticos, detalhando o emprego das rotinas ACD.

Referências

[1] Poleto, F.Z. (2006). *Análise de Dados Categorizados com Omissão*. Tese de Mestrado, IME - Universidade de São Paulo (http://www.teses.usp.br/teses/disponiveis/45/45133/tde-04122007-192457/).

I Encontro Português de Biometria & I Encontro Luso-Galaico de Biometria

Poster

Análise de dados de NGS para o estudo da expressão diferencial de mutações no gene SETD2 das células do carcinoma renal

Catarina Almeida
DEIO, Faculdade de Ciências da Universidade de Lisboa, catarinaalmeida@fm.ul.pt

Lisete Sousa
DEIO e CEAUL, Faculdade de Ciências da Universidade de Lisboa, lmsousa@fc.ul.pt

Ana Rita Grosso
IMM, Faculdade de Medicina da Universidade de Lisboa, agrosso@fm.ul.pt

Palavras–chave: *Next generation sequencing* (NGS), Carcinoma renal, Dados de contagem, Distribuição binomial negativa, Sobredispersão.

Resumo: Estudos recentes identificaram a ocorrência de mutações somáticas no gene SETD2 no cancro renal (Dalgliesh *et al.*, 2010). O objetivo deste trabalho é a comparação de linhas celulares de cancro renal com o gene SETD2 mutado e normal, com o propósito de avaliar o papel do gene SETD2 como supressor tumoral na resposta aos danos do DNA. A comparação será feita a nível de: genes, *splicing* alternativo e lincRNA (*large intergenic non-coding* RNA). Os dados do transcriptoma foram obtidos com recurso à técnica de NGS (*Next Generation Sequencing*) RNA-seq.(de Almeida *et al.*, 2010) Esta é uma técnica de sequenciação recente em que o DNA é fragmentado numa biblioteca de pequenos segmentos codificantes (*reads*) que são uniformemente e correctamente sequenciados em milhões de reacções paralelas. Estas *reads* são depois alinhadas com um genoma de referência. Consequentemente, o alinhamento destas sequências de bases (*reads*) vai revelar a sequência completa de cada cromossoma na amostra de DNA (Nagalakshmi, Waern e Snyder, 2010).
Os dados foram analisados recorrendo a diferentes programas disponíveis para análise de sequências genéticas: bibliotecas do R (através do Bioconductor - EDASeq, edgeR, DESeq), Cufflinks e MISO. Estes programas analisam o transcriptoma, identificando os vários genes, quantificando os níveis de expressão de cada gene, normalizando-os e comparando os níveis de expressão de cada gene sob duas condições experimentais distintas. Os dados de que dispomos são números inteiros positivos, correspondendo a contagens. Cada contagem diz respeito ao número de *reads* que alinham com um determinado gene, reflectindo assim o nível de expressão desse gene mediante determinada condição (Anders e Huber, 2012).
As metodologias estatísticas utilizadas por estes programas na detecção de genes com expressão diferencial são variadas. A maioria dos métodos considera que as contagens seguem uma distribuição Binomial Negativa (EDAseq, DESeq, edgeR), embora a distribuição Normal truncada também seja utilizada (Cufflinks). A comparação dos níveis de expressão dos genes sob as duas condições experimentais (cancro renal com gene mutado/gene não mutado) pode ser feita recorrendo ao teste da razão de verosimilhanças (edgeR, Cufflinks, MISO) ou a um teste exacto análogo ao teste exacto de Fisher, embora adaptado para dados caracterizados pela existência de sobredispersão (Robinson e Smyth, 2008) (DESeq e edgeR).

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito dos projectos PEstOE/MAT/UI0006/2011 (CEAUL) e PTDC/MAT/118335/2010. As autoras agradecem a Silvia Carvalho e a Sergio de Almeida a produção e cedência dos dados.

Referências

[1] de Almeida, S.F., Grosso, A.R., Koch, F., Fenouil, R., Carvalho, S., Andrade, J., Levezinho, H., Gut, M., Eick, D., Gut, I., Andrau, J.C., Ferrier, P., Carmo-Fonseca, M. (2011). Splicing enhances recruitment of methyltransferase HYPB/SETD2 and methylation of histone H3 Lys36. *Nature Structural Molecular Biology* 18, 977–983.

[2] Anders, S., Huber, W. (2012). Differential expression analysis for sequence count data. *Genome Biology* 11, R106.

[3] Dalgliesh, G.L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C., Teague, J., Andrews, J., Barthorpe, S., Beare, D., Buck, G., Campbell, P.J., Forbes, S., Jia, M., Jones, D., Knott, H., Kok, C.Y., Lau, K.W., Leroy, C., Lin, M.L., McBride, D.J., Maddison, M., Maguire, S., McLay, K., Menzies, A., Mironenko, T., Mulderrig, L., Mudie, L., O'Meara, S., Pleasance, E., Rajasingham, A., Shepherd, R., Smith, R., Stebbings, L., Stephens, P., Tang, G., Tarpey, P.S., Turrell, K., Dykema, K.J., Khoo, S.K., Petillo, D., Wondergem, B., Anema, J., Kahnoski, R.J., Teh, B.T., Stratton, M.R., Futreal, P.A.(2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* 463, 360–363.

[4] Nagalakshmi, U., Waern, K., Snyder, M. (2010). RNA-Seq: A Method for Comprehensive Transcriptome Analysis. *Current Protocols in Modecular Biology* Supl.89, 4.11.1–4.11.13.

[5] Robinson, M.D., Smyth, G.K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332.

Poster

Análise de sobrevivência com informação incompleta nas covariáveis - Estudo de simulação

Luís Antunes
Instituto Português de Oncologia do Porto, Instituto de Saúde Pública da Universidade do Porto, *luis.antunes@ipoporto.min-saude.pt*

Bernard Rachet
London School of Hygiene and Tropical Medicine, *Bernard.Rachet@lshtm.ac.uk*

Maria José Bento
Instituto Português de Oncologia do Porto, *mjbento@ipoporto.min-saude.pt*

Denisa Mendonça
Instituto Ciências Biomédicas Abel Salazar, Instituto de Saúde Pública da Universidade do Porto, *dvmendon@icbas.up.pt*

Palavras–chave: Análise de sobrevivência, Dados omissos, Simulação.

Resumo: A existência de dados omissos em dados na área da saúde, é uma realidade com a qual um bioestatístico se confronta com uma regularidade superior à desejada. No caso concreto de dados de registos oncológicos de base populacional, é frequente encontrar informação em falta em factores de prognóstico importantes como o estadiamento da doença. Esta falta de informação pode levar a que os resultados da análise, que se efectua a esses dados, sejam envieados, especialmente se o mecanismo de omissão não for completamente aleatório.
Pretendeu-se avaliar o desempenho da imputação múltipla como abordagem para lidar com a existência de dados omissos nas covariáveis duma análise de sobrevivência, através de um estudo de simulação, para diferentes proporções de omissão. Foi utilizada como base para o estudo de simulação, uma amostra de dados de sobrevivência correspondente a doentes diagnosticados com tumores gástricos. Os dados foram disponibilizados pelo Registo Oncológico Regional do Norte (RORENO).
Para cada conjunto de dados de sobrevivência simulados, procedeu-se da seguinte forma: eliminação de uma proporção escolhida de casos seguindo padrões de omissão semelhantes aos observados nos dados reais; imputação dos valores omissos usando imputação múltipla; análise de sobrevivência dos dados completados; combinação das diferentes estimativas seguindo as regras de Rubin; comparação dos valores obtidos com os valores reais conhecidos.
A extensão da doença é um dos factores de prognóstico para o qual a proporção de casos omissos é normalmente elevada. O seu valor é completamente definido pelo valor das três variáveis T(tumor), N(nódulos linfáticos) e M(metastização). A ausência do conhecimento da variável T ou da variável N, impede a atribuição do valor da extensão. Observa-se, nos casos reais, que para uma certa proporção de casos apenas uma ou duas destas variáveis se encontra omissa. Utilizando apenas a variável extensão, perde-se a informação disponível nas variáveis T ou N que poderia ser conhecida. Pretendeu-se comparar o comportamento do algoritmo de imputação múltipla em duas situações: imputação directa da variável extensão da doença, sem utilização da informação do TNM e imputação das três variáveis seguida de atribuição do valor da extensão nos dados imputados. A distribuição dos valores imputados das duas formas foi comparada com as distribuições reais, assim como os resultados obtidos no modelo de sobrevivência.

Poster

Desigualdades socioeconómicas na sobrevivência de doentes diagnosticados com tumores do estômago e bexiga na Região Norte de Portugal

Luís Antunes
Instituto Português de Oncologia do Porto, Instituto de Saúde Pública da Universidade do Porto, *luis.antunes@ipoporto.min-saude.pt*

Bernard Rachet
London School of Hygiene and Tropical Medicine, *Bernard.Rachet@lshtm.ac.uk*

Maria José Bento
Instituto Português de Oncologia do Porto, *mjbento@ipoporto.min-saude.pt*

Denisa Mendonça
Instituto Ciências Biomédicas Abel Salazar, Instituto de Saúde Pública da Universidade do Porto, *dvmendon@icbas.up.pt*

Palavras–chave: Análise de sobrevivência, Factores socioeconómicos, Cancro.

1 Introdução

A análise de sobrevivência de dados de registos de cancro de base populacional é uma importante ferramenta de apoio à decisão. Permite a avaliação dos cuidados de saúde prestados à população coberta por esses mesmos registos e permite a avaliação de heterogeneidades no acesso a esses cuidados. Diferentes estudos têm demonstrado a existência de associação entre as condições socioeconómicas e a sobrevivência de doentes oncológicos. Estas foram já reportadas para países como Inglaterra, Estados Unidos, Austrália, entre muito outros [1]. Para doentes residentes em Portugal não existem, no entanto, resultados publicados sobre esta avaliação.

2 Objectivos

Descrever a sobrevivência de doentes diagnosticados com tumores malignos do estômago ou tumores malignos da bexiga, na Região Norte de Portugal, durante o período 2000-2006. Estudar a associação entre alguns indicadores socioeconómicos e a sobrevivência desses doentes.

3 Material e métodos

Foram incluídos na análise todos os doentes diagnosticados no período de interesse com tumores malignos do estômago ou bexiga, residentes na Região Norte de Portugal e registados pelo Registo Oncológico Regional do Norte (RORENO). A condição socioeconómica de cada doente foi atribuída com base em variáveis ecológicas apenas, visto esta informação não estar disponível a nível individual. O nível geográfico utilizado para esta atribuição foi a freguesia (população mediana: 745). Os indicadores utilizados (nível de escolaridade, analfabetismo, desemprego) foram disponibilizados pelo Instituto Nacional de Estatística e baseam-se na informação obtida nos Censos de 2001 e 2011. Consideraram-se ainda dois indicadores compostos, um indicador de ruralidade e um indicador de acessibilidade a bens e serviços. A sobrevivência relativa foi estimada usando o método Ederer II. O efeito dos factores de prognóstico foi avaliado, estimando Razões de Excesso de Risco (RER), através de modelos paramétricos fléxiveis. Estes permitem uma modelação mais adequada da função de risco de base usando splines cúbicas [2].

4 Resultados

No período de diagnóstico considerado, foram registados 7820 doentes com cancro do estômago e 3630 doentes com cancro da bexiga. A sobrevivência relativa aos 5 anos foi de $33,8\%$ para os tumores do estômago e $73,7\%$ para os tumores da bexiga. A sobrevivência foi significativamente superior nas mulheres em relação aos homens, tanto para os tumores do estômago como para os da bexiga (RER ajustado para a idade: 0.81 e 0.84, respectivamente). Resultados preliminares sugerem que a sobrevivência de doentes residentes em áreas com o menor nível educional e em áreas com o maior índice de ruralidade é significativamente inferior à sobrevivência dos doentes residentes nas restantes áreas.

5 Discussão

Os resultados sugerem que doentes provenientes de áreas mais desfavorecidas apresentam um pior prognóstico, para ambos os tumores analisados. Este pior prognóstico poderá estar relacionado com tendência a diagnósticos da doença em fases mais avançadas. A existência duma grande proporção de informação em falta no estadiamento da doença, não permitiu a validação dessa hipótese. Apesar da dimensão mediana das freguesias ser relativamente baixa, algumas freguesias urbanas apresentam um número de habitantes elevado (acima dos 40 mil), o que poderá ter levado a uma subestimação das desigualdades socioeconómicas na sobrevivência.

Referências

[1] Woods, L.M., Rachet, B., Coleman, M.P. (2006). Origins of socio-economic inequalities in cancer survival: a review. *Annals of Oncology* 17(1), 5–9.

[2] Nelson, C.P., Lambert, P.C., Squire, I.B., Jones, D.R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 26(30), 5486–5498.

Poster

Valor prognóstico de biomarcadores no carcinoma da mama masculina

Mariana Aparício
Instituto Superior Técnico, *marianamefaparicio@gmail.com*

Giovani L. Silva
Instituto Superior Técnico, *gsilva@math.ist.utl.pt*

António E. Pinto
Instituto Português de Oncologia de Lisboa, *aepinto@ipolisboa.min-saude.pt*

Palavras–chave: Cancro da mama masculina, Biomarcadores, Análise de sobrevivência.

Resumo: O cancro da mama masculina é uma doença rara correspondendo a cerca de 1% dos casos de cancro da mama e a menos de 1% das neoplasias no homem [1]. Embora partilhe algumas similaridades com o cancro da mama feminina, existem diferenças acentuadas na sua incidência, idade de aparecimento, prognóstico e sobrevivência. Contudo, uma vez que é uma doença rara, existe pouca informação acerca da sua etiologia. Para que se possa inferir sobre o tipo de terapia adequada, a resposta do tumor a um dado tratamento e a sobrevivência global ou de remissão da doença, é de extrema importância o aumento do conhecimento dos factores de prognóstico e preditivos do cancro da mama no homem [2, 3].
Assim, a partir dos dados de 166 doentes tratados no Instituto Português de Oncologia de Lisboa, durante o período de 1970 a 2013, pretende-se investigar o valor prognóstico de variáveis clínicas e patológicas (idade, localização do tumor, antecedentes familiares, tipo histológico, grau de diferenciação, tamanho do tumor, envolvimento ganglionar e estádio da doença), variáveis moleculares (ploidia do DNA, receptores hormonais de estrogénio e progesterona e proteína ErbB-2) e variáveis relacionadas com a terapêutica (tipo de tratamento) em relação ao intervalo livre de doença (recidivas) e sobrevivência global (morte por doença).
Para esse efeito, faz-se uma análise estatística que consiste, inicialmente, de um estudo da associação de cada par de covariáveis (testes de independência de Pearson e de Fisher). Posteriormente, estimam-se as curvas de sobrevivência para cada uma das variáveis categorizadas usando o estimador não-paramétrico de Kaplan-Meier e faz-se o respectivo teste Log-rank para comparação das curvas. Por fim, procede-se à análise de sobrevivência com base no modelo semi-paramétrico de Cox, visando determinar o valor prognóstico dos biomarcadores estatisticamente significativo nos tempos de sobrevivência e de remissão do carcinoma da mama masculina.

Agradecimentos

Giovani Silva foi parcialmente financiado pelo projeto Pest-OE/MAT/UI0006/2011.

Referências

[1] Miao, H., Verkooijen, H.M., Chia, K.S., Bouchardy, C., Pukkala, E., Laronningen, S., Mellemkjær, L., Czene, K., Hartman, M. (2011). Incidence and Outcome of Male Breast Cancer: An International Population-Based Study. *Journal of Clinical Oncology* 29, 4381-4386.

[2] Giordano, S.H., Cohen, D.S., Buzdar, A.U., Perkins, G., Hortobagyi, G.N. (2004). Breast carcinoma in men: A population-based study. *Cancer* 101, 51-57.

[3] Cutuli, B., Le-Nir, C.C., Serin, D., Kirova, Y., Gaci, Z., Lemanski, C., De Lafontan, B., Zoubir, M., Maingon, P., Mignotte, H., de Lara, C.T., Edeline, J., Penault-Llorca, F., Romestaing, P., Delva, C., Comet, B., Belkacemi, Y. (2010). Male breast cancer. Evolution of treatment and prognostic factors. Analysis of 489 cases. *Critical Reviews in Oncology/Hematology* 73, 246-254.

Poster

Mapa polínico atmosférico da região de Évora - 2001 a 2007

Ricardo Assunção
Faculdade de Medicina Veterinária da Universidade Lusófona de Humanidades e Tecnologias e Universidade de Évora, ric.assuncao@gmail.com

Manuela Oliveira
Centro de Investigação em Matemática Aplicada (CIMA) e Departamento de Matemática da Universidade de Évora, mmo@uevora.pt

Fernando Afonso
Centro de Investigação Interdisciplinar Egas Moniz (CIIEM), d8718@alunos.uevora.pt

Elsa Caeiro
Sociedade Portuguesa de Alergologia e Imunologia Clínica, elcaeiro@yahoo.com

Rui Brandão
Instituto Ciências Agrárias e Ambientais Mediterrânicas (ICAAM), Universidade de Évora, ruibrand@uevora.pt

Maria Luísa Lopes
Hospital de Santa Luzia, Elvas, luisalopes@sapo.pt

Palavras–chave: Aerobiologia, Polinose, Évora, Contagens Polínicas.

Resumo: Os sintomas de alergia aos pólens por parte de doentes com polinose são alvo do estudo apresentado. Os doentes são oriundos da região de Évora e os dados clínicos foram recolhidos no Hospital do Espírito Santo de Évora. Os dados foram recolhidos nos anos de 2001 a 2007 para os meses de Março a Junho. Os sintomas manifestados pelos doentes, cujo quadro clínico foi determinado por testes alergológicos, foram diariamente e ao longo do mesmo período registados. A série relativa aos dados polínicos foi construída com base em registos diários obtidos no mesmo período na estação de recolha de Évora integrada na Rede Portuguesa de Aerobiologia (RPA).

1 Introdução

A alergia a pólen é uma causa frequente de manifestações alérgicas. Atualmente mais de um terço da população portuguesa sofre de pelo menos uma doença alérgica, ou seja, estas doenças afectam cronicamente mais de 3 milhões de portugueses: cerca de 30% da população tem queixas actuais de rinite e cerca de 10% tem asma. Reconhece-se que para os pacientes alérgicos, o início dos sintomas está relacionado com a concentração de grãos de pólen na atmosfera. Assim, o conhecimento detalhado do mapa polínico de cada região é fundamental na abordagem, quer diagnóstica quer terapêutica, do doente com polinose. Nesse sentido, a Sociedade Portuguesa de Alergologia e Imunologia Clínica (SPAIC) criou e tem vindo a promover a Rede Portuguesa de Aerobiologia (RPA). Criada em 2002 a RPA estabelece a ponte entre biólogos e imunoalergologistas de vários pontos do país, e é actualmente constituída por 9 centros de monitorização e 8 hospitais ou centros de imunoalergologia. No presente estudo os dados polínicos são provenientes da estação de recolha de Évora e foram analisados pelo laboratório de Palinologia do Departamento de Biologia da Universidade de Évora e dizem respeito a concentrações de 17 tipos polínicos medidos em grãos de pólen/m3. As medições do teor de pólen atmosférico na região Évora foram realizadas com um colector de impacto volumétrico de tipo "Hirst"(Recording Burkard Seven Day Volumetric Spore Trap). Os dados clínicos correspondem ao grau de severidade de 9 sintomas, associados aos diagnósticos de

Rinite e/ou Asma. Estes dados são provenientes das consultas externas do Hospital do Espírito Santo de Évora e dizem respeito a 102 pacientes, (42 do sexo masculino e 60 do sexo feminino). O diagnóstico clínico de pacientes foi realizado por testes cutâneos em Prick-modificado, no mesmo Hospital. Os dados, polínicos e clínicos, dizem respeito aos anos de 2001 a 2007 para os meses de Março a Junho.

2 Métodos e resultados

No âmbito do presente trabalho pretende-se conhecer o conteúdo polínico atmosférico da região de Évora e a sua evolução quantitativa e qualitativa, determinando os principais tipos polínicos presentes, avaliando a variação anual e interanual e estabelecendo o respetivo calendário polínico. Pretende-se ainda estabelecer uma relação entre os tipos de pólen detetados na atmosfera e os dados clínicos obtidos. No período de estudo, as concentrações polínicas totais mais elevadas registaram-se em 2003 com 68793 grãos de pólen/m^3 de ar enquanto as mais baixas se verificaram em 2002, com 43974 grãos de pólen/m^3 de ar. Em 2004 registou-se 46226 grãos de pólen/m^3, sendo as concentrações em 2005, 2006 e 2007 de 49411, 63429 e de 59982 grãos de pólen/m^3, respetivamente. Quando analisados os valores percentuais por tipo de pólen, verificou-se que o tipo polínico mais frequente pertencia à família Poaceae (27,93%), seguido pelos géneros Quercus sp. (25,08%) e Olea (10,17%). O tipo de pólen menos representado foi o pertencente à família Myrtaceae (0,17%). Verificou-se a existência de diferenças estatisticamente significativas entre algumas concentrações polínicas e vários anos de estudo. As concentrações mais elevadas de pólenes verificaram-se no mês de Maio, com exceção do ano 2005, em que ocorreram em Abril. De forma a estabelecer e/ou testar as possíveis relações entre os dados recolhidos, foram realizados testes de qui-quadrado entre a gravidade dos sintomas e meses e sexo. Estimou-se os valores de razão das chances (*odds-ratio*) e de risco relativo entre sexo e gravidade dos sintomas.

3 Conclusão

No espectro polínico verificou-se a dominância de pólenes pertencentes a plantas da família Poaceae, também conhecidas como gramíneas. As gramíneas constituem um elemento habitual na paisagem urbana, bem como nas zonas limítrofes da mesma. Os géneros Quercus, Olea e Platanus também se destacaram pela sua ocorrência, bem como a família Urticaceae. Comparativamente a outras regiões do país, o espectro polínico é semelhante. Os testes de independência de Qui-quadrado para as variáveis sexo e meses versus a presença/ausência de sintomas permitem rejeitar a hipótese nula de independência com um nível de significância de 0,001. Pelos valores de risco relativo e razão das chances calculados, verifica-se que a probabilidade de um paciente do sexo feminino com o sintoma "rinorreia", quando comparado com um paciente do sexo masculino, aumenta gradualmente à medida que o grau de severidade dos sintomas aumenta de moderada a forte e a muito forte. Este comportamento é também verificado para os outros sintomas.

Referências

[1] Brito, F., Gimeno, P., Carnés, J., Fernández-Caldas, E., Lara, P., Alonso, A., García, R., Guerra, F. (2010). Grass Pollen, Aeroallergens, and Clinical Symptoms in Ciudad Real, Spain. *Journal of Investigational Allergoly Clinical Immunology* 20(4), 295–302.

[2] Dopazo, A., Aira, M., Armisén, M., Vidal, C. (2002). Relationship of clinical and aerobiological pollen data in the north-west of Spain. *Allergolia et Immunopatholia* 30(2), 74–78.

[3] Caeiro, E., Brandão, R., Carmo, S., Lopes, L., Almeida, M., Gaspar, A., Oliveira, J., Todo-Bom, A., Leitão, T., Nunes, C. (2007). Rede Portuguesa de Aerobiologia: Resultados da monitorização do pólen atmosférico (2002-2006). *Revista Portuguesa de Imunoalergologia* 15(3), 235–250.

Póster

Predicción del origen tuberculoso del derrame pleural mediante los modelos aditivos generalizados robustos

Laura Calaza Díaz
Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, *laura.calaza@usc.es*

Mónica López Ratón
Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, *monica.lopez.raton@usc.es*

Carmen Cadarso Suárez
Unidad de Bioestadística, Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, *carmen.cadarso@usc.es*

Francisco Gude Sampedro
Unidad de Epidemiología Clínica. Complejo Hospitalario Universitario de Santiago de Compostela, *francisco.gude.sampedro@sergas.es*

Luis Valdés Cuadrado
Servicio de Neumología. Complejo Hospitalario Universitario de Santiago de Compostela, *luis.valdes.cuadrado@sergas.es*

Palabras clave: Generalized additive models, Robustness, Smoothing, Pleural effusion, Tuberculosis.

Resumen: En numerosas ocasiones, en la práctica clínica, el interés radica en predecir la presencia de una determinada enfermedad o patología en función de una serie de covariables o predictores. El efecto de dichas covariables sobre el riesgo de enfermedad puede no ser lineal. Los modelos aditivos generalizados (GAMs, Hastie y Tibshirani, 1990) constituyen una herramienta de regressión flexible que permite expresar el efecto no lineal de una covariable contínua en la respuesta y pueden ser aplicados a una gran variedad de datos, como por ejemplo, datos binarios o de conteo (em general, datos pertenecientes a la familia exponencial). Dadas las covariables $x_1, \cdots, x_p$ se formula el modelo GAM como:

$$\eta = X^*\theta + \sum_{j=1}^{p} f_j(x_j)$$

donde $\eta = g^{-1}(\mu)$ es la respuesta transformada, com g la función *logit* y μ la media de la respuesta, $X^*\theta$ corresponde a la parte estrictamente paramétrica del modelo y $f_j(x_j)$ el efecto parcial suave (desconocido) de x_j.

Con frecuencia, en la práctica los datos reales pueden contener valores atípicos ("outliers") y, en estos casos, sería adecuada una estimación robusta de los modelos GAM. Recientemente se han propuesto en la literatura varias técnicas de robustez para estes modelos (Alimadad y Salibian-Barrera, 2011; Croux *et al.*, 2011; Wong, Yao y Lee, 2013). En este trabajo nos centraremos en la metodología propuesta por Wong, Yao y Lee (2013), donde se utilizan ecuaciones de estimación robustas mediante diferentes suavizadores no paramétricos de f_j (como por ejemplo los P-splines (Eilers y Mark, 1996) o los *thin plate regression splines* (Wood, 2003) entre otros) y métodos automáticos para la selección del grado de suavización. En dicho artículo se proponen procedimientos de selección del parámetro de suavización basados en la idea de Validación cruzada ("Cross-Validation", CV) y en al Criterio de información generalizado ("Generalized information criterion", GIC; Konishi y Kitagawa, 1996).

Las técnicas robustas descritas se han aplicado a datos reales en el área de Neumología. Con este motivo, se ha utilizado una base de datos correspondiente a 971 pacientes ingresados en el Servicio de Neumología del Complejo Hospitalario Universitario de Santiago (Galicia), en el período de tiempo comprendido entre Enero de 2004 y Diciembre de 2010. Todos ellos presentaban un derrame pleural cuya causa es frecuentemente, difícil de determinar. Se sabe que niveles elevados de adenosin-desaminasa (PFADA) y porcentajes altos de linfocitos (PFLYM) en el líquido pleural se encuentran asociados a la tuberculosis (Valdés *et al.*, 2010). El objetivo del estudio es el de evaluar si estos dos marcadores son útiles para predecir el origen tuberculoso del derrame pleural en estos pacientes. Debido a la existencia de outliers en los datos correspondientes a PFADA, se ha utilizado la metodología propuesta por Wong, Yao y Lee (2013). Se ha considerado como base de suavización los P-splines y la selección de los parámetros de suavización se ha realizado en base al GIC. La implementación del modelo se llevó a cabo mediante la utilización del paquete en R *robustgam* [7]. Los resultados obtenidos mostraron una elevada capacidad de predicción de los marcadores PFLYM y PFADA del derrame pleural por causa tuberculosa. Además, se puso de manifiesto la necesidad de la utilización de los modelos GAM robustos para no extraer conclusiones erróneas en la práctica clínica bajo la presencia de datos outliers.

Agradecimientos

Este trabajo fue parcialmente financiado por el proyecto MTM2011-28285-C02-00 del Ministerio español de Innovación y Ciencia.

Referencias

[1] Alimadad, A., Salibian Barrera, M. (2011). An outlier-robust fit for generalized additive models with applications to disease outbreak detection. *Journal of the American Statistical Association* 106, 719–731.

[2] Croux, C., Gijbels, I., Prosdocimi, I. (2011). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics* 68, 31–44.

[3] Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science* 11(2), 89–121.

[4] Hastie, T.J., Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

[5] Konishi, S., Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* 83, 875-890.

[6] Wong, R.K.W., Yao, F., Lee, T.C.M. (2013). Robust estimation for generalized additive models. *Journal of Graphical and Computational Statistics*, to appear.

[7] Wong, R.K.W. (2013). *robustgam: Robust Estimation for Generalized Additive Models*. R package version 0.1.6., http://CRAN.R-project.org/package=robustgam.

[8] Valdés, L., San José, M.E., Pose, A., Gude, F., González-Barcala, F.J., Álvarez-Dobaño, J.M., Sahn, S.A. (2010). Diagnosing tuberculous pleural effusion using clinical data and pleural fluid analysis: A study of patients less than 40 years-old in an area with a high incidence of tuberculosis. *Respiratory Medicine* 104(8), 1211–1217.

[9] Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society B* 65, 95–114.

Poster

Estudo do perfil sensorial de vinhos do Porto

Elisete Correia
CM - Universidade de Trás-os-Montes e Alto Douro, ecorreia@utad.pt
Bebiana Monteiro
IBB/CGB - Universidade de Trás-os-Montes e Alto Douro, bmonteiro@engenheiros.pt
Alice Vilela
IBB/CGB - Universidade de Trás-os-Montes e Alto Douro, avimoura@utad.pt

Palavras–chave: Vinho do Porto, Análise sensorial, Descritores de vinhos do Porto, Análise descritiva quantitativa (ADQ), CTPCA, MANOVA.

Resumo: O vinho do Porto é um vinho fortificado tipicamente português, produzido exclusivamente na região demarcada do douro. O vinho do porto requer especial atenção, quer pela sua história quer pela internacionalização do mercado, contribui consideravelmente para a economia local. A análise sensorial tem vindo a ser amplamente aplicada em várias indústrias, sendo considerada a "ferramenta analítica"mais importante na investigação e desenvolvimento de novos produtos e no controlo de qualidade. A aplicação desta técnica na indústria do vinho do Porto tem como função melhorar a caracterização e o conhecimento das preferências do consumidor (perspectiva comercial e marketing) e a consequente produção dos diferentes estilos e categorias especiais do vinho do Porto que satisfaça e exceda as expectativas do público-alvo. Assim, foi objetivo deste trabalho a identificação, seleção de descritores e o estudo do perfil sensorial de diferentes marcas e estilos de Vinhos do Porto. Foram usados 19 vinhos do Porto das categorias Ruby, Branco e Tawny, de diferentes estilos e de marcas comerciais disponíveis no mercado, para o levantamento de descritores de vinhos do Porto. A escolha dos descritores para cada estilo do vinho do Porto foi realizada utilizando dois métodos: lista pré-estabelecida desenvolvida pelo IVDP, IP para a categoria Branco; e escolha livre para as restantes categorias (Ruby e Tawny). Aos provadores foi pedido que, com base na listagem fornecida (categoria Branco), descrevessem as suas sensações de acordo com os parâmetros organoléticos: aparência (cor e limpidez), aroma, sabor, flavor, sensações e fim-de-boca.

Devido à natureza das variáveis (variáveis ordinais) envolvidas no estudo para a caracterização sensorial dos vinhos do porto recorreu-se a uma análise em componentes principais categórica (CATPCA), que permitiu a identificação de grupos de vinhos de uma forma intuitiva, bem como a identificação dos descritores que possibilitam a discriminação entre os grupos e cada uma das marcas comerciais de vinhos do Porto entre si. Para a implementação da CATPCA foram especificadas inicialmente 5 dimensões, e com base na percentagem de variância total optou-se por se selecionar duas componentes por serem capazes de explicar mais de 80% os dados originais. Das 19 marcas estudadas, conclui-se que dentro de cada categoria, várias marcas têm atributos comuns e outras são caracterizadas por descritores que as diferenciam significativamente das restantes.

Póster

Control de Calidade en Aerobioloxía: Comparación de diferentes métodos de lectura de láminas

Tomás R. Cotos-Yáñez
Dpto. de Estatística e I.O. da Universidade de Vigo, cotos@uvigo.es
Francisco J. Rodríguez-Rajo
Dpto. Bioloxía Vexetal e Ciencias do Solo da Universidade de Vigo, javirajo@uvigo.es
Ana Pérez-González
Dpto. de Estatística e I.O. da Universidade de Vigo, anapg@uvigo.es
Victoria Jato
Dpto. Bioloxía Vexetal e Ciencias do Solo da Universidade de Vigo, vjato@uvigo.es

Palavras clave: Aerobioloxía, Estimación nonparametrica da densidade, Error de estimación.

Resumo: En monitorización de polen aerobiolóxico os métodos mais frecuentes usados na lectura de láminas están baseados na selección dalgunhas filas ou columnas da lámina ou ben selección aleatoria de campos, as cales representan unha pequena proporción da lámina enteira. Neste estudio evaluamos os diferentes métodos de contar o número de graos de polen dunha lámina e cuantificamos o erro cometido por cada un deles con respecto ao total. Ademáis proponse unha selección, tanto para os barridos lonxitudinais como transversais, que minimizan o erro cuadrático medio ou o erro absoluto relativo medio. Para elo, usamos 113 láminas recollidas no ano 2008 en Ourense (NW España). Realizouse unha comparativa entre os tres métodos, 4 barridos lonxitudinais, 12 barridos traversais e as 493 celas aleatorias. Análise estatístico dos resultados obtidos reflicte diferencias significativas entre as 3 metodoloxías.

1 Introdución

En estudios Aerobiolóxicos o uso de metodoloxía estandar permite comparar os resultados obtidos en diferentes áreas xeográficas. Hirst spore traps (Hirst [2]) é o máis común dispositivo usado para determinar o contido do polen no aire. O aire é succionado nun fluxo de $10L$ *per min* e os graos de polen impactan contra unha lámina Melinex. Xeralmente a cantidad de concentración de polen dase pola lectura dunha porcentaxe da superficie total da cinta Melinex na cal os graos de polen están adheridos. O resultado exprésase como graos de polen por metro cúbico de aire aspirado. Diferentes métodos de selección da sub-mostra a ler foron propostos para determinar a concentración media diaria de polen, sendo a sub-mostra por barridos lonxitudinais e os barridos trasversais os mais frecuentes, con mais do 95 % das redes de monitorización europeas. Un terceiro método, sub-mostra por celas aleatorias, úsase en poucas estacións (Galán *et al.* [1]).
O obxetivo deste traballo é probar se existen diferencias na distribución dos barridos lonxitudinais e transversais, cales son os barridos óptimos en cada caso, e evaluar as posibles diferencias entre os resultados obtidos polos tres métodos de contaxe usados en estudios aerobiolóxicos.

2 Material e métodos

Foron analizadas no ano 2008 en Ourense (NW-España) lecturas de 113 mostras aerobiolóxicas. A metodoloxía usada foi proposta pola Rede Aerobiolóxica Española (Galán *et al.* [1]). Para cada mostra contáronse tanto o polen total como a cantidade por tipo de especie (*Alnus*, Cupressaceae, *Fraxinus*, *Betula*, *Pinus*, *Platanus*, *Populus*, *Quercus*, *Olea*, *Plantago*, Poaceaea, *Castanea*,

Urticaceae) no total da superficie exposta e seguindo as tres técnicas de contaxe: catro barridos lonxitudinais, doce barridos transversais and 493 celas aleatorias. A superficie total examinada foi $14 \times 44 mm^2$ ($616 mm^2$). A porcentaxe da área examinada varía levemente segundo o método usado, 12,86 % para os 4 barridos lonxitudinais, 12,27 % para os 12 barridos transversais e 12,73 % para os 493 celas aleatorias.
Para chequear diferencias significativas entre a distribución de graos de polen para cada barrido lonxitudinal e transversal, realizouse o seguinte test de hipóteses:

$$\begin{array}{lll} H_0: & f_1^l = f_2^l = ... = f_I^l & \qquad H_0: \quad f_1^t = f_2^t = ... = f_J^t \\ H_1: & f_i^l \neq f_m^l & \qquad H_1: \quad f_u^t \neq f_v^t \end{array} \tag{1}$$

con $f_i^l, i = 1, 2, ..., 31$ and $f_i^t, i = 1, 2, ..., 98$ funcións de densidade do total de graos de polen por barrido lonxitudinal e transversal, respectivamente. Para o contraste (1), consideramos o test suave de k-mostras proposto en Martínez-Camblor and Uña-Álvarez [3]. Obtivéronse uns p-valores bootstrap $< 0{,}001$, polo tanto a selección de barridos lonxitudinal ou transversal non debe facerse aleatoriamente. A continuación deseñamos un mecanismo de obtención dos 4 barridos lonxitudinais e os 12 transversais que minimizan o erro cadrático medio e o erro relativo medio:

$$\text{RE: } I_0^{re} = \arg\min_{\text{I}} \frac{1}{113} \sum_{h=1}^{113} \frac{|T(x_I) - T_h|}{T_h} \qquad \text{SE: } I_0^{se} = \arg\min_{\text{I}} \frac{1}{113} \sum_{h=1}^{113} (T(x_I) - T_h)^2$$

con $I = (i, j, k, l) \in 1, 2, ..., 31, i \neq j \neq k \neq l$ ou $I = (i_1, i_2, ..., i_{12}) \in 1, 2, ..., 98$ no caso lonxitudinal ou transversal respectivamente, $T_I(\bullet)$ a correspondente estimación do número de graos de polen e T_h o número real diario de graos de polen.
Por último comparamos estas técnicas de contaxe coa selección aleatoria de celas a través dun modelo de análise da varianza con medidas repetidas, cun p-valor $< 0{,}001$.

3 Principais conclusións

- A distribución da variable número de graos de polen sobre barridos lonxitudinais e transversias non é uniforme, e polo tanto precísase unha adecuada selección dos mesmos para disminuir o erro cometido.
- O subconxunto óptimo acadado para os barridos lonxitudinais reducese en erro relativo promedio en case 45 %, e en erro cuadrático promedio en mais do 73 %. O subconxunto óptimo para os barridos transversais reducese en case o 23 % para o erro relativo promedio, e por mais do 50 % no erro cuadrátivo promedio.
- O estimador diario de graos de polen obtido pola lectura de barridos transversais difire significativamente do estimador pola lectura de barridos lonxitudinais e de celas aleatorias.

Agradecementos

Este traballo foi parcialmente financiado polo proxecto MTM2011-23204 do Ministerio Español de Ciencia e Innovación (fondos FEDER incluido).

Referencias

[1] Galán, C., Cariñanos, P., Alcázar, P., Domínguez-Vilches, E. (2007). *Spanish Aerobilogy Network: Management and Quality Manual*. Servicios de Publicaciones. Universidad de Condoba.

[2] Hirst, J.M. (1952). Changes in atmospheric spore content: diurnal periodicity and the effects of weather. *Transactions of the British Mycological Society* 36(4), 375–393.

[3] Martínez-Camblor, P., de Uña-Álvarez, J. (2013). Studying the bandwidth in k-sample smooth tests. *Computational Statistics* 28(2), 875–892.

Poster

História familiar de eventos cardiovasculares, rigidez arterial e pressão arterial central: Estudo de determinação do risco cardiovascular da população de Guimarães e Vizela, incluindo a prevalência de rigidez arterial e envelhecimento arterial precoce

Pedro G. Cunha
Centro Hospitalar do Alto Ave, Guimarães, Instituto de Ciências da Vida e da Saúde (ICVS), Escola de Ciências da Saúde, Universidade do Minho, *pedrogcunha@netcabo.pt*

Pedro Oliveira
Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, *pnoliveira@icbas.up.pt*

Jorge Cotter
Centro Hospitalar do Alto Ave, Guimarães, Instituto de Ciências da Vida e da Saúde (ICVS), Escola de Ciências da Saúde, Universidade do Minho, *jorgecotter@gmail.com*

Isabel Vila
Centro Hospitalar do Alto Ave, Guimarães, *isabelgvila@gmail.com*

Nuno Sousa
Instituto de Ciências da Vida e da Saúde (ICVS), Escola de Ciências da Saúde, Universidade do Minho, *njcsousa@ecsaude.uminho.pt*

Palavras–chave: Biometria, Epidemiologia.

Resumo: Observamos 2123 indivíduos aleatoriamente selecionados da população de duas cidades adjacentes do Norte de Portugal (Guimarães/Vizela) para incluir uma coorte representativa da distribuição por faixa etária e por sexo. Avaliamos as suas características clínicas e metabólicas. Recolhemos a História Familiar (HF) relevante de Eventos Cardiovasculares (ECV) e medimos Velocidade de Onda de Pulso (VOP) bem como a Pressão Arterial Central (PAC). Consideramos existir uma HF Positiva (HFP) para ECV, quando um sujeito tinha: a) dois familiares de primeiro grau com história de ECV prévio; b) 1 familiar de primeiro grau com história de ECV prematuro. O nosso objectivo foi o de compreender se a existência de HFP para ECV influencia manifestações de rigidez arterial (RA) ou parâmetros como a PAC, aumentando o risco cardiovascular.
Documentamos a existência de forte HFP para ECV em 227 sujeitos (61.2% mulheres; idade média global de 65.5 anos); Estes sujeitos apresentavam os seguintes valores médios de variáveis hemodinâmicas centrais: VOP – 9.0 m/seg; PAC Sistólica (PACs) – 134.0 mmHg; PAC Diastólica (PACd) – 79.6 mmHg; Pressão de Pulso Central (cPP) – 54.5 mmHg; Índice de Aumentação (IA) – 34.1. Ao compararmos estes valores médios com os da restante população em estudo, verificamos a existência de diferenças significativas no que dizia respeito aos valores de VOP/PACs/PACd/cPP/IA. Quando analisamos as diferenças (respeitantes a estas mesmas variáveis) entre a população geral e a população com HFP, após estratificar ambas por faixas etárias de 10 anos, verificamos a presença consistente de diferenças apenas na faixa etária entre os 40 – 49 anos, para os parâmetros VOP/PACs/PACd/cPP/IA. A relevância destes achados é discutida no que à estratificação de risco cardiovascular diz respeito, em particular no seio da população portuguesa com elevada prevalência de acidentes vasculares cerebrais.

Agradecimentos

Este trabalho é apresentado em nome do Grupo de Estudos de Guimarães.

Póster

Estudo da mortalidade por suicidio en Galicia

María José Ginzo Villamayor
Universidade de Santiago de Compostela, mariajose.ginzo@usc.es

Rosa María Crujeiras Casais
Universidade de Santiago de Compostela, rosa.crujeiras@usc.es

María Esther López Vizcaíno
Instituto Galego de Estatística, esther.lopez@ige.eu

María Isolina Santiago Pérez
Dirección Xeral de Innovación e Xestión da Saúde Pública, MariaIsolina.Santiago.Perez@sergas.es

Palavras clave: Besag-York-Mollié, INLA, Suicidio.

1 Introdución

A Organización Mundial da Saúde (OMS) considera o suicidio como un problema de saúde pública (véxase WHO, 2012). Cada ano, falecen por suicido case un millón de persoas no mundo (arredor do 1.8 % das mortes), estimándose que a porcentaxe de suicidios acade o 2.4 % en 2020. En Galicia, entre 2000 e 2011, producíronse entre 200 e 300 suicidios anuais.
Entre os factores de risco que inflúen no suicidio, inclúense as enfermidades mentais, o consumo de drogas, enfermidades crónicas, os cambios bruscos no estilo de vida (como por exemplo, a perda de emprego, a separación da parella,...) ou a combinación de varios destes factores (WHO, 2012). Tradicionalmente, as taxas de suicidios son máis altas na xente maior, pero obsérvase un incremento nas taxas de suicidios en xente nova, debido a algúns dos factores de risco antes comentados. Con todo, o suicidio é un proceso complexo, no que interveñen non só factores psicolóxicos e biolóxicos, senón tamén culturais e sociais.
Neste traballo analizarase o patrón espacial, por concellos (véxase Figura 1), dos suicidios en Galicia para os períodos 2000-2003, 2004-2007 e 2008-2011, e a posible influencia de variables socioeconómicas. Na seguinte sección descríbense brevemente os datos dos que se dispón e a metodoloxía empregada.

2 Material e métodos

As defuncións por suicidio en Galicia no período 2000-2011, que corresponden aos códigos X60-X84 da 10ł Clasificación Internacional de Enfermidades (CIE-10), obtivéronse do Rexistro de Mortalidade de Galicia. Como poboación utilizouse o Padrón Municipal de Habitantes.
Como covariables analizáronse o índice de envellecemento da poboación, o paro rexistrado nas oficinas de emprego público, o grao de urbanización (porcentaxe de poboación que vive en zonas pouco poboadas), a remuneración de asalariados sobre a renda dispoñible bruta e unha variable climática: número medio de horas de luz. A fonte de datos para as variables socioeconómicas foi o Instituto Galego de Estatística, e os datos climáticos proceden de Meteogalicia.
Para analizar o patrón espacial dos suicidios en Galicia, considerando os efectos de distintas covariables, axustouse o modelo proposto por Besag et al. (1991), adaptado a este contexto. Un dos supostos deste modelo é que o log-risco pódese descompoñer como suma dunha compoñente espacial estruturada e un erro aleatorio, pero tamén se pode incluír o efecto suave dalgunha covariable. O axuste do modelo realizouse empregando o método baseado na aproximación por integradores de

Laplace anidados (integrated nested Laplace approximation, INLA), proposto por Rue et al. (2009). Como resultados, preséntanse os mapas das compoñentes estruturadas e realízase unha análise do efecto das covariables.

