Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral — Outono de 2009

Modelos Econométricos

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@fc.ul.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 1000 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

Este Boletim tem o apoio da FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

PRÉMIO ESTATÍSTICO JÚNIOR 2010

Candidaturas até
28 DE MAIO DE 2010

CONTACTOS

Sociedade Portuguesa de Estatística
Bloco C6, Piso 4 – Campo Grande
1749-016 Lisboa

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@fc.ul.pt

Com o apoio:

PORTO EDITORA

Editorial

… **"consolidado"**…

1. O Boletim SPE Outono de 2009 contém uma extensa secção "SPE e a Comunidade" desenvolvida em torno da utilização do software R com livre direito de utilização e que pode ser obtido através de http://www.r-project.org. Neste contexto se inserem os artigos de divulgação publicados neste Boletim onde os diversos autores convidados assumem esta informação conhecida pelo leitor. Assim, os textos especializados que se publicam - de nível médio / avançado - estão dirigidos a utilizadores habituais de R. Para uma introdução aprofundada ao tema pode ser, por exemplo, utilizado o livro de Luís Torgo que anunciamos na secção Livros Publicados. Os vários "assuntos R" abordados, obviamente, formam uma perspectiva "criada" pelo editor. Dada a extensão do campo de utilizadores, são bem vindas novas contribuições que serão incluídas em edições futuras daquela secção do Boletim.
2. Quando, em 2006, terminei os seis anos de direcção na SPE, tinha uma certeza: o "objectivo Boletim" sempre tão desejado no conteúdo programático, por diversas vicissitudes, não tinha sido completamente atingido. Reflecti sobre este assunto e, como é óbvio, concluí sobre a extrema importância desta publicação. Uma feliz (co)incidência em conversa com a direcção seguinte, em particular com o seu presidente, fez avançar e consolidar a iniciativa que se foi concretizando, com a preciosa ajuda dos sócios, dos leitores e, principalmente, dos autores que, generosamente, têm colaborado em cada edição. O editor é (apenas) aquele que estabelece as condições que levam a uma continuidade desta importante publicação da SPE. É o produtor. Como editor, pouco mais faço do que assegurar a publicação desta edição SPE que se pode inserir num projecto didáctico e de divulgação científica. De facto, o valor de cada Boletim é criado pelos seus autores! Como editor, responsabilizo-me por cumprir aquela função no âmbito desse projecto: aquele por conta de quem corre a actividade de produção, (também) do dinamismo e da força e vigor da SPE…

Desde há 3 anos temos evoluído. Graças à generosidade e ao empenho de um grande número de colaboradores, esta publicação pode-se considerar consolidada.
3. Sabemos que o Boletim SPE publica artigos científicos de divulgação. Porque não tem um painel de avaliadores, o Boletim SPE insere-se "apenas" no seu âmbito, isto é, uma publicação periódica sobre "o estado da arte" na Estatística.

É, além disso, como se sabe, uma publicação onde a actualidade e unidade nos temas é um objectivo. Passo a passo fomos percorrendo o caminho que nos trouxe até aqui, consolidando as várias secções e criando outras… É o que faremos a partir do próximo Boletim Primavera de 2010 onde incluiremos uma nova secção; com "novos doutores" a escreverem artigos de divulgação alguns anos após o doutoramento. Este novo projecto, chamemos-lhe Secção Pós - doc, além de consolidação científica também divulga eventuais "novos caminhos" no pós-doutoramento.
4. Na última Assembleia Geral ordinária da SPE, realizada em 17 de Março, várias intervenções dos sócios (uma vez mais!) salientaram a importância e a necessidade de um amplo debate de temas e questões fulcrais em qualquer das vertentes da actividade estatística e, em particular, sobre o Boletim. As propostas de trabalho incidem fundamentalmente sobre dois campos. Por um lado pode (deve!) fazer-se "o uso dos meios electrónicos" que permitem um "debate em tempo real" e, por outro, obviamente, "reflectir no Boletim". São propostas que devem merecer o maior apoio da comunidade. E, talvez desde logo, pela análise e crítica dessas propostas de intervenção. O Boletim está - e sempre esteve, como tem sido dito - aberto e desejoso dessa participação.

O Boletim tem bases para suportar novos desafios!
Consolidado está!

O tema central do próximo *Boletim* será *"Data mining" - Prospecção (Estatística) de Dados?*.

Mensagem do Presidente

Caros Colegas:

Realizou-se em Sesimbra, de 30 de Setembro a 3 de Outubro de 2009, o nosso XVII Congresso Anual. A Comissão Organizadora do XVII Congresso, presidida pelo nosso Colega João Tiago Mexia, cujo jubileu celebrámos recentemente, e vice-presidida pelo nosso Colega Manuel Esquível, é credora do nosso reconhecimento pelo excelente trabalho desenvolvido e pelo sucesso do Congresso, que bateu o record (186) de comunicações apresentadas, sinal do crescimento do trabalho produzido pela comunidade estatística portuguesa. Agradecemos também à Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, nossa anfitriã, aos membros da Comissão Executiva e Científica, aos presidentes das sessões de trabalho, aos oradores convidados, aos autores das comunicações e a todos os 273 participantes. Esta reunião deve o seu sucesso à união de esforços de todos estes protagonistas.

As Actas do XVI Congresso, realizado em Vila Real em 2008, foram distribuídas em Sesimbra, sendo esta uma boa ocasião para felicitar a Comissão Organizadora, da UTAD, presidida pela Colega Irene Oliveira, que assim concluiu a sua bem sucedida missão. Aos autores, aos avaliadores e aos editores, os nossos agradecimentos pelo seu contributo para este relevante marco da actividade científica desenvolvida em Portugal na área da Estatística.

O Prémio SPE 2009 desta vez premiou um trabalho científico da autoria de dois jovens investigadores, Miguel de Carvalho e Paulo Canas Rodrigues, o qual foi apresentado no XVII Congresso. Tivemos de novo os Prémios Estatístico Júnior, a que se candidataram um número record de trabalhos escolares dos ensinos básico e secundário, com a coordenação do Colega Russell Alpízar-Jara e da Comissão Especializada de Educação e o apoio da Porto Editora. Aos candidatos e aos júris de ambos os prémios, o nosso agradecimento. No Congresso foi também lançada uma obra comemorativa dos 10 anos do ALEA, comemoração a que a SPE assim se associou com grande júbilo.

E, não tarda muito, teremos o XVIII Congresso. A Faculdade de Ciências e Tecnologia da Universidade de Coimbra e o Instituto Politécnico de Viseu são os anfitriões, sendo a Comissão Organizadora presidida pelo Colega Paulo Eduardo Oliveira, com a Colega Carla Henriques como braço direito. É com grande satisfação que vemos, pela primeira vez na história dos nossos Congressos, uma instituição do ensino superior politécnico como co-organizadora. Estão desde já convidados para S. Pedro do Sul, local onde se realiza o Congresso.