3 Resultados

No período 2000-2011 producíronse en Galicia 3.644 defuncións por suicidio, das que o 74 % ocorreron en homes. A distribución dos suicidios, por período e grupo de idade preséntase na táboa 1. En tódolos períodos, a maior porcentaxe de suicidios obsérvase no grupo de 65 anos en diante (en torno ao 40 %), seguido polo de 16 a 49.

	Grupos de idade				
Período	Menos de 16 anos	16-49	50-64	Máis de 64 anos	Total
2000-03	8	426	242	494	1170
2004-07	8	466	290	468	1232
2008-11	0	420	319	503	1242
Total	16	1312	851	1465	3644

Cadro 1: Suicidios por grupos de idade, para os períodos considerados.

O patrón espacial de mortalidade por suicidio, sen ter en conta o efecto de ningunha covariable, preséntase na Figura 1. Nos tres períodos analizados as provincias de Lugo e Coruña son as que teñen un risco maior de mortalidade por suicidio, aínda que se observan diferenzas entre os períodos.

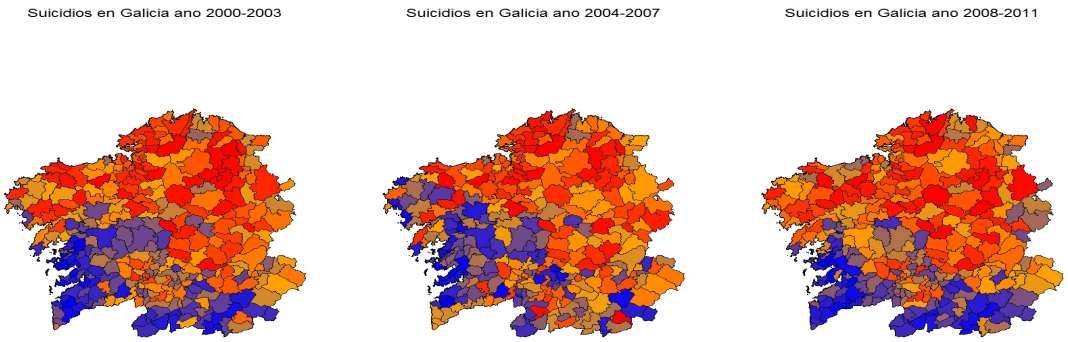

Figura 1: Distribución xeográfica das taxas de mortalidade por suicidio). Esquerda: 2000-2003. Centro: 2004-2007. Dereita: 2008-2011.

Agradecementos

María J. Ginzo e Rosa M. Crujeiras agradecen o apoio dos proxectos MTM2008-3010 do Ministerio de Ciencia e Innovación e a Rede IAP P7/06 de Belgian Science Policy.

Referencias

[1] Besag, J., York, J., Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1–59.

[2] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B* 71, 319–392.

[3] World Health Organization (WHO) (2012). Public Health Action for the Prevention of Suicide.

Poster

Definição operacional de doença periodontal na prática clínica e em epidemiologia

J.A. Lobo Pereira
Faculdade de Medicina Dentária da Universidade do Porto, *lobopereiramail@gmail.com*
Teresa Oliveira
Universidade Aberta e Centro de Estatística e Aplicações da Universidade de Lisboa, *toliveir@uab.pt*
Antonio Costa
Universidad Nacional de Educación a Distancia, *acosta@mat.uned.es*
Luzia Mendes
Faculdade de Medicina Dentária da Universidade do Porto, *luziacmmendes@hotmail.com*

Palavras–chave: Doença periodontal, Classificação, Epidemiologia, Clínica, Risco.

Resumo: As doenças periodontais constituem um grupo de patologias inflamatórias, de etiologia predominantemente bacteriana, que afetam os tecidos de suporte dentário e cuja expressão clínica varia desde a inflamação reversível dos tecidos gengivais até à sua destruição irreversível, que em última instância conduz à perda dentária [1]. Este conceito é suficientemente lato para permitir a compreensão do fenómeno patológico e abarcar as suas múltiplas formas de apresentação clínica porém não é operacional quando se pretende estudar a sua epidemiologia nem permite descriminar clinicamente os diferentes casos.

Nos estudos publicados ao longo dos anos as definições operacionais de "doença periodontal" têm variado, não existindo uma definição consensual [2] e polivalente apesar de terem sido criados vários sistemas de classificação sucessivamente modificados [3].

Com o objectivo de mostrar a importância da definição de "caso" em epidemiologia/saúde pública, apresentamos os resultados de um estudo realizado por nós, no qual foram obtidas diferentes prevalências resultantes da aplicação de distintas definições de caso, assim como as respetivas curvas ROC (receiver operating characteristic), também foram avaliadas as alterações da força da associação (*odds ratio*, *i.e.*, razão das chances) entre as variáveis independentes e a variável dependente (caso) num modelo de regressão logística. No contexto da prática clínica a classificação da condição periodontal deve adequar-se ao risco de progressão da doença e perda dentária, assim como ao plano de tratamento a elaborar. As evidências acumuladas sugerem-nos que a definição operacional de caso deve ser adequada ao contexto do estudo (epidemiológico ou clínico).

Referências

[1] Pihlstrom, B.L., Michalowicz, B.S., Johnson, N.W. (2005). Periodontal diseases. *Lancet* 366, 1809–20.

[2] Page, R.C., Eke, P.I. (2007). Case definitions for use in population-based surveillance of periodontitis. *Journal of Periodontology* 78(7 Suppl), 1387–99.

[3] Armitage, G.C. (1999). Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology* 4, 1–6.

Poster

Income from influenza in the pandemic period in Hospital Universitario A Coruña

Beatriz López-Calviño
Unidad de Epidemiología Clínica y Bioestadística, CHUAC A Coruña, beatriz.lopez.calvino@sergas.es

María José Pardo Landrove
Servicio de Medicina Preventiva, CHUAC A Coruña., maria.jose.pardo.landrove@sergas.es

Sonia Pértega-Díaz
Unidad de Epidemiología Clínica y Bioestadística, CHUAC A Coruña, sonia.pertega.diaz@sergas.es

Teresa Seoane-Pillado
Unidad de Epidemiología Clínica y Bioestadística, CHUAC A Coruña, maria.teresa.seoane.pillado@sergas.es

Salvador Pita-Fernández
Unidad de Epidemiología Clínica y Bioestadística, CHUAC A Coruña, salvador.pita.fernandez@sergas.es

José Manuel Suárez Lorenzo
Servicio de Medicina Preventiva, CHUAC A Coruña, jose.manuel.suarez.lorenzo@sergas.es

María Fernández Albalat
Servicio de Medicina Preventiva - CHUAC A Coruña, maria.fernandez.albalat@sergas.es

Keywords: Influenza A (H1N1), Time-series, Joinpoint regression.

1 Introduction

In June 2009, the World Health Organization (WHO) raised its alert to the highest level, phase 6. The end of the pandemic was declared in August 2010.
The pandemic that began in March 2009 was caused by a virus of influenza A H1N1 that represents a quadruple reassortment of two swine strains and a human strain, and a strain of bird flu, the largest proportion of the virus genes were from swine flu.

Objectives

- Determine the clinical-therapeutic patients diagnosed with influenza A (H1N1) and associated variables on admission to intensive care unit.
- Study time trend in hospitalization rates by epidemiological week.

2 Methods

- **Scope:** A Coruña University Hospital (Spain)
- **Period:** July 2009 to January 2010
- **Design:** Retrospective cohort observational study.

- **Inclusion criteria:** Patients of any age diagnosed with pandemic influenza A (H1N1) virus that meet hospitalization criteria: being an infuenza clinical case, having risk factors for complicated influenza, and laboratory diagnostic confirmation (real-time PCR).
- **Sample:** n=169 (precision=±7.6%, alpha-level=95%).
- **Measurements:** Demographic, clinical and treatment characteristics.
- **Statistical analysis:** mutiple logistic regression and joinpoint regression
- **Ethical and legal aspects:** CEIC Galicia (2012/260)

3 Results

Mean age was 35.0 ± 20.1 years, 52.1% were women. 82.8% presented some risk factor for complicated influenza, the most common were: asthma (24.9%), hemoglobinopathy and/or anemia (24.9%), pregnancy (13%), smoking (13%), obesity (12.4%), active immuno deficiency (12.4%).
Most cases were community-acquired (98.2%), with an average hospital stay of 8.3 ± 14.5 days (median: 5 days).
The most common symptoms were: fever (96.4%), cough (72.8%), myalgia (43.2%), dyspnea/difficulty breathing (36.1%).
Of the patients were included, 89.9% underwent chest radiography: 33.7% showed signs of pneumonia, with bilateral infiltrates in 33.3% of cases and unilateral infiltrates in 66.7%.
87.67% received antiviral treatment and 72.8% antibiotic treatment. The average time between symptom onset and treatment was 4.0 ± 3.4 days (median: 3 days).
11.8% of patients required admission to the ICU, with an average stay of 12.5 ± 12.0 days (median: 10 days).
3.6% of patients died: 66.7% female, mean age 51.0 ± 16.1 years (median: 50.5 years).
Variables independently associated with ICU admission were: smoking ($p = 0.021$, $OR = 6.912$, 95% $CI = 1.346$ to 35.497), active immuno deficiency ($p = 0.036$, $OR = 15.712$, 95% $CI = 1.190$ to 207.550) and bilateral pneumonic infiltrates ($p = 0.011$, $OR = 7.508$, 95% $CI = 1.596$ to 35.317).
By studying the rate of income epidemiological week (EW) are objectified two changes in the trend (EW: 39 and 44). There is a gradual increase in the number of admissions to week 39 (average percent change, $APC = 11.10$), increasing significantly between weeks 39 and 44 ($APC = 58.57$). From week 44, the number of admissions fell significantly ($APC = -43.14$).

4 Conclusions

The number of admissions increased until week 44, decreasing significantly from it.The variables associated with entering ICU were: smoking, having active immunodeficiency bilateral pneumonic infiltrates.

References

[1] Dongmei, Q., Kate, K., Tomomi, M., Tomokaka, S. (2009). A Joinpoint regression analysis of long-term trends in cancer mortality in Japan (1958-2004). *International Journal of Cancer* 124(2), 443–448.

[2] Kim, H., Fay, M., Feuer, E., Midthune, D. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* 19(3), 335–351.

Poster

Avaliação dos cuidados de saúde: implementação de valores *target*

Bruno Marques
DEIO/CEAUL/FCUL - Universidade de Lisboa & Siemens S.A., brunosm87@gmail.com

Palavras–chave: Valores *target*, Opinião de especialistas, Beta-pert, Decisão.

Resumo: As metodologias mais utilizadas para a avaliação dos cuidados de saúde baseiam-se na comparação dos desempenhos individuais dos hospitais com o desempenho médio global. O presente trabalho tem como objectivo encontrar uma metodologia que permita corrigir as limitações inerentes a este tipo de comparação, através da introdução de valores de referência fixos e independentes dos dados, recorrendo à opinião de especialistas das diferentes áreas clínicas e à teoria de decisão.

1 Introdução

A avaliação dos cuidados de saúde nas diferentes áreas clínicas é efectuada com o recurso a indicadores de processo, consensuais entre os hospitais, sociedades científicas e alinhados com as melhores práticas internacionais, que avaliam a concordância de determinadas práticas com as respectivas *guidelines* e que são utilizados para comparar a proporção de intervenções adequadas realizadas por cada hospital com a proporção de intervenções adequadas de todos os hospitais. Por outras palavras, é realizada uma comparação entre o desempenho de cada hospital e o desempenho médio global, dando origem à classificação dos hospitais, usualmente feita em 3 níveis [1] - abaixo da média, dentro da média e acima da média.
O objectivo deste trabalho é encontrar uma metodologia alternativa que permita comparar os desempenhos hospitalares com valores de referência (*target*). Pretende-se ultrapassar as limitações inerentes à comparação de desempenhos individuais com o desempenho médio global, como a incapacidade para distinguir os hospitais com desempenhos diferentes quando o desempenho médio global é muito bom ou muito mau, utilizando a opinião de especialistas para modelar a proporção de referência e determinar os valores *target*.

2 Opinião de especialistas

O primeiro passo é determinar a distribuição para a proporção de referência (p) de intervenções adequadas num hospital, através da recolha da opinião de especialistas. Compreendendo que não é fácil ao especialista determinar uma distribuição para a proporção em estudo, isto pode ser atingido questionando o especialista sobre o valor máximo, mínimo e mais provável que essa proporção pode tomar. Utilizando os modelos Beta-Pert[2] que são definidos por estes 3 valores, obtemos a distribuição de probabilidade da proporção de referência correspondente à opinião de cada especialista. Pode-se obter a opinião de um ou mais especialistas, sendo que no último caso existem diferentes abordagens matemáticas [3, 4] e comportamentais [4] para combinar as diferentes opiniões. Este trabalho não tem como objectivo comparar as diferentes abordagens, sendo utilizado o método cumulativo [2, 3] que é um método prático de aplicar e que utiliza toda a informação disponível das várias opiniões apresentando resultados robustos.

Exemplo 2.1 *Um dos indicadores utilizados na área de Enfarte Agudo do Miocárdio verifica se os utentes recebem aspirina à chegada ao hospital quando têm diagnóstico principal de enfarte. Cada especialista seria questionado sobre 3 valores: Qual o valor mais optimista, pessimista e mais provável da proporção de utentes que recebem aspirina? (máximo, mínimo e moda)*

O passo seguinte é gerar amostras aleatórias com distribuição Beta-Pert (máximo, mínimo, moda), uma amostra por cada especialista, sendo utilizado o método cumulativo para combinar as diferentes opiniões. Neste método são calculados os percentis das várias amostras, sendo depois calculada a média (que pode ser ou não ponderada) de cada percentil, sendo que estas médias formarão a amostra da opinião combinada. Os valores *target* - inferior TI e superior TU - são determinados através de dois quantis da amostra combinada, consoante o grau de exigência que se pretenda.

3 Classificação

A classificação dos hospitais nos 3 níveis é determinada pela posição do intervalo de confiança para p, a proporção de intervenções adequadas de cada hospital, face aos valores $target$. Tendo em conta que X representa o número de intervenções adequadas em n oportunidades, $\hat{p} = \frac{X}{n}$ é um estimador da proporção de intervenções adequadas num hospital que, para n suficientemente grande, segue assimptóticamente uma distribuição $Normal\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.
Os limites do intervalo de $100(1-\alpha)\%$ confiança para p são definidos por:
$LI_H = \hat{p} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad LU_H = \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Formulando o problema de classificação como um problema de decisão estatística, a regra de decisão [5] será: Se $LI_H > TU \Rightarrow$ nível 3; se $LU_H < TI \Rightarrow$ nível 1; caso contrário $\Rightarrow$ nível 2.
Definindo a distribuição condicional aos diferentes estados do hospital com base na informação dos especialistas, pode calcular-se a probabilidade dos hospitais serem classificados nos 3 níveis e ainda o risco associado a cada decisão tomada.

4 Discussão final

Embora a nova metodologia minimize as limitações da metodologia actual, estas não devem ser directamente comparadas uma vez que assentam sobre conceitos diferentes: a actual compara o desempenho individual com o desempenho médio global, enquanto que a nova metodologia compara o desempenho individual com valores *target*, que dependem da opinião de especialistas, da forma como estas opiniões são combinadas e do grau de exigência pretendido.

Agradecimentos

Este trabalho foi parcialmente financiado pela FCT projecto PEst-OE/MAT/UI0006/2011, e corresponde a uma parte do trabalho de estágio realizado na Siemens S.A., sob a orientação interna da Engenheira Inês Sá Dantas e orientação externa da Professora Doutora Antónia Amaral Turkman.
À Doutora Filipa Matos Baptista e à Engenheira Filipa Costa pelo tempo e apoio disponibilizado.

Referências

[1] The Joint Commission International (2010). Healthcare Professional Quality Report User Guide.

[2] Vose, D. (1996) *Risk analysis: A quantitative guide*. John Wiley & Sons, Chichester.

[3] Stark, K.D.C., Horst, H.S., Kelly, L. (2000). Combining expert opinions: a comparison of different approaches. *9th International Symposium on Veterinary Epidemiology and Economics*.

[4] Clemen, R., Winkler, R. (1999). Combining Probability Distributions From Experts in Risk Analysis. *Risk analysis* 19(2), 187–203.

[5] Paulino, C.D., Amaral Turkman, A., Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

Poster

Distribuição espacial de casos de dengue nos estados brasileiros

Natália da Silva Martins
ESALQ - Universidade de São Paulo, nsmbarreto@gmail.com
Paulo Justiniano Ribeiro Junior
Universidade Federal do Paraná, paulojus@ufpr.br

Palavras–chave: Correlação espacial, Índice de Moran, Dengue.

Resumo: O presente estudo objetivou avaliar a distribuição espacial da taxa de dengue nos estados brasileiros, utilizando o Índice de Moran. Com os resultados obtidos constatou-se que não há uma correlação da distribuição espacial dos casos de dengue. Verificou-se que os estados mais afetados são: Acre, Alagoas, Roraima, Goiás, Mato Grosso do Sul, Mato Grosso e Rio Grande do Sul, observando-se a necessidade da implementação de políticas públicas voltadas ao controle e prevenção da doença.

1 Introdução

A dengue é uma enfermidade causada por um vírus, sendo esta a arbovirose (doença provocada por um vírus que é essencialmente transmitido por artrópodes) mais comum que atinge o homem, [1].
De acordo com [1], o aumento de ocorrência da dengue constitui um crescente objeto de preocupação para a sociedade e para as autoridades de saúde, em decorrência das dificuldades enfrentadas para controlar as epidemias produzidas por esse vírus e ampliação da capacidade de atendimento dos serviços de saúde.
Diversos fatores vêm contribuindo com esses aumentos nos registros de casos de dengue. Segundo [2], o rápido crescimento urbano propicia uma fonte de indivíduos suscetíveis e infectados concentrados em áreas restritas, e este fato, associado às condições precárias de saneamento básico, moradia inadequada e fatores educacionais proporcionam condições ecológicas favoráveis à transmissão do vírus da dengue pelo mosquito *Aedes aegypti*.
Embora estes fatores facilitantes estejam presentes nos mais distintos locais, pesquisadores vêm observando que a distribuição geográfica da dengue tem sido considerada desigual entre os países, e dentro dos próprios países, [3]. Perante tal observação, a avaliação da distribuição espacial da dengue pode proporcionar a geração de hipóteses explicativas sobre os crescentes números de casos.
Deste modo, este estudo visa avaliar a distribuição espacial das taxas de incidência de dengue nos estados brasileiros, verificando como essas se encontram correlacionadas no espaço georreferenciado.

2 Materiais e métodos

Os dados utilizados neste estudo são referentes às taxas de incidências da dengue em casos por 100.000 habitantes, reminiscentes ao ano de 2010, disponibilizados pelo DATASUS no site http://www.datasus.gov.br.
Para avaliar como as taxas estão correlacionados no espaço georeferenciado, foram utilizadas técnicas de mensuração de autocorrelação espacial determinada por meio do índice de Moran global e local, a matriz de vizinhança adotada foi a *queen*, a qual define como conjunto de vizinhos o caminho que a rainha faz no tabuleiro de xadrez, [4].
As analises foram realizadas no software Geoda, que é um software livre, disponibilizado no site http://geodacenter.asu.edu.

3 Resultados e conclusão

O Índice de Moran é um coeficiente muito útil para medir a correlação espacial, sendo que este mede a relação do desvio padronizado da variável taxa de dengue em uma área i com o desvio padronizado das áreas vizinhas para a mesma variável. Este índice resultou em um valor de -0,08. Como sua distribuição é desconhecida, o mais comum é a utilização de uma pseudo-significância para testar sua significância, em que são geradas diferentes permutações dos valores de atributos associados às regiões. O valor-p produzido por meio da utilização dessas permutações mostrou-se não significativo a um nível de 5% de significância, revelando que a taxa de dengue em um estado não está correlacionado espacialmente com os estados vizinhos.
Observando-se a Figura 1, da distribuição espacial da taxa de dengue, é possível verificar que os estados brasileiros com as maiores incidências da doenças são: Acre, Alagoas, Roraima, Goiás, Mato Grosso do Sul, Mato Grosso e Rio Grande do Sul.

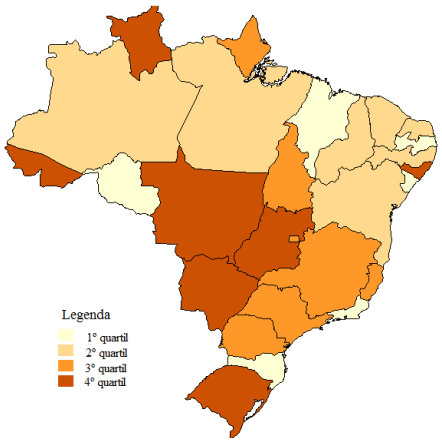

Figura 1: Mapa da distribuição espacial da taxa de dengue nos estados brasileiros

Os resultados obtidos por este estudo levantam a necessidade de estudos específicos que esclareçam questões relativas aos fatores envolvidos na transmissão da dengue e a implementação de políticas públicas voltadas ao controle e prevenção da mesma nas regiões mais afetadas.

Referências

[1] Holmes, E.C, Bartley, L.M, Garnet, G.P. (1998). The emergence of dengue past, present and future. In: Krause, R.M. (eds.): *Emerging Infectors*, 301–325. London: *Academic Press*.

[2] Costa, A.I.P., Natal, D. (1998). Distribuição espacial da dengue e determinantes socioeconômicos em localidade urbana no sudeste do Brasil. *Revista de Saúde Pública* 32, 232–237.

[3] Teixeira, M.G., Costa, M.C.N., Barreto, M.L., Mota, E. (2005). Dengue e febre hemorrágica do dengue no Brasil: que tipo de pesquisas a sua tendência, vigilância e experiências de controle indicam ser necessárias? *Caderno de Saúde Pública* 21, 1307–1315.

[4] Câmara, G., Carvalho, M.S., Cruz, O.G., Correa, V. (2003). *Análise espacial de dados geográficos*. INPE, São José dos Campos.

Poster

Inferência no modelo doença-morte

Luís Meira-Machado
Universidade do Minho, lmachado@math.uminho.pt
Artur Araújo
Universidade do Minho, b5498@math.uminho.pt
Jacobo de Uña-Álvarez
Universidade de Vigo, jacobo@uvigo.es

Palavras–chave: Função de incidência cumulativa, Modelos multiestado, Probabilidades de transição.

Resumo: Os modelos multiestado podem ser utilizados com sucesso na análise de dados de sobrevivência onde se identificam vários eventos, por exemplo, para descrever as diferentes fases na progressão da doença de um doente. O modelo doença-morte desempenha um papel central na teoria e prática dos modelos multiestado. Muitas das bases de dados de sobrevivência podem ser reduzidas a essa estrutura. Nestes modelos um objetivo importante é a modelação das intensidades de transição, mas os investigadores também têm interesse em apresentar os resultados de uma forma simples e resumida. Para este efeito, podem ser apresentadas as estimativas de probabilidades de transição entre estados, as probabilidades de ocupação, as funções de incidência cumulativa e a distribuição do tempo de permanência em cada estado. Neste trabalho apresentamos métodos de estimação para todas estas quantidades no contexto do modelo de doença-morte. A inclusão de covariáveis nestes estimadores também é considerada. Os métodos propostos são ilustrados com dados reais.

1 Introdução

Em estudos longitudinais médicos, os doentes podem observar vários eventos num determinado período de follow-up. A análise destes estudos pode ser realizada com sucesso pelos modelos de multiestado (Meira-Machado et al. 2009). Estes modelos podem ser considerados uma generalização da análise de sobrevivência onde a sobrevivência é o resultado final de interesse, mas onde estados intermediários são identificados. Um desses modelos é o modelo de doença-morte que é totalmente caracterizado por três estados e três transições entre eles (Figura 1). Por exemplo, em estudos de cancro mais do que um evento final pode ser definido como 'recidiva local', 'metástases à distância' e 'morte'. Na versão irreversível deste modelo, os indivíduos começam no estado 'saudável' e, posteriormente transitam para o estado intermédio 'doente' ou para o estado absorvente 'morte'. Muitas bases de dados de sobrevivência podem ser reduzidas a esta estrutura genérica.

Um dos objetivos principais em aplicações clínicas de modelos de multiestado é a estimação de probabilidades de transição. Estas quantidades têm proporcionado um crescente interesse pois elas permitem efetuar previsões a longo prazo do processo. Nos últimos anos várias contribuições foram feitas para a estimação destas quantitades. Meira-Machado et al. (2006) introduziram um substituto para o estimador de Aalen-Johansen (Aalen and Johansen, 1978) no contexto do modelo doença-morte não Markoviano. Meira-Machado et al. (2006) demonstram que o novo estimador pode ser mais eficiente quando o pressuposto de Markov não se verifica. Contudo, quando a percentagem de censura é elevada o estimador proposto apresenta uma grande variabilidade. Neste trabalho introduzimos novos estimadores para estas quantidades que têm por base os estimadores

propostos em Meira-Machado et al. (2006). Resultados obtidos em vários estudos de simulação sugerem que os novos estimadores têm um melhor comportamento.

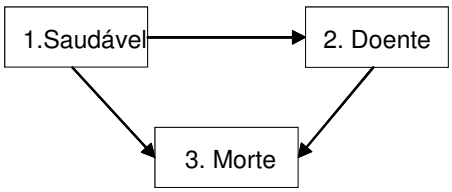

Figura 1: Modelo doença-morte

Também introduzimos estimadores para outras quantidades (probabilidades de ocupação, função de incidência cumulativa e distribuição do tempo de permanência em cada estado) no contexto de um modelo de doença-morte. Propomos também metodos para incorporar covariáveis nestes estimadores (Meira-Machado, de Uña-Álvarez and Datta (2012)).

Agradecimentos

Este trabalho foi financiado por fundos FEDER pelo "Programa Operacional Factores de Competitividade - COMPETE"e por fundos Portugueses pela FCT - "Fundação para a Ciência e a Tecnologia", projeto PTDC/MAT/104879/2008 e Est-C/MAT/UI0013/2011. Também agradecemos financiamento dos projetos MTM2008-03129 e MTM2011-23204 (FEDER incluido) do Ministério da Ciência e Tecnologia e Inovación de Espanha e 10PXIB300068PR da Xunta de Galicia.

Referências

[1] Aalen, O., Johansen, S. (1978). An emperical transition matrix for nonhomogeneous Markov and chains based on censored observations, *Scandinavian Journal of Statistics* 5, 141–150.

[2] Meira-Machado, L.F., de Uña-Álvarez, J., Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illness-death model. *Lifetime Data Analysis*, 12, 325-344.

[3] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K. (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, 18, 195 - 222.

[4] Meira-Machado, L., de Uña-Álvarez, J., Datta, S. (2012). Conditional Transition Probabilities in a non-Markov Illness-death Model. *Discussion Papers in Statistics and Operation Research nž 11/03, 2011. Department of Statistics and Operations Research, University of Vigo. http* : *//webs.uvigo.es/depc*05*/reports/*

Poster

Estimação da função condicional de tempos de sobrevivência sucessivos

Luís Meira-Machado
Departmento de Matemática e Aplicações, Universidade do Minho, *lmachado@math.uminho.pt*
Ana Moreira
Departmento de Matemática e Aplicações, Universidade do Minho, *a.moreira.cris@gmail.com*

Palavras–chave: Função distribuição condicional, Função de sobrevivência condicional, Tempos consecutivos, Estimador Kaplan-Meier.

Resumo: A Análise de Sobrevivência tem como principal objetivo analisar uma determinada ocorrência durante um período de tempo, ou seja, o tempo decorrido desde um evento inicial até à ocorrência de um determinado evento de interesse. No entanto, nem sempre é possível observar o evento de interesse devido, à perda de follow-up, abandono de estudo, etc. Nestas situações possuímos informação incompleta e as observações são designadas de observações censuradas. A análise de sobrevivência pode ser descrita pelo processo de Markov com dois estados, 'vivo' e 'morto' e uma única transição entre eles. Em estudos longitudinais médicos, os doentes podem experimentar vários eventos num determinado período de acompanhamento. Nestes estudos, os tempos entre dois estados consecutivos (e tempos consecutivos) são muitas vezes de interesse e lidam com problemas que tem recebido muita atenção recentemente. Nos últimos anos contribuições significativas tem sido desenvolvidas relacionadas com este tópico. Problemas de interesse incluem a estimação da função bivariada de sobrevivência, distribuição marginal e função de distribuição condicional do segundo tempo dado o primeiro tempo.
O objectivo deste trabalho é a estimação da função de distribuição condicional e função de sobrevivência condicional do segundo tempo dado o primeiro tempo. Diferentes abordagens serão consideradas para estimar estas quantidades, todas elas baseadas no estimador da função de sobrevivência de Kaplan-Meier. Neste trabalho comparamos vários estimadores e investigamos o desempenho destes através de estudos de simulação. As metodologias propostas são ilustradas recorrendo a dados reais. Por exemplo podemos calcular o número de meses que o paciente permanece sem desenvolver a segunda recorrência dado que já permaneceu um certo número de meses sem desenvolver a primeira recorrência. Iremos comparar ainda resultados entre as nossas abordagens usando o package do software R **survivalBIV** e o package **bwsurvival**.

Agradecimentos

Os autores agradecem a recepção de apoio financeiro do Ministério Português da Ciência, Tecnologia e Ensino Superior sob a forma de subvenções PTDC/MAT/104879/2008 e SFRH/BD/62284/2009. Este trabalho é financiado por Fundos FEDER através do Programa Operacional Factores de Competitividade COMPETE e por Fundos Nacionais através da FCT- Fundação para a Ciência e a Tecnologia no âmbito do projecto PEst/MAT/UI0013/2011 e pelo Centro de Matemática da Universidade do Minho.

Referências

[1] Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *The Annals of Statistics* 16, 1475–1489.

[2] Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481.

[3] Lin, D. Y., Sun, W., Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* 86, 59–70.

[4] Moreira, A., Machado, L. (2012). survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring. *Journal of Statistical Software* 46, 1–16.

[5] Serrat, C., Gómez, G. (2007). Nonparametric bivariate estimation for successive survival times. *SORT* 31, 75–96.

[6] Wang, W., Wells, M.T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika* 85, 561–572.

Poster

A influência da idade da mãe no desempenho da escala CRIB: curva ROC condicionada

Filipa Mourão
Instituto Politécnico de Viana do Castelo, fmourao@estg.ipvc.pt

Ana Cristina Braga
DPS - EE, Universidade do Minho, acb@dps.uminho.pt

Pedro Oliveira
ICBAS, Universidade do Porto, pnoliveira@icbas.up.pt

Palavras–chave: ROC (Receiver Operating Characteristic), ROC condicionada, CRIB (Clinical Risk Index for Babies), kernel.

Resumo: Em diagnósticos médicos, a análise ROC, em particular a curva ROC (Receiver Operating Characteristic), é uma técnica muito utilizada para avaliar o desempenho de um teste, sendo a AUC (Area Under the Curve) o índice mais usado. Por definição, a curva ROC é a representação gráfica, no plano unitário, dos pares de valores de sensibilidade ou Fração de Verdadeiros Positivos (FVP) e (1-especificidade) ou Fração de Falsos Positivos (FFP), ordenadas e abcissas, respectivamente, obtidos ao considerar todos os possíveis valores de corte da escala e que proporciona uma representação global da exactidão dessa escala. Uma curva ROC é deste modo uma descrição empírica da capacidade da escala poder discriminar entre dois estados (anormal, normal) na qual cada ponto traduz um compromisso diferente entre FVP e FFP obtido, por exemplo, pela adopção de valores de corte diferentes [2]. No entanto, na maior parte dos testes de diagnóstico, uma ou mais covariáveis, sejam elas contínuas ou categóricas, existem associadas à variável diagnóstico. A informação contida nesta covariável pode aumentar o poder discriminante da curva ROC e influenciar a respectiva AUC [1]. Ilustraremos, neste trabalho, através da curva ROC empírica e da curva ROC suavizada, aplicando o método kernel [3], como a idade da mãe pode influenciar o desempenho da escala CRIB (Clinical Risk Index for Babies) na discriminação entre bebés com risco de falecimento (anormais) e de sobrevivência (normais).

Referências

[1] López-de-Ullibarri, I., Cao, R., Cadarso-Suárez, C., Lado, M.J. (2008). Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics & Data Analysis* 52, 2623–2631.

[2] Metz, C.E. (1986). Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. Multiple Regression Analysis: Applications in the Health Sciences. *American Institute of Physics* 13, 365–384.

[3] Peng, L., Zhou, X. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference* 118, 129–143.

Poster

Exame à medula óssea e reações adversas

Sara Narciso
DEIO, Faculdade de Ciências da Universidade de Lisboa, saradiasnarciso@gmail.com
Fernanda Diamantino
DEIO e CEAUL, Faculdade de Ciências da Universidade de Lisboa, mfdiamantino@fc.ul.pt
Ana Batalha Reis
Hospital S. Francisco Xavier, breis.ana@gmail.com

Palavras–chave: Aspiração da medula óssea, Biópsia óssea, Tabelas de contingência.

Abstract: Foi realizado um questionário pelo Serviço de Hematologia e pelo Laboratório de Hematologia do Hospital São Francisco Xavier (HSFX), aos doentes submetidos a Punção Aspirativa da Medula Óssea e Biópsia Óssea, de modo a identificar problemas relacionados com estes exames. Este trabalho tem como objetivos avaliar o tipo e a incidência de reações adversas, identificar fatores de risco e avaliar a forma como decorreu o exame, de modo a permitir detetar eventuais problemas e a sua consequente correção para o futuro.

1 Introdução

O estudo da Medula Óssea (MO) é essencial na avaliação das doenças hematológicas e no esclarecimento de muitas patologias não hematológicas, sendo normalmente constituído por dois exames, a Punção Aspirativa da Medula Óssea/Aspiração Medular (AM) e a Biópsia Óssea (BO). Estes dois exames complementam-se, permitindo uma avaliação completa da MO, mas nem todas as situações têm indicação para fazer ambos os exames.
O estudo da MO tem um papel fundamental no diagnóstico, monitorização e estadiamento de muitas doenças, fornecendo informações sobre o estado e a capacidade de produção das células sanguíneas, sobre a presença de infiltrados celulares patológicos, sobre a presença de microrganismos infeciosos e outras informações relevantes. Este estudo tem indicação quando o paciente se enquadra em alguma das seguintes situações: Anemia grave, Leucemia aguda, Síndroma Mielodisplásica, Leucemia Mielóide Crónica, Policitémia Vera, Mielofibrose e Trombocitemia Essencial, Mieloma Múltiplo, Neutropenia, Trombocitopenia, Linfomas, Carcinomas, e em determinadas infeções. [1]
As reações adversas provenientes da Punção Aspirativa da Medula Óssea e da Biópsia Óssea embora raras, são reconhecidas, e algumas originam quadros clínicos graves. [3] As reações adversas mais relevantes descritas na literatura são a ocorrência de hemorragia, hematoma, infeção e outras. [2]
A equipa de médicos do Serviço de Hematologia e do Laboratório de Hematologia do HSFX efetuou um questionário no ano de 2010, que foi aplicado a todos os doentes que efetuaram estes exames. Neste questionário foram avaliadas a ocorrência de hematoma, hemorragia, infeção e outras reações adversas. Embora não se encontre referência relevante na bibliografia relativamente à avaliação da dor, provavelmente por não ser considerada uma reação adversa, mas naturalmente decorrente do próprio exame, os autores consideraram importante estudar a sua incidência durante e após o exame. Os dados obtidos com este questionário foram analisados de forma a permitir identificar fatores de risco associados a estas complicações, de modo a tentar prevenir problemas no futuro.

2 Uma pequena discussão final

Em 2010, no Serviço de Hematologia e no Laboratório de Hematologia do HSFX, 256 pacientes realizaram a AM e a BO e 1 doente apenas realizou AM. Foram detetados 63 (24%) doentes que referiram dor e 17 (6,6%) as seguintes reações adversas: 8 tiveram dor e hematoma, 1 teve hemorragia e hematoma, 1 teve dor, hemorragia e hematoma, 1 teve dor, hematoma e infeção, 5 tiveram apenas hematoma e 1 teve apenas hemorragia. A ocorrência mais sentida pelos doentes foi a dor após o exame. Com este trabalho procurou-se encontrar fatores de risco (IMC, diagnóstico inicial, sexo, idade, entre outros) que possam estar ligados à presença de reações adversas.

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto PEst-OE/MAT/UI0006/2011.

Referências

[1] Bain, B.J. (2001). Bone marrow aspiration. *Journal of Clinical Pathology* 54, 657–663.

[2] Bain, B.J. (2005). Bone marrow biopsy morbidity: review of 2003. *Journal of Clinical Pathology* 58, 406–408.

[3] Malempati, S., Joshi, S., Lai, S., Braner, D.A.V., Tegtmeyer, K. (2009). Bone marrow aspiration and biopsy. *The New England Journal of Medicine* 361, e28.

Poster

Modelos de idade-período-coorte para a projeção da incidência de cancro: novas abordagens

Joana Oliveira
Escola de Ciências, Universidade do Minho, joana_oliveira@hotmail.com
Clara Castro
Registo Oncológico Regional do Norte (RORENO), clara.castro@ipoporto.min-saude.pt
Luís Meira Machado
Dep. Matemática e Aplicações e CMAT, Universidade do Minho, lmachado@math.uminho.pt

Palavras–chave: Epidemiologia, Estatística, Projeções, Cancro.

Resumo: As projeções de incidência de uma doença são efetuadas com objetivos científicos e administrativos, sendo de fulcral importância que estas estimativas sejam precisas e o mais próximo possível dos valores reais futuros. Para este efeito, existem diversos métodos estatísticos que modelam a evolução da incidência da doença ao longo do tempo. No âmbito da epidemiologia do cancro salientam-se os modelos de Poisson [1], modelos simples que extrapolam a tendência observada, e modelos APC (age-period-cohort) [2] que incorporam os efeitos da idade, período e coorte nas previsões. Neste trabalho pretendeu-se utilizar uma nova abordagem descrita na literatura para a projeção da incidência de cancro, que consiste no recurso a modelos APC que utilizam splines cúbicas restritas [3]. A qualidade de ajustamento destes modelos foi comparada com a de modelos de regressão de Poisson, vulgarmente utilizados na realização de projeções a curto prazo. A aplicação prática dos modelos incidiu sobre os cancros do sistema digestivo na Região Norte de Portugal, nomeadamente o cancro colo-rectal, o cancro do estômago e o cancro do esófago, diagnosticados entre 1994 e 2008. Estes tumores foram escolhidos por apresentarem diferentes padrões de incidência, permitindo assim a realização de uma análise mais abrangente e a avaliação da utilidade dos diversos modelos em diferentes contextos.

Referências

[1] Hakulinen T, Dyba T. (1994). Precision of incidence predictions based on Poisson distributed observations. *Statistics in Medicine* 13, 1513–1523.

[2] Møller, B., Fekjær, H., Hakulinen, T., Sigvaldason, H., Storm, H.H., Talbck, M., Haldorsen, T. (2003). Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Statistics in Medicine* 22, 2751–2766.

[3] Rutherford, M.J., Thompson, J.R., Lambert, P.C. (2012). Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *International Journal of Biostatistics* 8(1), 33.

Poster

VIH/SIDA Estimação de probabilidades de atrasos e sub-notificação

Alexandra Oliveira
ESTSP-IPP, Faculdade de Ciências da Universidade do Porto, aao@estsp.ipp.pt
Joaquim Pinto da Costa
Faculdade de Ciências da Universidade do Porto, jpcosta@fc.up.pt
Ana Rita Gaio
Faculdade de Ciências da Universidade do Porto, argaio@fc.up.pt

Palavras–chave: Epidemiologia, VIH/SIDA, Estimação, Probabilidades, Atrasos, Sub-notificação.

Resumo: A epidemia do Vírus de Imunodeficiência Humana (VIH)/ Síndrome de Imunodeficiência Adquirida (SIDA) é um sério problema de saúde pública em Portugal. A situação epidemiológica observada a 31 de Dezembro de 2011 foi de 41035 casos de infecção notificados nos diferentes estádios de desenvolvimento [4]. Esta prevalência é uma das mais elevadas da europa[1].
Neste país, todos os estádios de desenvolvimento da doença (assintomático, complexo relacionado com a SIDA e SIDA) devem ser notificados assim como qualquer alteração ocorrida, incluindo a morte (Portaria N. 103/2005 de 25 janeiro) [2] ao Centro de Vigilância Epidemiológica de Doenças Transmissíveis (CVEDT). Esta notificação é feita desde a identificação da doença em Portugal e tornou-se obrigatória no ano de 2005. No momento do diagnóstico, os médicos devem preencher o formulário oficial e enviar para o CVEDT. Com base nestas recolhas são produzidos relatórios sobre o situação epidemiológica observada. Assim, qualquer atraso na recepção dos casos implica uma descrição menos fiável da situação epidemiológica. Para além deste problema, o sistema Português sofre de outro problema amplamente reconhecido, o problema da sub-notificação [2].
Este trabalho tem como principal objectivo o ajuste do número de casos de SIDA notificados relativamente aos atrasos de notificação e à sub-notificação. No caso do primeiro problema, foram comparadas diferentes metodologias para modelar eventos de contagens com excessos de zeros e heterocedasticidade. No caso do segundo problema, utilizou-se a metodologia descrita em [5].

Referências

[1] CNSida, Coordenação Nacional para a Infecção VIH/SIDA (2006). VIH / sida.*http://www.sida.pt*

[2] Mauch, S. (2009). Situational Assessment of the HIV/AIDS Notification System - A Portuguese Experience. *National Coordination For HIV Infection.*

[3] Amaral, J.A., Pereira, E.P., Paixão, M.T. (2005). Data and projections of HIV/AIDS cases in Portugal: an unstoppable epidemic? *Journal of Applied Statistics* 32(2), 127–140.

[4] Núcleo de Vigilância Laboratorial de Doenças Infecciosas. Unidade de Referência e Vigilância Epidemiológica. Departamento de Doenças Infecciosas do INSA Programa Nacional para a Infeção VIH/SIDA (2012). Infeção VIH/SIDA: a situação em Portugal a 31 de dezembro de 2011. *Instituto Nacional de Saáde Doutor Ricardo Jorge, IP*, 143.

[5] Fader, P., Hardie, B. (2000). A note on modelling underreported Poisson counts. *Journal of Applied Statistics* 27(8), 953–964.

Poster

Regressão quantílica para dados longitudinais: uma aplicação na área da Medicina

Ana Luisa Papoila
Faculdade de Ciências Médicas da Universidade Nova de Lisboa, Centro de Investigação do Centro Hospitalar de Lisboa Central, CEAUL, ana.papoila@fcm.unl.pt

Marta Alves
Centro de Investigação do Centro Hospitalar de Lisboa Central, marta.l.alves@gmail.com

Daniel Virella
Unidade de Cuidados Intensivos Neonatais, Hospital de Dona Estefânia, Centro de Investigação do Centro Hospitalar de Lisboa Central, danielvirella@oninetspeed.pt

Andreia Mascarenhas
Unidade de Cuidados Intensivos Neonatais, Hospital de Dona Estefânia, amascarenhas22@gmail.com

Teresa Neto
Faculdade de Ciências Médicas da UNL, Unidade de Cuidados Intensivos Neonatais, Hospital de Dona Estefânia, mariateresaneto@sapo.pt

Palavras–chave: Regressão quantílica, Curvas de crescimento, Saturação de oxigénio, Recém-nascidos de termo.

Resumo: Nos estudos longitudinais em que são registadas várias medidas para cada indivíduo, ao longo do tempo, as metodologias estatísticas a aplicar devem tomar em consideração a estrutura de autocorrelação existente entre as observações repetidas. Assim sendo, a variabilidade intraindividual deve ser considerada por forma a evitar o enviesamento das estimativas dos parâmetros dos modelos que se ajustem aos dados.
São exemplo deste tipo de estudos, aqueles que visam ajustar curvas que permitam descrever padrões de evolução e, ainda, identificar variáveis que expliquem essa mesma evolução. As curvas de crescimento e a regressão quantílica constituem a habitual escolha metodológica que permitirá caracterizar as alterações na variável resposta ao longo do tempo e determinar os factores que as originam. De facto, a utilização da regressão quantílica tem vindo a aumentar, pois permite descrever a distribuição da variável resposta em situações de grande assimetria, nas quais assumir a normalidade é questionável e a mediana constituirá, seguramente, uma medida de tendência central mais adequada e informativa.
Este estudo descreve a metodologia descrita por Geraci e Bottai (2007), que propõem um modelo de regressão quantílica condicional assumindo uma distribuição de Laplace Assimétrica para a distribuição da resposta contínua e incluindo, além de vários preditores, um *intercept* aleatório.
Para ilustração deste método, é utilizado um conjunto de dados de saturação de oxigénio de recém-nascidos de termo nos primeiros 30 minutos de vida, medida na sala de partos por oximetria de pulso transcutânea contínua, com o objectivo de caracterizar a variação deste parâmetro durante o período precoce de transição da vida fetal para a vida extra-uterina, em crianças aparentemente saudáveis, de modo a melhor orientar as decisões clínicas na sala de partos. Os dados foram analisados com o pacote do R *lqmm* (Geraci, 2011).

Agradecimentos

Este estudo foi parcialmente financiado por Fundos Nacionais através da FCT − Fundação para a Ciência e Tecnologia no âmbito do projecto PEst-OE/MAT/UI0006/2011.

Referências

[1] Geraci, M., Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* 8, 140–154.

[2] Geraci, M. (2011), *lqmm*: Linear Quantile Mixed Models, R package version 1.0.

Poster

Análise semiparamétrica bayesiana da diversidade do repertório de recetores de células T

Nuno Sepúlveda
London School of Hygiene and Tropical Medicine, U.K., *nuno.sepulveda@lshtm.ac.uk*

Carlos Daniel Paulino
IST, Universidade Técnica de Lisboa, *dpaulino@math.ist.utl.pt*

Michele Guindani
Department of Biostatistics, U.T. M.D. Anderson Cancer Center, Houston, TX, USA, *mguindani@mdanderson.org*

Peter Müller
Department of Mathematics, University of Texas at Austin, Austin, TX, USA, *pmueller@math.utexas.ed*u

Palavras–chave: mistura por processo Dirichlet, estimação da diversidade, contagens de sequências, recetor de célula T.

Resumo: Muitos conjuntos de dados gerados de técnicas de sequenciação correntes podem ser sumariados por uma distribuição empírica de abundância de sequências, como acontece em Imunologia na análise da diversidade dos recetores de linfócitos T. Esta distribuição é usualmente caraterizada por muitas sequências registadas com baixas frequências e poucas sequências altamente abundantes. Propõe-se aqui um modelo semiparamétrico bayesiano capaz de lidar com uma tal sobredispersão e de estimar o número de sequências distintas não observadas. O seu uso é ilustrado com a análise de um conjunto de dados de recetores de células T de estirpes de ratinhos respeitante a uma pesquisa sobre diabetes. O desempenho da abordagem proposta é também comparado com o de um elenco de contrapartidas paramétricas usadas em problemas deste género por meio de um estudo de simulação. Os resultados concernentes à estimação da diversidade demonstram a superioridade do modelo semiparamétrico proposto sobre as alternativas paramétricas em termos do viés de estimativas pontuais e cobertura de intervalos de credibilidade HPD.

Agradecimentos

Nuno Sepúlveda e Carlos Daniel Paulino foram parcialmente financiados pela fundação portuguesa FCT, através do projeto Pest-OE/MAT/UI0006/2011. Nuno Sepúlveda foi ainda apoiado por uma bolsa da Foundation for the National Institutes of Health (grant ref. #566) e Wellcome Trust (grant ref. 077383/Z/05/Z) através de Grand Challenges in Global Health Initiative. Peter Müller foi parcialmente apoiado pela bolsa NIH/R01 CA075981.

Poster

Ajuste da função logística a dados de crescimento

Glauber Márcio Silveira Pereira
Curso de Biometria-IB-UNESP-Botucatu-Brasil, *glauber@ibb.unesp.br*
Lídia Raquel de Carvalho
Curso de Biometria-IB-UNESP-Botucatu-Brasil, *lidiarc@ibb.unesp.br*
Martha Maria Mischan
Curso de Biometria-IB-UNESP-Botucatu-Brasil, *mmischan@ibb.unesp.br*

Palavras–chave: Regressão não linear, Função logística, Heterocedasticidade.

Resumo: As curvas de crescimento têm diversas aplicações de suma importância em várias áreas, em particular a função logística também tem sido bastante utilizada. O objetivo desta pesquisa foi o estudo da função Logística com ajustes em três estruturas: modelo de efeitos fixos, modelo com ponderação e modelo de efeito misto, a dados de peso de frutos de laranjeiras doces *Citrus sinensis* (L.) Osbeck em um experimento com cinco copas enxertadas em porta-enxertos situado na Fazenda Experimental Lageado em Botucatu. Para comparação dos modelos foram utilizados os critérios: Quadrado médio dos resíduos, Critério de informação de Akaike, Critério de Informação Bayesiano de Schwarz, teste de Breusch Pagan, teste de Durbin Watson e coeficiente de determinação. Concluiu-se que o modelo com ponderação foi o melhor para resolver os problemas de heterocedasticidade e autocorrelação.

1 Introdução

As curvas de crescimento têm diversas aplicações de suma importância em várias áreas, em particular a função logística também tem sido bastante utilizada.
Mazzini e colaboradores (2003) ajustaram as funções de Brody, Gompertz, Logística, Richards e von Bertalanffy a dados de crescimento de bovinos Hereford. Foram obtidos ajustes de curvas individuais para os animais em dois diferentes modelos: não-ponderado e ponderado . O melhor modelo foi o ponderado pelo inverso da variância dos pesos. As funções que apresentaram melhor ajuste foram as de von Bertalanfry e Gompertz, seguidas da Logística.
Silva, Aquino e Oliveira (2001) ajustaram a funções de crescimento de Brody, Logística, Gompertz, Richards e Bertalanffy para descrever o crescimento de 542 animais da raça Nelore. As funções foram ajustadas através dos mínimos quadrados generalizados para modelos de regressão não-linear com erros auto-regressivos de primeira ordem. Concluíram que as funções de Bertalanffy, Gompertz e Logística superestimaram o peso inicial e subestimaram o peso adulto dos animais, e que as funções de Brody e Richards apresentaram um melhor comportamento em relação às demais.

2 Utilização de estruturas

A função utilizada foi a Logística:

$$y = \frac{\alpha}{(1 + \exp(-(\beta + \gamma x)))}$$

sendo y a observação no tempo x, α a distância entre as duas assíntotas, β , um parâmetro de posição e γ está relacionado com a taxa de crescimento da função.

No modelo ponderado a função utilizada foi a mesma acima citada e foi empregado o método dos quadrados mínimos sendo que a ponderação foi feita pelo inverso da variância dos valores em cada tempo, empregando-se a opção WEIGHT do procedimento MODEL do SAS (SAS,1995).
O modelo misto foi o seguinte:

$$y_{i,j} = F(x_i, \theta) + \Delta_j g(x_i, \theta) + \xi_{i,j}, \text{ com } g(x_i, \theta) = \frac{F(x_i, \theta)}{\alpha}$$

onde Δ_j = efeito aleatório do j-ésimo indivíduo, com distribuição normal $(0, \sigma^2_\Delta)$ e $\xi_{i,j}$ = erro aleatório, independente, com distribuição normal $(0, \sigma^2_\xi)$, independente de Δ_j.

3 Uma pequena discussão final

Havia seis repetições para cada porta-enxerto em cada copa, porém foi feito um ajuste único para que fosse possível comparar com o modelo misto. Ficaram, portanto, vinte e cinco repetições e quando foi feito o ajuste para o modelo sem ponderação, foram detectados problemas em dezassete repetições da seguinte forma: em sete delas havia heterogeneidade de variâncias, em quatro havia heterogeneidade de variâncias e presença de autocorrelação residual, em cinco havia autocorrelação e uma repetição não convergiu. O modelo com ponderação corrigiu treze das dezessete repetições e o modelo misto corrigiu apenas cinco, porém, destas cinco somente em duas ele foi o melhor modelo, de acordo com os critérios utilizados.
Concluímos portanto que neste caso o modelo com ponderação foi o melhor para resolver os problemas de heterocedasticidade e autocorrelação.

Agradecimentos

Este trabalho foi financiado pela CAPES.

Referências

[1] Mazzini, A.R.A., Muniz, J.A., Aquino, L.H., Silva, F.F. (2003). Análise da curva de crescimento de machos Hereford. *Ciência e Agrotecnologia* 25(5), 1105–1112.

[2] Silva, F.F., Aquino, L.H., Oliveira, A.I.G. (2001). Influência de fatores genéticos e ambientais sobre as estimativas dos parâmetros das funções de crescimento de gado nelore. *Ciência e Agrotecnologia* 25(5), 1195–1205.

Poster

Biomonitorização ambiental em Portugal continental - uma análise espacial e temporal

Helena Piairo
Universidade do Minho, *helena.piairo@gmail.com*
Raquel Menezes
Universidade do Minho, *rmenezes@math.uminho.pt*
Inês Sousa
Universidade do Minho, *isousa@math.uminho.pt*

Palavras–chave: Biomonitorização, Estatística Multivariada, Geoestatística.

Resumo: Determinados organismos vivos apresentam capacidades de acumulação de elementos químicos e uma sensibilidade à poluição e, por este motivo, é possível utilizá-los para quantificar a concentração de poluentes nos organismos - Biomonitorização [1]. Assim, a biomonitorização consiste na determinação da concentração dos metais pesados presente nos organismos, que será um indicador da concentração dos mesmos no ambiente, tornando-se uma metodologia bastante simples e com uma elevada resolução espacial e temporal, que proporciona uma interpretação relevante em termos de efeitos sobre o ambiente.
Assim, os resultados da realização deste sistema de biomonitorização tornam possível a aplicação de metodologias estatísticas que permitam a identificação das interações do organismo com os parâmetros ambientais, ao longo do espaço e do tempo.
Neste trabalho, pretende-se fazer um estudo estatístico, através da análise univariada e multivariada, dos dados amostrados através de uma rede de biomonitorização que abrange Portugal continental. Os dados foram obtidos ao longo de quatro campanhas, nomeadamente em 1992, 1997, 2002 e 2006. Nas duas primeiras campanhas, foram recolhidas amostras de musgo em 173 e 177 localizações, respetivamente, definidas numa escala nacional mas intensificadas em grandes zonas urbanas e industriais (distritos de Lisboa e Aveiro). Na terceira campanha, recolheram-se em 151 localizações, cujas amostras se concentraram numa grelha mais fina na região de Aveiro. Por último, a quarta campanha, com características distintas das anteriores, apenas considerou 98 localizações delimitadas pela Região do Centro, uma das regiões da NUT II. Os detalhes sobre o processo de amostragem e as técnicas de medição utilizadas estão descritas em [1] e [2].
É objetivo deste estudo o reconhecimento das relações lineares existentes entre os conjuntos múltiplos de dados a que temos acesso, com o intuito de encontrar fontes de poluição que estejam relacionadas com as concentrações dos metais pesados nos musgos. Para alcançar tal objetivo, utilizar-se-ão métodos de análise multivariada, tais como a Análise em Componentes Principais e a Análise Fatorial, descritas em [3] e [4], cujos resultados interessa comparar ao longo das várias campanhas. Pretende-se, ainda, aplicar técnicas de modelação geoestatística, descritas em [5], sobre os *scores* obtidos através da Análise Fatorial.

Agradecimentos

Os dados analisados no presente estudo foram cedidos pelo Centro de Biologia Ambiental da Universidade de Lisboa, parceiro do projeto *Modelos conjuntos para processos espaço-temporais e respetivo desenho amostral, em Ciências do Ambiente e Saúde*, PTDC/MAT/112338/2009, financiado por fundos nacionais através da FCT, no qual este trabalho de investigação se integra.

Referências

[1] Figueira, R., Sérgio, C., Ramalho, C., Sousa, A.J. (2002). Distribution of trace metals in moss biomonitors and assessment of contamination sources in Portugal. *Environmental Pollution* 118, 153–163.

[2] Martins, A., Figueira, R., Sousa, A.J., Sérgio, C. (2012). Spatio-temporal patterns of Cu contamination in mosses using geostatistical estimation. *Environmental Pollution* 170, 276–284.

[3] Everitt, B.S., Dunn G. (2001). *Applied Multivariate Data Analysis*. Second edition, Arnold, London.

[4] Morrison, D. (2005). *Multivariate Statistical Methods*. Fourth edition, McGraw-Hill, New York.

[5] Diggle, P., Ribeiro, P., (2007). *Model-based Geostatistics*. Springer, USA.

Poster

O modelo de Crámer-Lundberg e a probabilidade de ruína

Celine Queirós
Universidade do Minho, celinequeiros@portugalmail.pt

Patrícia Gonçalves
Universidade do Minho, patg@math.uminho.pt

Irene Brito
Universidade do Minho, ireneb@math.uminho.pt

Palavras–chave: Modelo de Crámer-Lundberg, Probabilidade de ruína, Teorema fundamental do risco.

Resumo: A atividade seguradora, nas últimas décadas, está a ganhar terreno e a ocupar um lugar de destaque na sociedade. Os seguros oferecem proteção contra riscos de acidentes de viação, de acidentes pessoais, de acidentes de trabalho, de incêndios, climatéricos, de crédito, entre muitos outros. Quando um evento imprevisível ocorre solicitam-se os serviços de uma seguradora. Esta, através de um seguro transfere o risco de perdas financeiras em troca de um montante ou prémio. A probabilidade de ruína é uma preocupação constante para o negócio da atividade seguradora, ou seja, é a probabilidade de uma seguradora ficar sem reserva ou capital suficiente para pagar a indemnização de um determinado sinistro. Para a sobrevivência da mesma, é necessário avaliar o capital inicial investido e o cálculo do valor do prémio. O estudo da probabilidade de ruína de uma seguradora arruinar é feito usando o processo de reserva que se designa por $U(t), t \geq 0$. Apresenta-se assim o modelo clássico de Crámer-Lundberg:

$$U(t) = u + ct - S(t), t \geq 0,$$

onde $U(t)$ é designado de reserva de uma seguradora no instante $t \geq 0$, para fazer face a determinado risco, $u = U(0)$ representa o capital inicial e assume-se que os prémios são recebidos (por unidade de tempo) uma taxa constante que se denota por c. O processo $S(t) = \sum_{i=0}^{N(t)} X_i,\ t \geq 0$ representa o valor das indemnizações agregadas relativas ao intervalo de tempo $(0, t]$, onde $X_0 = 0$, para $i = 1, ..., N(t)$, X_i representa o montante da i-ésima indemnização cujo sinistro ocorreu no intervalo de tempo $(0, t]$ e $N(t)$ representa o número de sinistros ocorridos no intervalo de tempo $(0, t]$. Este é um modelo probabilísta, de tal forma que para $i = 1, ..., N(t)$ X_i são variáveis aleatórias com certa distribuição e $N(t)$ é um processo estocástico de contagem. A escolha para a distribuição das variáveis X_i e do processo $N(t)$, dependem da escolha do tipo de seguros a considerar.

A probabilidade de ruína ocorre quando $U(t)$ for negativo, num certo instante de tempo. De uma forma mais geral, diz-se que ocorreu ruína se o processo de reserva é inferior a uma barreira de ruína, num certo instante ou instantes de tempo. Desta forma, define-se o primeiro instante de ocorrência de ruína, denotado por T, como $T = \min\{t, t \geq 0\ e\ U(t) < 0\}$ Pressupõe-se que T é infinito se a reserva for não negativa para todo $t \geq 0$, ou seja $U(t) \geq 0$ para todo $t \geq 0$ o que significa a ausência de ruína para qualquer instante t. Assim, a probabilidade de ruína para t tempo contínuo e horizonte infinito é dada por $\psi(u) = P(T < \infty)$, partindo de uma reserva inicial u. Define-se ainda a probabilidade de sobrevivência, ou seja, a probabilidade de não ocorrer ruína em tempo contínuo e horizonte infinito como $\delta(u) = 1 - \psi(u)$.

Em muitas situações não é possivel obter o valor exacto da probabilidade de ruína, consegue-se apenas a sua quantificação de forma aproximada. Neste modelo, considera-se que $S(t)$ é um

processo de Poisson composto, ou seja $N(t)$ é um processo de Poisson. Pelo Teorema fundamental do Risco, tem-se que para um processo de risco Poisson composto com capital inicial $u > 0$ e sendo R o coeficiente de ajustamento, a probabilidade de ruína é dada por

$$\psi(u) = \frac{e^{-Ru}}{E\left[e^{-RU(T)}|T < \infty\right]}.$$

Ressalva-se que no caso de acontecer a ruína, esta ocorre como consequência da ocorrência de uma indemnização. Este resultado propõe o cálculo da probabilidade de ruína sem especificar o tempo em que ela ocorre, mas em que número de indemnização.
Num modelo de risco em cálculo atuarial, é preciso especificar a distribuição para o número de indemnizações e para a severidade das indemnizações. Na prática, e quando se implementa o modelo, não há conhecimento da distribuição destas quantidades e tem-se a necessidade de enquadrar os dados da amostra ao modelo teórico. Neste trabalho apresenta-se um estudo com base em diferentes distribuições, que usualmente são utilizadas para descrever o montante de indemnizações particulares, em diferentes ramos de seguros. Posteriormente é feita uma análise de uma base de dados reais de uma seguradora com atividade no mercado português, aplicando os conceitos teóricos mencionados anteriormente, de forma a dar uma visão da aplicabilidade desse modelo no contexto real.

Referências

[1] Burnecki, K. Misiorek, A., Weron, R. (2005). Loss Distributions. In Cizek, P., Haerdle, W., Weron, R. (eds.): *Statistical Tools for Finance and Insurance*, 289–317, Springer-Verlag, Berlin.

[2] Centeno, M. (2003). *Teoria de Risco na Actividade Seguradora*. Celta Editora.

[3] Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M. (2009). *Modern Actuarial Risk Theory: using R*. Springer.

[4] Reis, A. (2001). Teoria da Ruína. CEMAPRE-ISEG no 17/TA, Lisboa.

[5] Reis, A. (2002). O número de indemnizações até à ruína e recuperação. CEMAPRE-ISEG, Lisboa.

Poster

Modelagem do tempo de vida de pacientes com leucemia utilizando a distribuição geométrica beta generalizada semi-normal

Thiago Gentil Ramires
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, thiagogentil@usp.br
Edwin Moises Marcos Ortega
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, edwin@usp.br

Palavras–chave: Análise de sobrevivência, Distribuição geométrica beta generalizada semi-normal

Resumo: Com o avanço tecnológico aprimorado, diferentes comportamentos do tempo de vida vem sendo estudados, e com isso, é necessário a criação de modelos mais flexíveis para melhor ajustar esses comportamentos. Neste trabalho, foi utilizado a distribuição geométrica beta generalizada semi-normal para a modelagem do tempo de vida de pacientes diagnosticados com leucemia, a qual se mostrou mais flexível para a modelagem dos tempos.

1 Introdução

Em estatística, uma das areas que mais cresceu nos últimos anos foi a de análise de sobrevivência, fato este, evidenciado por sua quantidade de aplicações nas mais diversas áreas de pesquisa. Com o avanço tecnológico aprimorado, diferentes comportamentos do tempo de vida vêm sendo estudados, proporcionando a criação de modelos mais adequados. Alguns desses novos modelos compõe da classe das distribuições geométricas como: exponencial geométrica (EG) [1], Weibull geométrica (WG) [2], beta-Weibull geométrica (BWG) [4], entre outros. O novo modelo utilizado neste trabalho, é composto pela transformação geométrica aplicada à distribuição beta generalizada semi-normal, nome curto em inglês (BGHN) [5], composta pelas distribuições beta e generalizada Semi-Normal(GHN).

2 Distribuição Geométrica Beta Generalizada Semi-Normal

A distribuição geométrica beta generalizada semi-normal, nome curto em inglês (BGHNG), composta por cinco parâmetros, sendo eles α, θ, a, b e p, tem a função densidade de probabilidade descrita por:

$$f(x;\alpha,\theta,a,b,p) = \frac{\left(\frac{\alpha}{x}\right)\left(\frac{x}{\theta}\right)^{\alpha} e^{\left[-\frac{1}{2}\left(\frac{x}{\theta}\right)^{2\alpha}\right]}\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a-1}\left\{1-\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]\right\}^{b-1}}{2^{\frac{1}{2}-b}(1-p)^{-1}\sqrt{\pi}B(a,b)\left\{1-p\left[1-I_{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1}(a,b)\right]\right\}^{2}}, \quad x>0. \quad (1)$$

em que θ,a,$b \geq 0$ são parâmetros de forma, $\alpha \geq 0$ e $0 \leq p \leq 1$ são parâmetros de escala, Φ(x) é a função densidade acumulada da distribuição normal e $I_{(.)}$ representa a função beta incompleta.

3 Aplicação a dados de leucemia

Os dados utilizados no estudo provêm de uma pesquisa realizada por Feigl e Zelen [3] sobre o tempo de vida de pacientes com leucemia, composto por 33 pacientes, os quais o tempo de vida (X) em semanas foi observado. As estimativas dos parâmetros do modelo e sub-modelos e seus respectivos desvios padrões são apresentadas na Tabela 1.

Tabela 1: Estimativas dos parâmetros dos modelos para os dados de leucemia.

Model	α	θ	a	b	p	AIC	CAIC	BIC
BGHNG	7.2631	132.2900	0.1231	0.1907	0.8749	309.50	311.70	317.00
	(0.3662)	(27.3855)	(0.0251)	(0.4127)	(0.1580)			
BGHN	0.3229	73.8221	3.2298	3.3246	0	314.90	316.30	320.90
	(0.2922)	(74.3400)	(4.5373)	(1.1058)	-			
GHN	0.6209	46.5178	1	1	0	311.50	311.90	314.50
	(0.0921)	(10.1120)	-	-	-			

O teste da razão de verossimilhança, apresentado na Tabela 2, foi utilizado para verificar a qualidade do moledo comparado com alguns de seus sub-modelos, em que os resultados mostram claramente ao nível de 5% de significância que o modelo (BGHNG) é o mais apropriado para modelagem dos tempos de vida dos pacientes no estudo.

Tabela 2: Teste da razão de verossimilhança.

Carbon	Hipóteses	Estatística w	p-value
BGHNG vs BGHN	$H_0 : p = 0$ vs $H_1 : H_0$ é falso	7.40	0.0065
BGHNG vs GHN	$H_0 : p = 0$ e $a = b = 1$ vs $H_1 : H_0$ é falso	8.0	0.0460

A qualidade do ajuste também foi obtida por métodos gráficos, construindo o histograma da distribuição dos tempos junto com as distribuições de probabilidade ajustada do modelo e sub-modelos apresentadas na Figura 1.

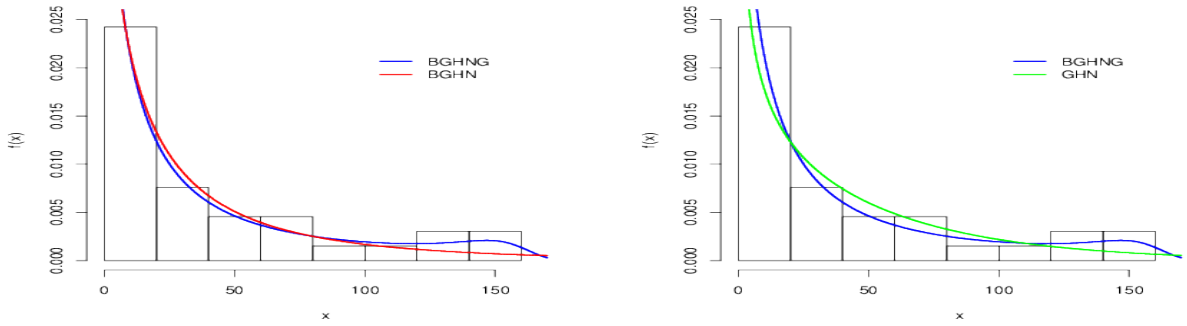

Figura 1: Estimativas dos modelos BGHNG, BGHN e GHN.

Referências

[1] Adamidis, K., Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statistics & Probability Letters* 39, 35–42.

[2] Barreto-Souza, W., Morais, A.L., Cordeiro, G.M. (2010). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation* 60, 35–42.

[3] Feigl, P., Zelen, M., (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* 21, 826–837.

[4] Cordeiro, G.M., Silva, G.O., Ortega, E.M.M. (2011). The beta-Weibull geometric distribution. *Statistics* DOI:10.1080/02331888.2011.577897

[5] Pescim, R.R., Demétrio, C.G.B., Cordeiro, G.M., Ortega, E.M.M., Urbano, M.R. (2010). The beta generalized half-normal distribution. *Computational Statistics and Data Analysis* 54(4), 945–957.

Poster

Influence of human mitochondrial DNA haplogroups on risk of osteoarthritis progression: an interval-censored approach

Ignacio Rego-Perez
INIBIC-Hospital Universitario A Coruña, ignacio.rego.perez@sergas.es

Angel Soto-Hermida
INIBIC-Hospital Universitario A Coruña, angel.soto.hermida@sergas.es

Sonia Pertega-Diaz
Unidad de Epidemiologia Clinica y Bioestadistica. CHU A Coruña, sonia.pertega.diaz@sergas.es

Mercedes Fernandez-Moreno
INIBIC-Hospital Universitario A Coruña, mercedes.fernandez.moreno@sergas.es

Juan Fernandez-Tajes
INIBIC-Hospital Universitario A Coruña, juan.fernandez.tajes@sergas.es

Eugenia Vazquez-Mosquera
INIBIC-Hospital Universitario A Coruña., maria.eugenia.vazquez.mosquera@sergas.es

Estefania Cortes-Pereira
INIBIC-Hospital Universitario A Coruña, estefania.cortes.pereira@sergas.es

Sara Relaño-Fernandez
INIBIC-Hospital Universitario A Coruña, sara.MA.relano.fernandez@sergas.es

Natividad Oreiro-Villar
INIBIC-Hospital Universitario A Coruña, natividad.oreiro.villar@sergas.es

Carlos Fernandez-Lopez
INIBIC-Hospital Universitario A Coruña, jesus.fernandez.lopez@sergas.es

Francisco Blanco-Garcia
INIBIC-Hospital Universitario A Coruña, francisco.blanco.garcia@sergas.es

Beatriz Lopez-Calviño
INIBIC-Hospital Universitario A Coruña, beatriz.lopez.calvinho@sergas.es

Teresa Seoane-Pillado
INIBIC-Hospital Universitario A Coruña, teresa.seoane.pillado@sergas.es

Keywords: Osteoarthritis, Progression, Survival, Interval-censoring.

Abstract: An interval-censored data analysis was performed to analyze the influence of mtDNA haplogroups on the risk of progression in patients with knee osteoarthritis. Results show that MtDNA haplogroups are associated with the osteoarthritis progression. Iidentification of haplogroups could allow a more personalized monitoring of patients with this disease

1 Introduction

The Osteoarthritis Initiative (OAI) is a multicenter project designed to identify risk factors associated with knee osteoarthritis. This project has created a public database that includes patients' clinical assessments, radiological images and biological samples. The aim of this study was to determine the influence of human mitochondrial DNA haplogroups (mtDNA) on the risk of osteoarthritis progression in a well-characterized cohort of patients from the OAI.

2 Methods

Design: Cohort study of n=891 Caucasian patients with knee osteoarthritis from the progression subcohort of the OAI project.
Measures: Age, gender, body mass index (BMI). The mtDNA haplogroup was determined for each patient. Patients were prospectively evaluated at 12, 24, 36 and 48 months.
Outcomes: Osteoarthritis progression, according to different criteria: i) Kellgren-Lawrence (KL) grading scale: increase of at least one grade in any knee, ii) Degree of joint space narrowing: an increase >=0.5 on the OARSI scale, iii) Presence of osteophytes in the medial tibial compartment, iv) Subcondral sclerosis in the medial tibial compartment
Analysis: Interval-censored data analysis methods were used [4, 2, 1]. Turnbull's extension of the Kaplan-Meier curve was used to estimate the cumulative probability of progression over time, according to haplogroups [5]. An extended Cox proportional hazard model using the iterative convex minorant algorithm was used for multivariate analysis [3]. Statistical significance was tested by means of confidence intervals for the risk ratios (RR), obtained using the bootstrap methodology (normal approximation, percentile method and adjusted bootstrap percentile). Statistical analysis were performed using R 2.10.0, in addition to the Icens (6), intcox and boot packages.

3 Results

Mean age was 61.9 (SD=9.4), 52.1% were women. Prevalence of different mtDNA haplogroups was: H (38.3%), U (17.7%), J (10.0%), T (9.5%), K (7.9%), others (16,6%).
The cumulative probability of KL progression at 48 months was significantly lower for patients with T haplogroup (22.0%) than for patients with the most frequent haplogroup H (42.5%) (RR = 0.49, 95% CI = 0.25 to 0.83). According to the joint space narrowing, carriers of haplogroup T also showed a lower probability of progression that carriers of haplogroup H (19.7% vs. 32.6%) (RR = 0.56, 95% CI = 0.29 to 0.99). Similar results were obtained for the presence of osteophytes (29.0% vs. 48.0%) (RR = 0.61, 95% CI = 0.34 to 0.99) and in the case of subchondral sclerosis (25.0% vs. 39.8%) (RR = 0.55; 95% CI = 0.29 to 0.90).

4 Conclusions

MtDNA haplogroups are associated with the risk of osteoarthritis progression. Early identification of these haplogroups could allow a more personalized monitoring of patients with this disease.

References

[1] Gomez, G., Calle, M.L., Oller, R., Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling* 9(4), 259–297.

[2] Lindsey, J.C., Ryan, L.M. (1998). Tutorial in biostatistics methods for interval censored data. *Statistics in Medicine* 17, 219–238.

[3] Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval censored data. *Journal of Computational and Graphical Statistics* 8, 109–120.

[4] Sun, J. (2006). *The statistical analysis of interval-censored failure time data.* Springer, New York.

[5] Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B* 38, 290-295.

Poster

Variabilidade espacial de cálcio no solo da região de Unaí-MG, Brasil

Ana Julia Righetto
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, ajrighetto@usp.br
Luiz Ricardo Nakamura
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, lrnakamura@usp.br
Diogo Néia Eberhardt
Escola Superior de Agricultura "Luiz de Queiroz", Univer. São Paulo, eberhardt.diogo@gmail.com
Paulo Justiniano Ribeiro Junior
Universidade Federal do Paraná, paulojus@ufpr.br
Roseli Aparecida Leandro
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, raleandr@usp.br

Palavras–chave: Função Matérn, Geoestatística, Krigagem.

Resumo: Este trabalho utilizou técnicas da geoestatística para analisar a variabilidade espacial do cálcio em uma região localizada em Unaí, noroeste do Estado de Minas Gerais, Brasil.

1 Introdução

Geoestatística é um ramo da geografia matemática e da estatística, que une o conceito de variáveis aleatórias com o conceito de variáveis regionalizadas, gerando um conceito de funções aleatórias. Neste ramo, as técnicas que mais se destacam são krigagem e a simulação estocástica e, com o auxílio dessas e outras técnicas, é possível calcular um valor de uma dada região em estudo, para cada centro da célula de uma malha tridimensional, valor este condicionado aos dados amostrados e uma função de correlação espacial entre estes dados. Neste trabalho foi utilizado o conceito de krigagem na construção de um mapa da região de Unaí-MG, Brasil, para analisar a variabilidade espacial do cálcio (Ca) no solo.

2 Material e métodos

A área em estudo é um quadrado de 80 m×80 m na Faculdade Juvêncio Ferreira Martins Agrícola de Unaí (16°32'26"S, 46°50'44"W, altitude de 600 m). As amostras de solo foram recolhidas, em 2010, por meio de uma grade de 5 m×5 m, a uma profundidade de 0-20 cm, resultando em um total de 240 amostras. Das variáveis coletadas neste trabalho, aqui apresenta-se o estudo espacial da variável Ca.
Para a aplicação da análise geoestatística, foram calculados o variograma e a razão de dependência espacial (RD), mais informações em [1]. Foram ajustados dois modelos, um considerando a função cúbica e outro a função Matérn, que posteriormente foram comparados por meio dos seguintes critérios: RD, AIC e BIC. Os modelos supracitados são dados, respectivamente por (1) e (2):

$$\gamma(\mathbf{h}) = \begin{cases} 1 - \left(7\left(\frac{h}{\phi}\right)^2 - 8,75\left(\frac{h}{\phi}\right)^3 + 3.5\left(\frac{h}{\phi}\right)^5 - 0,75\left(\frac{h}{\phi}\right)^7\right), \text{ quando} \quad h < \phi; \\ 0, \text{ caso contrário} \end{cases} \tag{1}$$

$$\gamma(\mathbf{h}) = \left(\frac{1}{2^{(k-1)}\Gamma(k)}\right)\left(\frac{h}{\phi}\right)^k K_k\left(\frac{h}{\phi}\right) \tag{2}$$

no qual $\Gamma(.)$ é a função gama, K_k é a função Bessel de ordem k, $||\mathbf{h}||$ é a distância euclidiana entre duas localizações e ϕ é o parâmetro de dependência espacial.

3 Resultados e discussões

Após o teste de Shapiro-Wilk, ao nível de 5% de significância, foi necessária a transformação da variável Ca para sua raiz quadrada, com o intuito de atender a pressuposição de normalidade da variável. Após a transformação, uma análise exploratória dos dados foi realizada por meio do pacote geoR [3], disponível no programa R [2], por meio da Figura 1 (a) e (b). O gráfico superior a

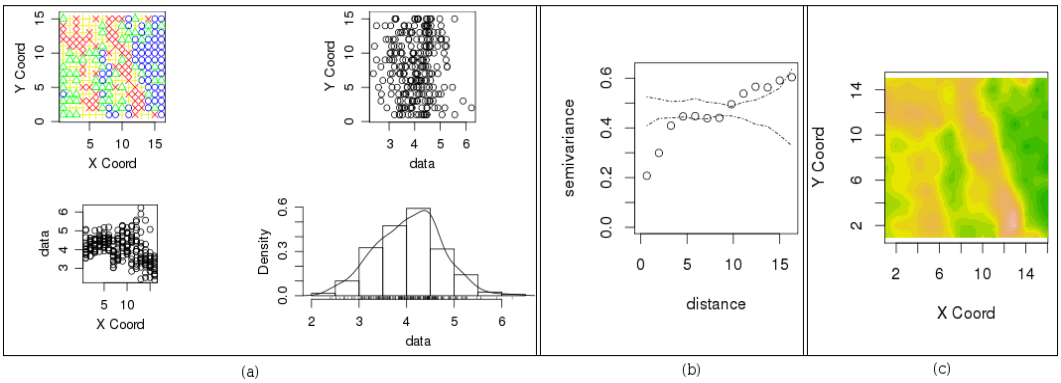

Figura 1: (a) Gráficos da variável Cálcio para análise exploratória; (b) Envelope do variograma da variável Cálcio; (c) Krigagem da variável Cálcio

esquerda, apresentado na Figura 1 (a), categoriza os dados nos quartis amostrais das observações, em que os símbolos " +", "Δ", " o", "x", nesta ordem, indicam os quartis amostrais. Essa imagem confirma a idéia da existência de padrão espacial nos dados, uma vez que existem conglomerados das categorias. O mesmo é constatado pela observação da Figura 1 (b): como existem pontos fora do envelope, pode-se afirmar que existe esta dependência. Ainda na Figura 1 (a), pode-se observar o gráfico da coordenada Y pelos dados (canto superior direito); no canto inferior esquerdo, o gráfico dos dados pela coordenada X; e, por último, no canto inferior direito, o histograma da variável transformada.
Com o auxílio das funções cúbica e Matérn, estimou-se os parâmetros da variância (σ^2), do efeito pepita (τ^2), média (β_0) e da dependência espacial (ϕ), dados na Tabela 1, bem como os avaliadores de qualidade de ajuste RD, AIC e BIC.

Tabela 1: Estimativas dos parâmetros dos modelos propostos e avaliadores de qualidade de ajuste

Modelo	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\phi}$	$\hat{\beta}_0$	RD	AIC	BIC
Cúbica	0,28	0,18	6,18	4,01	38,93	374,41	388,33
Matérn	0,39	0,10	2,29	3,96	20,06	374,74	388,66

Por meio da Tabela 1, observa-se que os valores de AIC e BIC para ambos os modelos são muito próximos, porém, o critério RD retorna que a função Matérn se sobressai, uma vez que apresenta dependência espacial forte (a função cúbica apresenta apenas dependência moderada) [1]. Assim, o modelo escolhido foi o que utiliza a função Matérn e com ele realizou-se a krigagem na área de estudo (Figura 1 (c)), onde pode-se observar que no canto direito da área está concentrado maiores valores de cálcio e no canto esquerdo valores mais baixos (vermelho) desta variável.

Referências

[1] Diggle, P.J., Ribeiro Jr., P.J. (2007). *Model-Based Geostatistics*. Springer, New York.

[2] R Core Team. (2013). *R: A language and enviroment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Disponível em: http://www.R-project.org.

[3] Ribeiro Jr., P.J., Diggle, P.J. (2011). geoR: A package for geostatistical analysis. *R-NEWS* 1 (2), 14–18.

Poster

Bivariate kernel smoothers. Applications in thoracic aorta pathology.

Javier Roca-Pardiñas
Department of Statistics and Operations Research, University of Vigo, Spain, *roca@uvigo.es*
Francisco de Asis Lopez Alvarez
Department of Statistics and Operations Research, University of Vigo, Spain, *franasisloal@gmail.com*
Pablo G. Tahoces
Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, Spain, *pablo.tahoces@usc.es*
Juan A. Martinez-Mera
Centro de Investigación en Tecnologías de la Información (CITIUS), University of Santiago de Compostela, Spain, *juanantonio.martinez@usc.es*
Jose M. Carreira
University Hospital Complex of Santiago de Compostela (CHUS), Spain, *josemartin.carreira@usc.es*

Keywords: Biometrics, Nonparametric regression, Radiology, Medical imaging.

Summary: Aneurysms are the most common pathology that affects the thoracic aorta. They often require surgery due to high risk of rupture, with the consequent death of the patient. Surgical intervention is indicated when maximum diameters exceed 6 cm in the ascending aorta or 7 cm in the descending aorta (Elefteriades, 2002). Hence, the precise description of the thoracic aorta is crucial for the diagnosis of this type of lesion.
The aim of the study is to develop robust method for measuring the calibre of the thoracic aorta to detect the presence of abnormalities. For this purpose, a database of 2206 CT (Computed tomography) images of 5 patients from the Department of Radiology of the University Hospital of Santiago de Compostela, were employed.
The method involves several steps such as the automatic segmentation of the aorta from CT slices, the calculation of the centre line of the vessel to determine the normal planes of the structure, the smoothness of the data volume to improve the accuracy of the method and the calculus of the diameters.
In order to obtain an accurate reconstruction of the whole volume of the aorta, we propose an adaptation of the local lineal kernel bivariarte smoothers (Ruppert and Wand, 1994). Such non-parametric regression techniques allow for a more flexible fit of real data than do the parametric regression techniques usually used. Bootstrap methods (Efron and Tibshirani, 1993) were used to draw inferences from the diameter of the aorta sections. The use of the bootstrap implies a high computational cost, since as it is necessary to repeat the estimation operations several times. Consequently, recourse to some computational acceleration technique is fundamental to ensure that the problem can be addressed adequately in practical situations. Here we have used binning techniques (Fan and Marron, 1994) to speed up the process.

Acknowledgments

The authors gratefully acknowledge the financial support from European Regional Development Fund (ERDF-/FEDER) under the Project CN2012/151, the project MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included) and Xunta de Galicia (10 PXIB 300 068 PR).

References

[1] Elefteriades, J.A. (2002). Natural history of thoracic aortic aneurysms: indicators for surgery, and surgical versus nonsurgical risks. *The Annals of Thoracic Surgery* 74, 1877–1880.

[2] Efron, E., Tibshirani, R.J. (1993). *An introduction to the Bootstrap*. Chapman and Hall, London.

[3] Fan, J., Marron, J. (1994). Fast implementation of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* 3, 35–56.

[4] Ruppert, D., Wand, M.P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics* 22(3), 1346–1370.

Poster

Ajuste de modelos platô de resposta em dados de crescimento de bovinos da raça nelore do estado de Minas Gerais

Tânia Jussara Silva Santana
Universidade Federal de Lavras (UFLA) e Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), *tanjussara@gmail.com*

João Domingos Scalon
Departamento de Ciências Exatas, Universidade Federal de Lavras (UFLA), *scalon@dex.ufla.br*

Isabel Cristina Costa Leite
Universidade Federal de Lavras (UFLA) e Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), Brasil, *isaleite@ifba.edu.br*

Azly Santos Amorim de Santana
Universidade Federal de Lavras (UFLA) e Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), Brasil, *azlly@hotmail.com*

Ângela Cristina da Fonseca Mirante
Universidade Federal de Lavras (UFLA) e Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA), Brasil, *angelamirante@ifba.edu.br*

Palavras–chave: Modelo não linear, Gado de corte, Métodos iterativos de estimação.

Resumo: Um modelo platô de resposta (Response Plateau Model) é caracterizado por duas fases. A primeira descreve o processo biológico através de uma reta, parábola ou um modelo exponencial antes do platô. Na segunda fase o modelo assume um valor constante igual a P, isto é, a partir deste ponto estabiliza-se. Este trabalho teve como objetivo estudar aspectos estatísticos envolvidos no ajuste de modelos platô de resposta em dados de crescimento de bovinos nelores criados a pasto em rebanhos de Minas Gerais. A variável analisada para o ajuste dos modelos foi o "peso médio dos animais em Kg". Os dados foram fornecidos pela ABCZ (Associação Brasileira de Criadores de Zebus). A análise foi realizada com o software R (R Development Core Team) 2013, utilizando a função nls do pacote stats, a partir dos valores da variável resposta nas diferentes idades dos bovinos, do peso ao nascer (Pn) até os 600 dias de idade (P600). A qualidade dos ajustes dos modelos polinomial quadrático (MPQ) e não linear exponencial (MNLE), ambos com platô, foi avaliada considerando o coeficiente de determinação (R^2), o quadrado médio do erro (QME) e o critério de informação de Akaike (AIC). Ambos os modelos mostraram-se adequados para estudar dados de crescimento de bovinos da raça nelore do estado de Minas Gerais. O MPQ e MNLE são recomendados para tal estudo por possuírem uma característica de utilização prática, uma vez que indica a idade ótima estimada do animal no ponto inicial do platô.

Agradecimentos

Agradecimentos à Associação Brasileira dos Criadores de Zebu (ABCZ), Programa de Apoio a Núcleos de Excelência (PRONEX), Fundo Multilateral de Investimento do Banco Interamericano de Desenvolvimento (Fumin/BID), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Poster

O papel do peso das caudas na utilização de testes quantitativos compostos

Rui Santos
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon, rui.santos@ipleiria.pt

Miguel Felgueiras
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon,
CIIC — Computer Science and Communications Research Centre of Polytechnic Institute of Leiria, mfelg@ipleiria.pt

João Paulo Martins
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon, jpmartins@ipleiria.pt

Palavras–chave: Testes compostos, Média, Extremos, Peso das caudas.

Resumo: A aplicação de testes conjuntos em análises clínicas, ou em amostragem de aceitação, permite poupar muitos recursos. Todavia, o recurso a este tipo de testes deve ser efetuado com precaução, de forma a evitar aumentar significativamente a probabilidade de má classificação. Este trabalho utiliza o peso das caudas da distribuição que carateriza a substância em análise como um indicador da adequação da aplicação de testes compostos.

1 Introdução

Em análises clínicas, bem como na amostragem de aceitação no controlo de qualidade, efetuam-se usualmente testes quantitativos individuais com o objetivo de identificar se a quantidade da substância em análise, no indivíduo testado, é superior (ou inferior) a determinado limiar ξ predefinido, que separa os testes negativos (indivíduos classificados como saudáveis) dos testes positivos (indivíduos classificados como infetados ou defeituosos). Nestes casos podem ser aplicados testes conjuntos, misturando-se o sangue proveniente de n indivíduos distintos e, desta mistura, retirando uma amostra para análise. Assim sendo, considerando que a quantidade em análise de cada indivíduo é descrita por uma distribuição $\mathbf{D}$ conhecida, no caso discreto podemos caraterizar a quantidade em análise na amostra combinada recorrendo a modelos hierárquicos e, no caso contínuo, utilizando a média como um valor aproximado. Caraterizada a distribuição desta quantidade, o objetivo dos testes conjuntos é identificar se o seu máximo (mínimo) é superior (inferior) ao limiar ξ, de forma a identificar se no grupo há algum indivíduo infetado.

A aplicação de testes conjuntos pode permitir poupar muitos recursos, possibilitando a diminuição do número de testes necessários a serem efetuados. Todavia, a sua utilização pode também aumentar significativamente a probabilidade de erros de classificação, que podem ser mensurados recorrendo aos usuais conceitos de sensibilidade (probabilidade de um teste positivo num indivíduo infetado) e especificidade (probabilidade de um teste negativo num indivíduo saudável). Por conseguinte, é vital identificar as caraterísticas que permitam distinguir os casos para os quais a aplicação de testes conjuntos é adequada, sem correr o risco de existência de má classificação com probabilidade demasiado elevada.

Santos, Pestana e Martins [3] determinaram a sensibilidade e a especificidade em testes compostos, bem como na aplicação da metodologia de Dorfman [1], em função da sensibilidade e da especificidade do teste individual. Investigaram, desta forma, a influência da rarefação (diluição da substância em análise quando diferentes amostras são misturadas) nos erros de classificação dos testes compostos. Martins, Santos e Sousa [2] propuseram duas metodologias para a realização de testes conjuntos para as quais determinaram, via simulação, a sensibilidade e a especificidade. Santos, Felgueiras e Martins [4], considerando observações independentes e identicamente distribuídas provenientes de diversas distribuições, determinaram a correlação entre a média e o máximo de forma a inferirem (via simulação) acerca da sensibilidade e da especificidade obtida em testes conjuntos. Deste modo concluíram, de uma forma geral, que para taxas de prevalência baixas, dimensões do grupo pequenas e distribuições **D** com a cauda direita (esquerda) pesada no caso de máximos (mínimos) os testes conjuntos podem ser aplicados com uma baixa probabilidade de erro de classificação. Neste trabalho iremos incluir o peso das caudas da distribuição **D** nesta análise, de forma a incluir um índice que nos permita identificar as distribuições para as quais os testes conjuntos podem ser aplicados sem aumento significativo da probabilidade de má classificação.

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst-OE/MAT/UI0006/2011 e pelo Instituto Politécnico de Leiria.

Referências

[1] Dorfman, R. (1943). The detection of defective members in large populations, *The Annals of Mathematical Statistics* **14**, 436–440.

[2] Martins, J.P., Santos, R., Sousa, R. (2013). Testing the maximum by the mean in quantitative group tests. *Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies* (aceite para publicação).