Através de e-mails e da nossa página web, estais certamente a par do que se vai passando na SPE e na comunidade estatística. Limito-me, assim, a dar-vos apenas informação sobre o progresso de dois temas pendentes. Um é o do pagamento das quotas por débito em conta, introduzido pela primeira vez este ano como modalidade alternativa, mais cómoda, de pagamento. Esta modalidade está disponível para todos os interessados e já houve um número razoável de aderentes, aos quais pedimos compreensão por algum atraso no débito em conta relativamente à data prevista (tratando-se do ano de arranque, houve que desenvolver e validar procedimentos informáticos complexos que demoraram mais que o previsto). Outro tema é o acordo de várias sociedades estatísticas europeias, entre as quais a SPE, com a Springer para uma nova série internacional de publicações, que incluem as Actas dos Congressos. Já há um acordo de princípio sobre as questões principais, faltando a Springer apresentar o texto do contrato, o que se espera suceda muito em breve.

E é altura de me despedir até ao próximo Boletim, o da Primavera de 2010, com uma saudação muito cordial.

• XVII Congresso SPE

Sesimbra: À pesca de estatísticos e dos riscos que eles estudam

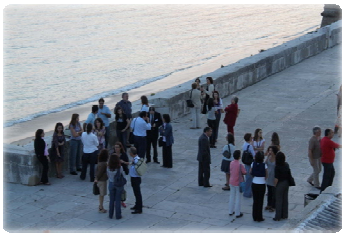
No dia XXX do mês IX do ano MMIX teve início o 17º Congresso da Sociedade Portuguesa de Estatística na bela cidade de Cempsibriga (burgo da tribo de Sesim), actualmente conhecida por Sesimbra. É uma *vila* que viveu ao longo dos anos da pesca, mas que está cada vez mais dependente economicamente do turismo, em que primam os bons restaurantes com dieta à base de peixe (o sonho de qualquer criança e de muitos adultos!). Do tempo dos mouros e mouras pouco resta para além de alguns calhaus no castelo, pois D. Afonso Henriques com a ajuda dos cruzados francos em 1165 achou por bem trazer para a Coroa Portuguesa todos os vinhedos de Palmela e das Terras do Sado! Bom gosto nunca fez mal a ninguém e fica sempre bem na Realeza!

A responsabilidade da organização do Congresso esteve a cargo da FCT / UNL, designadamente da Comissão Organizadora Local, presidida pelos colegas João Tiago Mexia e Manuel Esquível e composta ainda pelos colegas Frederico Caeiro, Isabel Natário e João Lita da Silva. De entre as múltiplas escolhas em Sesimbra, a organização do Congresso brindou-nos com a melhor: Sesimbra Hotel & SPA. Um local único, uma vista única, e na verdade um tempo único.

Os trabalhos começaram com o mini-curso sobre *Análise de Sobrevivência* leccionado pelas colegas Cristina Rocha e Ana Luísa Papoila, contando com mais de 70 participantes. O excelente livro do mini-curso bem como a exposição permitiu melhor lidar com os efeitos competitivos entre o Congresso e a praia em frente! Passou a doer menos deixar o terraço para assistir a mais uma sequência de apresentações. Sim, quem sobreviveu ao mini-curso – em que a cada 4 palavras uma era "morte", "morreu", "sobreviveu", "sobrevivência", "exposto" – ficou preparado para tudo: curado ou mesmo imune!

Deu-se em seguida a Abertura Oficial dos trabalhos do Congresso, tendo como oradores principais os Presidentes do Congresso e da Comissão Organizadora. Houve ainda a apresentação do livro *Um mundo para conhecer os números* que comemora o 10º aniversário do projecto ALEA (www.alea.pt). Seguiu-se a primeira Sessão Plenária intitulada *The comparison of maximum likelihood and PLS estimators for structural equation modeling. A simulation with customer satisfaction data* da responsabilidade do colega Manuel Vilares da UNL, que deu o mote para o arranque dos trabalhos.

Seguidamente teve lugar a atribuição do Prémio SPE 2009, tendo este ano sido contemplado o trabalho dos colegas Miguel de Carvalho e Paulo Canas Rodrigues com o título *Métodos de Imputação Recorrente: Análise Espectral Singular com Valores Omissos*. Este foi apresentado pelo colega Miguel de Carvalho.



Na Fortaleza de Santiago em Sesimbra, a Organização brindou-nos com uma prova de Moscatel! O Moscatel era um verdadeiro néctar dos deuses, mas os doces eram igualmente magníficos!

Passado este primeiro dia ainda de aquecimento, os trabalhos do Congresso começaram na manhã seguinte a todo o vapor! Leia-se às 9:00, quando o sol que entra pelo terraço dos quartos convida a mais um belo dia de praia ….

O número *record* de congressistas inscritos ascendeu aos 273, estando o seu contributo científico consubstanciado em 118 comunicações orais e 69 comunicações em formato poster. Isto para além dos 5 Conferencistas Convidados: um colega Russo, um colega Polaco, um colega do Reino Unido e dois colegas Portugueses. A Figura abaixo representa a diversidade de palavras-chave encontradas nos resumos das comunicações orais e em formato poster. Uma imagem neste caso contém (quase) mil palavras e fala de per si.

Palavras-chave nas comunicações orais e em poster - "Um tesauro visual" por Olga Bessa Mendes, *Livro de Programa e Resumos do XVII Congresso SPE, p. 229*.

"Morte", "vida", "sobrevivência" e "exposição" voltaram à frequência usual neste tipo de eventos. É verdade que a primeira apresentação da primeira sessão paralela do primeiro dia, prometia a cura (condicional a ter sobrevivido ao mini-curso)!
Seleccionar entre as estimulantes cinco sessões paralelas exigiu algum planeamento de modo a evitar-se que o momento da decisão não fosse posterior ao término das mesmas. Felizmente, as salas eram próximas e os moderadores foram férreos na boa sincronização das sessões paralelas. Apesar da oferta em geral suplantar a procura, houve casos em que a sala falhou o critério de suficiência. Foi o caso da apresentação do colega João Branco *Suficiência: tanto barulho para quê?*, pois se o colega tivesse que assistir de pé à apresentação também faria barulho, não???

A sessão de posters foi localizada de forma estratégica e central (entre as salas das comunicações orais e as doses diárias de cafeína/teína) o que permitiu uma excelente exposição aos mesmos.

No final da manhã do segundo dia de trabalhos teve lugar a segunda Sessão Plenária, apresentada pelo colega Stalislaw Mejza, da Poznan University of Life Sciences, Polónia, intitulada *Younden square with split units*.