[3] Santos, R., Pestana, D., Martins, J.P. (2013). Extensions of Dorfman's Theory. P.E. Oliveira *et al.* (eds.): *Recent Developments in Modeling and Applications in Statistics, Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies*, Springer, 179–189

[4] Santos, R., Felgueiras, M., Martins, J.P. (2013). Known mean, unknown maxima? Testing the maximum knowing only the mean, *Communications in Statistics – Simulation and Computation, Special Issue – Joint Meeting of y-BIS and jSPE* (aceite para publicação).

Poster

ANOVA não-paramétrica de dados de contagem com medidas repetidas e com excesso de zeros

Soane Mota dos Santos
Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil, soanems@ime.usp.br
Julio da Motta Singer
Instituto de Matemática e Estatística, Universidade de São Paulo, Brasil, jmsinger@ime.usp.br

Palavras–chave: Distribuição binomial negativa, Distribuição Poisson, Distribuições inflacionadas de zeros, Modelos mistos.

Resumo: A análise de dados com medidas repetidas possui diversos desafios, especialmente quando os dados são contagens e têm excesso de zeros. Embora procedimentos paramétricos sejam bastante utilizados com essa finalidade, na prática desconhecemos a verdadeira distribuição dos dados e o uso de técnicas não-paramétricas propostas por Brunner, Munzel e Puri (1999), constituem uma alternativa. Neste trabalho, comparamos a ANOVA não-paramétrica para medidas repetidas com análises paramétricas baseadas em modelos mistos. Esses modelos requerem a especificação da estrutura de covariância intraunidades amostrais e a inclusão de um parâmetro de sobredispersão. A ANOVA não-paramétrica, por outro lado, utiliza os postos das observações para comparar os efeitos dos fatores na distribuição da variável resposta, podendo ser aplicada a dados contínuos, discretos, categóricos e mesmo dicotômicos. Por intermédio de estudo de simulação, examinamos as taxas estimadas de erro Tipo-I e do poder em diferentes cenários. Simulamos dados com as distribuições Poisson, Binomial Negativa e Poisson inflacionada de zeros (ZIP) com proporção de zeros de 10%, 30%, 50% e 70%. Em cada caso consideramos três padrões de médias, nomeadamante ($\mu_1 = \mu_2 = \mu_3 = 10$), ($\mu_1 = 10, \mu_2 = 12, \mu_3 = 14$) e ($\mu_1 = \mu_2 = 10, \mu_3 = 25$) e correlação intraunidades amostrais igual a 0,60. Geramos 400 amostras de tamanhos n = 5, 10, 20, 50 e 100 em cada cenário, todas sob planejamentos balanceados. A probabilidade de erro Tipo-I estimada sob a ANOVA não-paramétrica é ligeiramente maior que a probabilidade do erro Tipo-I obtida quando o modelo utilizado na análise é aquele usado na geração dos dados, mas tende a se igualar ao valor nominal quando o tamanho da amostra aumenta. Além disso, apresenta probabilidade de erro Tipo-I estimada menor quando comparada com aquela obtida sob análise baseada em modelos mistos com distribuições diferentes das utilizadas para gerar as amostras. O poder estimado da ANOVA não-paramétrica tende a ser menor que o poder obtido quando o modelo utilizado na análise é aquele usado na geração dos dados, principalmente quando a proporção dos zeros é grande (maior do que 40%). A ANOVA não-paramétrica é uma alternativa conveniente para dados de contagem com excesso de zeros e com medidas repetidas, mas deve ser utilizada com cautela quando a proporção de zeros é muito alta. Uma função programada em R para sua implementação computacional pode ser encontrada em http://www.ime.usp.br/∼jmsinger/anova_npar.zip.

Referências

[1] Brunner, E., Domhof, S., Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York: Wiley.

[2] Brunner, E., Munzel, U., Puri, M.L. (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis* 70, 286–317.

[3] Singer, J.M., Poleto, F.Z., Rosa, P. (2004). Parametric and nonparametric analyses of repeated ordinal categorical data. *Biometrical Journal* 46, 460–473.

Poster

Validacão das medidas de curvatura, de um modelo não linear para crescimento em altura de eucalyptus, através da reparametrização

Romulo Barbosa Veloso
Universidade Estadual de Montes Claros, *romulo.veloso@unimontes.br*
Natalino Calegário
Universidade Federal de Lavras, *calegari@dcf.ufla.br*

Palavras–chave: Modelo não linear, Validação, Florestas.

Resumo: A estimativa de crescimento e de produtividade são imprescindíveis para tomadas de decisão no meio florestal. O uso de modelos não lineares para a estimativa dessas variáveis é extremamente indicado pois nestes os parâmetros em geral tem significado biológicos. Um método eficiente de validação para modelos não lineares é o da aproximação linear das ditas medidas de curvatura que intrínseca e a devido ao efeito dos parâmetros, as quais são consideradas inválidas quando assumem valores superiores a 0,3 e nestes casos podemos através de reparametrização diminuir o efeito paramétrico, tornando o novo modelo utilizável. Este trabalho apresenta a reparametrização de um modelo de altura para eucalipto.

1 Introdução

A teoria assintótica mostra que o estimador de mínimos quadrados torna-se cada vez menos viesado, cada vez mais normalmente distribuído e aproxima-se da variância mínima, à medida que o tamanho da amostra tende para infinito. No entanto, não existem regras para saber se o tamanho da amostra é grande o suficiente para que sejam válidas as propriedades assintóticas. Bates e Watts (1980) apresentaram as medidas de não linearidade intrínseca e não linearidade devido ao efeito dos parâmetros que permitem avaliar se o modelo para a amostra em uso atende as condições necessárias para aplicação dos mínimos quadrados. Bates e Watts (1980) observaram que a curvatura existente do modelo em relação a um conjunto de dados pode ser separada em duas componentes chamadas de curvatura intrínseca e curvatura devido à parametrização. Ratkowsky (1983) apresenta um estudo detalhado ratificando as principais conclusões da metodologia proposta por Bates e Watts(1980) para esta averiguação. Empiricamente o valor de 0,3 é assumido como determinador para as medidas de curvatura aceitáveis. Ratkowsky (1983) afirma ainda que a medida de curvatura intrínseca é impossível de ser modificada, contudo a medida de curvatura devido ao efeito dos parâmetros pode ser afetada por uma reparametrização. Schabenberger e Pierce (2001) discutem esta proposta de reparametrização que é obtida isolando o parâmetro responsável pelo excesso da medida para valor médio e valor da variável conhecidos, e assim é recolocado no modelo. Os dados utilizados neste trabalho são provenientes do inventário de Eucalyptus sp. com idades entre 2 e 12 anos, na cidade de Luminárias, estado de Minas Gerais, latitude de 21,274962ºS e longitude de 44,584422ºW, entre 880 e 1.001 m de elevação. As temperaturas médias anuais variam entre 14ºC e 26ºC, precipitação média anual de 1.385 mm e clima tipo Cwa (Veiga, 1975). Ajustou-se o modelo não linear

$$hd = \theta_1 + \theta_3^{(-\text{idade})}\, e^{(\theta_2 + \theta_3^{(-\text{idade})})} (1)$$

onde hd é altura total e θ_i, $i = 1, 2, 3$, são os parâmetros e idade é a idade do eucalipto dada em meses. A tabela 1 apresenta os valores de parâmetro obtidos e suas respectivas medidas de curvatura além do viés de cada parâmetro que foi obtido por simulação Monte Carlo acusando

Parâmetros	**Valor do parâmetro**	**Medidas de curvatura**		Viés (%)
		Intrínseca (IN)	Devido ao efeito dos parâmetros (EP)	
θ_1	18,506679			O,193822
θ_2	2,715303	0,023582	1,329608	1,839043
θ_1	1,039186			0,223421

Tabela 1: Parâmetros e medidas de curvatura

o parâmetro θ_2 como o causador do excesso na medida de não linearidade devido ao efeito do parâmetro. O valor aceitável de viés adotado foi de 1% .
O modelo foi reparametrizado fazendo-se

$$\theta_2 = \ln(\frac{\theta_1 - \mu^\star}{\theta_3^{-x^\star}}) \quad \text{para } x^\star = 60 \text{ meses e } \mu^\star = 18,506679$$

onde $x^\star$ é a variável idade e $\mu^\star$ é o valor médio para este valor de $x^\star$. O modelo foi reajustado e verificou-se as medidas de curvature assumindo os valores $IN = 0,02673$ e $EP = 0,199873$ com $\theta_1 = 19,211929$ e $\theta_3 = 1,074533$.

Referências

[1] Bates, D.M., Watts, D.G. (1980). Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society B* 42, 1–25.

[2] Ratkoswisky, D.A. (1983). *Nonlinear Regression Modeling, a Unified Practical Approach.* New York: M. Decker, 276p.

[3] Shabenberger, O., Pierce, F.J. (2002). *Comtemporary Statistical Models for the Plant and Soil Sciences.* Florida, CRC, 738p.

Poster

Avaliação de mecanismos de imputação múltipla na análise de sobrevivência com dados omissos

Ana Margarida Vinhas
Dep. Matemática e Aplicações, Universidade do Minho, *anamargaridavinhas@hotmail.com*
Luís Antunes
Instituto Português de Oncologia do Porto, Instituto de Saúde Pública da Universidade do Porto, *luis.antunes@ipoporto.min-saude.pt*
Inês Sousa
Dep. Matemática e Aplicações e CMAT, Universidade do Minho, *isousa@math.uminho.pt*

Palavras–chave: Dados omissos, Sobrevivência, Cancro.

Resumo: O objectivo principal deste trabalho é fazer um estudo de simulação para verificar qualidade de estimadores em modelo de sobrevivência, quando são utilizados métodos de imputação para dados faltantes nas covariáveis [1]. Este estudo de simulação é feito sob vários cenários de proporção de dados faltantes, proporção de dados censurados e mecanismos de dados faltantes.
Para o estudo de simulação foi utilizada uma base de dados fornecida pelo Registo Oncológico do Norte (RORENO), relativa a casos de tumores do colón.
Os parâmetros de sobrevivência utilizados para o estudo de simulação foram obtidos por ajustamento de um modelo Weibull aos dados para que depois fosse possível fazer a estimação destes parâmetros com base no modelo Weibull utilizado. Depois do ajustamento do modelo aos dados originais e da estimação dos parâmetros foi fixada a proporção de dados omissos pretendida nas covariáveis estudadas em cada um dos dois mecanismos de dados faltantes utilizados no estudo.
Foram utilizados os seguintes mecanismos de omissão [2]:

- MCAR (missing completely at random) que consiste em apagar aleatoriamente observações nas covariáveis em estudo. Neste mecanismo as proporções na omissão dos dados foram fixadas entre 10% e 90%.
- MAR (Missing at Random) que consiste em apagar observações nas covariáveis em estudo dependendo de informação observada. A probabilidade de omissão na covariáveis foi calculada segundo um modelo de regressão logístico. As proporções totais na omissão dos dados foram fixadas em 5%, 10%, 20%, 50% e 70%.

Com as proporções de dados omissos fixadas foram simulados os dados de sobrevivência (tempos de sobrevivência e respectivas censuras) também com base no modelo Weibull ajustado à base de dados original para depois apagar observações nas covariáveis em estudo utilizando os mecanismos de omissão enunciados acima: MCAR E MAR. Depois de fazer as imputações foram analisados os dados completos (base de dados sem imputações) e os dados completados (dados com imputações) e ajustado um modelo de Cox a cada um destes conjuntos de dados. No final foi calculada a média do viés e da variância das estimativas para fazer a representação gráfica destas medidas.
Os resultados provisórios sugerem que para o mecanismo MCAR a média da variância é mais baixa nos dados completados e que a média do viés das estimativas está perto de zero. No mecanismo MAR a média da variância também é mais baixa nos dados completados apesar de aumentar com a proporção de dados omissos, a média do viés das estimativas aumenta também com a proporção de dados omissos.

Referências

[1] Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

[2] Little, R., Rubin, D. (2002). *Statistical Analysis with Missing Data*, Second Edition, Wiley-Interscience.

Plenary Session

Good confidence intervals for categorical data analyses

Alan Agresti
Department of Statistics, University of Florida, U.S.A., aa@stat.ufl.edu

Keywords: Score test-based confidence interval, Pseudo-score confidence inference.

Summary: This talk surveys confidence intervals that perform well for estimating parameters used in categorical data analysis. Considerable research has now shown that intervals resulting from inverting score tests perform much better than inverting Wald tests and often better than inverting likelihood-ratio tests. For some models, ordinary score-test-based inferences are impractical, such as when the likelihood function is not an explicit function of the model parameters. For such cases, we propose pseudo-score inference based on a Pearson-type chi-squared statistic. For small samples, "exact" methods are conservative inferentially, but inverting a score test using the mid-P value provides a sensible compromise. Finally, we briefly review a different pseudo-score approach that approximates the score interval for proportions and their differences with independent or dependent samples by adding pseudo data before forming simple Wald confidence intervals.

Oral Communication

Survival and longitudinal analysis of breast cancer at Braga's hospital

Inês Sousa
Departamento de Matemática e Aplicações e CMAT, Universidade do Minho, *isousa@math.uminho.pt*

Ana Borges
Centro para Inovação e Investigação em Ciências Empresariais e Sistemas de Informação, Instituto Politécnico do Porto, *aib@estgf.ipp.pt*

Luís Castro
Unidade de Senologia, Hospital de Braga, *luis.castro@hospitaldebraga.pt*

Keywords: Breast cancer, Cox model, Flexible parametric survival model, survival model.

1 Introduction

The present work regards to breast cancer patients of the Senology Unit of Braga's Hospital. The two main aims of this work are: to describe the survival rate of breast cancer patients; and to describe the longitudinal progression of oncological markers in follow-up of the same patients. The data are being collected directly from the medical records of each patient, listed in the computer system of Braga's Hospital.

We start with a preliminary data analysis of 451 female patients, diagnosed with a malignant tumour in the corresponding period of 1993 until 2012, whose age at the time of diagnosis varies between 20 to 89 years. The reference date of March 1, 2013 was chosen as the end of study.

Subsequently we performed a survival analysis in order to describe the survival rate of these patients, as a function of possible risk factors. For that we fit a Flexible Parametric Survival Model to estimate hazard ratios over time since diagnosis by a set of statistical significant covariates. Recurrence, Type of surgical treatment, Neoadjuvant treatment and Triple negative breast cancer have a significant statistical effect on the global survival rate of the patients of this Hospital. Results were also compared to those obtained when adjusting to the well known Cox proportional hazard model, and were quite similar for both models. Moreover, a longitudinal model with random effects at individual level was fitted to oncological markers that are measured at after diagnosis appointments.

2 The data

From the information gathered in the medical reports we were able to collect 46 variables that are grouped into two categories: (i) the explanatory variables at individual level, that are a group of demographic characteristics that include a set of prognostic factors; (ii) and a second group of variables, explanatory variables at tumour level, that include characteristics of the tumour, some of them important prognostic factors already reported in the literature ([4],[5]).

The 451 female patients diagnosed with a malignant tumor, whose age at the time of diagnosis varies between 20 and 89 years, corresponds to a total number of 467 cases analysed, since 34 bilateral breast cancer cases were treated as independent. The total number of deaths from breast

cancer is 47. The response variable is time from date of diagnosis to death from breast carcinoma, or right censored date if lost to follow-up, or last observed date of March 1, 2013.

3 Statistical models

We fit the survival data with a flexible parametric survival model (FRPM), proposed by Royston and Parmar (2002), to estimate hazard ratios over time since diagnosis by a set of statistical significant covariates. For comparison, we have also adjusted the data with the well known Cox proportional hazards model [1] (CPHM). The CPHM defines the probability of survival, under the assumption of proportional hazards, as a function of time t, for a vector of covariates x_i, where the hazard function is $h_i(t|x_i) = h_0(t)\exp(x_i\beta)$. The FRPM comes from the family of functions that are based on transformation of the survival function by a link function $g(\cdot)$, $g[S(t|x_i)] = g[S_0(t)] + x_i\beta$. Since we are interested in estimating hazard ratios, and as Royston and Parmar [6] suggest, we use natural cubic splines to model $g[S_0(t)]$, $g(x;\theta) = \log\left(\frac{x^{-\theta}-1}{\theta}\right)$. Thus, the model transformation may be written as it follows: $g[S(t|x_i)] = \log[H(t|x_i)] = \eta_i = s(\log(t)|\gamma, k) + x_i\beta$. Since both CPHM and FRPM rely on the assumption of proportional hazards we preformed a statistical test proposed by Grambsch and Therneau [3] based on the Schoenfeld residual calculation.

For a longitudinal analysis of oncologial markers we have performed classical models as defined in Diggle et al [2], with random effects and a Gaussian process.

4 Results

Although many of 46 variables obtained are recognized as potential prognostic factors only four are shown to have a statistically significant effect on survival of the 451 patients of the Unity of Senology of Braga's Hospital, namely: Recurrence (with vs without); Neoadjuvant Treatement (with vs without); Type of surgery (withou surgery vs conservative surgery vs mastectomy) and Triple Negative (yes vs no). The resulting adjusted FRPM, with 0 knots, reveals that the patients with recurrence, submitted to neoadjuvant treatment, submitted to no surgery and with triple negative breast cancer, are the ones with higher risk of dying from breast cancer.

Acknowledgments

The first author acknowledges partial financial support from the project PTDC/MAT/112338/2009 (FEDER support included) of the Portuguese Ministry of Science, Technology and Higher Education.

References

[1] Cox, D.R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society*, series B, 34, 87–220.

[2] Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S. (2002). *Analysis of Longitudinal Data.* Oxford University Press, London, 2nd edition.

[3] Grambsch, P., Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–26.

[4] MacMahon, B. (2006). Epidemiology and the causes of breast cancer. *Int. Jour. Cancer* 118, 2373–2378.

[5] Pinheiro, P.S., Tyczynski, J.E., Bray, F., Amado, J., Matos, E., Parkin, D.M. (2003). Cancer incidence and mortality in Portugal. *European Journal of Cancer* 39 (17), 2507–20.

[6] Royston, P., Parmar, M.K.B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21, 2175–2197.

Comunicação Oral

Análise conjunta bayesiana de dados longitudinais e de sobrevivência com fracção de cura espacial e sua aplicação ao estudo do VIH

Rui Martins
Centro de Investigação Interdisciplinar Egas Moniz da Escola Superior de Saúde Egas Moniz e Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), ruimartins@ymail.com

Giovani Loiola da Silva
Departamento de Matemática do Instituto Superior Técnico da Universidade Técnica de Lisboa e Centro de Estatística e Aplicações da Universidade de Lisboa, gsilva@math.ist.utl.pt

Valeska Andreozzi
Centro de Estatística e Aplicações da Universidade de Lisboa, valeska.andreozzi@fc.ul.pt

Palavras–chave: Análise conjunta, Dados longitudinais e de sobrevivência, Facção de cura.

Resumo: A investigação relativa à análise conjunta de dados longitudinais e de sobrevivência tem evidenciado melhorias nos resultados face a uma análise em separado. A rápida evolução da medicina e das ciências da saúde tem permitido tornar crónicas algumas doenças que até aqui eram fatais, pelo que uma parte substancial dos pacientes não chegará a experimentar o evento de interesse. Modelos de sobrevivência capazes de acomodar estes indivíduos são um auxílio importante no prognóstico de doenças potencialmente terminais. Este trabalho propõe um quadro bayesiano que modela associações espaciais dos dados de sobrevivência por área usando uma classe geral de modelos de cura propostos por [2], em conjunto com uma modelação longitudinal da contagem de linfócitos T CD4$^+$.

1 Introdução

Com o advento da terapia HAART (Highly Active Antiretroviral Therapy) tem-se constatado, através de investigações recentes, que os indivíduos com VIH/SIDA têm uma esperança de vida aproximadamente igual à dos indivíduos não infectados [5]. Podemos então supor que, no limite, alguns dos pacientes infectados morrerão por diversas causas, mas não devido ao VIH/SIDA. Na literatura de análise de sobrevivência estes indivíduos são ditos *sobreviventes de longa duração*, *curados* ou *imunes*.
Estudos ligados à SIDA centram-se na análise de dados envolvendo o tempo de vida até ao evento de interesse, *e.g.* morte, (dados de sobrevivência) e medidas repetidas (dados longitudinais) em cada indivíduo. Nesta situação, uma análise separada dos dois tipos de dados não tem em conta a dependência entre as respectivas variáveis resposta. Para resolver esta questão, modelos baseados na verosimilhança conjunta das respostas longitudinal e de sobrevivência têm surgido na literatura ([1], [3],[4]). Propomos aqui um modelo bayesiano hierárquico para este efeito, obtendo as estimativas dos parâmetros de interesse através de métodos de Monte Carlo via Cadeias de Markov (MCMC).

2 Modelação conjunta

Suponhamos que temos N indivíduos e K regiões com n_k pacientes na k-ésima região, $k = 1, \ldots, K$. Seja Y_{ikj} a medida repetida obtida no j-ésimo tempo para o indivíduo i que mora na região k e T_{ik} o tempo até ao evento de interesse do indivíduo ik (tempo de sobrevivência), $j = 1, \ldots, n_{ik}$, $i = 1, \ldots, n_k$. Para analisar conjuntamente as respostas longitudinal e de sobrevivência usaremos um processo latente bivariado de média nula no instante t, denotado por $W_{ik}(t) = (W_{1ik}(t), W_{2ik}(t))$. As

medidas repetidas e o tempo de sobrevivência são considerados independentes condicionalmente ao processo de ligação $W_{ik}(t)$ e às covariáveis observadas $(\mathbf{x}_{1ik}, \mathbf{x}_{2ik})$. Portanto, o modelo longitudinal é descrito por uma estrutura linear de componentes observadas e não observadas, *i.e.*,

$$Y_{ikj} = \mu_{ik}(t_{ikj}) + W_{1ik}(t_{ikj}) + \epsilon_{ikj}, \tag{1}$$

onde $\mu_{ik}(t_{ikj}) = \mathbf{x}_{1ik}^T(t_{ikj})\boldsymbol{\beta}_1$ e $W_{1ik}(t_{ikj})$ representam, respectivamente, as covariáveis observadas ("efeitos fixos") e não observadas ("efeitos aleatórios") e ϵ_{ikj} os erros de medida, com $\epsilon_{ikj} \sim N(0, \sigma_\epsilon^2)$. No modelo de sobrevivência vamos incluir uma fracção de sobreviventes de longa duração e fragilidades espaciais. Para tal considere-se que cada indivíduo está sujeito aos efeitos de um factor latente, $M_{ik} \sim Bernoulli(\theta_{ik})$, onde θ é a probabilidade de o factor latente se activar e causar a morte ao paciente. Desta forma, a função de sobrevivência para a população é dada por

$$S_p(t) = 1 - \theta + \theta S(t) \tag{2}$$

onde $S(t)$ é a função de sobrevivência para os indivíduos não-curados e $1-\theta$ representa a fracção de cura da população. Esta fracção de cura pode ser suposta comum a toda a população ou variar de região para região. Vamos supor que a função de risco associada a $S(t)$ é uma distribuição Weibull, $\mathcal{W}(\rho, \eta(t))$, e introduzimos as covariáveis através do parâmetro de escala, $\eta_{ik}(t) = \exp\{\mathbf{x}_{ik}^\top \boldsymbol{\beta}\}$. Por forma a permitirmos que o risco de morte seja uma função do processo Gaussiano latente e de fragilidades espaciais vamos escrever

$$h_{ik}(t) = \rho t^{\rho-1} \exp\{\mathbf{x}_{2ik}^\top(t)\boldsymbol{\beta}_2 + W_{2ik}(t) + Q_k\}. \tag{3}$$

onde Q_k representa um efeito aleatório específico espacial. Espera-se que áreas vizinhas possuam fragilidades espaciais semelhantes (autocorrelação espacial), pelo que será usado um modelo CAR para lidar com tal situação. Do ponto de vista bayesiano isso significa introduzir uma distribuição *a priori* para lidar com a dependência espacial, a qual produz um efeito de "aproximar" a variabilidade espacial de uma região relativamente à dos seus vizinhos.

3 Conclusão

A modelação conjunta bayesiana apresenta melhorias significativas na previsão do tempo de sobrevivência mediano dos indivíduos quando comparado com os modelos separados. Doenças crónicas, como neste momento é o VIH/SIDA são um campo ideal de aplicação de modelos de sobrevivência com fracção de cura.

Agradecimentos

Este trabalho foi parcialmente financiado pelos projectos da Fundação para a Ciência e a Tecnologia (FCT) PTDC/MAT/118335/2010 e Pest-OE/MAT/UI0006/2011.

Referências

[1] Brown, E.R., Ibrahim, J.G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* 59, 686–693.

[2] Cooner, F., Banerjee, S., Carlin, B.P., Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association* 102, 560–572.

[3] Rizopoulos, D., Verbeke, G., Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* 66, 20–29.

[4] Tsiatis, A.A., Davidian, M.D. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* 14, 809–834.

[5] Van Sighem, A., Gras, L., Reiss, P., Brinkman, K., de Wolf, F. (2010). Life expectancy of recently diagnosed asymptomatic HIV-infected patients approaches that of uninfected individuals. *AIDS* 24, 1527–1535.

Comunicação Oral

Modelação conjunta de dados longitudinais e eventos competitivos em doentes renais crónicos

Laetitia Teixeira
Programa Doutoral em Matemática Aplicada–Instituto de Ciências Biomédicas Abel Salazar e Faculdade de Ciências, Universidade do Porto, *laetitiateixeir@gmail.com*

Anabela Rodrigues
Departamento de Nefrologia, Centro Hospitalar do Porto–Hospital Geral de Santo António e Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, *asr.cbs@sapo.pt*

Inê Sousa
Departamento de Matemática e Aplicações, Universidade do Minho, *isousa@math.uminho.pt*

Denisa Mendonça
Departamento de Estudo de Populações—Instituto de Ciências Biomédicas Abel Salazar e Instituto de Saúde Pública, Universidade do Porto, *dvmendon@icbas.up.pt*

Palavras–chave: Modelação conjunta, dados longitudinais, dados de sobrevivência, riscos competitivos

Resumo: A modelação conjunta de dados longitudinais e de dados de sobrevivência é uma técnica de modelação estatística recomendável quando na investigação são gerados dados longitudinais com medidas repetidas de uma variável resposta em diversos pontos no tempo e dados relativos a eventos onde o tempo de recorrência ou de término de determinados eventos são registados [1]. No presente trabalho, a modelação conjunta foi aplicada em dados de doentes renais crónicos a realizarem diálise peritoneal como método de substituição da função renal. Os dados deste estudo possuem a particularidade de serem constituídos por dados longitudinais, resultado das avaliações periódicas da condição renal do doente, e por dados de sobrevivência correspondente ao tempo até ao evento de interesse na presença de riscos competitivos.

1 Introdução

Em diversas áreas de investigação, como a saúde, economia ou demografia, alguns estudos têm como especificidade permitir a observação de dois tipos de outcome: (1) dados longitudinais e (2) dados de sobrevivência. Os dados longitudinais correspondem ao conjunto de medidas repetidas de variáveis respostas de interesse em diversos momentos de avaliação. Dados de sobrevivência correspondem ao tempo decorrido desde a entrada no programa em estudo até à ocorrência do evento de interesse, na presença de riscos competitivos.
Estes dois tipos de outcomes são muitas vezes analisados separadamente utilizando dois modelos distintos, um modelo para o outcome longitudinal (como por exemplo, um modelo linear misto) e um modelo para o outcome tempo até ao evento de interesse na presença de riscos competitivos (como por exemplo, modelo de Cox por causa-específica) [2]. No entanto, na presença destes dois outcomes, é aconselhável a utilização de um modelo conjunto para dados longitudinais e dados de sobrevivência quando o interesse é avaliar a inter-relação entre estes dois tipos de outcomes [3].

2 Aplicação: Diálise peritoneal

Um exemplo típico da relevância da abordagem de modelação conjunta, e a principal motivação deste estudo, é a avaliação de programas de diálise peritoneal. Ao longo do programa de diálise

peritoneal, os doentes renais crónicos são monitorizados periodicamente ao longo de todo o tempo de tratamento e dado que a recuperação da função renal é muito rara poderão experienciar apenas um dos possíveis eventos: morte, transferência para hemodiálise ou transplante renal. A ocorrência de um destes eventos impede a observação de outro evento, tratando-se por isso de um problema de eventos competitivos. Por exemplo, um doente que sai do tratamento de diálise peritoneal devido transplante renal não pode experienciar o evento morte em diálise peritoneal. Deste modo, na presença de riscos competitivos deverá ser utilizada metodologia que tem em conta a presença destes eventos.

A modelação conjunta tem como principais objectivos: (1) compreender os padrões de mudança intra-sujeito nos outcomes longitudinais e/ou (2) caracterizar a relação entre esses outcomes longitudinais e o tempo até ao evento de interesse (ou o tempo até outro risco competitivo) [4]. Neste estudo, pretende-se por exemplo avaliar conjuntamente o tempo até transplante renal (considerando a transferência para hemodiálise e a morte como riscos competitivos) e outros indicadores da condição da função renal (como por exemplo albumina, fósforo) como outcomes de medidas repetidas. Outras variáveis medidas na baseline, como por exemplo sexo, idade e diabetes, poderão também ser consideradas. Foi usado o software `R`, em geral, e o pacote `JM`, em particular, para estimar os parâmetros da modelação conjunta dos dados longitudinais e dos dados de sobrevivência na presença de riscos competitivos.

3 Conclusão

Os modelos conjuntos para dados longitudinais e dados de sobrevivência são particularmente relevantes em estudos clínicos onde indicadores longitudinais podem estar altamente associados com o tempo até ao evento de interesse. Estes modelos permitem analisar dados complexos e, por isso, a sua importância tem vindo a ser cada vez mais reconhecida.

Referências

[1] Henderson, R., Diggle, P., Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* 1(4), 465–480.

[2] Rizopoulos, D. (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data, *Journal of Statistical Software*, 35(9), 1 - 33.

[3] Williamson, P.R., Kolamunnage-Dona, R., Philipson, R., Marson, A.G. (2008). Joint modelling of longitudinal and competing risks data. *Statistics in Medicine*, 27(30), 6426 - 6438.

[4] Tsiatis, A.A., Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, 14(3), 809 - 834.

Oral Communication

Using independent information in distance sampling surveys to account for animal density gradients with respect to transects

Regina Bispo
ISPA, Instituto Universitário, Lisboa, Portugal; Bio3 - Estudos e Projectos em Biologia e Valorização de Recursos Naturais, Lda, Portugal; Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal, *rbispo@ispa.pt*

Tiago A. Marques
Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK; Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal, *tiago@mcs.st-and.ac.uk*

Stephen T. Buckland
Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK, *steve@mcs.st-and.ac.uk*

Brett Howland
Fenner School of Environment and Society, ANU College of Medicine, Biology & Environment, The Australian National University, Canberra, Australia, *brett.howland@anu.edu.au*

Keywords: Line transects, Density estimation, Assumption violation, GPS data, Kangaroos.

Summary: Distance sampling is one of the most widely used methods for estimating animal abundance, and perhaps one of the most abused. There are a number of assumptions which are often overlooked, and this leads to potential bias. Here we address the issues which arise from a non random location of samplers which might result in a pattern in animal distribution with respect to the transects. Examples include locating samplers along roads, rivers or shorelines. Based on Marques *et al.* (2013) we extend the conventional likelihood to include a model for the distribution of animals available to be detected, instead of assuming a distribution known by design. We consider an example from a kangaroo population for which true abundance is known. We estimate the animal distribution with respect to the roads using a sample of GPS collared animals, which allows to disentangle the otherwise joint information on detectability and availability information collected from the actual survey along tracks. We show that the actual density of animals is higher away from the tracks, which ignoring the non-random location of transects, leads to overestimation of detection probability and the corresponding underestimation of density. By including the GPS information and jointly estimating the density gradient and detection function the density estimates appear to be unbiased.

Acknowledgments

TAM and RB research is partially supported by Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the project PEst-OE/MAT/UI0006/2011. We would like to thank ACT Government Ecologists Dr Don Fletcher and Claire Wimpenny for providing GPS collar data, and Peter Hann for volunteering his time to conduct kangaroo surveys.

References

[1] Marques, T.A., Buckland, S.T., Bispo, R., Howland, B. (2013). Accounting for animal density gradients using independent information in distance sampling surveys. *Statistical Methods and Applications* 22, 67–80.

Comunicação Oral

Critério para descarte de animais em remodelação cardíaca

Renan Mercuri Pinto
Instituto de Biociências de Botucatu, Universidade Estadual Paulista, São Paulo, *renanmp@ibb.unesp.br*

Antônio Carlos Cicogna
Faculdade de Medicina de Botucatu, Universidade Estadual Paulista, São Paulo, *cicogna@fmb.unesp.br*

Carlos Roberto Padovani
Instituto de Biociências de Botucatu, Universidade Estadual Paulista, São Paulo, *bioestatística@ibb.unesp.br*

Palavras–chave: Componentes principais, Remodelação cardíaca, Homogeneização.

Resumo: O presente trabalho utilizou-se do conhecimento de estatística, em particular, a técnica de análise multivariada envolvendo a estrutura de variabilidade conjunta das variáveis biológicas para estabelecer um critério de descarte de animais. A construção do critério, objetiva melhorar a qualidade de homogeneização de animais e assegurar com máxima acurácia a inclusão de animais semelhantes e pequena frequência de descartes, motivando maximizar o tamanho do lote homogêneo para posterior submissão ao delineamento experimental por processo simples de casualização dos tratamentos.

1 Introdução

Normalmente, por desconhecimento estatístico, muitos pesquisadores optam por utilizar métodos empíricos ou subjetivos para a tomada de decisão. Um exemplo muito comum ocorre na homogeneização de amostras, que é fator decisivo para a randomização em pesquisas envolvendo animais como unidade experimental. Nessa situação, tem-se o hábito errôneo de fazê-la de maneira fragmentada ou intencional, ou seja, utilizar variáveis convenientes para classificar o lote como homogêneo. Fator que, além de resultar numa homogeneização viesada e inadequada, favorece a possibilidade de descartar animais por um simples valor espúrio do que por semelhança ou dessemelhança biológica. Um processo de homogeneização biológico deve considerar conjuntamente todas as variáveis mensuradas, pois estas são possivelmente correlacionadas e, a melhor forma de entender o comportamento animal está expressa no conjunto de todas as características do animal.
Neste enfoque de raciocínio, o Grupo de Pesquisa em Cardiologia Experimental da Faculdade de Medicina, Universidade Estadual Paulista (UNESP), Botucatu - São Paulo - Brasil, em seu expediente comum de produção de conhecimento científico, busca dentro da linha de Remodelação Cardíaca estabelecer critério objetivo e com alta acurácia para a homogeneização de animais induzidos à Estenose Aórtica (EAo).
Sob o aspecto da junção da estatística com as ciências biológicas e da saúde buscou-se um modelo estatístico multivariado de homogeneização de animais que considere descartar o menor número possível e preserve a máxima semelhança biológica entre os não descartados.

2 Material e métodos

O modelo estatístico desenvolvido para verificar a homogeneidade do lote de ratos (pesquisa aprovada pelo Comitê de Ética em Experimentação Animal da Faculdade de Medicina da UNESP, sob

número 850/2010), envolve simultaneamente 34 variáveis e leva em consideração toda a estrutura de variação existente nos dados, ou seja, a variação dentro das variáveis (intravariabilidade) e a variação entre as variáveis (intervariabilidade).
Como as variáveis possuíam diferentes unidades de medida, foram inicialmente padronizadas e, na sequência tiveram sua matriz de dispersão submetida à análise de componentes principais na busca dos eixos descritores para o sistema de homogeneização. Os descritores consistem no conjunto de eixos que determinam o sistema cartesiano para identificação dos "outliers" (animais cujos resultados experimentais estão com alta probabilidade de ocorrência de valores espúrios à população de origem). Para a determinação do número de eixos foram considerados os k primeiros componentes principais associados aos maiores autovalores de R que foram selecionados pelo diagrama de autovalores.
Estabelecido os eixos descritores, constroi-se a região de 95% de confiança a partir do centroide dos dados no sistema cartesiano gerado pelos eixos e, em seguida, todos os ratos do lote foram alocados no novo sistema. Animais que pertencessem ao interior da região são considerados homogêneos para o processo de casualização, no sentido contrário, descartados do experimento. No lote considerado para o estudo da estenose aórtica, 8 animais foram descartados do experimento.

3 Discussão final

Quando utiliza-se os procedimentos sem qualquer preocupação com a estrutura geral de variabilidade o número de descartes tem relevância em comparação com os 8 descartes realizados no critério apresentado no texto. Fato que corrobora com as expectativas apresentadas quando da origem das discussões envolvendo experimentos já realizados pelo grupo de Pesquisa em Remodelação Cardíaca.

Agradecimentos

Este trabalho é financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e desenvolvido no Programa de Pós-Graduação em Biometria, do Instituto de Biociências de Botucatu, com apoio do Grupo de Pesquisa em Cardiologia Experimental, da Faculdade de Medicina da UNESP do campus de Botucatu - São Paulo, Brasil.

Referências

[1] Cicogna, A.C., Okoshi, M.P., Okoshy, K. (2000). História natural da remodelação miocárdica: da agressão aos sintomas. *Revista da Sociedade de Cardiologia do Estado de São Paulo* 10, 8–16.

[2] Johnson, R.A., Wichern, D.W. (2002). *Applied Multivariate Analysis*. 5th ed. New Jersey: Prentice-Hall, 767p.

[3] Morrison, D.F. (2004). *Multivariate Statistical Methods*. 4th ed. New York: Duxbury Press, 498p.

[4] Sandaniello, V.L.M., Padovani, C.R. (2010). Construção de índice percentil de status de desenvolvimento sustentável de assentos rurais utilizando procedimento estatístico multivariado. *Energia na Agricultura, Botucatu* 25, 137–153.

[5] Silva, N.R., Padovani, C.R. (2006). Utilização de componentes principais em experimentação agronômica. *Energia na Agricultura, Botucatu* 21(4), 98–113.

Oral Communication

Simulation model of salmonella dynamics on a farrow-to-finish herd

Carla Correia-Gomes
ERU - Scotland's Rural College, *carla.gomes@sruc.ac.uk*

Theodoros Economou
CEMPS - University of Exeter, *t.economou@exeter.ac.uk*

Trevor Bailey
CEMPS - University of Exeter, *t.c.bailey@exeter.ac.uk*

João Niza-Ribeiro
ICBAS-UP & Instituto de Saúde Pública da Universidade do Porto, *nizaribeiro@gmail.com*

Keywords: Simulation model, Salmonella, Pigs.

Summary: A simulation model was built to simulate the Salmonella spp. infection dynamic on an average farrow-to-finish pig herd in Portugal. This stochastic model links production steps with infection parameters. The model allows to draw conclusions about which parameters influence more the outcomes and in the future to test the cost benefit of some control measures.

1 Introduction

Salmonella spp. is one of the major causes of food-borne outbreaks in the world (the second cause in Europe) [1]. Therefore Salmonella spp. control was considered necessary at European Union level. In practice, however, the control of this agent has proved to be difficult and expensive at the farm level [2]. Consequently the evaluation of the efficiency of control strategies for this agent has become an important and stringent issue, as stated in recent reports [3]. Modelling the dynamics of Salmonella spp. in pigs can become useful when assessing alternative control strategies. Susceptible - Infectious - Resistant (SIR) models are attractive tools to help in assessing the disease dynamics. The SIR model describes the dynamic of different states of individuals in the population. The variables in the system are given by the three compartments: group of susceptible (S), group of infectious (I) and group of resistant (R). The aim of this study was to develop a simulation model that would simulate the disease spread in an average farrow-to-finish Portuguese swine herd, and therefore could be used to test the efficiency of control measures.

2 Materials and methods

A simulation model that describes what happens on the herd in terms of production management was linked to an infection model that describes the dynamic of the disease spread on the herd. The model simulates a farrowing-to-finish herd in which batch farrowing is applied to sows, leading to batch management of pigs. The sow reproduction cycle is divided in three stages (mating, gestation and farrowing period) corresponding to the occupation of three different types of rooms. The pig growth is divided on three stages (sucking, post-weaning and fattening period) corresponding to the occupation of three different types of rooms. The modelling unit is the batch (for sows and pigs). In the model, batches of sows are groups of sows (the same number per batch) that are mated at the same time. One week interval between two successive batch mating was considered. Each batch of pigs was composed by the litters of the batch of sows. All animal leave the room they occupied simultaneously unless for the sows at gestation room when abortion occur. This production model describes the evolution of the number of animals within each batch. The time step is one week.

The model is mainly stochastic. The stochastic steps try to simulate the variability associated with biological phenomena of mortality, culling, insemination failure, abortion and litter size. It was used the binomial distribution to drawn the number of animals for each production process.
The infection model was based in a Susceptible-Infectious-Resistant model for Salmonella. It was considered the direct transmission between the pigs in the batch and also the indirect transmission via contaminated floor, rodents, etc. The binomial distribution was used to simulate the transition between susceptible and infectious state and from infectious to carrier state. For the transition between carrier state and infectious and carrier state and susceptible Poisson distributions were used. For the transition between susceptible and infectious a cohort time-dependent random effect was added to the transmission parameter. With this cohort time-dependent random effect we capture the dependent structure of the spreading of infection within cohorts where the velocity of infection is dependent of the number of susceptible and infectious in the previous time step.
The model was built and run on R free software (CRAN project, www.r-project.org/), this was run long enough (500,000 iterations) to ensure convergence on the final results. For starting the model we had to allocate an initial status for the sows/gilts at mating for the first batch and after that the model ran and filled the status for each sows of the following batches. The results that were saved in each simulation were the proportion of sows alive in each room, the proportion of sows pregnant at the end of mating and gestation room, and the proportion of sows/pigs in the infection states at the end of each room. A sensitivity analysis of the model was performed, where all the production parameters and infection parameters were increased and decreased by 50% and the result were compared with the unchanged parameter.

3 Results and discussion

The infection status of the pigs at fattening is highly dependent of what happens in the maternity (the piglets' protective factor and the transmission rate parameter from S to I). The infection status of the sows at farrowing is highly dependent of the initial status of the sows and the transmission parameters (from S to I, from I to R and from R to S). This model allows to draw conclusions about which parameters influence more the outcomes and in the future to test the cost benefit of some control measures.

References

[1] EFSA, ECDC (2012). The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2010, *EFSA Journal* 10, 2597.

[2] Hurd, H.S., Enoe, C., Sorensen, L., Wachman, H., Corns, S.M., Bryden, K.M., Grenier, M. (2008). Risk-based analysis of the Danish pork Salmonella program: past and future, *Risk Analysis* 28, 341–351.

Sessão Plenária

Epidemias de gripe: três problemas, três abordagens

M. Lucília Carvalho
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), mlcarvalho@fc.ul.pt

Palavras–chave: Gripe, Nowcasting, Cadeias de Markov escondidas não-homogéneas, Modelos de espaço de estados.

Resumo: A gripe é uma infeção respiratória aguda que em Portugal, e mais geralmente no hemisfério norte, é responsável por epidemias que se desenvolvem durante o outono e o inverno, com um impacto muito importante na saúde humana que se traduz muitas vezes por excessos de mortalidade e um considerável aumento da necessidade de cuidados de saúde, nomeadamente hospitalizações frequentemente em unidades de cuidados intensivos.
Além da perda de vidas humanas e dos custos financeiros dos serviços de prestação de cuidados de saúde, que podem ser, por si só, muito elevados, existem outros encargos sociais diretos e indiretos provocados pelo absentismo ao trabalho e às atividades escolares que, nos casos muito graves, podem mesmo conduzir a ruturas no fornecimento de bens e serviços.
Outro aspeto que convém realçar, motivo de grande preocupação das autoridades de saúde mundiais, é o caráter global que as epidemias de gripe podem rapidamente alcançar devido à atual celeridade de contacto entre populações de diferentes países e continentes e à facilidade de propagação da doença em grandes concentrações urbanas.
Neste contexto, tornou-se indispensável implementar Sistemas de Vigilância epidemiológica da Gripe (SVG), a nível nacional e internacional, com o objetivo de fornecer às autoridades de saúde informação para avaliações de risco atualizadas que permitam uma correta aplicação de medidas de controlo e mitigação das epidemias e suas consequências.
Apesar do intenso esforço de recolha de informação por parte dos SVG oficiais e de alguns sistemas alternativos, subsistem algumas falhas, nomeadamente no que diz respeito ao registo da mortalidade atribuível diretamente à gripe e também às causadas por demoras importantes nos processos de recolha, tratamento e análise dos dados, que tornam necessária a aplicação de métodos estatísticos bastante sofisticados para estimar alguns dos principais parâmetros de avaliação da evolução e dos efeitos destas epidemias.
O objetivo desta comunicação é apresentar resumidamente três artigos que resultaram do trabalho realizado desde 2009, sobre epidemias de gripe, pelo grupo de investigação constituído por Baltazar Nunes, Isabel Natário e M. Lucília Carvalho, sendo que os dois primeiros artigos foram elaborados no âmbito da tese de Doutoramento do primeiro dos autores citados.
Estes três artigos ilustram metodologias de abordagem de outros tantos problemas que surgem como consequência das falhas atrás apontadas.
O primeiro deles dedica-se à estimação do excesso de mortalidade provocado pelas epidemias de gripe, tendo-se identificado as diferenças entre os principais métodos que envolvem processos de estimação aplicados a séries temporais das quais são excluídos os períodos onde existe evidência de ocorrência de epidemia de gripe, i.e., séries temporais interrompidas. Isto permitiu enquadrar todos estes métodos numa única classe em que as características de cada um deles são especificadas através de variações em apenas três atributos essenciais, o que simplifica enormemente a sua descrição e comparação da respetiva aplicabilidade e resultados que fornecem.
Conclusões sobre as vantagens e desvantagens dos diferentes métodos desta classe foram obtidas por aplicação do flubase, pacote de rotinas em R em que se implementaram estes métodos, à série semanal do número de mortes por pneumonia e gripe em Portugal de 1980-81 a 2003-04.
O segundo artigo trata o problema de "nowcasting" (previsão a muito curto prazo) do estado de

um surto de gripe (epidémico ou não epidémico) e da correspondente taxa de incidência da doença através da aplicação de um modelo de cadeias de Markov escondidas não homogéneas em que as probabilidades de transição elementares entre os estados da cadeia de Markov são modeladas através de uma função logística de covariáveis dependentes do tempo. Como no caso anterior, as vantagens e desvantagens do método são ilustradas através de uma aplicação do método aos dados fornecidos pelo Instituto Nacional de Saúde Dr. Ricardo Jorge (INSA) referentes à incidência de gripe entre as semanas 40/2008 e 16/2011 e aos valores das covariáveis utilizadas.
Finalmente, apresenta-se uma segunda versão do terceiro artigo que trata da estimação da taxa de incidência de doenças aparentadas com a gripe, ou em inglês, *Influenza Like Illnesses* (ILI), através de um modelo de espaço de estados. A estimação é feita com base na informação recolhida por um sistema não oficial, o GRIPENET, em que voluntários comunicam semanalmente a presença ou ausência de sintomas de ILI. A metodologia proposta tenta superar as fragilidades que o facto da participação ser voluntária introduz na aleatoriedade dos dados, bem como na distinção entre verdadeiros e falsos positivos.
Uma bibliografia mais extensa sobre os temas desta comunicação, bem como sobre os diferentes modelos e técnicas estatísticas utilizadas, pode ser encontrada em cada um dos artigos que estão abaixo referidos.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação Nacional de Ciência e Tecnologia através do projeto PEst-OE/MAT/UI0006/2011.

Referências

[1] Nunes, B., Natário, I., Carvalho, M.L. (2011). Time series methods for obtaining excess mortality attributable to influenza epidemics. *Statistical Methods in Medical Research* 20, 331–345.

[2] Nunes, B., Natário, I., Carvalho, M.L. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in Medicine* 32(15), 2643–2660.

[3] Natário, I., Carvalho, M.L. (2009). Addressing the problem of lack of representativeness on syndromic schemes. *Discussion es Mathematicae - Probability and Statistics* 29, 169–183.

Mesa-redonda

Registos clínicos

Resumo: Os registos clínicos são a memória de um médico! Permitem arquivar as impressões subjetivas e os dados objetivos, servindo de suporte clínico e legal e constituem uma base indispensável no processo de formação e investigação em Medicina.
Os médicos têm procurado encontrar uma forma organizada e racional de executar os seus registos. De facto só um registo organizado pode ser considerado um documento científico.
Torna-se necessária uma codificação dos registos, o seu armazenamento em suporte digital e uma fácil atualização desta informação (e um update dinâmico e amigável da informação).
Na área da Oncologia, o Registo Oncológico é fundamental para inferir sobre, por exemplo, as taxas de incidência e de mortalidade nos diversos segmentos da população.
A base de dados de um doente oncológico é inicialmente decorrente do primeiro contacto entre o doente e a instituição, incluindo quatro itens: história da doença atual, história passada (pessoal, familiar, social), revisão por sistemas e exame físico.
Nesta mesa redonda pretende-se discutir, entre outros, os seguintes tópicos:

- Podemos ir mais longe na recolha de dados com o objetivo de se conseguir caracterizar o doente oncológico nas mais diversas vertentes, nomeadamente a ambiental, espacial, económica e socio-cultural?
- Poderemos, de alguma maneira, seguir o doente oncológico mesmo quando este passa de um sistema para outro? Por exemplo ambulatório para internamento ou vice-versa?
- Para obter mais informação, nomeadamente sociocultural, é necessário o consentimento do doente? Qual o papel do profissional de saúde nesta questão? Qual o papel do estatístico?
- Como integrar modelos estatísticos nas plataformas informáticas para ajudar na tomada de decisão? Nomeadamente para o conhecimento das tendências de evolução da prevalência e incidência das neoplasias?

Oradores:

- Maria de Fátima de Pina
  INEB - Departamento de Epidemiologia Clínica, Medicina Preditiva e Saúde Pública, Faculdade de Medicina, ICBAS - Instituto de Ciências Biomédicas Abel Salazar & ISPUP - Instituto de Saúde Pública da Universidade do Porto, *fpina@med.up.pt*
- Pilar Gayoso-Diz
  Hospital Clínico Universitario de Santiago de Compostela & Comité Ético de Investigación Clínica de Galicia (CEIC), *pilar.gayoso.diz@sergas.es*
- Javier Muñiz-García
  Medicina Preventiva y Salud Pública, Universidad de A Coruña & Instituto Universitario de Ciencias de la Salud, Hospital Marítimo de Oza, A Coruña & Registro español de Trasplante Cardíaco y Registro de tumores post trasplante, *javier.muniz.garcia@sergas.es*
- Paulo Costa
  Empresa ST+I, Unipessoal, Lda, Portugal, *pcosta@sti.pt*

Organizadores:

- María del Carmen Iglesias-Pérez
  Departamento de Estadística e I.O., Universidad de Vigo - Galicia, *mcigles@uvigo.es*

- Cecília Azevedo
  CMat - Centro de Matemática da Universidade do Minho - Portugal, *cecilia@math.uminho.pt*

Oral Communication

Survival of branching processes through mutations

Maria Conceição Serra
Center of Mathematics, Minho University, mcserra@math.uminho.pt
Serik Sagitov
Chalmers University of Technology, Sweden, serik@chalmers.se

Keywords: Bienaymé-Galton-Watson branching processes, Multitype, Extinction, Mutation.

Summary: Imagine a population of viruses trying to establish itself in a new environment. Suppose the currently dominating type is nearly critical, in that its mean offspring number is close to one. One can think of two main factors which may lead to survival of this population: reproductive success or an advantageous mutation (a mutation producing new type of particles forming a strictly supercritical process). While a reproductive success is possible in the slightly supercritical case, 'survival due to an advantageous mutation' is the only way to escape extinction for a slightly subcritical branching system.
The typical survival scenarios of such branching processes can be studied in terms of the so-called skeleton trees formed by lineages characterized by an appropriate signature of future reproduction success. In this work we propose an alternative approach of defining a skeleton that relies on a random marking of the lineages in the family tree of a Bienaymé-Galton-Watson process. The skeleton is then defined as the subtree formed by the infinite lineages together with the marked lineages.
In this work we show that, if marking is rare, such skeletons are approximated by birth and death processes which can be subcritical, critical or supercritical. We also obtain the limit skeleton for a sequential mutation model [2] and compute the density function for the time to escape from extinction.

Acknowledgments

This work was partially financed by FEDER Funds through "Programa Operacional Factores de Competitividade - COMPETE" and by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia", within the Project Est-C/MAT/UI0013/2011.

References

[1] O'Connell, N. (1993). Yule process approximation of the skeleton of a branching process. *Journal of Applied Probability* 41, 725–729

[2] Sagitov, S., Serra, M.C. (2009). Multitype Bienaymé-Galton-Watson processes escaping exctinction. *Advances in Applied Probability* 41, 225–246.

Comunicação Oral

Modelos mecanísticos de sobrevivência

Francisco Louzada
Universidade de São Paulo, Brasil, louzada@icmc.usp.br

Palavras–chave: Causas competitivas latentes, Esquemas de ativação, Fração de cura, Modelos de sobrevivência.

Resumo: O trabalho consiste na apresentação do estado-da-arte na construção de modelos mecanísticos de sobrevivência com diferentes esquemas de ativação para acomodar dados de tempo de sobrevivência na presença de causas competitivas latentes e fração de cura. Todos os procedimentos apresentados são ilustrados por aplicações reais oriundas de pesquisas e consultorias na área médica.

1 Gênese

Historicamente modelos de sobrevivência compreendem uma das principais ferramentas de suporte a análise de dados relacionados a tempo até a ocorrência de um determinado fenômeno. O desenvolvimento de tais modelos baseia-se na construção de um procedimento formal para descrever quais características dos indivíduos estão, efetivamente, relacionadas com o seu risco de sofrer o evento de interesse, bem como qual a intensidade e direção desse relacionamento.
Em termos evolutivos, geralmente novos modelos de sobrevivência são propostos com a idéia básica relacionada à incorporação de modelos existentes em uma estrutura mais geral e flexível. Além de produzir um melhor ajuste aos dados, estes permitem a determinação do modelo mais apropriado para um conjunto de dados em particular, e estudo de má especificação de modelo. Do ponto de vista prático, existem motivações e características que podem e/ou devem ser incorporadas na modelagem. Por exemplo, sabemos que a ocorrência de um evento de interesse pode ser causada por uma, dentre várias causas competitivas; que essas causas podem ser latentes, no sentido de que não se conhece o número de causas e nem mesmo se observa o tempo de sobrevivência associado a cada causa; que um número determinado de causas é necessário para ativar a ocorrência do evento de interesse; que uma proporção de indivíduos pode não ser susceptível a ocorrência do evento; que existem características disponíveis do tratamento, como o número de doses, o tempo entre as doses e a eficiência de cada dose.
Neste contexto surgem então os modelos mecanísticos de sobrevivência, capazes de acomodar características físico-químicas e estruturais observadas na prática. Além de ser uma característica marcante, a vantagem destes modelos se traduz em uma maior capacidade preditiva da modelagem e em um alinhamento maior da díade prática/teoria.

Agradecimentos

Este trabalho foi parcialmente financiado pelo CNPq, Brasil.

Comunicação Oral

A distribuição de Birnbaum-Saunders gerada pela lei Logística

Emilia Athayde
Universidade do Minho, mefqa@math.uminho.pt
Assis Azevedo
Universidade do Minho, assis@math.uminho.pt

Palavras–chave: Distribuição de Birnbaum-Saunders generalizada, lei logística, taxa de falha.

Resumo: O modelo de Birnbaum-Saunders (BS) consiste numa transformação (não negativa) de uma variável aleatória (v.a.) normal padrão Z. O modelo de Birnbaum-Saunders generalizado (GBS) deriva de substituir Z por qualquer v.a. simétrica. Neste trabalho consideramos o modelo GBS gerado pela distribuição logística, a que chamamos BS-L. Provamos que a taxa de falha é uma função crescente para $t < t_0$ e decrescente para $t > t_0$, para algum $t_0 > 0$. Estimamos os parâmetros do modelo e o ponto de viragem t_0 pelos métodos da verosimilhança máxima e dos momentos modificado. Por último, aplicamos o modelo a dados reais de sobrevivência.

1 Introdução

O modelo de Birnbaum-Saunders [3] consiste numa transformação (não negativa) de uma variável aleatória normal padrão Z, sendo essa transformação e a sua inversa dadas respetivamente por

$$T = \beta \left[\frac{\alpha}{2} Z + \sqrt{\left(\frac{\alpha}{2} Z\right)^2 + 1} \right]^2 \quad \text{e} \quad Z = \frac{1}{\alpha} \left[\sqrt{\frac{T}{\beta}} - \sqrt{\frac{\beta}{T}} \right],$$

que inclui um parâmetro de forma, α ($\alpha > 0$), e um parâmetro de escala, β ($\beta > 0$). Trata-se de uma distribuição unimodal, com assimetria positiva e bastante abrangente em curtose. Recentemente demonstrou-se [2, 5] que a taxa de falha $h_T(t) = \frac{f_T(t)}{1-F_T(t)}$ é uma função IBT (*inverse bathtub*), i.e., é crescente para $t < t_0$ e decrescente para $t > t_0$. Este ponto t_0 é um indicador importante em análise de sobrevivência, pois representa um ponto de viragem no comportamento da taxa de mortalidade.

A classe de distribuições Birnbaum-Saunders generalizada, introduzida por Díaz-Garcia e Leiva [4], consiste na generalização da distribuição anterior, percorrendo Z a classe das v.a.'s simétricas (em relação a zero). Este novo modelo apresenta maior flexibilidade, permitindo uma modelação ainda mais vasta em termos de curtose, incluindo distribuições quer unimodais quer bimodais, e admitindo ainda outras formas para a taxa de falha para além da forma IBT [4, 6, 7]. O caso da distribuição BS-t, correspondente a uma distribuição t de Student para Z, já se encontra estudado quanto às propriedades da taxa de falha [1].

Neste trabalho consideramos Z com distribuição logística, dada pela função distribuição

$$F(t) = \frac{1}{1+e^{-t}}, -\infty < t < +\infty$$

e tomamos a correspondente GBS, a que chamamos BS-L. Estabelecemos resultados sobre a taxa de falha e sobre estimação dos parâmetros α e β, terminando com uma aplicação a dados reais.

2 Resultados principais

Nesta secção apresentamos os principais resultados obtidos para a distribuição BS-L.

Prova-se em primeiro lugar que $\lim_{t\to 0} h_T(t) = 0$ e $\lim_{t\to\infty} h_T(t) = 0$. Seguidamente, estabelece-se que os pontos de viragem da taxa de falha da BS-L são as soluções em t ($t > 0$) da equação (tomando $\beta = 1$, sem perda de generalidade)

$$1 + \exp\left(\frac{1}{\alpha}\frac{t-1}{\sqrt{t}}\right) = \frac{1}{\alpha\sqrt{t}}\frac{(t+1)^2}{t+3}$$

e prova-se que a equação anterior tem uma única solução positiva, levando ao resultado que segue.

Proposição 2.1 *A taxa de falha da BS-L é uma função IBT.*

Tal como no caso da BS e da BS-t, o ponto de viragem da taxa de falha da BS-L não admite uma forma explícita. Para $\alpha > 1.4$ obtém-se a aproximação $(-0.3235 + 3.1683\ \alpha)^{-2}$ para este ponto, em função de α (no caso $\beta = 1$).

A estimação dos parâmetros α e β pelo método da máxima verosimilhança (ML), a partir de uma amostra aleatória $(t_1, ..., t_n)$ resume-se a resolver o sistema

$$\begin{cases} \alpha = \left(\frac{1}{n}\sum v_i\left(\frac{t_i}{\beta} + \frac{\beta}{t_i} - 2\right)\right)^{1/2} \\ \beta = \left(\frac{\frac{1}{2\alpha^2}\sum v_i t_i}{\frac{1}{2\alpha^2}\sum \frac{v_i}{t_i} - \sum \frac{1}{t_i+\beta} + \frac{n}{2\beta}}\right)^{1/2} \end{cases}$$

onde $v_i = \frac{\tanh(b_{t_i}/2)}{b_{t_i}}$. Desenvolveu-se um processo iterativo no R para obtenção da solução. O método dos momentos modificado, baseado no valor médio de T e de $1/T$, foi também utilizado.

3 Aplicação a dados reais

Aplicou-se o modelo BS-L aos dados tratados por Kundu, Kannan e Balakrishnan [5], de tempos de sobrevivência numa amostra de 72 porquinhos-da-índia infetados com o bacilo de Koch, tendo sido estimados os parâmetros, a taxa de falha e o ponto de viragem. O modelo BS-L superou o modelo BS, usando os critérios da distância de Kolmogorov-Smirnov e da log-verosimilhança.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho através do financiamento plurianual da FCT.

Referências

[1] Azevedo, C., Leiva, V., Athayde, E., Balakrishnan, N. (2012). Shape and change point analyses of the Birnbaum-Saunders-t failure rate and associated estimation. *Comput. Statist. Data Anal.* 56, 3887–3897.

[2] Bebbington, M., Lai, C., Zitikis, R. (2008). A proof of the shape of the Birnbaum-Saunders hazard rate function. *Math. Scientist* 33, 49–56.

[3] Birnbaum, Z.W., Saunders, S.C. (1969). A new family of life distributions. *J. Appl. Probab.* 6, 319–327.

[4] Díaz-Garcia, J.A., Leiva, V. (2005). A new family of life distributions based on elliptically contoured distributions. *J. Statist. Plann. Infer.* 128, 445–457.

[5] Kundu, D., Kannan, N., Balakrishnan, N. (2008). On the hazard function of Birnbaum-Saunders distribution and associated inference. *Comput. Statist. Data Anal.* 52, 2692–2702.

[6] Leiva, V., Riquelme, M., Balakrishnan, N., Sanhueza, A. (2008). Lifetime analysis based on the generalized Birnbaum-Saunders distribution. *Comm. Statist. Data Anal.* 52, 2079–2097.

[7] Sanhueza, A., Leiva, V., Balakrishnan, N. (2008). The generalized Birnbaum-Saunders distribution and its theory, methodology and applications. *Comput. Statist. – Theory and Methods* 37, 645–670.

Oral Communication

Survival of colorectal cancer patients with unknown censorship

Beatriz López-Calviño
Unidad de Epidemiología Clínica y Bioestadística-CHUAC A Coruña, beatriz.lopez.calvino@sergas.es

Ricardo Cao-Abad
Universidad de A Coruña, ricardo.cao@udc.es

Ewa Strzalkowska-Kominiak
Universidad de A Coruña, strzalkowska@udc.es

Sonia Pértega-Díaz
Unidad de Epidemiología Clínica y Bioestadística-CHUAC A Coruña, sonia.pertega.diaz@sergas.es

Salvador Pita-Fernández
Unidad de Epidemiología Clínica y Bioestadística-CHUAC A Coruña, salvador.pita.fernandez@sergas.es

Teresa Seoane-Pillado
Unidad de Epidemiología Clínica y Bioestadística-CHUAC A Coruña, maria.teresa.seoane.pillado@sergas.es

Keywords: Biometry, Epidemiology, Demography.

Summary: Information bias can occur in medical research, due to the cause of death is not always known and therefore, there is a lack of information on whether the observed lifetime is censored or not. Moreover, it may happen that the censorship indicator is lost.
The main aim of this study is to analyze the survival of colorectal cancer in two cohorts of A Coruña Hospital, by the estimates of the survival function with censoring indicators missing randomly proposed by Wang and Ng (2008).

1 Introduction

In prognostic studies, there may be a bias in the estimation of the cause-specific survival when considered as censorship deaths from causes other than the disease of interest.Another limitation is that you need to have the cause of death is not always available. May cause a bias in the estimation of survival.
We study estimators of the survival function with censoring indicators missing at random data of Wang and Ng (2008), with the objective of determining the colorectal cancer-specific survival in two cohorts of patients.

2 Methods

Follow-up ambispective study of two cohorts incident cases of colorectal cancer diagnosed in the A Coruña Hospital.

- Cohort 1 (n=1464): period: 1994-2000 (Follow-up: 6.0 ± 4.9 years. 42.0% patients with cause of death unknown)

- Cohort 2 (n=1323): period: 2006-2012 (Follow-up: 1.6 ± 1.1 years. 3.5% patients with cause of death unknown)

We compared the colorectal cancer-specific survival according to several methods:

- Kaplan-Meier estimator on observations with known cause of death (deleted from the analysis patients with unknown cause of death).
- Kaplan-Meier estimator, assuming as cancer-related deaths of unknown cause.
- Presmoothed Kaplan-Meier estimators (Cao et al., 2005) and estimators with censoring indicators missing at random data (Wang and Ng, 2008). These estimates are based on estimating the probability that a patient has died of the tumor, either for all patients in the cohort (S_n^P presmoothed Kaplan-Meier estimator and $\hat{S}_{n,W}$ estimator proposed by Wang), or only for patients with unknown cause of death ($\hat{S}_{n,I}$, $\tilde{S}_{n,I}$ estimators proposed by Wang).

3 Results

For both cohorts the mean age was 62.2 ± 11.5 years and 70.0 ± 11.1 years. Men (55.3% and 60.9%). Most frequent tumor grade T3, N0 and M0. By tumor exitus 28.4% and 25.7%, respectively.
Compared to the Kaplan-Meier estimator on their complete observations, the Kaplan-Meier considering missing data as uncensored underestimates survival. Increasing this bias by increasing the percentage of lost 3.5% to 42.0%.
Regardless of the percentage of cases with unknown censorship in both cohorts were similar results among the estimators $\hat{S}_{n,I}$ and $\tilde{S}_{n,I}$ of Wang, who are coming to Kaplan-Meier with complete observations, being further from the solution obtained with S_n^P presmoothed Kaplan-Meier estimator and $\hat{S}_{n,W}$ of Wang.

4 Conclusions

In analyzing cancer-specific survival, the Kaplan-Meier estimator can be used only in the case of complete observations. Wang estimators, that estimate censorship only if missing data, obtained results close to Kaplan-Meier for complete observations, performing better with a lower rate of lost.

References

[1] Cao, R., López-de-Ullibarri I., Janssen, P., Veraverbeke, N. (2005). Presmoothed Kaplan-Meier and Nelson-Aalen estimators. *Journal of Nonparametric Statistics* 17(1), 31–56.

[2] Kaplan, E.L., Meier, P. (1958). Nonparametric estimation form incomplete observations. *Journal of the American Statistical Association* 53, 457–481.

[3] Wang, Q., Ng, K.W. (2008). Asymptotically efficient product-limit estimators with censoring indicators missing at random. *Statistica Sinica* 18, 749–768.

Comunicação Oral

Modelos de Sobrevivência Weibull modificado aplicados a dados de câncer de mama

Gleici Perdona
FMRP-USP, pgleici@fmrp.usp.br

Francisco Louzada
ICMC-USP, louzada@icmc.usp.br

Cleyton Zanardo
NAP-HC Barretos, bioestatistica@hcancerbarretos.com.br

Hayala Cavenague
Saude em Comunidade-DMS-USP, haycavenague@hotmail.com

Palavras–chave: Câncer de Mama, Fração de Cura, Modelos de sobrevivência, Weibull modificada.

Resumo: Neste trabalho discutimos a família de modelos de risco Weibull modificado para dados de longa duração aplicado a problemática do câncer.
O câncer de mama é a neoplasia maligna mais incidente entre as mulheres no mundo todo. Apesar da importância desta neoplasia, existem poucos levantamentos sobre o problema no Brasil, Cobre et al. (2012). A família de modelos considerados é bastante flexível e baseada no modelo de mistura de Bergson e Cage (1952) e Lai et al. (2003) entretanto formulada via função de risco. Esta família apresenta diversos casos particulares, permitindo acomodar diferentes formas para a função de risco, monótona e não monótonas e também incorpora na sua formulação um parâmetro que denota a presença de longa duração nos dados. Esta formulação também permite a determinação do modelo mais apropriado entre os casos particulares, para um conjunto de dados particular. Considerando T uma variável aleatória não negativa que representa o tempo de vida de um indivíduo, de uma população. Segundo, Lawless (2003), a função de risco no tempo t, é definido como $h(t) = \lim_{\Delta t \to 0} Pr(t < T < t + \Delta t \mid T \geq t)/(\Delta t) = f(t)/S(t)$. A classe baseada na função de risco, Perdona e Louzada-Neto (2011), é dada por:

$$h(t; p, \theta, \nu) = \frac{p\theta \frac{\partial[1-g(t;\nu)]}{\partial t}}{1 - p[1 - g(t;\nu)]^{\theta}} [1 - g(t;\nu)]^{\theta-1}, \quad (1)$$

onde a função de risco $g(t;\nu)$ é positiva, monótona decrescente, ν é um vetor de parâmetros de $g(.)$, θ é parâmetro de forma e p parâmetro entre zero e um $(0 < p < 1)$, denotando a proporção de tempos de longa duração, Maller e Zhou (1996).
A vantagem de (1) é a caracterização da modelagem por uma função, g(), genérica, acomodando, portanto, um largo espectro de modelos de risco.
Procedimentos baseados em máxima verossimilhança são abordados, e estudos de simulação foram desenvolvidos para verificar propriedades frequentistas dos procedimentos de inferência adotados. Estes procedimentos apresentam facilidade em termos computacionais via utilização de pacotes estatísticos. Apesar de os procedimentos de maximização poderem ser executados para resolver o sistema de equações não-lineares dadas pelas derivadas parciais do logaritmo da função de verossimilhança, $l(p, \alpha, \beta, \lambda)$, com respeito aos parâmetros, na nossa experiência o método Newton-Raphson puro tem sido muito susceptível a falha na convergência, onde optamos por considerar o algoritmo BFGS para computar as estimativas de máxima verossimilhança (MLEs) no software R via função "optim".
Os estudos de simulações realizados, para examinar a probabilidade de cobertura para intervalos assintóticos para os parâmetros do modelo e para selecionar ajuste de casos particulares (Critérios

AIC e BIC), mostraram que a probabilidade de cobertura para todos os parâmetros é próxima da nominal para tamanhos moderados, decrescendo para 90% a medida que o tamanho da amostra diminui e a quantidade de censura aumenta e que ambos os critérios (AIC e BIC) apresentam dificuldade de identificar o modelo correto, particularmente quando comparado com o modelo Weibull de longa duração e quando o tamanho da amostra é pequeno e a censura aumenta.
Aplicamos a metodologia proposta a um conjunto de dados de câncer de mama, Cobre et al. (2012), que consiste em uma amostra de 40 mulheres com diagnóstico de carcinoma ductal invasivo tratadas na região sudoeste do Brasil, Ribeirão Preto, SP, no HCFMRP-USP. Os resultados mostraram-se favoráveis ao modelo completo.

Agradecimentos

Este trabalho é financiado pelo projeto FAPESP N.2011/000180-3.

Referências

[1] Berkson, J., Gage, R.P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **52**, 501–515.

[2] Cobre, J., Perdona, G.S.C., Louzada, F., Peria, F. (2012). A mechanistic breast cancer survival modeling through the axillary lymph node chain. *Statistics in Medicine*, In press.

[3] Lai, C.D., Min X., Murthy, N.P. (2003). A modified Weibull distribution. *IEEE Transactions on Reliability* **52**, 33–37.

[4] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley.

[5] Maller, R.A., Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. New York: John Wiley.

[6] Perdona, G.S.C., Louzada, F. (2011). A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics* **38**, 1395–1405.

Comunicação Oral

Comparação de teste de rastreio de glaucoma primário de ângulo aberto na diabetes tipo 2, utilizando curvas ROC

Ana C. Braga
Universidade do Minho, *acb@dps.uminho.pt*

Hugo Frade
Universidade do Minho, *hugoecfrade@gmail.com*

Lígia Figueiredo
Centro Hospitalar de Vila Nova de Gaia/Espinho EPE, *ligia_figueiredo@hotmail.com*

Dália Meira
Centro Hospitalar de Vila Nova de Gaia/Espinho EPE, *daliamartinsmeira@gmail.com*

Palavras–chave: curva ROC, AUC (area under the curve), Comp2ROC, GPAA (glaucoma primário de ângulo aberto).

Resumo: O glaucoma primário de ângulo aberto (GPAA) é uma das principais causas de cegueira [4]. Os fatores de risco comumente aceites para GPAA incluem idade avançada, pressão intraocular elevada, história familiar positiva para glaucoma e raça [3].
Os indivíduos com diabetes mellitus tipo 2 (DM2) apresentam maior risco de ter glaucoma de ângulo aberto (GPAA) do que aqueles sem DM2. Assim, os programas de rastreio de retinopatia diabética poderão ser uma excelente oportunidade para a triagem de GPAA e implementação de testes adicionais.
Foi efetuado um estudo transversal retrospetivo de indivíduos com DM2 com avaliação de retinopatia diabética no período de 2008 a 2010 no Centro Hospitalar de Vila Nova de Gaia. Os indivíduos que apresentaram tela positiva foram encaminhados para a digitalização da camada de fibras nervosas da retina com aparelho GDx$^{\text{TM}}$. O diagnóstico de GPAA foi realizado com base numa análise oftalmoscópica computadorizada e estática e com base nesta, os olhos foram classificados como saudáveis ($n_N = 85$) e com glaucoma definitivo ($n_A = 37$).
A comparação dos indicadores de diagnóstico foi efetuada através do *Comp2ROC* [2].
A versão 1.0 do *Comp2ROC* é o resultado da compilação de funções que implementam uma metodologia de otimização multi-objetivo para avaliar dois sistemas de diagnóstico baseados em índices de avaliação de desempenho das curvas ROC. O pacote permite desenhar curvas ROC no plano unitário, gráficos de distâncias e resultados provenientes da comparação das curvas. A aplicação foi desenvolvida em R (requer versão 2.15.1 ou superior e os pacotes `boot` e `ROCR`). Este pacote foi desenvolvido tendo por base a metodologia proposta por Braga et al [1].
Das medidas do GDx$^{\text{TM}}$avaliadas, o NFI (Nerve Fiber Indicator) revelou ser o indicador com maior capacidade descriminante para o diagnóstico de GPAA ($AUC = 0,9251, SE = 0,028$).

Referências

[1] Braga, A.C. , Costa, L. , Oliveira, P. (2013). An alternative method for global and partial comparison of two diagnostic system based on ROC curves. *Journal of Statistical Computation and Simulation* 83(2), 307–325.

[2] Braga, A.C., Frade, H. (2013). Comp2ROC: R Package to Compare Two ROC Curves. In Mohd Saberi Mohamad, Loris Nanni, Miguel P. Rocha and Florentino Fdez-Riverola (eds.): *7th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2013)*, Advances in Soft Computing, in press.

[3] Chopra, V., Varma, R., Francis, B.A., Wu, J., Torres, M., Azen, S.P. (2008). Type 2 diabetes mellitus and the risk of open-angle glaucoma the Los Angeles Latino Eye Study. *Ophthalmology* 115(2), 227–32.

[4] Leske, M.C. (1983). The epidemiology of open-angle glaucoma: a review. *American Journal of Epidemiology* 118(2), 166–91.

Comunicação Oral

Adesão ao tratamento em doenças crônicas

Fernando Gomes
Estudante do Mestrado em Estatística do Departamento de Matemática e Aplicações da Universidade do Minho, parentegomes@hotmail.com

Cecília Azevedo
CMat - Centro de Matemática da Universidade do Minho, cecilia@math.uminho.pt

Palavras–chave: Adesão, Programação em R, Regressão logística, SIDA.

Resumo: A não adesão por parte dos pacientes a tratamentos tem vindo a ser estudada em várias doenças, com destaque para as doenças crónicas como, por exemplo as seguintes Ã¡reaspatologias, Oncologia, Reumatologia, Anti-neoplásicos, Fibrose Quística, Diálise Peritoneal, SIDA, Insuficiência Crónica e Transpelante Renal, Insuficiência Renal Crónica $Epoetina + FerroIV$, Esclerose Múltipla, Profilaxia Rejeição Aguda Transplante Hepático Alogénico, Hepatite B, Hipertensão Arterial Pulmonar, entre outras. A questão da adesão/não adesão ao tratamento é muito importante e preocupante pois pode expôr o paciente a risco de vida.
Os profissionais de saúde em geral e os médicos em particular devem encontrar soluções para aumentar esta taxa de adesão. Com o nosso trabalho pretendemos ajudar o profissional de saúde na tomada de decisão através da implementação de um modelo probabilístico na aplicação informática que este utiliza no seu serviço.
Neste trabalho pretendemos identificar os fatores que influenciam, ou mais influenciam, a não adesão a determinado tipo de tratamento. A patologia com que iniciamos o estudo é a SIDA mas pretendemos alargar o estudo a outro tipo de patologias. Fazemos comparações entre duas regiões de Portugal distintas social e economicamente.

1 Introdução

Na literatura médica existem várias interpretações de adesão que vão desde o conceito tradicional referente ao paciente que cumpre regras, obedecendo às prescrições médicas, até ao entendimento de que adesão é um processo que envolve uma decisão autónoma da pessoa.
Neste trabalho é fundamental definir o critério de adesão. Em nenhuma das duas bases de dados a que temos acesso existe registo explícito sobre se o paciente está a aderir, ou não, ao tratamento.
O tratamento para a doença crónica em análise exige que o doente se dirija ao serviço de saúde para levantar, periodicamente, a sua medicação. O doente não pode estar nem um dia sem medicação. Também não pode ter, na sua posse, medicação em excesso. O médico prescreve a quantidade de medicamentos de modo a garantir que, na data da próxima consulta, o doente não tenha medicamentos sobrantes nem em falta.
É através dos registos de variáveis como "Data que o médico indica para o levantamento do total da medicação até à próxima consulta", "Data de marcação da próxima consulta", "Data efetiva de levantamento de medicamentos", "Quantidade de medicamentos que o paciente leva (podem ser vários tipos de medicação)" que tentamos obter valores para a variável ADESÃO.
A definição desta variável a partir de outras (referidas anteriormente) conduz a uma maior complexidade do problema. Não se trata somente de implementar rotinas em R que percorram toda a base de dados (a base de dados teste tem 10 000 registos/observações) mas a deteção de várias incongruências nos registos dos pacientes que levam a uma constante visualização local dos registos para corrigir os programas.

2 A variável ADESÃO

Como foi já referido a base de dados teste tem cerca de 10 000 registos/observações.
Estes registos não correspondem ao número de pacientes em análise porque um paciente tem, em geral, vários registos /observações em que alguns destes são efetuados ao longo do tempo (dados longitudinais). Um paciente pode ter várias observações num só instante de tempo enquanto outros pacientes podem ter várias observações em vários instantes (medidas repetidas).
Para conhecer a taxa de adesão dos doentes ao tratamento e se esta é função de variáveis conhecidas (cujas observações estão disponíveis) assim como o peso de cada uma destas variáveis na variável ADESÃO usamos um modelo de regressão logística.

3 Algumas variáveis explicativas

O conhecimento da população é fundamental. A partir dos dados devemos explorar previamente algumas das principais características do doente, tais como, distância entre a localidade em que vive ao Serviço de Saúde em que tem que fazer o tratamento, sexo, idade, estado civil, entre outras. O cruzamento de variáveis é também muito importante. Os dados de que dispomos revelam-nos que o número de homens com a doença é sensivelmente o dobro do número de mulheres e que a maioria dos doentes são solteiros. A idade destes doentes, independentemente do género, tem uma distribuição que parece ser normal, com um coeficiente de assimetria de, aproximadamente, 0.27 e de achatamento de 3.75, sendo rejeitada contudo a hipótese de normalidade quando utilizado o teste de Shapiro-Wilk. Tal poderá dever-se ao facto de se estar a trabalhar com um número muito elevado de pacientes. De qualquer maneira a idade dos doentes apresenta bastantes outliers apesar de serem todos moderados. De entre os doentes que constituem esta população, menos de 6% apresentam também uma patologia secundária. Esta patologia é um fator com 3 níveis mas a Hepatite C é a mais significativa. Verifica-se que a presença de patologias secundárias é independente da idade. Quando se restringe o universo aos doentes que apresentam a patologia secundária Hepatite C, verifica-se que a idade dos pacientes é normalmente distribuida e que a idade média dos mesmos se situa no intervalo [40.17, 43.99] com 95% de confiança. Muito mais há a dizer e a fazer sobre os dados que dispomos. A análise apresentada ilustra uma ínfima parte de tudo o que pode ser conseguido.

Agradecimentos

Este trabalho foi parcialmente financiado pela ST+I, Serviços Técnicos de Informática.

Referências

[1] Diggle, P.J., Heagerty, P., Liang, K-Y, Zeager, S.L. (2002). *Analysis of Longitudinal Data*, Oxford.

[2] Everitt, B.S., Hothorn T., (2005). *A handbook of statistical analysis using R.*

[3] Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression.* New York: John Wiley Sons.

[4] Vittinghoff, E., Glidden, D., Shiboski, S., Mcculloch, C. (2005). *Regression Methods in Biostatistics - Linear, Logistic, Survival, and Repeated Measures Models.* USA: Springer.

Comunicação Oral

Avaliação de imagens radiográficas de sementes usando ICA

Isabel Cristina Costa Leite
Instituto Federal da Bahia, *isaleite@ifba.edu.br*
Thelma Sáfadi
Universidade Federal de Lavras, *safadi@dex.ufla.br*
Maria Laene Moreira de Carvalho
Universidade Federal de Lavras, *mlaene@gmail.com*

Palavras–chave: Análise de Imagem, Qualidade de Sementes, Raio-X, ICA.

Resumo: Embora subjetivo, o uso de imagens de raios X de sementes é uma ferramenta importante na análise da qualidade de lotes de sementes. Este trabalho teve como objetivo aplicar a análise de componentes independentes (ICA) no processamento automático das imagens radiográficas de sementes. Foram processadas imagens de 445 sementes de girassol (*Helianthus annuus L.*) e realizado teste de germinação com as mesmas. A técnica ICA foi implementada com o uso do algoritmo FastICA que decompôs as imagens de raios X em imagens-base independentes. A partir das características extraídas pela ICA foi usada a análise discriminante como técnica de classificação das sementes segundo diferentes níveis de qualidade física. A classificação obteve um acerto global de 82% e ao se avaliar a classificação das imagens associando com os resultados do teste de germinação pode-se considerar que a aplicação dos métodos propostos é viável.

1 Introdução

O uso de imagens de raios X na análise de sementes teve início em 1953 com Simak e Gustafsson que o utilizaram na avaliação de sementes de espécies florestais. É recomendado pela Associação Internacional de Análise de Sementes (ISTA, 2011) e é um método que permite uma avaliação rápida e não destrutiva, a partir da visualização das estruturas internas da semente, diferenciando sementes bem formadas de sementes vazias, com danos mecânicos ou com ataque de insetos. Contudo, a avaliação das imagens radiográficas ainda está sujeita à subjetividade do analista. Esta subjetividade pode ser minimizada com o uso de técnicas de processamento automático das imagens, nas quais a análise realizada por softwares auxilia na avaliação do analista.
Neste trabalho propõe-se a utilização da análise de componentes independentes (ICA), método que surgiu em meados dos anos 80 com Herault, Jutten e Ans e foi posteriormente desenvolvido como técnica de processamento de sinais. O uso da ICA neste contexto visa a redução de dimensão, ao buscar uma projeção linear dos dados observados (imagens de raios X) em uma base com dimensão reduzida de imagens estatisticamente independentes, proporcionando assim uma representação maximamente informativa e parcimoniosa dos dados originais (Fiori, 2003).
A partir da representação das imagens obtida pela ICA é utilizada a Análise Discriminante (AD) como técnica de classificação das sementes segundo diferentes níveis de qualidade física.

2 Metodologia

A análise de componentes independentes (ICA) é uma técnica estatística que procura revelar a estrutura interna de um conjunto de dados multivariados, decompondo-o numa base de componentes que sejam, o máximo possível, estatisticamente independentes entre si e não gaussianos (Hyvärinen, Karhunen e Oja, 2001).

Neste sentido, a aplicação da ICA em imagens radiográficas de sementes parte do pressuposto de que cada imagem seja resultado da mistura de um conjunto de imagens-base que são informações independentes comuns a todas as radiografias de sementes de diferentes níveis de qualidade. Cada imagem-base ou componente independente (IC) contribui com algum tipo de informação, tal como forma, grau de preenchimento e diferente tipo de dano numa região específica da semente.
Dada a matriz $X = (x_1, x_2, ..., x_n)^T$, cujas linhas são vetores p-dimensionais constituídos dos pixels de cada imagem de semente, considera-se que esta matriz seja gerada pela mistura de n componentes independentes entre si, sendo o modelo ICA expresso da forma $X_{(nxp)} = A_{(nxn)}S_{(nxp)}$ em que A é a matriz dos coeficientes a_{ij} da combinação linear, dita matriz de mistura, e S é a matriz dos componentes independentes $\mathbf{s}_i$.
Objetivando a redução de dimensão, um número $k < n$ de ICs pode ser escolhido utilizando a análise de componentes principais (PCA) como pré-processamento para ICA, de forma que $X_{(nxp)} \approx A_{(nxk)}S_{(kxp)}$. Cada imagem x_i é decomposta em uma combinação linear de ICs (imagens-base), dada por $\mathbf{x}_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 + ... + a_{ik}\mathbf{s}_k$, para todo $i = 1, 2, \ldots, k$, de modo que esta passa a ser representada pelos coeficientes de cada componente independente na mistura.
A matriz de mistura A e a matriz dos componentes independentes S são estimadas por algoritmos que se baseiam na independência das variáveis e usam estatísticas de ordens superiores que maximizam a não gaussianidade dos ICs.
Foram radiografadas 445 sementes de girassol sem nenhum tipo de preparo especial, sendo 175 sementes classificadas como cheia, 140 classificadas como sementes com má formação ou dano grave e 130 classificadas com dano leve. Foram usadas 300 imagens de sementes (100 de cada classificação) como amostra de treinamento para uma regra discriminante que classifica uma nova imagem em um dos três grupos previamente definidos. As 145 imagens restantes foram submetidas a classificação a partir da função discriminate quadrática. Posteriormente foram estimados os acertos e erros de classificação.

3 Resultados e discussão

O grupo de sementes deformadas ou com danos graves mantém percentuais de acerto igual ou superior a 90%, independente do número da dimensão de entrada dos dados. De forma geral os melhores resultados de classificação ocorrem com o uso de 30 ICs.
Considera-se falso-positivo o erro de deixar de classificar um elemento no devido grupo de origem e falso-negativo o erro de classificar no grupo em questão elementos de outro grupo. Para 30 ICs , a maior taxa de erro é de falso-positivos na categoria sementes com danos leves que em sua maioria foram consideradas como sementes cheias.
O confundimento entre sementes cheias e com danos leves já era esperado, pela precisão das imagens, no entanto esse confndimento não interfere no potencial de germinação e sim na avaliação do vigor. A classificação de sementes com danos leves poderia levar a não se obter boas previsões em relação ao vigor, mas poderia servir como um indicativo do potencial germinativo do lote.

Agradecimentos

Este trabalho foi financiado pela FAPEMIG e CNPq.

Referências

[1] Fiori, S. (2003). Overview of independent component analysis technique with an application to synthetic aperture radar (SAR) imagery processing. *Neural Networks* 16(3-4), 453-467.

[2] Hyvärinen, A., Karhunen, J., Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons, Inc.

[3] ISTA (2011). International Rules For Seed Testing Association. Zurich.

Comunicação Oral

Adequação de modelos de classes latentes a planos experimentais relevantes no contexto biomédico

Ana Subtil
Faculdade de Ciências da Universidade de Lisboa e CEMAT, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal, *asubtil@ihmt.unl.pt*

Luzia Gonçalves
CEAUL e Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal, *luziag@ihmt.unl.pt*

Patrícia Bermudez
CEAUL e Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal, *pcbermudez@fc.ul.pt*

Palavras–chave: Modelo de classes latentes, Análise bayesiana, Teste de diagnóstico, Plano experimental.

Resumo:
Os testes de diagnóstico são ferramentas de grande utilidade no contexto das Ciências Biomédicas, pois têm como função indicar a presença ou ausência de doença (ou infecção). Saliente-se que a relevância clínica e a utilidade prática de cada teste devem basear-se na avaliação do seu desempenho, ou seja, da capacidade do teste discernir correctamente os indivíduos doentes dos não doentes (ou os infectados dos não infectados). Idealmente, as medidas de desempenho de um teste de diagnóstico são estimadas por comparação com um *gold standard*, i.e., um teste de referência com sensibilidade e especificidade perfeitas. Contudo, na grande maioria dos casos, não existe um *gold standard* para a doença em estudo ou a sua utilização é limitada por questões de custo, tempo ou ética.
Na ausência de um *gold standard*, o recurso ao Modelo de Classes Latentes (MCL) na análise de dados provenientes da aplicação de múltiplos testes de diagnóstico pode permitir estimar a prevalência da doença e as sensibilidades e especificidades dos testes. Assim, o MCL com uma variável latente binária, indicando o estado da doença, e variáveis manifestas binárias, traduzindo os resultados dos testes de diagnóstico, é amplamente utilizado na prática.
O interesse de investigar a variação da prevalência e do desempenho dos testes com factores externos é manifesto. Entre estes factores, refiram-se, como exemplo, a idade, a estação do ano na amostragem ou factores laboratoriais. Dependendo do plano amostral, da informação prévia disponível e das características dos factores externos, duas abordagens diferentes têm sido consideradas na literatura. Certos estudos admitem subpopulações diferentes, explorando as eventuais diferenças dos parâmetros entre as subpopulações e outros abordam a variação dos parâmetros como função de covariáveis consideradas relevantes.
Quando se pretende explorar as diferenças entre prevalências e/ou desempenho dos testes segundo uma variável que define naturalmente subpopulações distintas, é possível optar por uma amostragem estratificada, recolhendo amostras independentes nos diferentes estratos. Uma situação deste tipo é descrita em [1], onde se estuda a febre aftosa do gado no Cambodja, considerando cinco subpopulações distintas, correspondentes a cinco divisões administrativas. Também em [2] a dirofilariosis canina é estudada com base em amostras independentes recolhidas em três distritos de Portugal Continental, que definem naturalmente três subpopulações distintas. Mesmo que seja inconveniente ou impossível estratificar a população antes de recolher a amostra, Dohoo [3] argumenta que a pós-estratificação de uma população segundo um critério com significado prático pode ser aceitável. Este procedimento é adoptado num estudo sobre testes de diagnóstico da malária documentado em [4].