Na componente lúdica, a organização da SPE2009, brindou-nos com duas opções igualmente estimulantes: um passeio pelas Rotas dos Galeões do Sal ou em alternativa, para aqueles mais dados a actividades radicais, um 4x4 na Arrábida. Apesar de inicialmente previstos 42 congressistas, os radicais reduziram-se a 7 magníficos! Dos 6 *jeeps* previstos achámos que seria importante fazermos o trajecto em grupo: 3+4. Para além do conhecimento profundo e da simpatia dos guias, a paisagem natural da Arrábida era arrebatadora! Vimos os famosos golfinhos no estuário do Sado, as salgas de peixe romanas da península de Tróia, raríssimos exemplares de *olea silvestre* que tenta sobreviver no espaço

deixado pela sua versão comercial, *olea europaea*, a azeiteira! Visitámos as caves do Moscatel da J.M. Fonseca (felizmente nenhum dos sete tinha ao seu cuidado um *jeep*). Visitámos ainda a Olaria de S. Simão em Azeitão onde se produzem azulejos manualmente. Um dos sete aventurou-se na produção de um azulejo tal como eram produzidos no século XVII e cujo acabamento posterior seria "*majólica italiana*". Para além das vistas deslumbrantes, a proximidade aos conventos de El Carmen, e dos Franciscanos da Arrábida, vistas de cortar a respiração sobre o Portinho da Árrabida, a parte, sem dúvida, mais radical foi a caça à raposa (estava igualmente prevista a caça ao javali, mas tal não se proporcionou, para grande decepção do grupo). Vimos ainda uma rampa de lançamento de óvnis na Arrábida, muito utilizada também por praticantes de parapente! A visita terminou com uma ida ao Castelo de Sesimbra. Só faltou mesmo o javali para o programa ser completo e satisfazer as elevadas expectativas do grupo. Quem sabe numa próxima edição do Congresso nas proximidades da Capital teremos direito ao programa completo!

Aparentemente, o passeio de barco foi mais apelativo aos congressistas, uma vez que contámos com a presença de 92 participantes, repartidos por três barcos, dois dos quais à vela para os mais destemidos.

Felizmente as condições atmosféricas foram as ideais, uma ligeira brisa típica de alto mar e um sol brilhante no céu infinito que se confunde com o próprio mar. Ao longo de 4 horas fomos brindados com paisagens magníficas, possibilitando uma outra perspectiva de Sesimbra, da costa de Setúbal até ao Cabo Espichel (ou pelo menos era essa a intenção...), da península de Tróia e do rio Sado. As pequenas praias escondidas entre escarpas, só acessíveis por barco, que guardam mil histórias de mar, de pescadores, de baleeiros e de piratas, transportaram os participantes para um cenário paradisíaco e de fantasia. Alguns ainda levaram fato de banho, mas o prometido mergulho inesquecível ficou adiado para a SPE 2034!!!

E ao 3º dia, já todos nos acomodámos ao deslumbramento da vista dos terraços! Neste dia tivemos duas Sessões Plenárias. A terceira Sessão Plenária *On the probabilistic and algorithmic approaches to the concept of RANDOMNESS*, a cargo do colega Albert Shiryaev, do Steklov Mathematical Institute, Rússia. Aluno de doutoramento de A. N. Kolmogorov, Shiryaev começou a sua apresentação com um vídeo de tributo a um dos marcos da Matemática do século XX. No final da tarde teve lugar a quarta Sessão Plenária sobre *Stochastic multi-population mortality models,* cuja apresentação esteve a cargo do colega Andrew Cairns, da Heriot-Watt University – Edimburgo.

Um magnífico pôr de sol esperava os congressistas para um aperitivo no terraço oeste do Castelo de Palmela. Houve alguém que confundiu os fotógrafos penetras com o fotógrafo oficial. Depois de algumas tentativas lá se conseguiu tirar a fotografia de grupo. Seguiu-se o jantar no Claustro da Pousada, tendo a anTUNia (uma das Tunas da FCT/UNL) abrilhantado o serão. De destacar na foto as duas aniversariantes da noite, a quem se cantaram os parabéns.

O 4º e último dia começou psicologicamente mais cedo, em particular para os colegas com apresentações logo às 9:00! Mas houve solidariedade de todos, pois não nos pareceu que as salas estivessem mais vazias. A encerrar a apresentação de trabalhos teve lugar a quinta Sessão Plenária intitulada *Um modelo para problemas de estimação,* apresentada pelo colega Paulo Oliveira, da Universidade de Coimbra. Na Figura apresenta-se o slide até onde a classe modal da assistência conseguiu acompanhar a sua apresentação (no slide lê-se "Introdução")!

A Direcção da SPE, à semelhança das campanhas de marketing juvenil (seja abrir uma conta bancária ao primeiro aniversário ou tornar o filho sócio do maior clube de futebol do mundo), decidiu instituir os Prémios Estatístico Júnior, o que é naturalmente uma excelente ideia. Para além de criar um certo coleccionismo (neste caso de troféus), permite ainda dar um ar jovem ao último dia do Congresso e mostrar que a Estatística tem futuro! Houve, assim, mais uma sessão de entrega de prémios.

Por fim, houve a Sessão de Encerramento do Congresso. Agradeceu-se à Comissão Presente o excelente trabalho realizado. Os colegas João Tiago Mexia e Manuel Esquível fizeram chegar as flores recebidas aos membros femininos da organização. Agradeceu-se ainda à anterior Comissão Editorial das Actas do XVI Congresso a entrega das mesmas durante o XVII Congresso.

Finalmente, passou-se o testemunho aos colegas Paulo Oliveira e Carla Henriques, organizadores do XVIII Congresso SPE, no Hotel do Parque, nas Termas de São Pedro do Sul, uma organização conjunta do Departamento de Matemática da Universidade de Coimbra e do Instituto Politécnico de Viseu. Certamente este próximo ano o colega Paulo Oliveira estará mais ocupado com outro tipo de convergências que resultem numa comissão organizadora uniformemente distribuída em esforço e que em limite tenha um Congresso pelo menos tão bom como o deste ano!

São os nossos votos que os próximos relatores se divirtam tanto quanto nós, na sua missão "cliente-mistério" …

Fátima Salgueiro e José Gonçalves Dias

(ISCTE – IUL)

• Jubilação do Professor Tiago Mexia

No passado mês de Junho, jubilou-se o Professor João Tiago Mexia, que escolheu o tema "Modelos e Inferência - Caso Normal" como "última lição".
Celebrando a jubilação do Professor João Tiago Mexia, Professor Catedrático do Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, realizou-se um Workshop em Estatística.
Este evento contou com a participação de vários convidados nacionais e estrangeiros nomeadamente pessoas que no decurso da sua vida profissional se relacionaram mais de perto com o Prof. Mexia.

FR

• Prémios "Estatístico Júnior 2009"

A atribuição de prémios "Estatístico Júnior 2009" é promovida pela Sociedade Portuguesa de Estatística, com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística. Ao apelo para submissão de trabalhos correspondeu uma adesão bastante mais elevada do que em 2008, tendo sido recebidos 41 trabalhos na categoria Ensino Básico, envolvendo um total de 104 alunos, e 29 na categoria Ensino Secundário, envolvendo um total de 71 alunos.

A cerimónia de entrega dos Prémios Estatístico Júnior 2009, conforme estipulado no Regulamento, decorreu na Sessão de Encerramento do XVII Congresso Anual da Sociedade Portuguesa de Estatística, no dia 3 de Outubro de 2009, às 13 horas, nas instalações do Sesimbra Hotel & SPA, Sesimbra.

Excepcionalmente, este ano foi atribuída uma menção honrosa ao trabalho *"Futuro^3- Será que os jovens de hoje terão amanhã um futuro brilhante?"* autoria do aluno Carlos Moura Pereira Lucas Teixeira da Escola Básica D. Manuel I em Tavira, e orientado pela professora Maria Augusta Carvalho de Azevedo.

O Júri foi constituído pelos professores: Doutora Maria Eugénia Graça Martins (Presidente) e Doutora Luísa Canto e Castro de Loura do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa e Doutor Russell Alpizar-Jara do Departamento de Matemática da Universidade de Évora.

No final deste Boletim são apresentados os premiados.

A Direcção

• Prémio SPE 2009

Foram apresentados os seguintes trabalhos concorrentes ao Prémio SPE 2009:

- Séries temporais de memória longa com aplicações ao controlo motor - estudo de tarefas de tapping repetido, da autoria de Ana Maria Fité Alves Diniz.
- Método de imputação recorrente: análise espectral singular com valores omissos, da autoria de Miguel de Carvalho e Paulo Canas Rodrigues.

O júri, constituído por Maria Nazaré Mendes Lopes (Presidente), Paulo Rodrigues e Maria Eduarda Silva, atribuiu o prémio ao trabalho de Miguel de Carvalho e Paulo Canas Rodrigues.
O trabalho vencedor foi apresentado no primeiro dia do XVII Congresso da SPE. O respectivo Resumo é publicado na contra capa deste Boletim SPE.

FR

• Seminários patrocinados pela SPE

A Sociedade Portuguesa de Estatística propõe-se patrocinar a organização de seminários e de palestras apresentados por convidados de outras instituições nacionais e por convidados de instituições estrangeiras que estejam de passagem pelo país.
No primeiro caso, a ideia é a de contribuir para um maior contacto e colaboração entre investigadores das várias instituições nacionais.
No segundo caso, pretende-se que cientistas de instituições estrangeiras que estejam de visita a uma instituição nacional possam, durante a sua estadia, ir dar um seminário a outra instituição, deste modo permitindo um melhor conhecimento de diferentes equipas e possibilitando uma mais ampla colaboração. Solicitamos assim que os colegas que prevêem a visita de cientistas o façam saber a potenciais interessados; a SPE poderá naturalmente ajudar nesta divulgação.
O patrocínio da SPE inclui o anúncio de divulgação pelos sócios e um apoio financeiro às instituições interessadas em acolher um seminário, através de um subsídio máximo de 150 euros.
Os sócios interessados em usufruir desta iniciativa deverão contactar a SPE, candidatando-se a este tipo de apoio.

No ano de 2009, foi apoiado por este programa um Seminário do Departamento de Métodos Quantitativos do ISCTE, intitulado "Captura-Recaptura: Aplicações nas Ciências Sociais" e proferido pelo Prof. Doutor Russell Alpizar-Jara (Cima/Departamento de Matemática/Universidade de Évora), realizado no dia 24 de Abril.

A Direcção da SPE

• JOCLAD 2010

Caro(a) colega,

De 25 a 27 de Março de 2010, o ISCTE – Instituto Universitário de Lisboa recebe as XVII Jornadas de Classificação e Análise de Dados (JOCLAD 2010). O prazo limite para submeter propostas de trabalhos é 11 de Fevereiro de 2010. Mais informações estarão brevemente disponíveis em www.joclad2010.dmq.ibs.iscte.pt. Contamos com a sua presença nas Jornadas!!

Pela Comissão Organizadora,

José Gonçalves Dias (ISCTE-IUL)

• A SPE no Encontro Científico da Sociedade Italiana de Estatística

Por ocasião do septuagésimo aniversário da Sociedade Italiana de Estatística (SIS), realizou-se o congresso "Statistical Methods for the analysis of large data-sets", em Pescara – Itália, entre 23 e 25 de Setembro de 2009. Várias Sociedades de Estatística europeias, entre as quais a SPE, responderam ao convite do presidente da SIS para integrarem o programa científico com três comunicações na temática da conferência.

A SPE foi representada por:

- Antónia Amaral Turkman (Universidade de Lisboa), organizadora da sessão;
- Joaquim Pinto da Costa (Universidade do Porto), com a comunicação intitulada "A weighted principal component analysis and its applications to microarray data". Na sua comunicação descreveu um novo método de selecção de genes com expressão diferencial em *microarrays,* baseado em componentes principais ponderadas;
- Lisete Sousa (Universidade de Lisboa), com a comunicação "Proteomics: Predicting proteins structure", na qual reviu vários métodos de predição da topologia de proteínas transmembranares disponíveis na *Internet*, chamando a atenção para a importância da interacção entre a Estatística e áreas como a Biologia Molecular, Genética, Bioquímica e Bioinformática;
- Giovanni Silva (Universidade Técnica de Lisboa), com a comunicação "Modelling and analysis of forest fire data in Portugal", abordou o tema dos fogos em Portugal usando modelos lineares generalizados para modelar a proporção de área florestal ardida.

Embora não representando a SPE, esteve também presente na conferência Filipe Sousa, aluno finalista da licenciatura em Bioquímica da Universidade do Porto, que sente uma forte atracção pelas áreas de Estatística e suas aplicações à Bioquímica e Biologia Molecular. Como era de esperar, dado o tema da conferência, houve um número significativo de sessões dedicadas àquela temática.

A língua oficial do encontro foi o Inglês. Os conferencistas convidados foram Jerome H. Friedman (Stanford University), com o tema "Fast sparce regression and classification" e Marco Riani (University of Parma) com a comunicação "Problems and challenges in the analysis of complex data: static and dynamic approaches".

Para trás, além de uma interessante conferência, ficou um alegre convívio, e uma simpática cidade, Pescara, banhada pelo Mar Adriático.