No cenário de subpopulações independentes entre si, o modelo que emerge naturalmente é o produto de distribuições Multinomiais. A partir do modelo mais geral que admite todos os parâmetros diferentes entre as subpopulações, é possível introduzir restrições e assim explorar as diferenças e semelhanças entre subpopulações, em termos de prevalência e desempenho dos testes, tal como se encontra descrito em [2, 4].
Outra perspectiva com utilidade prática na análise do desempenho dos testes de diagnóstico consiste em admitir covariáveis e estudar o modo como a variação dos valores destas determina variações na prevalência e no desempenho dos testes. Em [5] propõe-se uma abordagem bayesiana admitindo covariáveis, em que é usada uma função de ligação logit para relacionar sensibilidades, especificidades e prevalência com um vector de covariáveis.
As duas abordagens anteriores surgem conjugadas em [6], onde são explorados MCL bayesianos que admitem a estratificação da população e a inclusão de covariáveis.
O objectivo do presente trabalho é adequar MCL bayesianos a outros cenários experimentais relevantes do ponto de vista biomédico. Assim, admitindo uma situação de estratificação, tem interesse investigar o problema da existência de dependências entre as subpopulações consideradas. Outro caso relevante do ponto de vista experimental diz respeito a estudos em que as amostras são recolhidas em diferentes momentos ao longo do tempo, podendo ser ou não constituídas pelos mesmos indivíduos.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação Nacional para a Ciência e Tecnologia (FCT), através dos projectos PTDC/SAU-SAP/113523/2009 e PEst-OE/MAT/UI0006/2011. Ana Subtil tem uma Bolsa de Doutoramento da FCT SFRH/BD/69793/2010.

Referências

[1] Bronsvoort, B., Toft, N., Bergmann, I., Sørensen, K., Anderson, J., et al. (2006). Evaluation of three 3ABC ELISAs for foot-and-mouth disease non-structural antibodies using latent class analysis. *BMC Veterinary Research*, 2.

[2] Gonçalves, L., Subtil, A., Brites, N., Oliveira, M.R., Alho, A.M., Meireles, J., Madeira de Carvalho, L.M., Belo, S. (2012). Bayesian Latent Class Models in Veterinary and Human Epidemiology. *46th Scientific Meeting of the Italian Statistical Society*, Rome, 22 June 2012.

[3] Dohoo, I. (2008). Quantitative Epidemology: Progress and Challenges. *Preventive Veterinary Medicine* 86, 260-269.

[4] Gonçalves, L., Subtil, A., Oliveira, M.R., Rosário, V., Lee, P., Shaio, M.-F. (2012). Bayesian latent class models in malaria diagnosis. PLoS ONE:e40630

[5] Martinez, E.Z., Louzada-Neto, F., Derchain, S.F.M., Achcar, J.A., Gontijo, R.C., Sarian, L.O.Z., Syrjänen, K.J. (2008). Bayesian Estimation of Performance Measures of Cervical Cancer Screening Tests in the Presence of Covariates and Absence of a Gold Standard. *Cancer Informatics* 6, 33–46.

[6] Pereira, G.A. (2011). *Avaliação de Testes de Diagnóstico na Ausência de Padrão de Ouro Considerando Relaxamento da Suposição de Independência Condicional, Covariáveis e Estratificação da População: Uma Abordagem Bayesiana*. Tese de Doutoramento da Universidade Federal de São Carlos, Brasil.

I Portuguese Meeting of Biometry & I Portuguese-Galician Meeting of Biometry

Oral Communication

Investigating the performances of classification techniques: an application to medical diagnosis data

Derya Ersel
Department of Statistics, Hacettepe University, Turkey, dtektas@hacettepe.edu.tr
Suleyman Gunay
Department of Statistics, Hacettepe University, Turkey, sgunay@hacettepe.edu.tr

Keywords: Data mining, Machine learning, Classification, Diagnosis.

1 Introduction

Making a diagnosis is a major problem in medicine. Although it is easy to obtain and access data about patients in medicine, there is not enough effective analysis tools to detect important relationships from this rich data and to make accurate diagnosis. Data mining is the process of discovering relationships, patterns and trends by analyzing large data sets with statistics, pattern recognition, machine learning, visualisation and database techniques. So, data mining techniques can be used to analyse medical data and to make accurate diagnosis based on patients' characteristics, symptoms and tests results [5, 6].
Data mining models are basically divided into descriptive and predictive models. Whereas the aim of descriptive models is to discover patterns that summarize relationships among variables in the data, the aim of predictive models is to predict the value of a response variable (Y) based on the given values of explanatory variables $(X_1, X_2, ..., X_p)$. In predictive models, a function $f = (\mathbf{X}; \theta)$ which estimates y values given vector of measured values $(\mathbf{X})$ and a set of estimated parameters (θ) is obtained [1, 2].
Classification models are predictive models with the categorical response variable. One of the most popular problems analysed with classification models is to provide a diagnosis based on patients' tests results. In classification models, Y is the class variable. The goal of classification models is to assign an object to the correct class depending on given values of $\mathbf{X}$ by modelling the boundaries between classes. There are many classification techniques each of which models decision boundaries with different ways [1].
In literature, there are three basic approaches to build classifiers; discriminative, regression and class conditional. In discriminative approach, class-conditional and posterior class probabilities for classes are not calculated, decision boundaries are modeled directly. In regression approach, posterior class probabilities $p(y_k|\mathbf{x})$ of classes are modeled and maximum $p(y_k|\mathbf{x})$ is used in prediction. In class-conditional approach, class-conditional distributions $p(\mathbf{x}|y_k, \theta_k)$ are modeled and posterior class probabilities $p(y_k|\mathbf{x})$ of classes are calculated by using $p(y_k)$ probabilities and Bayes rule [1]. In this study, classificiation techniques based on these approaches are introduced and they are used to analyse a medical data set to make accurate diagnosis. Besides, their performances are compared and the best classification technique over medical diagnosis data are investigated.

2 Classification techniques

Some classification techniques dealt within the context of this study are summarized below.
- *Decision trees*: Decision trees divide input space of a data set into mutually exlusive fields and assign each field a label that characterizes objects in these fields. A decision tree has a clear structure

and how decisions are made can easily be understood by following this structure. Decision tress consist of interval and external nodes connected with branches [3].
- *Bayesian network classifier*: Bayesian network are probabilistic graphical models that encode relationships between random variables in databases. Bayesian network classifiers learn conditional distributions of each variable $(X_i, i = 1, 2, ..., p)$ given class label from training data set. Then, $p(y_k|\mathbf{x})$ is calculated for each class by using Bayes Rule. Finally, the class which has the highest probability value $p(y_k|\mathbf{x})$ is predicted [3].
- *K-nearest neighbor classifiers*: These classifiers depend on similarity between samples. In this method, training samples are defined as n-dimensional numeric variables and each training sample represent a point in n-dimensional pattern space. When an unknown sample is entered to the system, k-nearest neighbour classifier searches the nearest k training samples to this unknown sample in this space [4].
- *Logistic discriminant analysis*: When class variable has only two categories, a popular method is logistic discriminant analysis based on regression viewpoint. In this method, posterior class probabilities $p(y_k|\mathbf{x})$ are predicted by using a logistic regression function [1].

3 Comparison of classification techniques

In practice, performances of classifiers vary according to nature of the problem. Escpecially in medical diagnosis applications, methods generating posterior class probabilities are preferred rather than methods generating only class labels. Although models based on class-conditional distributions provide an exact definition for each class, predicting class-conditional distributions may be difficult for high-dimensional problems. In this situation, discriminative classifiers may give better results. In general, models based on class-conditional distributions have more complex structure than regression models; and regression based models have more complex structure than discriminative models. However, complex models are more informative than simpler models [1].

References

[1] Hand, D., Mannila, H., Smyth, P. (2001). *Principles of Data Mining*. The MIT Press, Cambridge.

[2] Larose, D.T. (2004). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, New York.

[3] Othman, M.F., Shan Yau, T.M. (2007). Comparison of different classification techniques using weka for breast cancer. In Ibrahim, F., Abu Osman, N.A., Usman, J., Kadri, N.A. (eds.): *Biomed 06, IFMBE Proceedings* Vol 15, 520–523. Springer-Verlag, Berlin.

[4] Phyu, T.N. (2009). Survey of classification techniques in data mining. *MultiConference of Engineers and Computer Scientists 2009* Vol 1, IMECS 2009, Hong Kong.

[5] Soni, J., Ansari U., Sharma D., Soni, S. (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. *International Journal of Computer Applications* 17, 43–48.

[6] Tan, P., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley, Boston.

[7] Witten, I.H., Frank, E., Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elseiver, Burlington.

Plenary Session

Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models

Charmaine Dean
University of Western Ontario, Canada, cbdean@uwo.ca

Cindy Feng
University of Saskatchewan, Canada, cindy.feng@usask.ca

Keywords: Joint disease mapping, Common spatial factor model, Conditional autoregressive model, Markov chain Monte Carlo, Zero-inflated Poisson.

Summary: This talk discusses joint outcome modeling of multivariate spatial data, where outcomes include count as well as zero-inflated count data. The framework utilized for the joint spatial count outcome analysis reflects that which is now commonly employed for the joint analysis of longitudinal and survival data, termed shared frailty models, in which the outcomes are linked through a shared latent spatial random risk term. We discuss these types of joint mapping models and consider the benefits achieved through such joint modeling in the disease mapping context. We also consider the power of tests for common spatial structure across maps and develop recommendations on the sort of power achievable in some contexts, as well as overall recommendations on the utility of joint mapping. We illustrate the approaches in an analysis of lung cancer mortality as well as an ecological study of Comandra blister rust infection of lodgepole pine trees.

Oral Communication

Bayesian joint analysis of zero-inflated counting and severity data

Giovani Loiola da Silva
CEAUL and IST - Technical University of Lisbon, Portugal, *gsilva@math.ist.utl.pt*
Elizabeth Juarez-Colunga
CSPH - University of Colorado Denver, U.S.A., *elizabeth.juarez-colunga@ucdenver.edu*
Charmaine Dean
FS and DSAS - University of Western Ontario, Canada, *sciencedean@uwo.ca*

Keywords: Bayesian hierarchical models, Zero-inflated data, Joint analysis, Smoothing.

Summary: In longitudinal studies, it is often of interest to jointly model the frequency and intensity of events in order to explain mechanisms generating the recurrent event of interest. For example, in the study of earthquakes, it is important to develop a joint distribution of counts (number of earthquakes) and severities (earthquake magnitude level) in space. Dunson (2003) provided a landmark publication demonstrating the joint analysis of different responses applied to sociological and psychological contexts, whereas Herring and Yang (2007) proposed similar approaches for handling a terminating event and a longitudinal marker in a vaginal bleeding pregnancy study.
This work is motivated by the need for such joint analyzes in a trial involving participants who were healthy menstruating women prior to hysterectomy/ovariectomy for benign disease (Prior *et al.*, 2007). Weekly data provided the number of hot flushes over one year for individuals in each of two treatment arms, medroxyprogesterone acetate (MPA) and conjugated equine oestrogen (CEE). A primary aim was to investigate the effect of the treatments in reducing the event of interest (hot flush) and its severity.
In this analysis there is a relatively large number of zeros, leading to so-called zero-inflated data analysis. We propose a Bayesian joint analysis of counts and severity for longitudinal zero-inflated data in the context of a multivariate correlated model for the joint analysis of a combination of binary, ordinal, multinomial, discrete or continuous outcomes measuring the same underlying trend over time.
This joint modelling enables association between related outcomes and provides better power in identifying effects. It is also useful for providing an understanding of the mechanisms generating the outcomes. In the analysis, there was some difference between the two treatments only in the high severity; this was an important scientific observation. However, the proportion of zeros differed in the treatment arms.

Acknowledgments

Giovani Silva was partially supported by Pest-OE/MAT/UI0006/2011.

References

[1] Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association* 98, 555–563.

[2] Herring, A.H., Yang, J. (2007). Bayesian modeling of multiple episode occurrence and severity with a terminating event. *Biometrics* 63, 381–388.

[3] Prior, J.C., Nielsen, J.D., Hitchcock, C.L., Williams, L.A., Vigna, Y.M., Dean, C.B. (2007). Medroxyprogesterone and conjugated oestrogen are equivalent for hot flushes: a 1-year randomized double-blind trial following premenopausal ovariectomy, *Clinical Science* 112, 517–525.

Comunicação Oral

Modelos de regressão binária: que ligação escolher?

Isabel Natário
CEAUL, Departmento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516, Caparica, Portugal, icn@fct.unl.pt

Sílvia Shrubsall
CESUR, Instituto Superior Técnico - Technical University of Lisbon, Lisbon, Portugal, silviashrubsall@gmail.com

Palavras–chave: Regressão binária, Funções de ligação, Acidentes de viação urbanos.

Resumo:
Os modelos de regressão com dados binários começaram por se desenvolver na epidemiologia, mas rapidamente se estenderam a muitos outros campos do conhecimento [1]. Os modelos de regressão tipicamente procuram descrever e modelar as relações existentes entre uma variável dita resposta, Y, e um conjunto de outras variáveis ditas explicativas, $x_1, \ldots, x_k$, que sendo relacionadas com a resposta servem para ajudar a descrever os seus valores. A modelação é usualmente feita sobre a esperança condicionada $\mathrm{E}[Y|x_1, \ldots, x_k]$, que se relaciona com uma combinação das variáveis independentes, que no caso linear é $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, o chamado preditor linear. A forma da relação entre $\mathrm{E}[Y|x_1, \ldots, x_k]$ e o preditor η, $g(\mathrm{E}[Y|x_1, \ldots, x_k]) = \eta$ é chamada a função de ligação do modelo.
Quando a variável resposta é binária, *i.e*, quando segue uma distribuição Bernoulli(p), onde p é a probabilidade de sucesso da resposta, a esperança da resposta vale p, um valor em $]0, 1[$. Igualar tal quantidade a um preditor linear, que à partida pode tomar qualquer valor em $\mathbb{R}$, não fará muito sentido. Adicionalmente, os exemplo práticos revelam que a variação desta probabilidade com os valores das covariáveis tende a apresentar um gráfico semelhante a de uma função distribuição, sugerindo assim uma função de ligação para este caso baseada numa tal função. Várias foram as sugestões, sendo a mais utilizada a função de distribuição logística (resultando no modelo de regressão logística)

$$p = \frac{e^{\eta}}{1 + e^{\eta}} \Leftrightarrow ln\left(\frac{p}{1-p}\right) = \eta \Leftrightarrow g(p) = \eta,$$

e a segunda mais usada a função distribuição de uma variável aleatória normal reduzida (resultando no modelo de regressão probit):

$$p = \Phi(\eta) \Leftrightarrow \Phi^{-1}(p) = \eta \Leftrightarrow g(p) = \eta.$$

Outra função de ligação possível é a chamada complementar log-log, $ln(-ln(1-p))$, que é a função distribuição inversa da distribuição de valores extremos log-Weibull, com função distribuição

$$F(\eta) = 1 - e^{-e^{\eta}}.$$

Outras opções estão ainda disponíveis. Consequentemente, a pergunta que se põe a um investigador com um problema deste tipo em mãos é que função de ligação escolher? E o mais certo é ele acabar por usar a função de ligação logística, apenas porque é o que se faz. Mas será que a sua escolha é acertada? Será que é indiferente a função ligação que se toma?
Efetivamente, neste tipo de estudos, a escolha mais comum é a função de ligação logística, sendo as principais razões apontadas para tal escolha que o modelo resultante tem um tratamento matemático simples e flexível, com coeficientes de regressão que apresentam uma interpretação clínica simples. E na verdade, as transformações logística e probit são quase funções lineares uma da outra

para valores de p em $]0.1, 0.9[$, pelo que tendem a resultar em estimativas semelhantes na maior parte das situações. Mas o mesmo não sucede com outras funções ligação. O que acontecerá então nesses casos?
A má especificação da função ligação poder conduzir a enviesamentos consideráveis nos parâmetros de regressão e nas estimativas do valor esperado da resposta, apesar de haver pouca preocupação na escolha da função ligação [4]. É então importante avaliar se a escolha da função ligação é adequada, seja através do teste de Hosmer-Lemeshow [2] (com alguns problemas), ou usando a função desvio dos modelos para fazer comparações em termos das inferências e uma espécie de validação cruzada para comparar em termos das predições [6], ou utilizando o critério de Informação de Akaike [7], AIC, ou ainda embebendo a escolha da função ligação numa classe paramétrica de funções de ligação que se assume conter a função de ligação adequada e testar a sua adequabilidade - por exemplo [5]. Este trabalho sistematiza quais devem ser os passos a dar para, de uma forma expedita, o investigador saber decidir no seu caso concreto o que fazer. Estes passos são então dados numa análise de dados de acidentes de viação urbanos graves. Lisboa tem um número de acidentes por habitante superior à correspondente média das outras capitais europeias. Cerca de 11% destes acidentes podem ser classificados de graves por envolverem pelo menos uma morte ou um ferido muito grave (que tem de ser hospitalizado por pelo menos 24 horas após o acidente). Uma conjunto de dados excepcionalmente completo de cerca de 9300 acidentes de viação com vítimas reportados (994 graves e 8269 não-graves) entre 2004 e 2007, juntamente com outras caraterísticas de interesse associadas, está disponível, como produto do projeto SACRA, Spatial Analysis of Child Road Accidents (PTDC/TRA/66161/2006). As referidas caraterísticas cobrem uma larga gama de aspetos passíveis de influenciar a gravidade dos acidentes, desde variáveis específicas dos acidentes e variáveis específicas da via bem como variáveis de condições de trânsito. O objetivo é então escolher quais os fatores que mais podem contribuir para a gravidade do acidente e se as diferentes funções de ligação resultam em diferentes conclusões, explorando essas diferenças também ao nível das interpretações possíveis.

Agradecimentos

À Ana Rita Nunes pela análise inicial dos dados. A ANSR e o LNEC forneceram parte da informação contida nos dados. Este trabalho foi parcialmente financiado por fundos nacionais através da Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT, pelos projeto PEst-OE/MAT/UI0006/2011 e pelo projeto de investigação Spatial Analysis of Child Road Accidents (SACRA), PTDC/TRA/66161/2006

Referências

[1] Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition.* Wiley Series in Probability and Mathematical Statistics, New York.

[2] Hosmer, D.W., Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics . Theory and Methods* 9, 1043–1068

[3] Savolainena, P.T., Manneringb, F.L., Lord, D., Mohammed, A.Q. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention* 43, 1666–1676.

[4] Czado, C., Santner, T.J. (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference* 33, 213Ű-231.

[5] Aranda-Ordaz, F.J. (1981) On two families of transformations to additivity for binary response data. *Biometrika* 68, 357–364.

[6] Huettmann, F., Linke, J. (2003) Assessment of Different Link Functions for Modeling Binary Data to Derive Sound Inferences and Predictions. In V. Kumar et al. (Eds.) *ICCSA 2003* , LNCS 2669, 43–48.

[7] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716-726.

Comunicação Oral

Regressão quantílica bayesiana para proporções

Bruno Santos
Instituto de Matemática e Estatística, Universidade de São Paulo, bramos@ime.usp.br
Heleno Bolfarine
Instituto de Matemática e Estatística, Universidade de São Paulo, hbolfar@ime.usp.br

Palavras–chave: Regressão quantílica bayesiana, Proporções.

Resumo: A análise de regressão de proporções é importante em diversas áreas como biometria e epidemiologia. Porém, considerando os modelos de regressão beta como a principal forma de análise para esse tipo de dados, estuda-se somente a relação entre a média condicional da proporção de interesse e as variáveis explicativas. Nesse trabalho, sugerimos o uso da regressão quantílica bayesiana como uma análise mais completa entre a distribuição condicional da variável resposta no intervalo [0,1], a partir de um modelo que considera uma massa pontual nos pontos {0,1} e a distribuição de Laplace assimétrica no intervalo (0,1). Dessa maneira, é possível analisar não somente a média condicional, mas sim diferentes pontos da distribuição condicional da variável resposta, através dos diferentes quantis estimados pelo modelo de regressão quantílica.

1 Introdução

Considerando o método estatístico de análise de regressão, diversos problemas podem ser encontrados em que a variável dependente é obtida em forma de proporção (ver [1]). Utilizando os modelos de regressão beta propostos por [2] é possível estimar a relação entre a média dessa variável aleatória que apresenta valores no intervalo (0,1) e as variáveis explicativas, de acordo com o interesse do pesquisador. Inclusive, essa forma de análise que considera relações entre a média da variável resposta e as variáveis explicativas é bastante utilizada, tendo em vista a popularidade alcançada pelos modelos lineares generalizados [6].
Apesar disso, esse tipo de análise pode ser incompleta se considerarmos que essa relação pode variar para diferentes quantis da distribuição condicional da variável resposta. Nesse sentido, [4] sugerem estimar os efeitos das variáveis explicativas para os quantis condicionais da variável resposta, diferentemente dos métodos usuais. Do ponto de vista bayesiano, a primeira proposta desses modelos foi feita por [9]. Uma segunda proposta para esses modelos sob o paradigma bayesiano, e que consideraremos aqui, é a de [3], em que os autores consideram a distribuição de Laplace assimétrica para a variável resposta e especificam essa distribuição a partir de uma mistura de localização-escala da distribuição normal e da distribuição exponencial, calculando dessa maneira amostradores de Gibbs que podem ser utilizados para obtenção da distribuição *a posteriori* dos parâmetros da regressão quantílica.

2 Regressão quantílica bayesiana para proporções

Tendo em vista variáveis aleatórias observadas como proporções e definidas no intervalo [0,1], ou num subconjunto desse intervalo, como variável resposta na análise de regressão, um possível candidato para análise é o modelos beta inflacionado de zeros ou uns (ver [7]). Por outro lado, podemos estender a proposta de [5], em que os autores consideram para os dados uma densidade contínua censurada e uma massa pontual de probabilidade. Um caso especial desse modelo é o

modelo Tobit ([8]), quando a probabilidade da massa pontual é igual a zero e a censura nos dados é a esquerda.
Para a proposta de modelos de regressão quantílica devemos considerar a distribuição de Laplace assimétrica ([10]) para a densidade contínua, com o objetivo de avaliar os efeitos das variáveis explicativas nos quantis da variável resposta. Além disso, podemos considerar massas pontuais de probabilidade nos pontos zero e/ou um. E podemos associar essas massas pontuais a covariáveis através de funções de ligação, como a função de probabilidade acumulada da distribuição normal ou da distribuição logística.
Para completar a especificação do modelo é necessário definir distribuições *a priori* apropriadas para os parâmetros do modelo. Seguindo [3], para os parâmetros da parte contínua da distribuição Laplace assimétrica consideramos uma distribuição *a priori* normal. Também utilizamos a distribuição normal para a distribuição *a priori* dos parâmetros da massa pontual. Para o parâmetro de escala da distribuição utilizamos a distribuição gama invertida. É importante ressaltar que todos os hiperparâmetros dessas distribuições *a priori* são conhecidos.
Com essa especificação, é possível obter a distribuição *a posteriori* dos parâmetros a partir de algoritmos MCMC. Para a atualização dos parâmetros da probabilidade da massa pontual é necessário utilizar o algoritmo Metropolis-Hastings, enquanto que é possível obter um amostrador de Gibbs para a parte contínua do modelo no intervalo (0,1), de forma similar a [3].

Agradecimentos

Os autores agradecem a ajuda financeira da FAPESP, por meio dos projetos 2012/20267-9 e 2012/21788-2, e CNPq.

Referências

[1] Cribari-Neto, F., Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software* 34, 1–24.

[2] Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31, 799–815.

[3] Kozumi, H., Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* 81, 1565–1578.

[4] Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* 46, 33–50.

[5] Moulton, L., Halsey, N.A. (1995). A mixture model with detection limits for regression analysses of antibody response to vaccine. *Biometrics* 51, 1570–1578.

[6] Nelder, J.A., Wedderburn, W.M. (1972). Generalized linear models. *Journal of Royal Statistical Society, Series A* 135, 370–384.

[7] Ospina, R., Ferrari, S. (2012). A general class of zero-or-one inflated regression models. *Computational Statistics & Data Analysis* 56, 1609–1623.

[8] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36.

[9] Yu, K., Moyeed, J. (2001). Bayesian quantile regression. *Statistics & Probability Letters* 54, 434–447.

[10] Yu, K., Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods* 34, 1867–1879.

Comunicación Oral

Análisis cluster de curvas de regresión no parametrica en recursos marinos

Nora M. Villanueva
Departamento Estatística e I. O., Universidade de Vigo, *nmvillanueva@uvigo.es*
Marta Sestelo
Departamento Estatística e I. O., Universidade de Vigo, *sestelo@uvigo.es*
Javier Roca-Pardiñas
Departamento Estatística e I. O., Universidade de Vigo, *roca@uvigo.es*

Palabras clave: Regresión no parametrica, Cluster, Contraste, Bootstrap.

1 Introducción

Uno de los principales objetivos de la modelización estadística es conocer la dependencia de una variable Y con respecto a otra variable X. Este tipo de dependencia puede estudiarse a través de los modelos de regresión no paramétrica, donde la relación entre la variable respuesta y la variable explicativa es modelada sin especificar de antemano la función que las une. En el contexto de la inferencia estadística un problema importante es la comparación de dos o más grupos. Esta comparación puede realizarse a través del estudio de las curvas de regresión. Sean l vectores aleatorios independientes (X_j, Y_j) que satisfacen los siguientes modelos de regresión, para $j = 1, \ldots, l$

$$Y_j = m_j(X_j) + \varepsilon_j \tag{1}$$

donde m_j son funciones suaves (desconocidas), ε_j son los errores de media cero y las covariables tienen soporte común.
Del planteamiento del modelo en (1) surgen algunas cuestiones: ¿se pueden clasificar estas curvas en grupos o clusters? y de ser esto posible, ¿cuál es el número adecuado de ellos? En la literatura existen numerosas referencias sobre las técnicas de análisis cluster [1]. Este tipo de técnicas tratan de buscar grupos homogéneos de individuos en un conjunto de datos. Algunos ejemplos ampliamente conocidos y utilizados en numerosas aplicaciones son el algoritmo jerárquico de Ward [2] y el algoritmo k-means [3]. Sin embargo, el problema principal en estos algoritmos es la necesidad de especificar de antemano el número de clusters. El objetivo principal de este trabajo es clasificar las curvas de regresión en clusters, determinando el número, k, apropiado de ellos utilizando técnicas bootstrap [4].

2 Metodología

Para estimar las curvas de regresión se utilizaron suavizadores locales lineales tipo kernel [5, 6]. El parámetro de suavizado o ventana se seleccionó automáticamente por validación cruzada [7]. Esta técnica implica un alto coste computacional, por lo que se aplicaron técnicas de aceleración computacional o binning [8].
Para clasificar las curvas en grupos es necesario obtener una función $\pi : \{1, \ldots, l\} \longrightarrow \{1, \ldots, k\}$ que asocie cada curva m_j $(j = 1, \ldots, l)$ al cluster $\pi(j) \in \{1, \ldots, k\}$. Para ello, se propone el uso de uno de los algoritmos más conocidos dentro de la metodología cluster, el algoritmo k-means. Este algoritmo trata de dividir el conjunto de observaciones de entrada en grupos o clusters, en los cuales cada observación pertenece al grupo con la media más cercana. En este trabajo, en lugar de un conjunto de observaciones como entrada se consideraron las $\hat{m}_1, \ldots, \hat{m}_l$ curvas. Además, para determinar el número de clusters se desarrolló un contraste basado en bootstrap.

3 Aplicación a datos reales

Este estudio se llevó a cabo utilizando medidas de percebes, *Pollicipes pollicipes* (Gmelin, 1789), recogidos en cinco localidades de la costa Atlántica gallega: Laxe do Mouro, Punta Lens, Punta de la Barca, Punta del Boy y Punta del Alba [9]. Para cada una de estas localidades se estimó la relación entre dos variables biométricas que se corresponden con la longitud y la anchura de cada individuo: LT (longitud total) y RC (anchura rostro-carenal). En esta aplicación se pretende agrupar las distintas localizaciones geográficas según la morfología de esta especie. Una vez aplicado el contraste se obtuvo un número de clusters igual a dos. Laxe do Mouro, Punta Lens y Punta del Alba se agruparon en uno de los clusters, mientras que las localidades Punta de la Barca y Punta del Boy en el otro.

Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos MTM2008-03129, MTM2011-23204 (fondos FEDER incluídos) del Ministerio de Ciencia e Inovación de España y por el proyecto 10PXIB300068PR de la Xunta de Galicia.

Referencias

[1] Everitt, B.S. (1980). *Cluster Analysis*. Second Edition, Heinemann, London.

[2] Ward Jr., J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 301(58), 236–244.

[3] Macqueen, J.B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

[5] Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman and Hall, London.

[6] Wand, M.P., Jones, M.C. (1995). *Kernel smoothing*. Chapman and Hall, London.

[7] Golub, G., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2), 215–223.

[8] Fan, J., Marron, J. (1994). Fast implementation of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* 3, 35–56.

[9] Sestelo, M., Roca-Pardiñas, J. (2011). A new approach to estimation of the length?weight relationship of *Pollicipes pollicipes* (Gmelin, 1789) on the Atlantic coast of Galicia (Northwest Spain): some aspects of its biology and management. *Journal of Shellfish Research* 30(3), 939–948.

Comunicação Oral

Modelação flexível do problema de triagem via processos de Dirichlet dependentes

Sandra Ramos
Instituto Superior de Engenharia do Porto (ISEP) e Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), sfr@isep.ipp.pt

Maria Antónia Amaral Turkman
Faculdade de Ciências da Universidade de Lisboa (CEAUL) e Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), antonia.turkman@fc.ul.pt

Marília Antunes
Faculdade de Ciências da Universidade de Lisboa e Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), marilia.antunes@fc.ul.pt

Palavras–chave: Problemas de triagem, Modelos bayesianos não-paramétricos, Métodos MCMC.

Resumo:
O procedimento de triagem envolve a construção de uma região de especificação $C_{\mathbf{X}}$, no espaço $d-$dimensional, de modo a que um indivíduo futuro com um vector de características em $C_{\mathbf{X}}$ tenha maior probabilidade de ser identificado como um *sucesso* (a resposta Y pertence a uma região conhecida C_Y). Na abordagem preditiva bayesiana, a obtenção da região $C_{\mathbf{X}}$ é baseada num critério óptimo assente na maximização de $P(Y \in C_Y | X \in C_{\mathbf{X}}; D)$, restringida à classe das regiões $C_{\mathbf{X}}$ com probabilidade preditiva de triagem $\alpha = P(\mathbf{X} \in C_{\mathbf{X}}|D)$ fixa.
Habitualmente a obtenção de $C_{\mathbf{X}}$ é baseada na modelação paramétrica de $(Y, \mathbf{X})$. Contudo, a modelação em contexto paramétrico necessita da especificação de um certo número de pressupostos, cuja validação é por vezes difícil na prática. Neste trabalho relaxamos o pressuposto paramétrico propondo uma metodologia de triagem bayesiana não-paramétrica baseada em processos de Dirichlet dependentes. O modelo é ilustrado com aplicações em problemas de classificação supervisionada de dados de *microarrays*.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e Tecnologia - FCT: Projecto $PTDC/MAT/118335/2010$ e $Pest-OE/MAT/UI0006/2011$.

Comunicación Oral

Identificación de factores de riesgo de lesión en el fútbol profesional

María del Carmen Iglesias Pérez
Departamento de Estadística e I.O., Universidad de Vigo, mcigles@uvigo.es
Miguel Martínez González
Doctor en Ciencias de la Actividad Física y el Deporte, miguelmartinezg@gmail.com
Luis Casáis Martínez
Facultad de Ciencias de Educación y Deporte, luisca@uvigo.es
Marta Sestelo
Departamento de Estadística e I.O., Universidad de Vigo, sestelo@uvigo.es
Javier Roca Pardiñas
Departamento de Estadística e I.O., Universidad de Vigo, roca@uvigo.es

Palabras clave: Factor de riesgo, Lesión, Fútbol, Regresión logística, Regresión de Cox, Selección de variables.

Resumen: En este trabajo se investigan factores de riesgo de lesión en fútbol profesional mediante técnicas estadísticas multivariantes tales como la regresión logística, el análisis discriminante y la regresión de Cox. El elevado número de variables independientes respecto al tamaño muestral disponible y las correlaciones existentes entre las variables llevan a utilizar técnicas de reducción de la dimensión como Análisis de Componentes Principales (ACP) y métodos de selección de variables.

1 Introducción

A pesar de que son numerosas las referencias acerca de la epidemiología lesional y sus factores y mecanismos de producción en el fútbol, se carece de datos suficientes sobre el control del estado neuromuscular en futbolistas de élite y sobre su relación como indicadores de riesgo de lesión en esa misma población.
El objetivo de este trabajo es la identificación de factores de riesgo de lesión en el fútbol profesional, más concretamente de las áreas anatómicas de la articulación de la rodilla y de la musculatura del muslo (cuádriceps e isquiotibial). Entre las variables a investigar como factores de riesgo se han incluido variables que definen el estado y relación funcional de grupos musculares o articulaciones (rodilla) mediante tecnología actual y de vanguardia como la Tensiomiografía, así como variables obtenidas por distintos medios de la relación y simetría entre grupos musculares con una alta incidencia lesional en fútbol. En total se consideraron 57 variables independientes, medidas a 30 jugadores de un equipo profesional de 2ª división durante la temporada 2007/8. Como variables dependientes se definieron la variable dicotómica Lesión (si/no) y una variable que recogió los días transcurridos desde la valoración física hasta la ocurrencia de lesión, Días_supervivencia. La información recogida por [1] es de alto valor al proceder de una muestra de élite y difícil acceso.

2 Metodología y resultados

Para medir las variables independientes se utilizaron una serie de instrumentos suficientemente validados en la literatura y se siguieron cuidadosamente los protocolos establecidos en cada uno de ellos. Las mediciones se realizaron en 2 momentos puntuales de la temporada, al principio de la primera vuelta y al principio de la segunda vuelta de la competición de liga. El orden seguido

en la secuencia de las pruebas fue: Tensiomiografía (tono muscular en distintos grupos musculares de la pierna dominante y no dominante, 30 variables), Flexibilidad (4 variables), Test de Bosco (salto CMJ=CounterMovementJump) e Isocinético (fuerza, 20 variables). Además se consideraron las variables Lesión previa y Edad, sumando 57 variables independientes. La muestra de jugadores en la primera vuelta estuvo formada por los 28 jugadores de la plantilla oficial, mientras que en la segunda vuelta se recogieron datos de 24 de ellos y de 2 fichajes nuevos.
Para identificar los factores de riesgo de lesión en cada vuelta se utilizó regresión logística y análisis discriminante. Se efectuaron comparaciones de las variables por grupos (lesión/ no lesión) con la prueba t y la U de Mann-Whitney y también ACP para reducir el número de variables a seleccionar en el método de pasos sucesivos. Además, en regresión logística se utilizó el algoritmo propuesto en [2] para la selección de las variables predictoras. En la identificación de los factores de riesgo del tiempo (días) hasta la lesión se utilizó la regresión de Cox con variables dependientes del tiempo (al existir una medición intermedia) y la selección por pasos con el criterio AIC.
En la primera vuelta, los factores de riesgo seleccionados en las comparaciones por grupos fueron DMRectoFemoralDominante, TDRectoFemoralDominante y SimetriaRodillaDominante (Variables de Tensiomiografía; DM=Desplazamiento muscular, TD=Tiempo de reacción muscular). El análisis discriminante y la regresión logística con selección a partir de todas las variables no convergen, pero partiendo de las variables significativas en la comparación por grupos se seleccionaron: DMRectoFemoralDominante (Discriminante) y DMRectoFemoralDominante, SimetriaRodillaDominante y Lesión previa (Logística), siendo mayor el porcentaje de clasificación correcta de la última. El algoritmo tipo [2] seleccionó 5 variables: DMRectoFemoralDominante, TDVastoMedialDominante, TDBicepsFemoralDominante, CMJ y FlexibilidadQDominante. La regresión logística a partir de las componentes principales por ACP selecciona aquellas componentes con mayor peso en DMRectoFemoralDominante, TDRectoFemoralDominante y TDVastoMedialDominante.
En la segunda vuelta hubo 11 variables significativas en la comparación por grupos candidatas a entrar en el análisis discriminante y la regresión logística, que seleccionaron la SimetriaRodillaDominante (5 % nivel de entrada) y el TCRectoFemoralNoDominante, TCBícepsFemoralNoDominante (TC=tiempo de contracción) y CMJ (10 % nivel entrada). El algoritmo tipo [2] seleccionó 3 variables, siendo un modelo posible el formado por TCBicepsFemoralNoDominante, RatioFuncionalDominante_180 (Isocinético) y DMVastoLateralNoDom.
Considerando el tiempo de supervivencia hasta la lesión, la regresión de Cox eligió como variables explicativas del riesgo el CMJ, SimetríaRodillaDominante, TCRectoFemoralDominante y TDVastoMedialDominante entre una selección previa de 9 variables significativas en el modelo de Cox simple.
Finalmente, concluimos que la Tensiomiografía aporta información relevante para el riesgo de lesión así como el CMJ.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto MTM2011-23204 (fondos FEDER incluidos) del Ministerio de Ciencia de España y por el proyecto 10PXIB300068PR de la Xunta de Galicia (España).

Referencias

[1] Martínez, M. (2012). *Predictores de lesión artromuscular en futbolistas profesionales*. Tesis doctoral, Universidad de Vigo.

[2] Sestelo, M., Villanueva, N.M., Roca-Pardiñas, J. (2013). `FWDselect`: An `R` package for selecting variables in regression models. *Discussion Papers in Statistics and Operation Research*, 13/02.

Comunicação Oral

Arrow plot: um novo gráfico para a seleção de genes em dados de *microarrays*

Carina Silva-Fortes
Escola Superior de Tecnologia da Saúde de Lisboa e CEAUL, carina.silva@estesl.ipl.pt
Maria Antónia Amaral Turkman
Faculdade de Ciências da Universidade de Lisboa e CEAUL, antonia.turkman@fc.ul.pt
Lisete Sousa
Faculdade de Ciências da Universidade de Lisboa e CEAUL, lmsousa@fc.ul.pt

Palavras–chave: *Microarrays*, Estimador do núcleo, Curva ROC degenerada, Coeficiente de sobreposição.

Resumo: Um objetivo muito comum na análise de dados de *microarrays* é determinar que genes são diferencialmente expressos sob dois (ou mais) tipos de tecido ou sob amostras submetidas a diferentes condições experimentais. Sabe-se que as amostras biológicas são heterógeneas devido a vários fatores, como por exemplo, antecedentes genéticos e subtipos moleculares, os quais são, na maior parte das vezes, do desconhecimento do investigador. Por exemplo, em experiências que envolvam a classificação de tumores é importante que se identifiquem subtipos do cancro em investigação. Distribuições bimodais ou multimodais geralmente refletem a presença de misturas de subclasses. Consequentemente, pode haver genes que sendo diferencialmente expressos (DE) quando se tem em conta os diferentes subgrupos, não são identificados pelos métodos usualmente utilizados para selecionar genes DE. Neste trabalho propõe-se uma nova representação gráfica que não só permite identificar genes com regulação positiva e regulação negativa, mas também genes DE em subgrupos. Esta ferramenta baseia-se em duas medidas, nomeadamente na área abaixo da curva (AUC) *receiver operating characteristic* (ROC) e no coeficiente de sobreposição entre duas densidades (OVL). Os resultados indicam que a nova ferramenta, *Arrow plot*, apresenta um bom desempenho na seleção de genes com diferentes tipos de expressão diferencial, sendo flexível e útil na análise de perfis da expressão de genes em dados de *microarrays* [1].

Agradecimentos

Trabalho parcialmente financiado por fundos nacionais através da FCT no âmbito dos projetos PEst-OE/MAT_UI0006/2011 e PTDC/MAT/118335/2010 e da bolsa de doutoramento SFRH/BD/45938/2008.

Referências

[1] Silva-Fortes, C., Amaral Turkman, M.A., Sousa, L. (2012). Arrow plot: a new graphical tool for selecting up and down-regulated genes and genes differentially expressed on sample subgroups. *BMC Bioinformatics* 13, 147.

Comunicação Oral

Efeito do preprocessamento de dados na deteção de genes diferencialmente expressos

Adelaide Freitas
Departamento de Matemática & CIDMA, Universidade de Aveiro, adelaide@ua.pt
Sara Roque
Departamento de Matemática, Universidade de Aveiro, sara.roque86@gmail.com

Palavras–chave: Dados de *microarrays*, Correção de *background*, Normalização, Taxa de falsas descorbertas.

Resumo: Tomando diferentes bases de dados públicas, investigamos o efeito de 36 técnicas de preprocessamento de dados de *microarrays* na identificação de genes diferencialmente expressos. As técnicas de preprocessamento aqui consideradas resultam da combinação de 6 técnicas de correção de *background* (conhecidas na literatura por `none`, `substraction`, `minimum`, `half`, `edwards` e `normexp`, Freitas el al (2009)) com 6 técnicas de normalização (conhecidas na literatura por `none`, `Intensity Global loess`, `Intensity Local loess`, `Spatial Local loess`, `Intensity Global loess followed by Spatial Local loess`, e `Intensity Local loess followed by Spatial Local loess`, Wu et al (2005)). Para cada base de dados de *microarrays*, aplicamos a metodologia SAM (do inglês, *Significance of Analysis of Microarray*, Tusher et al (2001)) para identificar genes diferencialmente expressos e avaliamos resultados em função da taxa de falsas descobertas. Comparativamente, em termos dos métodos de preprocessamento considerados, evidenciam-se diferenças substanciais nos conjuntos de genes tido como diferencialmente expressos obtidos pelo SAM. Assim, na identificação de genes diferencialmente expressos, é altamente recomendado aplicar mais do que uma estratégia de preprocessamento de dados de *microarrays* e comparar resultados antes de qualquer conclusão e/ou análises subsequentes sobre os níveis de expressão dos genes.

Agradecimentos

Trabalho subsidiado pelo FEDER, através do COMPETE ("Programa Operacional Factores de Competitividade"), pelo CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações) da Universidade de Aveiro e FCT (Fundação para a Ciência e a Tecnologia), dentro do projecto PEst-C/MAT/UI4106/2011 com número COMPETE FCOMP-01-0124-FEDER-022690.

Referências

[1] Freitas, A., Castillo, G., São Marco, A. (2009). Effect of Background Correction on Cancer Classification with Gene Expression Data. In *Proceedings of AIME 2009, Lecture Notes in Artificial Intelligence*, 416–420, Springer Verlag.

[2] Tusher, V., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences*, 98, 9, 5116–5121.

[3] Wu, W., Xing, E., Myers, C., Mian, I.S., Bissel, M.J. (2005) Evaluation of normalization methods for cDNA microarray data by k-NN classification. *BMC Bioinformatics* 6:191.

Comunicação Oral

Dinâmica da população do tubarão *Centroscymnus coelolepis* nas águas continentais portuguesas

Ivone Figueiredo
Instituto Português do Mar e da Atmosfera (IPMA), Portugal, *ifigueiredo@ipma.pt*

Isabel Natário
Centro de Estatística e Aplicações da Universidade de Lisboa e Departamento de Matemáica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, *icn@fct.unl.pt*

Teresa Moura
Instituto Português do Mar e da Atmosfera (IPMA), Portugal, *tmoura@ipma.pt*

Lucília Carvalho
Centro de Estatística e Aplicações da Universidade de Lisboa, *mlucilia.carvalho@gmail.com*

Palavras–chave: *Centroscymnus coelolepis*, dinâmica populacional, modelo de espaço de estados.

Resumo: As pressões antropogénicas sobre os tubarões, particularmente as decorrentes da pesca, podem causar reduções na abundância e alterações na estrutura das suas populações. A grande susceptibilidade das diferentes populações de tubarões à pesca está muito associada com a complexidade dos seus ciclos de vida e à baixa produtividade das espécies.

Na costa continental portuguesa, a espécie *Centroscymnus coelolepis* Barbosa du Bocage & de Brito Capello, 1864, é capturada como espécie acessória da pescaria de palangre de profundidade. Neste trabalho, avalia-se a evolução da abundância de *C. coelolepis*, recorrendo a um modelo de espaço de estados. Estes modelos têm vindo a ser cada vez mais utilizados no estudos sobre a dinâmica de recursos pesqueiros [1].

No caso presente, o modelo é aplicado à subpopulação de fêmeas e inclui dois processos que decorrem em simultâneo: o processo não observável, que descreve a abundância, em número, desta subpopulação e o processo observável que se refere aos valores de sua captura. De acordo com a formulação do modelo a subpopulação de fêmeas é representada por um vetor de estados com cinco componentes, $\mathbf{n}_t$ para o ano t, expresso por $\mathbf{n}'_t = \big(n_{b,t}\ n_{j,t}(\overline{F})\ n_{a,t}(\overline{F})\ n_{j,t}(F)\ n_{a,t}(F)\big)$. A primeira componente, $n_{b,t}$, inclui os juvenis que no ano t não são recrutados à pesca. A segunda e terceira componentes representam a abundância das fêmeas que no ano t sobrevivem à pesca, no primeiro fêmeas juvenis, i.e., sem capacidade de se reproduzirem, $n_{j,t}(\overline{F})$ e no segundo fêmeas adultas, capazes de se reproduzirem, $n_{a,t}(\overline{F})$. As duas últimas componentes representam a abundância de fêmeas juvenis e adultas que no ano t foram pescadas, respectivamente $n_{j,t}(F)$ e $n_{a,t}(F)$.

Na definição do modelo assume-se que as fêmeas se comportam idêntica e independentemente umas das outras (hipótese IID). Assume-se ainda que processo de estados é um processo markoviano e que a evolução anual do processo é formulada em termos dos valores esperados condicionais $E[\mathbf{n}_t|\mathbf{n}_{t-1}] = P\,\mathbf{n}_{t-1}$, em que a matriz P descreve o efeito médio a longo prazo do processo estocástico.