Antónia Turkman e Lisete Sousa

• Workshop: Statistical Modelling: Challenges in Health (9-12 / Maio / 2010)

O workshop **StaM2010** tem como objectivo promover o encontro de investigadores interessados em estatística avançada aplicada a problemas desafiantes na área da Saúde. Pretende-se que o workshop sirva também para promover a partilha de conhecimento e experiência, bem como encorajar a cooperação entre participantes. Os principais temas do workshop são:

- Estatística Espacial em Saúde
- Análise de Sobrevivência
- Estatística em Genética
- Estatística em Biologia Molecular
- Bioinformática
- Modelos de classes latentes em Saúde.

Além de oradores convidados de renome, o workshop contará também com uma sessão de comunicações *poster*. Os participantes são convidados a expor os seus trabalhos em desenvolvimento e discuti-los com os convidados numa sessão intitulada "Statistical Clinics".

Página Web: *http://stam2010.fc.ul.pt*

Lisete Sousa

Modelos Econométricos

Breve Contributo para a História do Ensino de Econometria em Portugal[1]

José A. F. Machado, *jafm@fe.unl.pt*
Faculdade de Economia, Universidade Nova de Lisboa

J. M. C. Santos Silva, *jmcss@essex.ac.uk*
University of Essex e CEMAPRE

A econometria é um ramo relativamente novo da economia, podendo o seu nascimento como área autónoma ser associado à fundação da Econometric Society em 1930. A evolução da econometria tem sido enorme nas últimas décadas. Vários factores têm contribuído para esta evolução, mas não há dúvida de que os avanços técnicos têm sido um dos principais motores deste desenvolvimento. Em particular, os avanços da informática permitem hoje a existência de bases de dados impensáveis nos anos 30, bem como os meios de cálculo necessários ao seu tratamento adequado e rápido. As características específicas dos dados económicos, tipicamente não experimentais, levaram a que a econometria desenvolvesse muitas técnicas estatísticas novas, contribuindo assim para a sua crescente autonomização em relação a outras áreas da ciência em que o recurso à estatística matemática é igualmente intenso, como a biometria. A forma como a econometria tem sido ensinada reflecte tanto a juventude da área como os rápidos desenvolvimentos técnicos da segunda metade do Século XX. Neste trabalho faz-se um breve resumo da forma como o ensino da econometria evoluiu em Portugal.

A disciplina de econometria foi introduzida pela primeira vez no plano de estudos de um curso de economia de uma universidade portuguesa na sequência da reforma em 1949 do plano de estudos do então Instituto Superior de Ciências Económicas e Financeiras (ISCEF), tendo sido leccionada pela primeira vez no ano lectivo de 1952/53. Nesta reforma, tornaram-se obrigatórias para todos os estudantes da licenciatura em economia do ISCEF duas disciplinas de matemática. Esta alteração levou a que a disciplina de estatística pudesse pela primeira vez tratar a estatística matemática, abrindo assim caminho à nova disciplina semestral de econometria. Na nota em que Armando Gonçalves Pereira (1949) apresenta a revisão do plano de estudos pode ler-se: "Não nos podemos deixar de regozijar com a criação de um Curso de Econometria, matéria que se presta a estudos da maior relevância".

Pode ter-se uma ideia do conteúdo da disciplina de econometria então leccionada consultando o programa da disciplina que é apresentado por Manuel Jacinto Nunes (1953). Infelizmente, esse programa não apresenta nenhuma lista de bibliografia usada, pelo que é difícil conhecer com algum detalhe a forma como o curso foi leccionado.[2] Para se conhecer melhor o conteúdo dos cursos da época

[1] Este trabalho é uma versão resumida do artigo "50 Anos de Ensino de Econometria em Portugal", publicado pelos autores na revista *Economia* em 2002. Os autores agradecem aos colegas Nuno Crato, José António Girão, Bento Murteira e Carlos Bastien Raposo a disponibilidade para discutir a história do ensino da econometria em Portugal e as muitas informações prestadas. Este trabalho não teria sido possível sem a colaboração de Ana Amaral da Biblioteca do ISEG. Naturalmente, os autores são os únicos responsáveis pelas opiniões aqui expressas e por eventuais incorrecções.

[2] É interessante notar que o curso foi leccionado pelos professores Francisco de Paula Leite Pinto, Henri Guitton (Universidade de Dijon) e José de Castañeda (Universidade de Madrid), bem como pelo então assistente Manuel Jacinto Nunes.

pode recorrer-se à sebenta de econometria editada pela Associação Académica do ISCEF, com base nas lições de Bento Murteira no ano lectivo de 1955/56 (Murteira, 1956). Esta obra está dividida em cinco capítulos, com os seguintes temas: 1) O conceito e objectivos da econometria, 2) Elementos da teoria da procura, 3) Teoria clássica da produção, 4) Função consumo, e finalmente 5) Modelos macroeconómicos. Nos dias de hoje, a estrutura desta sebenta parece ser mais apropriada a uma disciplina de economia aplicada do que a um curso introdutório de econometria, reflectindo afinal aquela que é a génese da econometria como hoje a conhecemos. Aliás, a importância que se dá nestas lições à teoria económica está de acordo com o que era feito em livros de econometria da época, como sejam os de Gerhard Tinter (1952) e Lawrence Klein (1953).[3]

Outro aspecto que é notório nestes apontamentos é o facto de a calculatória necessária, nomeadamente à estimação do modelo de regressão linear múltipla pelo método dos mínimos quadrados, ter uma importância relativa muito pequena. No entanto, tal não é de estranhar se se recordar que as primeiras calculadoras electrónicas (de secretária!) foram introduzidas uma década mais tarde. Portanto, é natural que numa disciplina de licenciatura não houvesse lugar a esse tipo de preocupações, uma vez que os meios de cálculo ao dispor dos alunos não eram de todo adequados, nem sequer à estimação de modelos relativamente simples. Esta carência é colmatada com a apresentação dos resultados de inúmeros estudos empíricos, que são cuidadosamente discutidos e interpretados.

Numa análise mais pormenorizada das lições de Bento Murteira verifica-se que cerca de metade da disciplina era dedicada ao estudo de problemas de microeconometria, nomeadamente a estimação de curvas de Engel e de curvas de custos. Numa segunda parte, estudavam-se modelos para dados agregados (aquilo que hoje genericamente se designa por macroeconometria), dando-se especial atenção à função consumo, que merece um capítulo próprio, e aos modelos de equações simultâneas. Curiosamente, é este último capítulo, que segue de perto o manual de Lawrence Klein (1953), o que mais se aproxima de capítulos correspondentes em manuais de econometria actuais, incluindo a habitual discussão do problema da identificação e dos métodos de estimação com informação limitada e completa.

Cerca de dez anos depois de se ter iniciado o ensino da econometria em Portugal, foi publicada a primeira edição do livro Econometric Methods de Jack Johnston (1963), que marcou profundamente o ensino da econometria em todo o mundo, e que com a sua quarta edição (Johnston e DiNardo, 1997) continua a ser uma obra de referência em muitas universidades portuguesas e estrangeiras. Uma vez que no essencial a evolução do ensino da econometria foi marcada pelo estilo e substância das sucessivas edições deste livro, vale a pena analisar com um pouco de atenção esta obra.