Acresce que dada a complexidade da dinâmica populacional este processo é decomposto em subprocessos, definidos em função de etapas do ciclo de vida consideradas vitais para a espécie. Os subprocessos considerados são *S - sobrevivência à mortalidade natural*, *C - transição de classe de comprimento*, *B - nascimento* e *F - sobrevivência à mortalidade por pesca* o que conduz à decomposição $E[\mathbf{n}_t|\mathbf{n}_{t-1}] = FBCS\,\mathbf{n}_{t-1}$. Sob a suposição markoviana, o processo de estados é completamente definido se a sua distribuição no ano t condicionada aos estados no ano $t-1$ for conhecida, ou seja, $\mathbf{n}_t \overset{d}{=} H_t[\mathbf{n}_{t-1}]$. Considerando que os subprocessos se sucedem no tempo, sempre

pela mesma ordem, esta distribuição é ainda dividida em:

$$\mathbf{u}_t^S \stackrel{d}{=} \mathbf{H}_t^S[\mathbf{n}_{t-1}] \qquad \mathbf{u}_t^C \stackrel{d}{=} \mathbf{H}_t^C[\mathbf{u}_t^S] \qquad \mathbf{u}_t^B \stackrel{d}{=} \mathbf{H}_t^B[\mathbf{u}_t^C] \qquad \mathbf{n}_t = \mathbf{u}_t^F \stackrel{d}{=} \mathbf{H}_t^F[\mathbf{u}_t^B],$$

em que, para cada ano $t = 1, \ldots, T-1$, $\mathbf{u}_t^S$, $\mathbf{u}_t^C$, $\mathbf{u}_t^B$ e $\mathbf{u}_t^F$ representam os vetores de estados após a realização de cada um dos subprocessos. No primeiro subprocesso considera-se que a probabilidade de sobrevivência à mortalidade natural, $\phi_{S_{bja}}$, é semelhante para três componentes populacionais, donde:

$$\mathbf{u}_t^S \sim \mathbf{H}_t^S(\mathbf{n}_{t-1}) : \left(\begin{array}{lcl} u_{b,t}^S & \sim & \mathrm{Bi}\left(n_{b,t-1}, \phi_{S_{bja}}\right) \\ u_{j,t}^S & \sim & \mathrm{Bi}\left(n_{j,t-1}(\overline{F}), \phi_{S_{bja}}\right) \\ u_{a,t}^S & \sim & \mathrm{Bi}\left(n_{a,t-1}(\overline{F}), \phi_{S_{bja}}\right) \end{array} \right).$$

No subprocesso de transição de classe, C_b representa probabilidade de um juvenil não recruta transitar para a classe de juvenis passíveis de serem pescados e C_j representa a probabilidade de um juvenil transitar para a classe adulta. Assim:

$$\mathbf{u}_t^C \sim \mathbf{H}_t^C(\mathbf{u}_t^S) : \left(\begin{array}{lcl} u_{b,t}^C & \sim & \mathrm{Bi}\left(u_{b,t}^S, 1 - C_b\right) \\ u_{j,t}^C & = & Y[u_{j,t}^S] + \left(u_{b,t}^S - u_{b,t}^C\right), \text{ with } Y[u_{j,t}^S] \sim \mathrm{Bi}\left(u_{j,t}^S, 1 - C_j\right) \\ u_{a,t}^C & = & u_{a,t}^S + \left(u_{j,t}^S - X[u_{j,t}^S]\right) \end{array} \right).$$

Após o subprocesso de nascimento, o vetor da população é representado por:

$$\mathbf{u}_t^B \sim \mathbf{H}_t^B(\mathbf{u}_t^C) : \left(\begin{array}{lcl} u_{b,t}^B & = & u_{b,t}^C + X[u_{a,t}^C], \text{ with } X[u_{a,t}^C] \sim \mathrm{Bi}\left(f u_{a,t}^C, p_B\right) \\ u_{j,t}^B & = & u_{j,t}^C \\ u_{a,t}^B & = & u_{a,t}^C \end{array} \right),$$

em que f é o número máximo de embriões por ninhada, $u_{a,t}^C$ é o número de fêmeas adultas e p_B é a probabilidade de um embrião dar origem a uma fêmea que sobreviva.
Finalmente, após a realização do subprocesso de pesca, o vetor da população é:

$$\mathbf{n}_t = \mathbf{u}_t^F \sim \mathbf{H}_t^F(\mathbf{u}_t^B) : \left(\begin{array}{lcl} n_{b,t} & = & u_{b,t}^B \\ n_{j,t}(\overline{F}) & \sim & \mathrm{Bi}(u_{j,t}^B, 1 - \phi_{jt}) \\ n_{a,t}(\overline{F}) & \sim & \mathrm{Bi}(u_{a,t}^B, 1 - \phi_{at}) \\ n_{j,t}(F) & = & u_{j,t}^B - n_{j,t}(\overline{F}) \\ n_{a,t}(F) & = & u_{a,t}^B - n_{a,t}(\overline{F}) \end{array} \right),$$

em que ϕ_{jt} and ϕ_{at} são, respectivamente, a probabilidade de um juvenil e de um adulto serem pescados no ano t.
O processo observacional, sendo uma função estocástica de estados desconhecidos, é representado por $\{y_t, t = 0, 1, \ldots, T\}$, em que y_t corresponde ao número total de fêmeas pescadas no ano t, que se admite ser observado com erro. No caso presente $y_t|\mathbf{n}_t$ tem a distribuição normal

$$y_t|\mathbf{n}_t \sim \mathrm{N}\left(n_{j,t}(F) + n_{a,t}(F), \psi^2(n_{j,t}(F) + n_{a,t}(F))\right).$$

O modelo da evolução dos processos observacional e de estados pode assim ser descrito por

$$g_0(\mathbf{n}_0; \Theta), \qquad g_t(\mathbf{n}_t|\mathbf{n}_{t-1}; \Theta); \qquad f_t(y_t|\mathbf{n}_t; \Theta), \qquad \Theta = (\phi_{S_{bja}}, C_b, C_j, p_B, \phi_j, \phi_a, \psi),$$

$$g_t(\mathbf{n}_t|\mathbf{n}_{t-1}; \Theta) = \int_{\mathbf{u}_t^F} \int_{\mathbf{u}_t^B} \int_{\mathbf{u}_t^C} \int_{\mathbf{u}_t^S} g^S(\mathbf{u}_t^S|\mathbf{n}_{t-1}; \Theta)\, g^C(\mathbf{u}_t^C|\mathbf{u}_t^S; \Theta) g^B(\mathbf{u}_t^B|\mathbf{u}_t^C; \Theta) g^F(\mathbf{n}_t|\mathbf{u}_t^B; \Theta)\, \mathrm{d}\mathbf{u}_t^F\, \mathrm{d}\mathbf{u}_t^S\, \mathrm{d}\mathbf{u}_t^T\, \mathrm{d}\mathbf{u}_t^B.$$

A estimação é feita no contexto bayesiano, recorrendo ao método de amostragem-reamostragem, *sequential importance sampling* [2].

Referências

[1] de Valpine, P., Hilborn, R. (2005). State-space likelihoods for nonlinear fisheries time-series. *Canadian Journal of Fisheries and Aquatic Sciences* 62(9), 1937–1952.

[2] Liu, J., West, M. (2001). Combining parameter and state estimation in simulation-based filtering In Doucet A., Freitas N., Gordon N. (eds.): *Sequential Monte Carlo Methods in Practice*, 197–224, Springer.

Comunicação Oral

Ecologia das comunidades vegetais: utilização da distribuição multinomial numa perspetiva bayesianao

Luís Silva
CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Pólo dos Açores - Departamento de Biologia, Universidade dos Açores, 9501-801 Ponta Delgada, Portugal, *lsilva@uac.pt*

Palavras–chave: Ecologia Vegetal, Espetro de espécies, Estatística bayesiana, Multinomial.

Resumo: Analisa-se a utilização da distribuição multinomial em Ecologia no âmbito de uma abordagem inferencial bayesiana, em particular no caso da descrição e comparação de comunidades vegetais.

1 Introdução

A distribuição multinomial foi utilizada em ecologia [1,2] e noutras áreas [3,4,5,6,7]. Nas comunidades vegetais, é possível agrupar as espécies em categorias, de acordo com a sua origem, estatuto de conservação, forma de vida, entre outras. Para cada comunidade amostrada, as proporções de cada categoria têm que somar até à unidade, e cada espécie é exclusivamente atribuída a uma categoria, pelo que um modelo multinomial poderá ser adequado para as respetivas contagens. Ou seja, cada espetro de espécies [8] é representado por uma vetor aleatório multinomial. Analisámos o modelo multinomial com uma distribuição *a priori* Dirichlet [1,2,13], segundo uma abordagem bayesiana, com a aplicação WinBUGS [9], uma ferramenta utilizada em ecologia [1,2,10] e noutras áreas [11,12].

2 Métodos

Analisou-se a composição das comunidades vegetais em três situações: i) ao longo de um gradiente de altitude e entre a margem e a zona nuclear da vegetação [14]; ii) ao longo de trilhos, a diferentes altitudes e a diferentes distâncias perpendiculares ao trilho; e iii) ao longo de um gradiente de comunidades vegetais, sob diferentes níveis de intensidade de gestão e de distúrbio de origem humana [8].
Em todos os casos realizaram-se inventários florísticos, registou-se a abundância ou a percentagem de cobertura de cada uma das espécies, as quais foram categorizadas segundo o seu estatuto de conservação ou a sua forma de vida.
Em todos os casos utilizaram-se três cadeias de Markov e o modelo foi atualizado um número de vezes superior ao necessário para atingir a convergência, utilizando os critérios normalmente aceites [2,12]. Na estimativa dos parâmetros apenas se consideraram as amostras obtidas após a convergência. O DIC foi utilizado como medida de ajustamento dos modelos [15].

3 Resultados

Caso i) Independentemente da altitude, foram encontradas diferenças na composição florística entre a margem e o núcleo das comunidades. No que se refere à origem e à forma de vida, não se encontraram diferenças importantes na composição das comunidades ao longo do gradiente. Caso ii) Verificou-se uma maior contribuição de plantas endémicas na ilha das Flores, comparativamente

a São Miguel, e uma redução das espécies introduzidas com o aumento da altitude. Caso iii) Houve um efeito considerável do fator tipo de comunidade, e não do fator ilha. Verificou-se uma redução da componente endémica e nativa nos habitats sujeitos a um maior distúrbio antropogénico.

4 Discussão

A metodologia utilizada permitiu uma descrição e comparação eficaz da composição das comunidades estudadas. A produção de espetros de espécies, que podem ser representados graficamente, permite ilustrar a composição das comunidades. É de referir que, inicialmente, recorreu-se a testes mais convencionais que originaram resultados de difícil integração, ao contrário dos obtidos com a análise bayesiana.

Referências

[1] McCarthy, M.A. (2007). *Bayesian Methods for Ecology.* Cambridge University Press, Cambridge, 296 pp.

[2] King R., Morgan, B.J.T., Gimenez, O., Brooks, S.P. (2010). *Bayesian analysis for Population Ecology.* Chapman & Hall/CRC, Boca Raton, 442 pp.

[3] Boender, C.G.E., Rinnooy Kan, A.H.G. (1987). A multinomial Bayesian approach to the estimation of population and vocabulary size. *Biometrika* 74(4), 849–856.

[4] Vasko, K, Toivonen, H.T.T., Korhola, A. (2000). A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction. *Journal of Paleolimnology* 24(3), 243–250.

[5] Griffiths, T.L., Tenenbaum, J.B. (2002). Using Vocabulary Knowledge in Bayesian Multinomial Estimation. *Advances in Neural Information Processing Systems* 14, 1385–1392.

[6] Kazembe, L., Namangale, J. (2007). A Bayesian multinomial model to analyse spatial patterns of childhood co-morbidity in Malawi. *European Journal of Epidemiology* 22, 545–556.

[7] Calvo, E. (2009). The competitive road to proportional representation. *World Politics* 46(3), 167–174.

[8] Marcelino, J.A.P., Silva, L., Garcia, P.V., Weber, R., Soares, A.O. (2012). Using species spectra to evaluate plant community conservation value along a gradient of anthropogenic disturbance. *Environmental Monitoring and Assessment* (DOI: 10.1007/s10661-012-3019-9).

[9] Spiegelhalter, D.J., Thomas, A., Best, N.G. (2003). *WinBUGS User Manual, Version 1.4.* MCR Biostatistics Unit, Cambridge.

[10] Kéry, M. (2010). *Introduction to WinBUGS for Ecologists. A Bayesian Approach to Regression, ANOVA, Mixed Models and Related Analyses.* Academic Press, Elsevier, Burlington, 302 pp.

[11] Cowles, M.K. (2009). Review of WinBUGS 1.4. *The American Statistician* 58, 330–336.

[12] Christensen, R., Johnson, W., Branscum, A., Hanson, T.E. (2011). *Bayesian Ideas and Data Analysis. An Introduction for Scientists and Statisticians.* CRC Press, Taylor & Francis Group, Boca Raton, 498 pp.

[13] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London, 526 pp.

[14] Prestes, A.C.L., Magalhães, B.I., Xavier, E.R.N., Silva, L. (2012). Changes in plant community composition along an altitudinal gradient on a coastal protected area in the Azores. *Book of abstracts, FloraMac2012, Funchal, Portugal, 5-8 de Setembro de 2012* p. 111.

[15] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* 64, 583–639.

Comunicação Oral

Comparação dos desempenhos de semeadoras manuais por meio da distribuição triangular discreta generalizada

Silvio Sandoval Zocchi
Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, São Paulo, Brasil, sszocchi@gmail.com

Célestin C. Kokonendji
Universidade de Franche-Comté, Besançon, França, celestin.kokonendji@univ-fcomte.fr

Palavras–chave: Agricultura familiar, Agricultura de conservação, Distribuições discretas, Moda.

Resumo: Apresenta-se um nova metodologia para a avaliação e comparação de desempenhos de semeadoras manuais baseada na nova família de distribuições discretas triangulares apresentada por Kokonendji e Zocchi [1], ilustrada considerando os dados analisados por Molin et al [2].

1 Introdução

No Brasil, grande parte dos agricultores produzem diversas culturas em pequenas ou médias propriedades, com pouca tecnologia e mão-de-obra familiar. Dessas áreas, segundo Wall [3], adotam-se práticas de agricultura de conservação em aproximadamente 200.000 ha utilizando-se, para isso, semeadoras manuais ou tracionadas por animais. As primeiras têm sido produzidas por inúmeros fabricantes e exportadas especialmente para países africanos. Propõe-se aqui, uma nova metodologia de comparação baseada na nova família de distribuições discretas triangulares apresentada por Kokonendji e Zocchi [1], ilustrada considerando os dados analisados por Molin et al [2].

2 Material e métodos

Considere um experimento em que k semeadoras manuais são reguladas de modo a que caiam exatamente m sementes de cada vez (ou por golpe). Essas máquinas são, então, repetidamente utilizadas (n_j vezes, $j = 1, \ldots, k$) e ao final são computados os números de sementes que caíram por golpe da semeadora e suas frequências (ver Tabela 1 com dados parciais de Molin et al [2]).
Seja Y_j a variável número de sementes por golpe da j-ésima semeadora ($j = 1, 2, \ldots, k$) e considere-a como tendo distribuição triagular discreta generalizada com parâmetros m, a_1, a_2, h_{1j} e h_{2j} (ver definição e propriedades em Kokonendji e Zocchi [1]). Considere que y_{ij} é a i-ésima ($i = 1, 2, \ldots, n_j$) observação (assumidas independentes) de Y_j e denote $\mathbf{y} = (y_{ij})_{i,j}$, $\mathbf{h}_1 = [h_{11}, h_{12}, \ldots, h_{1k}]$ e $\mathbf{h}_2 = [h_{21}, h_{22}, \ldots, h_{2k}]$. Note que, na prática, recomenda-se considerar m fixo, no caso igual ao número ideal de sementes por golpe (no caso, $m = 2$) e escolher a_1 e a_2 de acordo com a amplitude de variação dos dados (no caso $a_1 = 2$ e $a_2 = 3$ ou $a_2 = 4$). Então, a partir das $n = n_1 + n_2 + \cdots + n_k$ fmp's individuais, o logaritmo da função de verossimilhança é dado por

$$l_{m,a_1,a_2}(\mathbf{h}_1, \mathbf{h}_2; \mathbf{y}) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \log f(y_{ij}; m, a_1, a_2, h_{1j}, h_{2j}) \tag{1}$$

com $f(y_{ij}; m, a_1, a_2, h_{1j}, h_{2j}) = \dfrac{\left[1-\left(\frac{m-y_{ij}}{a_1+1}\right)^{h_{1j}}\right]\mathbf{1}_{\aleph^*_{a_1,m}}(y_{ij}) + \left[1-\left(\frac{y_{ij}-m}{a_2+1}\right)^{h_{2j}}\right]\mathbf{1}_{\aleph_{m,a_2}}(y_{ij})}{(a_1+a_2+1) - (a_1+1)^{-h_{1j}} \sum\limits_{k=1}^{a_1} k^{h_{1j}} - (a_2+1)^{-h_{2j}} \sum\limits_{k=1}^{a_2} k^{h_{2j}}}$, em que $\aleph^*_{a_1,m} = \{m - a_1, \cdots, m-1\}$, $\aleph_{m,a_2} = \{m, \cdots, m + a_2\}$, e $\mathbf{1}_S(y)$ denota a função indicadora para qualquer

conjunto dado S que assume o valor 1 para $y \in S$ e 0, caso contrário. Logo, a comparação global do desempenho entre k semeadoras diferentes pode ser realizada por meio do teste de hipóteses

$$\begin{cases} H_0 : h_{11} = h_{12} = \cdots = h_{1k} = h_1 \ ; \ h_{21} = h_{22} = \cdots = h_{2k} = h_2 \\ H_1 : \text{ ao menos uma igualdade em } H_0 \text{ não vale.} \end{cases} \tag{2}$$

A partir de (1) pode-se, então, utilizar a estatística T para (2), dada por

$$T = 2\left\{l_{m,a_1,a_2}(\mathbf{h}_1, \mathbf{h}_2; \mathbf{y}) - l_{m,a_1,a_2}(h_1, h_2; \mathbf{y})\right\} \sim \chi^2_{2(k-1)}, \tag{3}$$

em que χ^2_d denota a distribuição qui-quadrado com d graus de liberadade. Note que acomparação das semeadoras duas a duas, caso seja de interesse, pode ser realizada considerando-se a hipótese de nulidade $H_0^* : h_{1j} = h_{1j^*} = h_1^* \ ; \ h_{2j} = h_{2j^*} = h_2^*$ para $j \neq j^*$.

Tabela 1: Frequências dos números de sementes de milho por golpe, para quatro semeadoras reguladas para cair 2 sementes por golpe, estimativas de máxima verossimilhança de h_{1j} e h_{2j} $(j = 1, \ldots, 4)$, seus erros padrões assintóticos (entre parênteses), valores dos logaritmos da função de verossimilhança considerando-se $m = 2$, $a_1 = 2$ e $a_2 = 3$, e resultados dos testes de hipóteses.

	N° de sementes por golpe								logaritmo da
Semeadora	0	1	2	3	4	5	$\widehat{h}_1$	$\widehat{h}_2$	verossimilhança
A	0	19	103	18	3	7	0,1280 (0,0338)	0,1220 (0,0277)	-159,22
B	2	26	70	48	4	0	0,2803 (0,0695)	0,3337 (0,0686)	-199,98
C	14	21	82	27	5	1	0,3319 (0,0782)	0,1824 (0,0410)	-198,22
D	7	19	82	42	0	0	0,2280 (0,0566)	0,2276 (0,0473)	-187,84
					Geral		0,2315 (0,0283)	0,2074 (0,0221)	-754,35
				Estatística			$T = 18{,}181$	valor-p = 0,0058	

Exemplo 2.1 *Para os dados apresentados na Tabela 1, rejeita-se a hipótese H_0 de que as semeadoras tenham desempenhos iguais considerando-se o nível de significância 5%. Do ponto de vista prático, valores de h_1 e h_2 menores correspondem a distribuições mais concentradas ao redor do alvo $(m = 2)$. Assim, nota-se, descritivamente, o melhor desempenho aparente da semeadora A.*

3 Considerações finais

A metodologia apresentada pode ser facilmente estendida a casos em que há, para cada fabricante, réplicas das semeadoras.
É comum agrupar os números de sementes em três classes: menor, igual ou maior do que o alvo. Esta prática, no entanto, não leva em consideração a distância estocástica entre cada ponto e o alvo, não sendo, portanto, recomendada.

Agradecimentos

Este trabalho foi parcialmente financiado pelo projeto FAPESP (Processo n° 2012/21389-0) e UMR 6623 CNRS-UFC.

Referências

[1] Kokonenji, C.C., Zocchi, S.S. (2010). Extensions of discrete triangular distribution and boundary bias in kernel estimation for discrete functions. *Statistics and Probability Letters* 80, 1655–1662.

[2] Molin, J.P., Menegatti, L.A.A., Gimenez, L.M. (2001). Avaliação do desempenho de semeadoras manuais. *Revista Brasileira de Engenharia Agrícola e Ambiental* 5, 339–343.

[3] Wall, P.C. (2007) Tailoring conservation agriculture to the needs of small farmers in developing countries: an analysis of issues. *Journal of Crop Improvement* 19, 137–155.

Comunicação Oral

Um modelo de Markov espaço-temporal não homogéneo para a ocorrência de Dengue

Marília Antunes
CEAUL e DEIO/FCUL, Universidade de Lisboa, Portugal, marilia.antunes@fc.ul.pt

Antónia Amaral Turkman
CEAUL e DEIO/FCUL, Universidade de Lisboa, Portugal, maturkman@fc.ul.pt

Kamil Feridun Turkman
CEAUL e DEIO/FCUL, Universidade de Lisboa, Portugal, kfturkman@fc.ul.pt

Marco A. Horta
Escola Nacional de Saúde Pública Sérgio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil, entomologica@gmail.com

Cristina Catita
Instituto Dom Luíz, Universidade de Lisboa, Portugal, cmcatita@fc.ul.pt

Palavras–chave: Biometria, Epidemiologia, Demografia, Estatística.

Resumo:
A Dengue é uma doença infecciosa causada por quatro tipos de vírus da dengue, transmitida pelo mosquito *Aedes*, vector da doença. Esta espécie de mosquito encontra-se principalmente durante e logo após um período chuvoso em áreas tropicais e subtropicais, predominantemente em áreas urbanas e semi-urbanas. O mosquito *Aedes* é particularmente bem sucedido na transmissão da Dengue uma vez que se alimenta quase exclusivamente em seres humanos, é activo durante o dia, e tem preferência por áreas urbanas, onde se reproduz em qualquer recipiente com água, como pratos de vasos ou o interior de pneus abandonados. A disseminação da Dengue é sensível às condições meteorológicas. Com efeito, sabe-se que as características biológicas do mosquito *Aedes* são directamente influenciadas pela temperatura e pela quantidade de precipitação. A Dengue está a expandir-se: em 2009, apareceu pela primeira vez em Cabo Verde, África, e, em 2012, na ilha da Madeira, Portugal. Dada a gravidade da doença, a detecção de um surto (e provável subsequente epidemia) e sua duração, são questões importantes. As autoridades de saúde consideram estar-se na presença de epidemia de Dengue quando ocorrem de mais de cinco casos por semana por 100000 habitantes.
O objectivo deste trabalho é detectar a ocorrência da doença através de um modelo de Markov espaço-temporal não homogéneo para a série de casos semanais de Dengue no período 2004 a 2009, em 121 sectores da cidade de Coronel Fabriciano - Minas Gerais, usando como covariáveis a temperatura máxima semanal e precipitação total, bem como algumas covariáveis sociodemográficas que caracterizam os sectores e que, sendo determinantes para a presença de mosquitos, influenciam a ocorrência e propagação da doença.

Agradecimentos

O trabalho de M. Antunes, A.A. Turkman e K.F. Turkman foi parcialmente financiado pelos projectos PTDC/MAT/118335/2010 e PEst-OE/MAT/UI0006/2011. O trabalho de M.A. Horta foi financiado por CAPES, FAPERJ.

Mesa-redonda

Os desafios actuais da Epidemiologia e a sua ligação à Biometria

Resumo: A Epidemiologia e a Estatística, e áreas afins, cruzam-se frequentemente enfrentando alguns desafios em comum e outros separadamente. Apesar do incentivo à multidisciplinaridade e ao diálogo entre as diversas áreas do saber, ainda permanecem algumas barreiras e dificuldades mesmo em áreas próximas.
Esta mesa-redonda tem como objectivos (i) discutir os alguns desafios enfrentados pela Epidemiologia, não só em Portugal e Galiza, mas também focando outros contextos europeus e africanos e (ii) explorar a ligação à Estatística e à Matemática que pode variar desde estatísticas e modelos simples a outros modelos com maior complexidade do ponto de vista matemático e computacional. A discussão terá uma duração aproximada de 80 minutos, contando com 3 intervenções orais de 12 minutos (que se espera que sejam mais interrogativas e em interacção com a plateia do que expositivas) e o restante tempo para o comentador e discussão geral.

Oradores:

- Paulo Ferrinho
  Unidade de Saúde Pública Internacional e Bioestatística & Centro de Malária e outras Doenças Tropicais, Instituto de Higiene e Medicina Tropical - Universidade Nova de Lisboa, *pferrinho@ihmt.unl.pt*

- Xurxo Hervada-Vidal
  Dirección Xeral de Saúde Pública e Planificación & Consellería de Sanidade & Xunta de Galicia, *xurxo.hervada.vidal@sergas.es*

- Helena Sofia Rodrigues
  Escola Superior de Ciências Empresariais do Instituto Politécnico de Viana do Castelo & Unidade de Investigação ALGORITMI da Universidade do Minho, *sofiarodrigues@esce.ipvc.pt*

Comentador:

- Vitor Rodrigues
  Faculdade de Medicina da Universidade de Coimbra, *vrodrigues@fmed.uc.pt*

Organizadores:

- Luzia Gonçalves
  Unidade de Saúde Pública Internacional e Bioestatística, IHMT-UNL & Centro de Estatística e Aplicações da Universidade de Lisboa - Portugal, *luziag@ihmt.unl.pt*

- María Esther López-Vizcaíno
  Instituto Galego de Estadística - Galicia, *esther.lopez@ige.eu*

Oral Communication

Assessing the limits of multiple imputation in tackling missing genotypes in a scenario of limited genetic information

Nuno Sepúlveda
London School of Hygiene and Tropical Medicine & Center of Statistics and Applications of University of Lisbon, nuno.sepulveda@lshtm.ac.uk

Alphaxard Manjurano
Kilimanjaro Christian Medical Centre, amanjurano@yahoo.co.uk

Taane Clark
London School of Hygiene and Tropical Medicine, taane.clark@lshtm.ac.uk

Eleanor Riley
London School of Hygiene and Tropical Medicine, eleanor.riley@lshtm.ac.uk

Chris Drakeley
London School of Hygiene and Tropical Medicine, chris.drakeley@lshtm.ac.uk

Keywords: Genetic association, Missing data, Multiple imputation, Full conditional specification.

Abstract: In genetic association studies the main goal is to identify the set of genetic markers highly correlated to a given phenotype. This task is usually performed under a generalised linear modelling approach, where each genetic marker and a set of confounders are considered as covariates, and the phenotype as the response variable. A difficulty that often occurs in practice is the presence of missing genotypes. The popular strategy of discarding individuals with incomplete data has proven to lead to biased estimates, decreased statistical power, and a waste of valuable and useful resources. Alternatively, multiple imputation has been proposed to deal with that problem [1]. The basic idea is to replace the missing genotypes by highly plausible values generated from the underlying linkage disequilibrium (LD) structure between markers. However, when the genetic markers are located far apart from each other, such as in candidate gene approach studies, the quality of imputed values may be compromised due to a weak LD structure. In this work we bring forward data from a candidate gene approach of malaria in Tanzania in order to assess the performance of multiple imputation in a setting where the genetic markers have limited information of each other. We focus our study on the α-thalassemia gene where a large proportion of individuals have missing genotype information due to sampling constraints, and where the genetic markers under analysis are not in strong LD with that locus. Using multiple imputation based on full conditional specification, where one can use genotype and phenotype data altogether [2], we perform a simulation study with the goal of estimating genotypic error rates and bias of different imputation alternatives. We finally show the effect of multiple imputation on the association strength and the underlying genetic effect of α-thalassemia gene on different malaria-related phenotypes.

Acknowledgments

Nuno Sepúlveda is partially funded by FCT, Portugal, through the project Pest-OE/MAT/UI0006/2011.

References

[1] Marchini, J., Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews in Genetics* 11, 499–511.

[2] Souverein, O.W., Zwinderman, A.H., Tanck, M.W.T. (2006). Multiple Imputation of missing genotype data for unrelated individuals. *Annals of Human Genetics* 70, 372–381.

Comunicação Oral

Análise bayesiana semiparamétrica de respostas binárias com uma covariável contínua sujeita a omissão informativa

Frederico Poleto
Instituto de Matemática e Estatística, Universidade de São Paulo, frederico@poleto.com
Carlos Daniel Paulino
Instituto Superior Técnico, Universidade Técnica de Lisboa, dpaulino@math.ist.utl.pt
Julio Singer
Instituto de Matemática e Estatística, Universidade de São Paulo, jmsinger@ime.usp.br
Geert Molenberghs
I-BioStat, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium, geert.molenberghs@uhasselt.be

Palavras–chave: Dados omissos; mecanismo de omissão informativa; MNAR; mistura por processo Dirichlet.

Resumo: Omissão em variáveis explicativas requer um modelo marginal para elas mesmo que o interesse se centre no modelo condicional das respostas dadas as covariáveis. Uma especificação incorreta de tais modelos marginais ou para o respetivo mecanismo de omissão pode conduzir a inferências enviesadas sobre os parâmetros de interesse. Em literatura já publicada usam-se para as covariáveis ou distribuições paramétricas com um mecanismo de omissão informativa (MNAR) ou modelos mais flexíveis semiparamétricos ou não paramétricos identificados com a suposição de omissão ao acaso (MAR).
Neste trabalho considera-se uma análise bayesiana de respostas binárias, combinando um modelo não paramétrico, baseado em mistura por um processo Dirichlet, para as covariáveis contínuas com um mecanismo de omissão MNAR. A ilustração é feita através de simulações e de um conjunto de dados reais sobre pacientes com suspeita de embolia pulmonar retirados de literatura médica.

Agradecimentos

Financiamento parcial de F.Poleto e J.Singer pela CAPES, FAPESP e CNPq, Brasil; C. D. Paulino pela FCT através do projeto Pest-OE/MAT/UI0006/2011; G. Molenberghs por IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

Comunicação Oral

Impacto de valores omissos em estudos epidemiológicos - Uma aplicação na modelação do Índice de Massa Corporal

Beatriz Preto Goulão
Faculdade de Ciências da Universidade de Lisboa, beatriz.goulao@gmail.com
Valeska Andreozzi
Centro de Estatística e Aplicações da Universidade de Lisboa, valeska.andreozzi@gmail.com
Patrícia de Zea Bermudez
Faculdade de Ciências da Universidade de Lisboa e CEAUL, pbermudez980@gmail.com

Palavras–chave: Modelação em Epidemiologia, Valores omissos, Índice de massa corporal.

Resumo: Os dados omissos são muito comuns em estudos clínicos e epidemiológicos [1]. Os métodos usados por diversos softwares para tratar este tipo de problema (por exemplo, a rejeição total dos registos com observações omissas nalguma das variáveis) não são satisfatórios [2]. De facto, se os valores omissos diferirem significativamente dos valores observados, então, não considerar os dados incompletos, poderá enviesar os resultados do estudo. Por outro lado, se as variáveis incluídas no modelo tiverem muitos dados omissos poderão ter de ser excluídas, diminuindo a eficiência das estimativas dos parâmetros dos modelos ajustados [3] e reduzindo também a potência de testes envolvidos [4]. Esta investigação pretende avaliar o impacto de valores omissos no estudo da associação do "número de anos de residência em Portugal" no "índice de massa corporal" (IMC) dos Imigrantes Brasileiros e Africanos no ano 2007.

1 Introdução

A utilização de imputação é apenas uma das formas possíveis de tratar o problema da existência de dados omissos. Efetivamente, a literatura é rica em métodos alternativos de tratamento desse tipo de problemas, os quais envolvem a utilização de todo o conjunto de dados - completo e omisso. Este trabalho centra-se na utilização de algumas técnicas de imputação. Estas dividem-se em: Imputação simples (IS) - Métodos dedutivos (usando respostas anteriores); Métodos determinísticos (por exemplo, substituindo os valores omissos pela média dos valores observados); Métodos estocásticos (*Hot-Deck*; Associação flexível; métodos regressivos com efeitos aleatórios); e Imputação múltipla (IM).
Os dados usados no presente estudo fazem parte de um estudo transversal intitulado **Acesso aos Cuidados de Saúde dos Imigrantes Brasileiros e Africanos**, tendo sido recolhidos em 2007. A partir destes dados, pretende-se estudar o impacto do número de anos de residência em Portugal no IMC desses imigrantes. As restantes variáveis (independentes) são: sexo, idade, escolaridade, número de refeições principais e intermédias e estado civil. A variável escolaridade é a que apresenta maior percentagem de dados omissos (7%). No presente estudo pretende-se avaliar o efeito da aplicação da IS na variável número de anos de escolaridade, enquanto variável independente, num modelo de regressão linear generalizado cuja variável resposta é o IMC.
Neste trabalho, os valores omissos da variável escolaridade serão imputados usando o *Predictive Mean Matching* (método de regressão com *Hot-Deck*) e a IS usando a mediana. Em seguida recorrer-se-á à imputação calculando um índice de propensão. Para tal, desenvolveram-se duas funções em R com o propósito de proceder à imputação dos valores da variável escolaridade, as quais foram desenvolvidas pela segunda autora deste trabalho.

2 Discussão e apresentação de alguns resultados

A tabela 1 apresenta a distribuição das variáveis na base de dados completa e a percentagem de dados omissos para cada variável. A tabela 2 mostra a distribuição da variável escolaridade, quando aplicados diferentes métodos de imputação: CC - Casos Completos, IS^a - IS por substituição pela mediana, IS^b - IS por *predictive mean matching* e IS^c - IS por índice de propensão.

Variável (% Dados omissos)	N = 1980
Sexo (0)	Feminino: 1058 (53,4%); Masculino: 922 (46,6%)
IMC (0)	25,07 ± 4,46 kg/m2
Idade (0)	35,1 ± 10,95 anos
Estado civil (0,3)	Solteiro: 744 (37,6%); Casado: 1100 (55,6%); Outro: 130 (6,6%)
Escolaridade (6,8)	9,21 ± 3,53 anos
Origem (0)	Africanos: 1080 (54,6%); Brasileiros: 705 (35,6%)
Anos de residência (1,2)	9,84 ± 8,14 anos
Nºrefeições principais (0,9)	<3: 537 (27,1 %); 3: 1426 (72,0%)
Nºrefeições intermédias (2,6)	0: 560 (28,2%); 1: 737 (37,2%); 2: 404 (20,4%); ≥ 3: 229 (11,6%)

Tabela 1: Caracterização da amostra e da percentagem de dados omissos por variável

	Escolaridade (CC)	Escolaridade (IS^a)	Escolaridade (IS^b)	Escolaridade (IS^c)
n (NAs)	1785 (135)	1904 (0)	1904 (0)	1904 (0)
Mínimo	1,000	1,000	1,000	1,000
1ºQuartil	7,000	7,000	6,803	6,000
Mediana	9,000	9,182	9,000	9,000
Média	9,207	9,182	9,098	9,091
3ºQuartil	19,000	12,000	12,000	12,000
Máximo	19,000	19,000	19,000	19,000

Tabela 2: Distribuição da variável escolaridade

Ajustaram-se modelos glm aos dados completos e aos três conjuntos de dados imputados. Verifica-se que as variáveis consideradas significativas são as mesmas nos 4 modelos. As variáveis explicativas importantes para explicar o IMC são a idade, o estado civil, os anos de permanência em Portugal e o número de snacks por dia. Este resultado evidencia o comportamento satisfatório das três metodologias de imputação. Porém, é necessário referir que o bom desempenho final se pode dever ao facto da % de valores omissos na variável escolaridade ser muito reduzida.

Agradecimentos

Este trabalho foi parcialmente financiado por PEst-OE/MAT/UI0006/2011 e PTDC/MAT/118335/2010. A recolha dos dados usados neste trabalho foi financiada pelo Alto Comissariado da Saúde e agradece à equipa do IMP responsável pelo projeto, aos entrevistadores, assim como aos participantes voluntários.

Referências

[1] Molenberghs, G., Kenward, M.G. (2007). *Missing Data in Clinical Studies*. England, Wiley.

[2] Gelman, A., Hill, J. (2007). *Data Analysis using Regression Multilevel/Hierarchical Models*. Cambridge University Press.

[3] Harrell, F. (2001). *Regression Modeling Strategies*. New York: Springer.

[4] Little, R., Rubin, D. (2002). *Statistical Analysis with Missing Data*. New York, Wiley.

Comunicação Oral

Combinação de valores de prova e valores de prova generalizados

Maria de Fátima Brilhante
Departamento de Matemática, Universidade dos Açores
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), fbrilhante@uac.pt
Dinis Pestana
FCUL/DEIO, Universidade de Lisboa
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)
Instituto de Investigação Científica Bento da Rocha Cabral, dinis.pestana@fc.ul.pt
Fernando Sequeira
FCUL/DEIO Universidade de Lisboa
Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), fjsequeira@fc.ul.pt

Palavras–chave: Ampliação computacional de amostras, uniformidade, contaminação, valores de prova generalizados

1 Introdução

Diz-se que rejeitar a hipótese nula é uma decisão forte, e mantê-la é uma decisão fraca. De facto, a manutenção da hipótese nula é, frequentemente, simples consequência da escassez de dados. A globalização dos resultados obtidos na cooperação (ou simples publicitação dos resultados) de diversas equipas de investigação pode permitir ultrapassar esse problema. Diversas formas de fazer uma síntese meta-analítica dos resultados podem, por outro lado, levar a uma conclusão global de resultados contraditórios.
A publicação em Medicina cada vez mais exige a disponibilização de dados, por forma a no futuro ser possível uma meta análise para determinar a magnitude de efeitos de diversos tratamentos. Por outro lado, ainda é comum nas publicações científicas apenas divulgar-se o valor de prova, havendo desde os anos 30 do século passado técnicas para os combinar.
Quando dispomos de valores de prova independentes, sob validade da hipótese nula temos uma amostra de uma população uniforme padrão. Usando a contaminação de uniforme padrão por uma Beta(2,1) ou por uma Beta(1,2), estudamos o efeito de ampliar computacionalmente amostras, com o intuito de incrementar a potência.

2 Valores de prova aleatórios e generalizados

A combinação de valores de prova, usando-os diretamente ou transformando-os (Pestana [3]) é simples, mas muitas vezes pouco eficaz, pois quando se tem como objetivo avaliar uma determinada hipótese nula, é muito pouco adequado assumir que é válida a correspondente hipótese combinada. Consequentemente, alguns dos valores de prova observados não devem ser uniformes. Daí que os conceitos de valores de prova generalizados e de valores de prova aleatórios para uma síntese possam ser úteis (veja-se Hartung *et al.* [1] e Kulinskaya *et al.* [2]).
Para além de discutirmos estes conceitos e exemplificar como se calculam valores de prova generalizados, analisamos a situação especial de considerar que em lugar de uniforme se tem uma combinação X_m de uniforme e Beta(2,1) ou Beta (1,2), mostrando que nesta classe de funções (m

é o coeficiente de mistura) se X_m e X_p forem independentes, então

$$\min\left\{\frac{X_m}{X_p}, \frac{1-X_m}{1-X_p}\right\} = X_{\frac{mp}{6}}$$

e consequentemente sempre mais próxima da uniforme X_0 (coincidindo com a uniforme se alguma das variáveis originais for uniforme).
Apresentamos também uma extensão para o caso de dependência auto-regressiva. Usamos aquele resultado para aumentar computacionalmente amostras de valores de prova, discutindo formas de o fazer que não piorem a potência do teste combinado.

3 Conclusão

Afinal a ampliação computacional das amostras pode não ser uma boa ideia, pois é possível que características estruturais do modelo – em particular, a entropia máxima da uniforme padrão, na classe das leis com suporte (0,1) – tenham um efeito perverso na potência dos testes que se pretende fazer.

Agradecimentos

Trabalho financiado por fundos nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito do projecto PEst-OE/MAT/UI0006/2011.

Referências

[1] Hartung, J., Knapp, G., Sinha, B.K. (2008). *Statistical Meta-Analysis with Applications*, Wiley, New York.

[2] Kulinskaya, E., Morgenthaler, S., Staudte, R.G. (2008). *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester.

[3] Pestana, D. (2011). Combining p-values. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, 1145–1147, Springer Verlag, New York.

Oral Communication

Adjusted p-values for SGoF multitesting procedure: Definition and properties

Irene Castro Conde
SiDOR Research Group, University of Vigo, Spain, *ircastro@alumnos.uvigo.es*
Jacobo de Uña Álvarez
SiDOR Research Group; Department of Statistics and OR, University of Vigo, *jacobo@uvigo.es*

Keywords: False discovery rate, Family-wise error rate, Multiple comparisons, Dependent tests.

Summary: In the paper Carvajal-Rodríguez, de Uña-Álvarez and Rolán-Álvarez (2009) [1] a new multitest correction named SGoF (from Sequential Goodness-of-Fit) was introduced; this method was extended to possibly correlated tests in de Uña-Álvarez (2012) [4] , who introduced the Beta-Binomial SGoF (BB-SGoF) procedure. Both SGoF and BB-SGoF have the property of increasing their statistical power when increasing the number of tests, which is very useful in omic sciences: genomics, proteomics, etc. because they typically involve the simultaneous testing of hundreds or thousands of hypotheses. Statistical properties and false discovery rate and power levels in practical settings for SGoF-type strategies were further investigated in de Uña-Álvarez (2011, 2012) [3, 4] and Castro-Conde and de Uña-Álvarez (2013) [2]. In this talk we introduce adjusted p-values for SGoF method and we investigate their properties. Time permitting, adjusted p-values for BB-SGoF will be presented and discussed too.

1 Introduction

Nowadays, there exist many statistical inference problems in areas such genomic and protenomics which involve the simultaneous test of hundreds or thousands of null hypotheses producing as a result a number of signicant p-values or effects. Moreover, these hypotheses may have complex and unknown dependence structure among themselves. See e.g. Dudoit and Van der Laan (2008) [5] for an introduction to this area.
One of the main problems in multiple hypotheses testing is that, if one does not take the multiplicity of tests into account, then the probability that some of the true null hypotheses are rejected may be overly large. So, in the multitesting setting, a specific procedure for deciding which null hypotheses should be rejected is needed.
In the paper Carvajal-Rodríguez, de Uña-Álvarez and Rolán-Álvarez [1] a new multitest correction named SGoF (from Sequential Goodness-of-Fit) was introduced; this method was extended to possibly correlated tests in de Uña-Álvarez [4], who introduced the Beta-Binomial SGoF (BB-SGoF) procedure. Both SGoF and BB-SGoF procedures use the p-values to decide which hypotheses are to be rejected. Let us define formally the concept of unadjusted p-value in this multitest setting (Dudoit and Van der Laan [5]).

Definition 1.1 (Unadjusted p-value) *The unadjusted p-value p_i, for the single test of null hypothesis H_{0i}, is defined as*

$$p_i \equiv inf\{\alpha \in [0,1] : \text{ Reject } H_{0i} \text{ at single test nominal level } \alpha\}, i = 1, ..., n.$$

That is, the unadjusted p-value p_i, for null hypothesis H_{0i}, is the *smallest nominal Type I error level* of the *single hypothesis testing procedure* at which one would reject H_{0i}. The smaller the unadjusted p-value p_i, the stronger evidence against the corresponding null hypothesis H_{0i}.
Specifically, null hypothesis H_{0i} is rejected at single test nominal Type I error level α if $p_i \leq \alpha$. But, as we have said, we need an a specific procedure for deciding which null hypotheses should

be rejected taking in account multiplicity so we can't use unadjusted p-values. For solving this problem, we introduce the adjusted p-values. The notion of p-value extends directly to multiple testing problems as follows:

Definition 1.2 (Adjusted p-value) *The adjusted p-value $\tilde{p}_i$, for the test of null hypothesis H_{0i}, is defined as*

$$\tilde{p}_i \equiv inf\{\alpha \in [0,1] : \text{ Reject } H_{0i} \text{ at multiple comparison procedure nominal level } \alpha\}$$

That is, the adjusted p-value $\tilde{p}_i$, for null hypothesis H_{0i}, is the smallest nominal Type I error level of the multiple hypothesis testing procedure at which one would reject H_{0i}.
Obtaining adjusted p-values consists of determining for each comparison the smallest level of significance that would result in the comparison being declared significant. As in single hypothesis tests, the smaller the adjusted p-value $\tilde{p}_i$, the stronger the evidence against the corresponding null hypothesis H_{0i}.
In this talk we introduce adjusted p-values for SGoF method and we investigate their properties. Time permitting, adjusted p-values for BB-SGoF will be presented and discussed too. These method have the particularity that they depends on two parameters, α which is the nominal level of the metatest and controls the FWER weakly and γ which represents the initial significance threshold for significant p-values. In practice, to calculate the adjusted p-values we will considerate that $\alpha = \gamma$ so we consider there exists only one parameter playing the role of nominal level.