O que distingue o livro de Jack Johnston da maioria dos seus antecessores é o facto de, como o seu nome indica, se centrar claramente nos métodos econométricos e não nas aplicações. De facto, este livro apresenta de uma forma sistemática o modelo de regressão linear múltipla, bem como uma série de tópicos que ainda hoje fazem parte da maioria dos cursos de econometria, como sejam os erros nas variáveis, autocorrelação, heterocedasticidade, multicolinearidade, variáveis artificiais e variáveis desfasadas. Naturalmente, tal como em manuais anteriores, continua a ser dada grande importância aos sistemas de equações simultâneas.

Contrariamente aos livros da década anterior, o livro de Jack Johnston foi escrito com o propósito de ser usado como manual em disciplinas de econometria das licenciaturas. O facto de se concentrar nos métodos econométricos permitia que numa disciplina anual de econometria fossem cobertos praticamente todos os tópicos tratados no livro, e que representavam grande parte dos métodos usados na altura. Sob este ponto de vista, este livro constituiu um grande progresso em relação ao que se passava anteriormente, não sendo de espantar que tenha ganho tanta popularidade. Naturalmente, esta alteração na forma de ensinar econometria teve custos.

Na sua edição de 1963, e contrariamente ao que era habitual, o livro Econometric Methods praticamente não apresenta exemplos realistas de aplicação da econometria. Como consequência, ainda

[3] Note-se no entanto que nestes manuais era possível encontrar tratados outros tópicos, como sejam a análise multivariada, a análise espectral, o cálculo numérico, e modelos de input-output, que não são de todo tratados nas lições de Bento Murteira, mas que também tipicamente não fazem parte dos manuais de econometria modernos.

que o assunto seja mencionado de passagem algumas vezes, este livro nunca discute a natureza dos dados usados pela econometria nem a forma como estes são recolhidos, quase não distinguindo entre dados seccionais e temporais (os dados de painel não ganharam importância senão alguns anos mais tarde). De facto, quase todo o livro apresenta os métodos econométricos com base na hipótese de regressores fixos, que é praticamente insustentável em econometria aplicada. É claro que do ponto de vista da leccionação dos métodos econométricos, esta hipótese é relativamente inócua uma vez que uma parte importante dos métodos pode ser aplicada tanto no caso de regressores fixos como no de regressores estocásticos.

A tendência para valorizar essencialmente a mecânica dos métodos econométricos em detrimento da interpretação dos procedimentos ganhou peso nos anos 70 com a publicação da segunda edição do livro de Jack Johnston e com o surgimento das calculadoras electrónicas portáteis. De facto, o manual de Johnston (1972)[4] já não inclui o capítulo sobre erros nas variáveis, sendo o tema remetido para uma modesta secção no capítulo sobre regressores estocásticos. Este era um tópico que tinha tido algum destaque nos primeiros cursos de econometria (veja-se Murteira, 1956) e era uma das poucas oportunidades que ainda restava para alertar os estudantes para os problemas que resultam da especificidade dos dados usados em econometria. Paralelamente, o surgimento de calculadoras electrónicas portáteis veio tornar possível que os estudantes pudessem estimar pequenos modelos. No entanto, do ponto de vista prático, esta possibilidade de estimar pequenos modelos econométricos não era muito importante uma vez que por essa altura teve início a comercialização dos primeiros programas informáticos dedicados à econometria, significando que qualquer estudo sério de econometria aplicada seria já feito com o recurso a meios de cálculo mais sofisticados.

Apesar das suas insuficiências, esta forma de ver o ensino da econometria perdurou, sendo reforçada pelo surgimento de outros manuais que mantinham esta separação entre, por um lado, os métodos e as técnicas da econometria, e, por outro, os dados dos quais dependem todos os resultados obtidos. Desta forma, pelo menos até aos finais dos anos 80, era enorme o peso da calculatória quer nos cursos quer nas provas de avaliação de econometria da maioria das universidades portuguesas, sendo dado relativamente pouco peso aos exemplos de aplicações reais.

Nos anos 90, com o rapidíssimo crescimento da capacidade de cálculo posta à disposição dos estudantes, a situação alterou-se qualitativamente. De facto, desde o final da década de 80 foi-se vulgarizando a utilização de programas informáticos de econometria nos cursos de licenciatura, proporcionando-se assim aos estudantes o contacto com ferramentas de trabalho poderosas e sofisticadas. Este avanço, que acompanhou o que se passava noutros países da Europa, permitiu dar uma formação muito mais adequada às necessidades dos futuros economistas, libertando a disciplina de econometria de uma parte substancial da calculaória até então indispensável.

Paralelamente aos avanços na capacidade de cálculo, a econometria teve um rápido desenvolvimento desde o final dos anos 70, o qual se tem acelerado progressivamente desde então. A título de exemplo, podem destacar-se o surgimento de temas novos como estimação semi e não paramétrica, o bootstrap, a cointegração, e mesmo a maior atenção dada aos testes de especificação, temas que tipicamente estavam ausentes dos manuais de econometria dos anos 80. O crescimento exponencial da variedade de técnicas usadas em econometria torna praticamente impossível que este desenvolvimento possa ser acompanhado por uma disciplina de licenciatura, especialmente numa época em que há tendência para a redução da duração dos cursos. Apesar das dificuldades, foi feito algum esforço neste sentido, tendo a maioria das disciplinas de econometria passado a integrar durante os anos 90 o tratamento de temas novos como a análise da estacionaridade de séries económicas e a cointegração.[5]

As disciplinas de Econometria leccionadas hoje nas licenciaturas aproximam-se da filosofia primordial patente nas licções de Bento Murteira. A grande acessibilidade de programas informáticos

[4] Este é um livro ao qual ambos os autores muito devem, e do qual guardam uma muito grata recordação, pois foi por ele que estudaram econometria durante as suas licenciaturas.

[5] No entanto, apesar de se ter tornado ainda mais necessário, continuou a ser dada relativamente pouca importância ao contexto estocástico em que se desenvolve o estudo dos modelos de regressão e à natureza dos dados económicos. (Veja-se Machado e Santos Silva, 2002, para uma discussão mais detalhada.)

de fácil utilização, o surgimento de manuais modernos como o de Jeff Wooldridge (2000), e o nascimento de uma activa comunidade de investigação na área, possibilitaram um enfoque maior nas aplicações, na formalização de modelos e na interpretação dos resultados das estimações. Curiosamente, em 50 ou 60 anos fechou-se um círculo.

Referências

Gonçalves Pereira, A. (1949). "O Instituto Superior de Ciências Económicas e Financeiras. A Recente Reforma de Estudo", *Economia e Finanças, Anais do Instituto Superior de Ciências Económicas e Financeiras*, 17, 445-447.

Johnston, J. (1963). *Econometric Methods*, Nova Iorque: McGraw-Hill.

Johnston, J. (1972). *Econometric Methods*, 2ed., Nova Iorque: McGraw-Hill.

Johnston, J. e DiNardo, J. (1997). *Econometric Methods*, 4ed., Nova Iorque: McGraw-Hill.

Klein, L.R. (1953). *A Textbook of Econometrics*, Evanston: Row, Peterson and Company.

Machado, J.A.F. e Santos Silva, J.M.C. (2002). "50 Anos de Ensino de Econometria em Portugal", *Economia*, 26, 95-112.

Murteira, B.F. (1956). *Econometria, 1º Curso*, Associação Académica do ISCEF, Lisboa.

Jacinto Nunes, M. (1953). "Lições de Econometria no Instituto Superior de Ciências Económicas e Financeiras", *Economia e Finanças, Anais do Instituto Superior de Ciências Económicas e Financeiras*, 21, 238-241.

Tinter, G. (1952). *Econometrics*, Nova Iorque: John Wiley & Sons.

Wooldridge, J.M. (2000). *Introductory Econometrics, A Modern Approach*, Cincinnati: South-Western College Publishing.

Séries Temporais: Evolução e Tópicos Recentes

Luís Catela Nunes, *lcnunes@fe.unl.pt*
Faculdade de Economia, Universidade Nova de Lisboa

Paulo M. M. Rodrigues, *pmrodrigues@bportugal.pt*
Banco de Portugal

1. Introdução

Historicamente pode dizer-se que a análise estatística de séries temporais se iniciou no início do século XX (Yule, 1927) e que atingiu a maturidade nos anos setenta aquando da publicação do famoso livro de G.E.P. Box e G.M. Jenkins, *Time Series Analysis: Forecasting and Control, San Francisco, 1970*.

Esta publicação foi importante, no sentido em que contribuiu com uma abordagem sistemática permitindo a aplicação de métodos de séries temporais para a previsão. Box e Jenkins (1970) popularizaram os modelos autoregressivos integrados de médias móveis, os famosos ARIMA(p,d,q),

$$(1-\phi_1 L-\phi_2 L^2- \ldots - \phi_p L^p)(1-L)^d z_t = c + (1-\theta_1 L-\theta_2 L^2- \ldots - \theta_q L^q)e_t$$

onde p, d e q assumem valores não negativos, c é uma constante e L é o operador de desfasamento temporal tal que $Lz_t = z_{t-1}$. Os dois polinómios em L, $(1-\phi_1 L-\phi_2 L^2- \ldots - \phi_p L^p)$ e $(1-\theta_1 L-\theta_2 L^2- \ldots - \theta_q L^q)$ não têm factores comuns e assume-se que as suas raízes caiem fora do circulo unitário. É frequente adicionalmente assumir-se que e_t segue uma distribuição Gaussiana. Esta abordagem assume que z_t é estacionário se d=0 e que z_t contém raízes unitárias quando $d\neq 0$. O sucesso deste modelo originou investigação substancial na área das séries temporais.

Inicialmente a análise das séries temporais dividia-se (e divide-se) em duas abordagens: a análise no domínio da frequência (veja Brillinger, 1975 e Priestley, 1981) e a análise no domínio temporal. A abordagem no domínio temporal utiliza a função de autocorrelação dos dados e modelos paramétricos (e.g. os ARIMA) para descrever a dependência dinâmica das séries enquanto que a abordagem no domínio da frequência se centra na análise espectral para a análise das séries temporais. Actualmente, a opção pela utilização de uma abordagem ou outra prende-se mais com aspectos de ordem pratica do que de ordem filosófica (à semelhança do que acontece com as abordagens Bayesiana e não Bayesiana).

Os avanços ao nível dos métodos e meios computacionais tiveram um impacto profundo sobre a análise de séries temporais. No contexto do que é designado por "análise tradicional" (c.f. Tsay, 2000) deram-se muitos desenvolvimentos importantes. Entre outros, a análise de outliers e a detecção de quebras estruturais passaram a fazer parte integrante do *kit* de diagnóstico do modelo, e foram desenvolvidos vários critérios de selecção para ajudar na escolha dos modelos; veja, entre outros, Akaike (1974) e Hannan (1980).

Em Economia, de acordo com Tsay (2000) a análise de series temporais é utilizada entre outros propósitos para, (a) estudar a estrutura dinâmica de um processo, (b) para investigar as relações dinâmicas entre variáveis, (c) para proceder ao ajustamento sazonal de dados económicos, (d) para melhorar a análise de regressão quando os erros se encontram correlacionados e (e) para produzir previsões pontuais ou intervalos de previsão.

As características das séries económicas e financeiras (assim como outras) têm originado desenvolvimentos importantes dos quais destacamos alguns nas secções seguintes.

2. Tópicos Recentes

2.1 Quebras Estruturais

Ao se considerarem modelos de séries temporais pode-se colocar sempre a questão da estabilidade dos parâmetros ao longo da amostra considerada. A existência de pelo menos uma alteração estrutural poderá resultar em erros de inferência e de previsão se tais quebras não forem devidamente tidas em conta. Nos últimos 50 anos a literatura sobre a estimação e teste de modelos com quebras estruturais teve uma grande expansão. Como ilustração do problema, considere-se o caso mais simples de uma série temporal que de acordo com a hipótese nula tem média e variância constantes ao longo do tempo, mas de acordo com a hipótese alternativa a sua média altera-se numa certa data não conhecida. Este é um problema não *standard* já que um dos parâmetros, a data da quebra, só é identificado sob a hipótese alternativa. Quandt (1958, 1960) propõe a utilização do teste *sup F*, ou *sup Wald*, que corresponde ao teste de rácio de verosimilhança avaliado na data de quebra que maximiza a função de verosimilhança (assumindo uma distribuição Gaussiana). Esta solução foi mais tarde utilizada em contextos mais gerais por Davies (1977).

Andrews (1993) considera também testes baseados no valor máximo dos testes de Wald e do multiplicador de Lagrange (LM) e mostra que estes são assimptoticamente equivalentes. Andrews (1993) também apresenta versões dos testes robustas à presença de autocorrelação e heterocedasticidade. Andrews e Ploberger (1994) desenvolvem testes óptimos, como o *exp Wald*, no sentido em que a potência média ponderada é maximizada.

Muitos outros testes alternativos foram apresentados por diversos autores como sejam os baseados em somas parciais dos resíduos (teste LM de Gardner, 1969) ou no máximo de somas parciais de resíduos recursivos (teste CUSUM de Brown, Durbin e Evans, 1975).