Acknowledgments

Work supported by the Grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science and Innovation. Support from the Xunta de Galicia Grant 10PXIB300068PR is also acknowledged.

References

[1] Carvajal-Rodríguez, A., de Uña-Álvarez, J., Rolán- Álvarez, E. (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics* 10, 209.

[2] Castro-Conde, I., de Uña-Álvarez, J. (2013). Performance of Beta-Binomial SGoF multitesting method for dependent gene expression levels: a simulation study. *Proceedings of Bioinformatics 2013 International Conference on Bioinformatics Models, Methods and Algorithms* (Pedro Fernandes, Jordi Sole-Casals, Ana Fred and Hugo Gamboa Eds.), SciTePress.

[3] de Uña-Álvarez, J .(2011). On the statistical properties of SGoF multitesting method. *Statistical Applications in Genetics and Molecular Biology* 10(1), 18.

[4] de Uña-Álvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. *Statistical Applications in Genetics and Molecular Biology* 11(3), 14.

[5] Dudoit, S., Van der Laan, M.J. (2008). *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Comunicação Oral

Influência do nível socioeconómico da região no risco de fratura do fémur proximal

Carla Oliveira
INEB - Instituto de Engenharia Biomédica, Universidade do Porto; Departamento de Epidemiologia Clínica, Medicina Preditiva e Saúde Pública, Faculdade de Medicina, Universidade do Porto; ISPUP - Instituto de Saúde Pública da Universidade do Porto, *carlaoliveir@gmail.com*

Denisa Mendonça
ICBAS - Instituto de Ciências Biomédicas Abel Salazar e ISPUP - Instituto de Saúde Pública da Universidade do Porto, *dvmendon@icbas.up.pt*

Maria de Fátima de Pina
INEB - Departamento de Epidemiologia Clínica, Medicina Preditiva e Saúde Pública, Faculdade de Medicina, ICBAS - Instituto de Ciências Biomédicas Abel Salazar e ISPUP - Instituto de Saúde Pública da Universidade do Porto, *fpina@med.up.pt*

Palavras–chave: Epidemiologia, Sistemas de Informação Geográfica (GIS), Estatística, Biometria.

1 Introdução

As fraturas do fémur proximal (FFP) são consideradas um grave problema de saúde pública e podem estar relacionadas com as condições sociais e económicas das populações. Alguns estudos têm mostrado que regiões mais desfavorecidas, com maior percentagem de indivíduos com má alimentação, atividade física insuficiente e deficiente acesso ao sistema tendem a ter taxas de incidência de FFP mais elevadas.

2 Objetivo

Avaliar a associação entre a incidência de FFP e o nível socioeconómico (SES) por município, em Portugal.

3 Material e métodos

Foram seleccionados os internamentos (de 2000 a 2008) do Registo Nacional de Altas Hospitalares, com diagnóstico de FFP (código 820.xx de ICD9.CM), causadas por impacto de baixa energia, em indivíduos com idade superior a 49 anos. Foram excluídos os casos de cancro do osso e readmissões para pós-tratamento. As taxas ajustadas para a idade e as especificas por idades das FFP foram calculadas com base nos Censos da população e estimativas populacionais para os anos intercensitários, para os 278 municípios de Portugal Continental. A padronização foi pelo método direto e a população europeia foi utilizada como população padrão. Os concelhos foram classificados em seis níveis SES sendo os desfavorecidos - nível 1 e os mais favorecidos - nível 6. Para estimar o risco relativo (RR) de FFP, e o correspondente intervalo de confiança de 95% (IC 95%), de acordo com categorias de SES utilizou-se o modelo aditivo generalizado de regressão de Poisson, tendo como variável dependente o número de FFP e estratificando por sexo. Os resultados foram ajustados para grupo etário (50-54, 55-59, ..., 84-85, 85+), ruralidade e ano civil do evento e mapeados usando Sistema de Informação Geográfica (GIS).

4 Resultados

Este estudo incluiu 76.542 FFP, sendo 59.269 (77,4%) mulheres. A idade média de ocorrência de FFP é significativamente ($p < 0,05$) mais elevada para as mulheres (81,1 ± 8,52 anos) do que para os homens (78,1 ± 10,09 anos). Ajustando para grupo etário, ruralidade e ano civil, o risco de fratura tende a diminuir à medida que aumenta o SES (Mulheres: RR_{1vs2} 0,97 (IC95%: 0,92–1,03); RR_{1vs3} 0,95 (0,91–1,00); RR_{1vs4} 0,94 (0,91–0,97); RR_{1vs5} 0,87 (0,83–0,91) e Homens: RR_{1vs2} 1,00 (0,91–1,1); RR_{1vs3} 1,00 (0,92–1,07); RR_{1vs4} 0,97 (0,91–1,02); RR_{1vs5} 0,88 (0,82–0,94)), exceto no SES mais alto (Mulheres: RR_{1vs6} 1,07 (1,04–1,11)) e Homens: RR_{1vs6} 1,02 (0,97–1,08)). O padrão descrito acima não é tão claro quando analisado a taxa específica para a idade. Foi encontrada uma interação significativa entre a idade e nível SES para ambos os sexos. A associação entre a incidência de FFP e o SES (re-agregadas em três grupos) foi inversa em idades mais jovens (Mulheres com idade <60: RR_{1vs2} 0,92 (0,73–1,17); RR_{1vs3} 0,92 (0,72–1,18); Homens com idade <60: RR_{1vs2} 0,84 (0,74–0,96); RR_{1vs3} 0,78 (0,68–0,9)) e direta em idades mais avançadas (Mulheres com idade ≥60: RR_{1vs2} 0,95 (0,89–1,01); RR_{1vs3} 1,08 (1,01–1,15); Homens ≥70: RR_{1vs2} 1,01 (0,94–1,1); RR_{1vs3} 1,09 (1–1,18)), sendo este facto mais evidente nos homens.

5 Conclusão

Existe alguma evidência de que a incidência de FFP pode estar relacionada com o contexto socioeconómico da região. A razão para o risco de FFP ser menor para as classes 2-5 quando comparadas com a classe 1 pode dever-se a um estilo de vida menos saudável, consequentemente, maior exposição a fatores de risco para a osteoporose. No entanto esta relação não se observa no SES mais alto, a razão pode prender-se por questões de sedentarismo nos indivíduos que vivem nas categorias extremas do SES e pelo tipo de atividade profissional. O trabalho manual é um fator protetor por promover uma maior atividade física. Assim, as áreas com SES mais alto podem ter maior proporção de indivíduos com ocupações "colarinho branco", que exigem menos esforço físico. A interacção encontrada, entre o SES e a idade, pode dever-se a desigualdades nas ações de saúde para prevenção da osteoporose e das fraturas, no entanto mais estudos precisam ser desenvolvidos para esclarecer os padrões encontrados.

Agradecimentos

Este trabalho foi financiado pelo projeto PTDC/SAU-EPI/115254/2009.

Referências

[1] Moreira, J., Pina, M.F. (2008). *Fracturas Osteoporóticas do Colo do Fémur em Portugal e seus Determinantes Socioeconómicos*. Porto: Faculdade de Engenharia da Universidade do Porto.

Comunicação Oral

Aplicação de modelos de equações estruturais na avaliação da qualidade de vida em pessoas com doenças metabólicas

Estela Vilhena
Instituto Politécnico do Cávado e do Ave, Barcelos, e ICBAS e ISPUP, Universidade do Porto, evilhena@ipca.pt

José Luís Pais Ribeiro
Faculdade de Psicologia e Ciências da Educação da Universidade do Porto e UIPES, Lisboa, jlpr@fpce.up.pt

Luísa Pedro
Unidade de Investigação em Psicologia e Saúde (UIPES) e Escola Superior de Tecnologia da Saúde de Lisboa (ESTeSL), IPL, luisa.pedro@estesl.ipl.pt

Isabel Silva
Universidade Fernando Pessoa, Porto, isabels@ufp.pt

Rute F. Meneses
Universidade Fernando Pessoa, Porto, rmeneses@ufp.edu.pt

Helena Cardoso
UMIB/ICBAS, Universidade do Porto, e Hospital Santo António/CHP, helenacardoso@icbas.up.pt

António Martins da Silva
UMIB/ICBAS, Universidade do Porto, e Hospital Santo António/CHP, ams@icbas.up.pt

Denisa Mendonça
Instituto de Ciências Biomédicas Abel Salazar (ICBAS) e Instituto de Saúde Pública da Universidade do Porto (ISPUP), Universidade do Porto, dvmendon@icbas.up.pt

Palavras–chave: Doenças metabólicas, Modelos de equações estruturais, Qualidade de vida.

Resumo: Com o objetivo de avaliar a Qualidade de Vida em pessoas com doenças metabólicas, foram aplicados Modelos de Equações Estruturais. O modelo teórico a testar pressupôs: 1) a perceção do estigma, o otimismo e o afeto positivo são preditores da qualidade de vida; 2) o otimismo exerce um efeito medidor entre a perceção do estigma/afeto positivo e a qualidade de vida. Controlando para variáveis sócio-demográficas e clínicas, verificou-se um impacto positivo de uma menor perceção do estigma, do otimismo e do afeto positivo na qualidade de vida destes doentes. Os resultados mostraram ainda que o otimismo exerce um efeito mediador entre o afeto positivo/perceção do estigma e o bem-estar geral e entre a perceção de estigma e a saúde mental.

1 Introdução

Os Modelos de Equações Estruturais - SEM (*Structural Equation Modeling*) [1] são considerados uma metodologia de análise estatística multivariada que permitem representar, estimar e testar modelos teóricos, que envolvem diversas relações entre variáveis (observadas e latentes), de forma a compreender os padrões de correlação/covariância entre estas. A Qualidade de Vida (QV) é um constructo composto por um número de fatores que contribuem para o bem-estar de um indivíduo e para o ajustamento a uma determinada doença. As doenças metabólicas, entre outras, são consideradas um dos problemas de saúde pública relevante. Após o diagnóstico, estes doentes são obrigados a viver com as limitações impostas pelas suas condições. Este trabalho, teve como objetivo avaliar um modelo hipotético, que consistiu na análise do impacto da perceção do estigma, otimismo e do

afeto positivo nas componentes da QV (bem-estar geral, saúde física e mental) e simultaneamente, na avaliação do efeito mediador [2] do otimismo, num grupo de doentes metabólicos.

2 Métodos

O estudo, transversal, incluiu uma amostra de 365 doentes, recrutados nos principais hospitais de Portugal. Os critérios de inclusão foram: idade superior ou igual a 18 anos, nível de escolaridade superior ou igual a 6 anos, diagnóstico de pelo menos 3 anos, vida estabilizada e não apresentar distúrbios psiquiátricos. Foi aplicado um questionário que incluía um conjunto de variáveis sócio-demgráficas, perceção de estigma, otimismo, afeto positivo e as componentes da QV. Os SEM foram aplicados para testar a qualidade do modelo teórico hipotético. As relações entre as variáveis foram estimadas usando o método de Máxima Verosimilhança. Para testar a adequação do modelo foram usados os índices CFI - *Comparative Fit Index* e o RMSEA - *Root Mean Error Approximation*. A análise foi efetuada usando o softwae EQS 6.1.

3 Resultados

Os resultados obtidos revelaram um ajustamento razoável do modelo, CFI=0.9 e RMSEA=0.053. Controlando para variáveis sócio-demográficas e clínicas, verificou-se que uma menor perceção do estigma, o otimismo e o afeto positivo exercem um impacto simultâneo, estatisticamente significativo, positivo, na qualidade de vida destes doentes. Os resultados evidenciaram também um efeito mediador do otimismo entre o afeto positivo/perceção de estigma e o bem-estar geral e entre a perceção de estigma e a saúde mental.

4 Conclusão

Os Modelos de Equações Estruturais são hoje considerados uma componente importante na análise multivariada, aplicados para abordar questões científicas complexas, que exigem a análise de múltiplas relações simultâneas, com múltiplas variáveis, incluindo latentes. Testa um conjunto de relacionamentos lineares através de um modelo que operacionaliza a teoria. Os resultados encontrados sugerem que uma menor perceção do estigma, uma atitude otimista, mais ativa e entusiástica podem facilitar o doente à sua nova condição de vida, atitudes, que por sua vez contribuirão para uma menor qualidade de vida.

Referências

[1] Tabachnick, B., Fidell, L. (1996). *Using Multivariate Statistics*. Third edition. New York: HarperCollins College Publishers.

[2] Baron, R.M., Kenny, D.A. (1986). The Moderator Mediator Variable Distinction in Social Psychological-Research - Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology* 51(6), 1173–1182.

Comunicación Oral

Las tablas de mortalidad de Galicia y la Región Norte de Portugal empleando el software R

María Martín Vila
Instituto Galego de Estatística, maria.martin@ige.eu

M. Esther Calvo Ocampo
Instituto Galego de Estatística, esther.calvo@ige.eu

Solmary Silveira Calviño
Instituto Galego de Estatística, solmary.silveira@ige.eu

Gael Naveira Barbeito
Dirección Xeral de Innovación e Xestión da Saúde Pública, Consellería de Sanidade, Estatistica.SP@sergas.es

M. Isolina Santiago Pérez
Dirección Xeral de Innovación e Xestión da Saúde Pública, Consellería de Sanidade, soly.santiago.perez@sergas.es

Carlos Iglesias Patiño
Instituto Galego de Estatística, carlos.iglesias@ige.eu

M. Esther López Vizcaíno
Instituto Galego de Estatística, esther.lopez@ige.eu

Palavras clave: Tablas de mortalidad, Demografía, Esperanza de vida, R.

Resumen: En este trabajo se utiliza el software R para el cálculo de las tablas de mortalidad de Galicia y la Región Norte de Portugal.

1 Introducción

La mortalidad es un fenómeno demográfico inevitable, porque todo individuo tendrá que experimentarlo; irrepetible, porque cada persona sólo lo experimentará una vez; e irreversible, porque supone un cambio de estado, de vivo a muerto, sin posibilidad de retorno al anterior. Estas tres características distinguen a la mortalidad con respecto a los otros fenómenos demográficos (natalidad y movimientos migratorios) y permiten análisis específicos, como por ejemplo la elaboración de tablas de mortalidad a partir de las probabilidades de ocurrencia del fenómeno. Estas tablas evitan la influencia de la estructura por edades de la población estudiada, lo que permite establecer comparaciones entre distintos grupos; otra de sus utilidades es que a partir de ellas se obtiene un indicador muy utilizado, la esperanza de vida (EV), tanto como indicador del nivel de envejecimiento de una población como indicador general de desarrollo de un colectivo.
Las tablas de mortalidad de período representan el fenómeno de la mortalidad de una cohorte ficticia como resultado de aplicar las condiciones de mortalidad vigentes en el momento de su elaboración a un colectivo hipotético a lo largo de su vida.
Hoy en día existen herramientas informáticas que permiten el cálculo de las tablas de mortalidad de manera sencilla. Entre ellas se encuentra EPIDAT, programa de libre distribución desarrollado por la Dirección Xeral de Innovación e Xestión da Saúde Pública de la Consellería de Sanidade (Xunta de Galicia) y la Organización Panamericana de la Salud (OPS-OMS). Este programa permite la elaboración de tablas de mortalidad completas y abreviadas por sexos para un período y espacio determinados.

En algunas ocasiones, como sucede a menudo en el ámbito de la estadística oficial, es necesario elaborar tablas de mortalidad para distintos espacios y distintos períodos simultáneamente. En estos casos, el uso de EPIDAT para su cálculo puede resultar una tarea tediosa, ya que es necesario replicar la ejecución para cada período y espacio. La utilización del software estadístico R permite superar esta desventaja y flexibilizar datos que en Epidat no son opcionales.

2 Objetivo

El objetivo de este trabajo es, por un lado, el desarrollo de un programa informático que permita la construcción simultánea de tablas de mortalidad completas o abreviadas para distintos períodos, espacios geográficos y condiciones iniciales; y por otro, el análisis y comparación de las condiciones de mortalidad de Galicia y del Norte de Portugal.

3 Material y métodos

Los datos de población necesarios para elaborar las tablas de mortalidad de Galicia proceden del Padrón Municipal de Habitantes (con fecha de referencia 1 de enero) y en el caso de Portugal se consideraron las *Estimativa de Populacao Residente por genero e idade ano a ano* pero del año inmediatamente anterior, pues la fecha de referencia de estas estimaciones es diciembre. Los datos de defunciones y nacimientos que se utilizaron son, en el caso de Galicia, los del Movimiento natural de la población publicados por el IGE y en el caso del Norte de Portugal los *Óbitos por Local de residência, Sexo e Idade* y *Nados-vivos por Local de residência da mãe (NUTS 2002) e Sexo* proporcionados por el INE.
Se programó una función en R que, a partir de estos datos, permite calcular todas las funciones de la tabla de mortalidad para los dos espacios geográficos, para los diferentes períodos de tiempo y para todas las variantes del sexo en una única ejecución.
La función programada permite introducir como argumento el promedio de años vividos de los que mueren con edad cumplida x (a_x) considerando por defecto los coeficientes de Coale-Demeny [1] del tipo de región correspondiente. La tasa de mortalidad infantil también se puede introducir como argumento en la función; de no hacerlo se supone desconocida y el programa la aproxima con los datos introducidos.
Con este programa se elaboraron las tablas de mortalidad completas y abreviadas para las regiones de Galicia y Norte de Portugal desde 2000, utilizando para cada período los datos agregados de tres años consecutivos. Para su cálculo no se introdujeron como argumentos ni los promedios de años vividos de los que mueren con edad cumplida x (a_x) ni las tasas de mortalidad infantil, por lo que se utilizaron las opciones que el programa elaborado en R proporciona por defecto.

4 Resultados

Tanto en Galicia como en el Norte de Portugal, la esperanza de vida al nacer presenta una tendencia creciente en el período estudiado. En todos los trienios, la EV al nacer de Galicia es superior a la del Norte de Portugal, aunque la diferencia tiende a reducirse. Este patrón de tendencia creciente con valores más altos en Galicia se observa en la EV a todas las edades, y tanto en hombres como en mujeres.

Referencias

[1] Coale, A.J., Demeny, P., Vaughan, B. (1983). Models of mortality and age composition. *En: Coale et al., editores. Regional model life tables and stable populations. 2 ed.*, 1–8.

Comunicação Oral

Modelação e projeção da incidência de cancro colo-rectal e do estômago no Sul de Portugal

Ricardo São João
ESTGS - Instituto Politécnico de Santarém & CEAUL, ricardo.sjoao@esg.ipsantarem.pt
Ana Papoila
FCM - Universidade Nova de Lisboa & CEAUL, ana.papoila@fcm.unl.pt
Bruno de Sousa
FPCE - Universidade de Coimbra & CEAUL, bruno.desousa@fpce.uc.pt
Ana Miranda
Registo Oncológico Regional (ROR-Sul), amiranda@ipolisboa.min-saude.pt

Palavras–chave: incidência, cancro, projeção, modelação.

1 Introdução

O efeito de mudanças estruturais da população modifica substancialmente o número esperado de casos futuros de cancro. Em Portugal, e na maior parte dos países desenvolvidos, a lenta diminuição da natalidade a par de um aumento da esperança de vida e de um ligeiro decréscimo na taxa de mortalidade terá importantes implicações nos próximos 50 anos[12]. De facto, estes factores terão um impacto nas doenças crónicas e degenerativas e também no aumento da procura de serviços de saúde e de outros recursos afectos a esta "nova futura" realidade. A generalidade dos cancros afecta sobretudo pessoas idosas. Estimativas mundiais indicam que cerca de 45% dos cancros ocorrem em idades superiores a 65 anos. Segundo a nossa pirâmide etária, este facto releva-se no mínimo preocupante.
Dentre as várias neoplasias existentes, merecem destaque as do cólon, do recto e do estômago. Na Europa, o cancro colo rectal é o segundo cancro mais comum sendo também o segundo mais mortífero[1]. A nível mundial, o cancro do estômago é o quarto mais comum sendo a segunda neoplasia com maior mortalidade[4, 5]. A par do cenário europeu e mundial, Portugal, através do registo de cancro de base populacional da região Sul (ROR-Sul), registou taxas elevadas de incidência e mortalidade associadas a estas neoplasias[13].
Na epidemiologia do cancro, estudos que permitam identificar, descrever e projectar taxas de incidência de neoplasias têm vindo a assumir um maior relevo[2]. A etiologia e os factores de risco em grande parte das neoplasias não são conhecidos, sendo relevantes para o seu estudo as seguintes variáveis temporais:idade do doente à data do diagnóstico; o período (data do diagnóstico) e a coorte de nascimento[3]. A projecção de tendências permite, então, quantificar o impacto do cancro no futuro e planear programas de controlo e prevenção[6]. Os modelos idade-período-coorte (Age-Period-Cohort models)[7, 8, 9, 10] têm sido utilizados como instrumento descritivo no comportamento de taxas de incidência permitindo a sua projecção num curto horizonte temporal. Por outro lado, modelos mais parcimoniosos, lineares no tempo mas não forçosamente nos parâmetros[11], revelam um bom ajuste quer na projecção de taxas de incidência de cancro quer do número absoluto de casos[6].

2 Objectivos

Com base no registo ROR-Sul, o presente estudo tem como objectivos:

- A modelação da incidência de cancro do cólon, recto e do estômago no período 1998-2006, com recurso aos modelos APC;
- A projecção da taxa de incidência no período 2007-2010 para as referidas neoplasias com base nos modelos APC e em modelos lineares, log-lineares e não lineares.

3 Resultados

Na modelação da incidência do cancro do cólon, estômago e recto, no período em estudo, foram considerados 15598, 9770 e 7380 casos respectivamente, com idades entre os 25 e 84 anos. No cancro do estômago, o modelo *age-drift* revelou o melhor ajuste face aos restantes, indiciando uma descida nas taxas de incidência de 3.8% por 10^5 pessoas-ano. No cancro do cólon, o modelo idade-coorte revelou o melhor ajuste; sendo portanto o efeito de coorte (*"generational influence"*) o mais significativo. No cancro do recto, o modelo APC revelou o melhor ajuste sendo os efeitos de período (*"secular influence"*) e coorte significativos. Para o período 2007-2010, as projecções apontam para um decréscimo nas taxas de incidência do cancro do estômago em oposição ao cancro do cólon e numa estabilização do cancro do recto. Em relação à incidência, verificou-se um desvio de 10% nas projecções face aos dados de 2007, disponibilizados pelo registo ROR-Sul no corrente ano.

Agradecimentos

Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projecto PEst-OE/MAT/UI0006/2011

Referências

[1] Ferlay, J., Autier, P., Boniol, M., Heanue, M., Colombet, M., Boyle, P. (2007). Estimates of the cancer incidence and mortality in Europe in 2006. *Annals of Oncology* 18(3), 581–592.

[2] Elaporte, R. (1993). How to improve monitoring and forecasting of disease patterns. *British Medical Journal* 307, 1573–1574.

[3] Hakulinen, T. (1996). The future cancer burden as a study subject. *Acta Oncologica* 35, 665–670.

[4] Parkin, D.M., Bray, F.I., Devesa, S.S. (2001). Cancer burden in the year 2000. The global picture. *European Journal of Cancer* 37 Suppl 8: S4–S66.

[5] Parkin, D.M. (2004). International variation. *Oncogene* 23, 6329–6340.

[6] Hakulinen, T., Dyba, T. (1994). Precision of incidence predictions based on Poisson distributed observations. *Statistics in Medicine* 13, 1513–1523.

[7] Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in Medicine* 26, 3018–3045.

[8] Carstensen, B., Keiding, N. (2005). Age-period-cohort models: statistical inference in the Lexis diagram. *Unpublished manuscript at www.biostat.ku.dk/ bxc/APC.*

[9] Clayton, D., Schifflers, E. (1987). Models for temporal variation in cancer rates. I: Age-period and age-cohort. *Statistics in Medicine* 6, 449–467.

[10] Clayton, D., Schifflers, E. (1987). Models for temporal variation in cancer rates. II: Age-period and age-cohort. *Statistics in Medicine* 6, 469–481.

[11] Dyba, T., Hakulinen, T., Päivärinta, L. (1997). A simple non-linear model in incidence prediction. *Statistics in Medicine* 16, 2297–2309.

[12] Ferlay, J., Bray, F., Pisani, P., Parkin, D.M. (2004). *GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide.* IARC-Scientific Publications, Lyon.

[13] Miranda, A., Pereira, A., Mesquita, C., Bastos, J., Ribeiro, M., Lunet, N. (2008). *Os 10 tumores mais frequentes na população portuguesa adulta, na região sul de Portugal no período 2000-2001.* Editado por ROR Sul.

Comunicação Oral

Técnicas de meta-análise na estimação de uma taxa de prevalência

João Paulo Martins
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon, jpmartins@ipleiria.pt

Miguel Felgueiras
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon,
CIIC — Computer Science and Communications Research Centre of Polytechnic Institute of Leiria, mfelg@ipleiria.pt

Rui Santos
School of Technology and Management, Polytechnic Institute of Leiria,
CEAUL — Center of Statistics and Applications of University of Lisbon, rui.santos@ipleiria.pt

Palavras–chave: Testes conjuntos, Meta-análise, Taxa de prevalência.

Resumo: Os testes conjuntos têm como principal objetivo a poupança de recursos, quando se pretende efetuar a identificação de todos os indivíduos infetados numa determinada população. Para este fim, seguindo a metodologia de Dorfman [2] e suas extensões (cf. Finucan [3], Kim et al. [6]), a população é dividida em grupos de n indivíduos aos quais é efetuado um teste conjunto. Se o resultado do teste conjunto for negativo, então nenhum elemento do grupo está infetado. No caso do teste ser positivo, então pelo menos um dos membros do grupo está infetado e, como tal, será necessário efetuar testes individuais para a identificação dos indivíduos infetados. Dependendo da taxa de prevalência p da infeção, a dimensão ótima n de cada grupo será aquela que minimiza o custo esperado dos testes uma vez que o custo de misturar de forma homogénea as amostras é, em geral, negligível (cf. Liu et al. [7]). Assim, a aplicação desta metodologia permite poupar muitos recursos (Dorfman [2]), havendo aplicações em diversas áreas, tais como o controlo de qualidade ou as análises clínicas (Boswell et al. [1]).

A investigação de metodologias que envolvam testes a amostras compostas tem sido muito ativa desde o trabalho seminal de Dorfman (Hughes-Oliver [5]). Além disso, o recurso a amostras compostas tem sido utilizado em problemas de classificação (identificação dos indivíduos infetados) bem como em problemas de estimação de uma taxa de prevalência p. Neste caso, a execução de testes individuais é apenas opcional na medida em que não é necessário identificar em concreto quais os indivíduos infetados. Aliás, o uso exclusivo de amostras compostas pode ser útil para garantir o anonimato dos resultados. Além disso, existem situações em que os estimadores obtidos pela utilização de testes conjuntos gozam de melhores caraterísticas que os tradicionais, obtidos com testes individuais (cf. Garner et al. [4], Loyer [8]).

O planeamento de uma metodologia de testes conjuntos pode ser delineado de diferentes formas (Kim et al. [6]), devido ao facto do resultado do teste poder estar errado. A sensibilidade e especificidade de um teste medem a qualidade do seu resultado. Em particular, a sensibilidade de um teste tende a decrescer como aumento do número de amostras misturadas. A escolha de uma determinada metodologia depende da quantidade de amostra de cada indivíduo disponível, da sensibilidade e especificidade do teste e do custo do processo em causa (Liu et al. [7]).

Note-se que os diferentes laboratórios produzem estimativas para a taxa de prevalência de uma dada infeção utilizando diferentes amostras bem como metodologias distintas, quer as baseadas em testes individuais, quer as baseadas em testes conjuntos (e, neste último caso, provavelmente utilizando diferentes dimensões para cada grupo). A análise à estimação da taxa de prevalência incluirá os resultados provenientes de testes conjuntos qualitativos, onde só é testada a presença

versus ausência de uma dada caraterística, e os provenientes de testes conjuntos quantitativos, onde se pretende testar se algum valor individual é superior (ou inferior) a um determinado patamar. Nesta apresentação, sugere-se a utilização de técnicas de meta-análise na combinação de diferentes estimativas de uma mesma taxa de prevalência. As possíveis diferenças entre as metodologias usadas nos diferentes estudos serão tidas em conta através do recurso a covariáveis. O peso atribuído a cada estimativa tem em conta a sensibilidade e a especificidade, tal como foi definido em Santos et al. [10], do processo subjacente. Para a obtenção de uma estimativa global é discutido o algoritmo recentemente apresentado por Martins et al. [9].

Agradecimentos

Este trabalho foi financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projeto PEst-OE/MAT/UI0006/2011 e pelo Instituto Politécnico de Leiria.

Referências

[1] Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P. (1996). Annotated bibliography of composite sampling, Part A: 1936–92. *Environmental and Ecological Statistics* 3, 1–50.

[2] Dorfman, R. (1943). The detection of defective members in large populations, *Annals of Mathematical Statistics* 14, 436–440.

[3] Finucan, H.M. (1964). The blood testing problem. *Applied Statistics* 13, 43–50.

[4] Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R. (1989). Probability Estimation With Sample Compositing Techniques. *Journal of Official Statistics* 5, 365–374.

[5] Hughes-Oliver, J.M. (2006). Pooling experiments for blood screening and drug discovery. In Dean A., Lewis S. (eds.): *Screening – Methods for Experimentation in Industry, Drug Discovery, and Genetics*, 48–68, Springer.

[6] Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C. (2007). Comparison of group testing algorithms for case identification in the presence of testing errors. *Biometrics* 63, 1152–1163.

[7] Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, T.C. (2011). Cost analysis in choosing group size when group testing for Potato virus Y in the presence of classification errors. *Annals of Applied Biology* 159, 491–502.

[8] Loyer, M.W. (1983). Bad Probability, Good Statistics, and Group Testing for Binomial Estimation. *The American Statistician* 37, 57–59.

[9] Martins, J.P., Felgueiras, M., Santos, R. (2013). Meta-analysis techniques applied in prevalence rate estimation. *Discussiones Mathematicae* (aceite para publicação).

[10] Santos, R., Pestana, D., Martins, J.P. (2013). Extensions of Dorfman's Theory. Oliveira P.E. et al. (eds.): *Recent Developments in Modeling and Applications in Statistics, Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies*, 179–189, Springer.

Método para decidir entre un evento compuesto o una de sus componentes como variable principal en un ensayo clínico. Plataforma web para facilitarlo

Guadalupe Gómez Melis
Universitat Politècnica de Catalunya, lupe.gomez@upc.edu

Palavras clave: Composite endpoints, Eficiencia relativa asintótica, Ensayos clínicos, Tiki Wiki.

1 Introducción

Al comparar dos grupos de tratamiento en un ensayo clínico aleatorizado (RCT) basándose en una variable principal definida como el tiempo hasta la realización de un evento, es común usar eventos compuestos. Por ejemplo, en pacientes con enfermedad de la arteria coronaria se usa la unión entre infarto de miocardio no fatal y la revascularización; en pacientes con cáncer inoperable se utiliza la progresión clínica o la muerte. Los principales argumentos clínicos para el uso de un evento compuesto (*composite endpoint, CE*) son que el *CE* podría (a) reflejar más adecuadamente los eventos más importantes asociados con la enfermedad que se está tratando, (b) aumentar la frecuencia de ocurrencia de la variable principal con la esperanza de aumentar la potencia para detectar diferencias entre uno y otro tratamiento y, (c) cuando la variable principal no es observable debido a un evento fatal (por ejemplo, la progresión de la enfermedad y la muerte), el uso de la supervivencia libre de progresión como variable principal en el RCT, evita problemas de interpretación asociados con el tiempo hasta la progresión de la enfermedad.
Ha habido un considerable debate sobre el uso de variables combinadas en cuanto a la interpretación de los resultados, pero muy poca discusión estadística sobre las ventajas y desventajas de la utilización de un *CE* frente a un subconjunto de sus componentes como variable principal en el ensayo clínico.
Varios autores han advertido contra el uso de variables combinadas basadas en motivos no estadísticos (entre otros Ferreira-González et al, 2007). En primer lugar, el uso de un *CE* en el que sus diferentes componentes son de importancia clínica muy diferente, puede ser problemático debido a que el tratamiento podría afectar beneficiosamente a solamente una de ellas, y por lo tanto, dar una impresión equivocada. En segundo lugar, un efecto de tratamiento (ya sea real o estadísticamente significativo) en el *CE* no implica necesariamente un efecto sobre cada componente (Gómez, 2011). Y, por último, la falta de efecto del tratamiento (real o estadísticamente significativo) en una componente no implica que no haya ningún efecto del tratamiento sobre las otras componentes. Dos argumentos estadísticos han sido propuestos abogando por el uso de un *CE*: primero, el uso del *CE* podría aumentar la eficiencia estadística, es decir, conducir a una prueba más potente para la eficacia del tratamiento. En segundo lugar, el uso de un *CE* puede reducir los problemas de multiplicidad asociados si se comparasen los grupos de tratamiento con respecto a cada uno de las componentes.
Gómez y Lagakos (2013) desarrollan una metodología estadística basada en la eficiencia relativa asintótica (ARE) que permite cuantificar el aumento de eficiencia derivado de ampliar una variable relevante para el estudio $\mathcal{E}_1$ (la muerte, por ejemplo) a un *CE*, $\mathcal{E}_*$, unión de $\mathcal{E}_1$ y $\mathcal{E}_2$ donde $\mathcal{E}_2$ es una variable secundaria (progresión clínica, por ejemplo). Su método basa la comparación entre los dos tratamentos en la prueba *logrank* y compara la eficiencia que se obtendria si se usase el logrank para T_1: tiempo hasta la aparición de $\mathcal{E}_1$, versus el logrank para $T_* = \text{mín}\{T_1, T_2\}$: tiempo hasta la aparición de $\mathcal{E}_*$. La expresión de la ARE depende de un pequeño conjunto de parámetros interpretables y en cierta medida anticipables similares a los que se usan de forma habitual para

el cálculo del tamaño muestral: (1) la frecuencia anticipada de $\mathcal{E}_1$ y $\mathcal{E}_2$ en el grupo de tratamiento control, (2) la magnitud del efecto del tratamiento previsto en $\mathcal{E}_1$ y $\mathcal{E}_2$ mediante las correspondientes razones de riesgo, y (3) el grado de dependencia entre T_1, y T_2. La cópula bivariada de Frank se utiliza para unir T_1 y T_2 y se explora la robustez del método con respecto a la elección de la cópula (Plana y Gómez, 2013).
Se demuestra asimismo que la interpretación habitual de la eficiencia relativa asintótica (ARE) como la relación recíproca de los tamaños de muestra necesarios para dos pruebas de hipótesis establecidas para las mismas hipótesis nula y alternativa, para alcanzar la mismo potencia al mismo nivel de significación, puede extenderse a dos pruebas de hipótesis para dos hipótesis nulas diferentes H_0 : efecto nulo del tratamiento medido en T_1 y H_0^*: efecto nulo del tratamiento medido en T_* y las correspondientes hipótesis alternativas (Gómez y Gómez-Mateu, 2013).

2 Plataforma Web y aplicación

Con el objetivo de transferir este método y ponerlo a disposición de investigadores, estamos construyendo una plataforma web flexible, interactiva y amigable. La construcción de esta plataforma utiliza el software de Tiki Wiki CMS / Goupware (Tightly Integrated Knowledge Infrastructure), escrito bajo la licencia GNU/LGPL que permite a los usuarios compartir y modificar su contenido libremente. Este software se ha ampliado con un plug-in que permite programar todos los cálculos en R y ejecutarlos internamente a través del servidor de red. El usuario no necesita conocimientos de R (ni disponer de su instalación) para obtener respuestas que le ayudarán a decidir qué variable principal es más eficiente entre un conjunto de posibles eventos.
Un estudio de caso en el área de investigación cardiovascular se utiliza para ilustrar el uso de la plataforma (Gómez, Gómez-Mateu y Dafni, 2013). Se verá cómo se pueden combinar diferentes *CE* y cómo se elige la variable combinada más eficaz como variable primaria del RCT.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto MTM2012-38067-C02-01.

Referencias

[1] Ferreira-González, I., Permanyer-Miralda, G., Busse, J.W., Bryant, D.M., Montori, V.M., Alonso-Coello, P., Walter, S.D., Guyatt, G.H. (2007). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology* 60, 651–657.

[2] Gómez, G. (2011). Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment. *Proceedings of the 26th International Workshop on Statistical Modelling*, 14–21.

[3] Gómez, G., Lagakos, S. (2013). Statistical considerations when using a composite endpoint for comparing treatment groups. *Statistics in Medicine* 32, 719–738.

[4] Gómez, G., Gómez-Mateu, M., Dafni, U. (2013). Informed choice of composite endpoints in cardiovascular trials. *Submitted*.

[5] Gómez, G., Gómez-Mateu, M. (2013). The asymptotic relative efficiency and the ratio of sample sizes when testing two different null hypotheses. *Submitted*.

[6] Plana, O., Gómez, G. (2013). Does the selection of the copula matter? Effect on the recommendation for the primary endpoint in a Randomized Clinical Trial. *Submitted*.

Índice de Autores

Afonso, Fernando, 8
Agresti, Alan, 59
Almeida, Catarina, 2
Alves, Marta, 34
Amaral Turkman, Maria Antónia, 101, 104, 112
Andreozzi, Valeska, 62, 116
Antunes, Luís, 4, 5, 57
Antunes, Marília, 101, 112
Aparício, Mariana, 7
Araújo, Artur, 25
Assunção, Ricardo, 8
Athayde, Emilia, 77
Azevedo, Assis, 77
Azevedo, Cecília, 73, 85

Bailey, Trevor, 69
Bento, Maria José, 4, 5
Bispo, Regina, 66
Blanco-Garcia, Francisco, 45
Bolfarine, Heleno, 97
Borges, Ana, 60
Braga, Ana Cristina, 29, 83
Brandão, Rui, 8
Brilhante, Maria de Fátima, 118
Brito, Irene, 41
Buckland, Stephen T., 66

Cadarso-Suárez, Carmen, 10
Caeiro, Elsa, 8
Calaza-Díaz, Laura, 10
Calegário, Natalino, 55
Calvo-Ocampo, María Esther, 126
Cao-Abad, Ricardo, 79
Cardoso, Helena, 124
Carreira, Jose M., 49
Carvalho, Lídia Raquel de, 37
Carvalho, Maria Laene Moreira de, 87
Carvalho, Maria Lucília, 71, 106
Casáis-Martínez, Luis, 102
Castro, Clara, 32
Castro, Luís, 60
Castro-Conde, Irene, 120
Catita, Cristina, 112
Cavenague, Hayala, 81
Cicogna, Antônio Carlos, 67
Clark, Taane, 114
Correia Gomes, Carla, 69
Correia, Elisete, 12
Cortes-Pereira, Estefania, 45
Costa, Antonio, 18
Costa, Joaquim Pinto da, 33
Costa, Paulo, 73
Cotos-Yáñez, Tomás R., 13
Cotter, Jorge, 15
Crujeiras-Casais, Rosa María, 16
Cunha, Pedro G., 15

De Uña-Álvarez, Jacobo, 25, 120
De Zea Bermudez, Patrícia, 89, 116
Dean, Charmaine, 93, 94
Diamantino, Fernanda, 30
Drakeley, Chris, 114

Eberhardt, Diogo Néia, 47
Economou, Theodoros, 69
Ersel, Derya, 91

Felgueiras, Miguel, 52, 130
Feng, Cindy, 93
Fernández-Albalat, María, 19
Fernández-López, Carlos, 45
Fernández-Moreno, Mercedes, 45
Fernández-Tajes, Juan, 45
Ferrinho, Paulo, 113
Figueiredo, Ivone, 106
Figueiredo, Lígia, 83
Frade, Hugo, 83
Freitas, Adelaide, 105

Gaio, Ana Rita, 33
Gayoso-Diz, Pilar, 73
Ginzo-Villamayor, María José, 16
Gomes, Fernando, 85
Gómez-Melis, Guadalupe, 132
Gonçalves, Luzia, 89, 113
Gonçalves, Patrícia, 41
Goulão, Beatriz Preto, 116
Grosso, Ana Rita, 2
Gude-Sampedro, Francisco, 10
Guindani, Michele, 36
Gunay, Suleyman, 91

Hervada-Vidal, Xurxo, 113
Horta, Marco A., 112
Howland, Brett, 66

Iglesias-Patiño, Carlos, 126
Iglesias-Pérez, María del Carmen, 73, 102

Jato, Victoria, 13
Juarez-Colunga, Elizabeth, 94

Kokonendji, Célestin C., 110

Leandro, Roseli Aparecida, 47
Leite, Isabel Cristina Costa, 51, 87
Lobo Pereira, José A., 18
Lopes, Maria Luísa, 8
López-Alvarez, Francisco de Asis, 49
López-Calviño, Beatriz, 19, 45, 79
López-Ratón, Mónica, 10

López-Vizcaíno, María Esther, 16, 113, 126
Louzada Neto, Francisco, 76, 81

Manjurano, Alphaxard, 114
Marques, Bruno, 21
Marques, Tiago A., 66
Martins, João Paulo, 52, 130
Martins, Natália da Silva, 23
Martins, Rui, 62
Martín-Vila, María, 126
Martínez-González, Miguel, 102
Martínez-Mera, Juan A., 49
Mascarenhas, Andreia, 34
Meira Machado, Luís, 25, 27, 32
Meira, Dália, 83
Mendes, Luzia, 18
Mendonça, Denisa, 4, 5, 64, 122, 124
Meneses, Rute F., 124
Menezes, Raquel, 39
Miranda, Ana, 128
Mirante, Ângela Cristina da Fonseca, 51
Mischan, Martha Maria, 37
Molenberghs, Geert, 115
Monteiro, Bebiana, 12
Moreira, Ana, 27
Moura, Teresa, 106
Mourão, Filipa, 29
Müller, Peter, 36
Muñiz-García, Javier, 73

Nakamura, Luiz Ricardo, 47
Narciso, Sara, 30
Natário, Isabel, 95, 106
Naveira-Barbeito, Gael, 126
Neto, Teresa, 34
Niza Ribeiro, João, 69

Oliveira, Alexandra, 33
Oliveira, Carla, 122
Oliveira, Joana, 32
Oliveira, Manuela, 8
Oliveira, Pedro, 15, 29
Oliveira, Teresa, 18
Oreiro-Villar, Natividad, 45
Ortega, Edwin Moises Marcos, 43

Padovani, Carlos Roberto, 67
Papoila, Ana Luísa, 34, 128
Pardo-Landrove, María José, 19
Paulino, Carlos Daniel, 1, 36, 115
Pedro, Luísa, 124
Perdona, Gleici, 81
Pereira, Glauber Márcio Silveira, 37
Pérez-González, Ana, 13
Pértega-Díaz, Sonia, 19, 45, 79
Pestana, Dinis, 118
Piairo, Helena, 39
Pina, Maria de Fátima de, 73, 122
Pinto, António E., 7
Pinto, Renan Mercuri, 67
Pita-Fernández, Salvador, 19, 79
Poleto, Frederico Zanqueta, 1, 115

Queirós, Celine, 41

Rachet, Bernard, 4, 5
Ramires, Thiago Gentil, 43
Ramos, Sandra, 101
Rego-Perez, Ignacio, 45
Reis, Ana Batalha, 30
Relaño-Fernández, Sara, 45
Ribeiro Junior, Paulo Justiniano, 23, 47
Ribeiro, José Luís Pais, 124
Righetto, Ana Julia, 47
Riley, Eleanor, 114
Roca-Pardiñas, Javier, 49, 99, 102
Rodrigues, Anabela, 64
Rodrigues, Helena Sofia, 113
Rodrigues, Vitor, 113
Rodríguez-Rajo, Francisco J., 13
Roque, Sara, 105

Sáfadi, Thelma, 87
Sagitov, Serik, 75
Santana, Azly Santos Amorim de, 51
Santana, Tânia Jussara Silva, 51
Santiago-Pérez, María Isolina, 16, 126
Santos, Bruno, 97
Santos, Rui, 52, 130
Santos, Soane Mota dos, 54
São João, Ricardo, 128
Scalon, João Domingos, 51
Seoane-Pillado, Teresa, 19, 45, 79
Sepúlveda, Nuno, 114
Sequeira, Fernando, 118
Serra, Maria Conceição, 75
Sestelo, Marta, 99, 102
Shrubsall, Sílvia, 95
Silva, António Martins da, 124
Silva, Giovani L., 7, 62, 94
Silva, Isabel, 124
Silva, Luís, 108
Silva-Fortes, Carina, 104
Silveira-Calviño, Solmary, 126
Singer, Julio da Motta, 1, 54, 115
Soto-Hermida, Angel, 45
Sousa, Bruno de, 128
Sousa, Inês, 39, 57, 60, 64
Sousa, Lisete, 2, 104
Sousa, Nuno, 15
Strzalkowska-Kominiak, Ewa, 79
Suárez-Lorenzo, José Manuel, 19
Subtil, Ana, 89

Tahoces, Pablo G., 49
Teixeira, Laetitia, 64

Turkman, Kamil Feridun, 112

Valdés-Cuadrado, Luis, 10
Vazquez-Mosquera, Eugenia, 45
Veloso, Romulo Barbosa, 55
Vila, Isabel, 15
Vilela, Alice, 12
Vilhena, Estela, 124
Villanueva, Nora M., 99
Vinhas, Ana Margarida, 57
Virella, Daniel, 34

Zanardo, Cleyton, 81
Zocchi, Silvio Sandoval, 110

NOTAS