Inicialmente, a caracterização da distribuição assimptótica de muitos dos testes propostos não era feita. Por exemplo, Quandt (1960) nota que a distribuição do teste *sup Wald* sob a hipótese nula não correspondia a uma distribuição de qui-quadrado (que se obteria caso a data de quebra fosse conhecida). Tal como no caso dos testes de raízes unitárias, a utilização do teorema do limite central funcional permite caracterizar de forma simples a distribuição assimptótica de todas as variantes destes testes como funcionais de movimentos Brownianos. Por exemplo em MacNeill (1974) a distribuição assimptótica do teste LM proposto por Gardner é caracterizada como o integral do quadrado de uma ponte Browniana no intervalo [0,1]. Outro exemplo é o teste CUSUM cuja distribuição assimptótica se pode representar como o supremo de um movimento Browniano devidamente normalizado.

O problema básico acima considerado foi generalizado ao modelo de regressão linear múltipla com várias quebras estruturais. Bai (1997) e Bai e Perron (2003) estudam as propriedades do estimador de mínimos quadrados e dos testes de quebras múltiplas, e ainda a inferência acerca das datas de quebras.

O caso em que se pretende estimar o número de quebras foi também estudado por vários autores. Liu *et al* (1997) e Yao (1988) propõem a utilização de critérios de selecção de modelos enquanto Bai e Perron (2003) consideram a utilização sequencial de testes de m quebras contra m+1 quebras, partindo do caso m = 0 até ocorrer uma rejeição.

A maioria dos testes de quebras estruturais são válidos apenas quando as séries temporais são estacionárias, excluindo como tal o caso de raízes unitárias. Considere-se o seguinte problema acerca do comportamento da tendência de uma série temporal: $y_t = a + b\ t + e_t$ em que os erros e_t podem ser auto-correlacionados, estacionários, I(0), ou não-estacionários com uma raíz unitária, I(1). Neste exemplo, os testes habituais sobre uma quebra estrutural do declive da componente de tendência determinística b têm taxas de convergência diferentes sob a hipótese nula de não existirem quebras nos casos I(0) e I(1). Só muito recentemente foram encontradas soluções para este problema. Perron e Yabu (2005) e Harvey, Leybourne e Taylor (2009) propõem testes de quebras estruturais que são

válidos e não requerem conhecimento prévio acerca da estacionariedade ou não dos erros.

O problema oposto, de testar uma raiz unitária, na eventual presença de uma quebra estrutural também levanta problemas. Perron (1989) mostra que a análise da função de autocorrelação ou a utilização de testes de raíz unitária tendem a concluir pela existência de raiz unitária mesmo quando as séries são estacionárias na presença de uma quebra na componente de média ou tendência determinística. Como solução, o autor propõe uma modificação aos testes de raízes unitárias do tipo Dickey-Fuller. Estes testes foram generalizados ao caso em que a data da quebra não é conhecida por Zivot e Andrews (1992) e a várias quebras por Ohara (1999). Mais recentemente, Kim e Perron (2009) apresentam uma solução para o caso em que se pretende testar uma raiz unitária sem necessidade de se saber se existe ou não uma quebra tanto sob a hipótese nula como alternativa.

A revisão apresentada acima, considera apenas modelos em que o número de quebras embora possa ser desconhecido é determinado *a priori*. Como tal, nestes modelos as quebras ocorrem de forma exógena e independente. Em alternativa poderá considerar-se que as alterações estruturais seguem elas próprias um processo estocástico gerado a partir de um determinado modelo, passando a ser possível por exemplo inferir sobre a possibilidade de ocorrência de quebras no futuro a partir da observação dos dados e eventuais quebras ocorridas no passado. Uma das abordagens mais populares consiste na modelização de vários regimes, não observados, como sendo gerados a partir de uma cadeia de Markov. Este modelo tornou-se popular na análise de ciclos económicos (Hamilton, 1989) uma vez que permite por exemplo definir à partida a existência de dois estados, um de "recessão" e outro de "expansão", em que a série temporal da actividade económica é gerada a partir de modelos auto-regressivos com parâmetros diferentes e em que a duração de cada regime é determinada por uma matriz de transição. Este modelo tem tido também bastante popularidade na caracterização da evolução da variância condicional de séries financeiras por exemplo em diferentes períodos de crise (Turner, Startz e Nelson, 1989).

De facto, existe uma classe geral de modelos com variáveis não observadas denominados de modelos de espaço de estados que têm tido uma grande aplicação em economia e finanças (ver por exemplo Harvey, 1990; e Hamilton, 1994). A principal ferramenta utilizada na estimação destes modelos é o filtro de Kalman, um procedimento recursivo que permite estimar as variáveis não observadas a partir da informação disponível em cada momento do tempo, e que permite construir a função de verosimilhança tendo em vista a estimação do parâmetros do modelo. Em Kim e Nelson (1999) são apresentados vários modelos de espaço de estados com mudanças de regime tanto numa perspectiva clássica como Bayesiana.

Outra abordagem que também permite gerar diferentes regimes são os modelos TAR (*threshold auto-regression*) popularizados por Tong (1990). Neste caso, os estados ou regimes são determinados por intervalos de valores que uma série temporal pode ir tomando ao longo do tempo.

Tal como no caso dos modelos com quebras estruturais, nestes dois modelos alternativos existe um problema de não identificação de parâmetros sob a hipótese nula (por exemplo os valores da matriz de transição no caso do modelo com mudanças de regime Markovianas) e que causa dificuldades na determinação do número de estados ou regimes. Uma revisão recente destes e outros modelos não lineares aparece em Teräsvirta (2006).

2.2 Raízes unitárias

No final da década de 70, economistas e econometristas reconheceram que estava a ser dada pouca atenção aos mecanismos de tendência dos dados. Esta avaliação gerou um conjunto substancial de desenvolvimentos resultando na modificação de como os estudos de séries temporais eram feitos. A investigação em não estacionaridade avançou significativamente com os contributos de Granger e Newbold (1974), Davidson *et al*. (1978), Hendry e Mizon (1978) e Phillips (1986), entre outros.

Para ilustrar o problema que as raízes unitárias originam, considere-se o seguinte processo autoregressivo de ordem 1,

$$x_t = x_{t-1} + e_t \quad (1)$$

onde e_t é ruído branco. Assumimos para simplificação da exposição que o valor inicial da série é zero, i.e., $x_0 = 0$. Este processo designa-se por passeio aleatório. Frequentemente também se utiliza a designação de série integrada de ordem 1 (I(1)), indicando a necessidade de considerar as primeiras diferenças para a obtenção de uma série estacionária (I(0)).

Uma característica interessante deste processo é o facto de contrariamente a um processo estacionário o impacto dos choques passados não diminuir com o passar do tempo. Em particular, reescrevendo x_t em função dos choques verifica-se que a importância dos choques passados é exactamente a mesma da de choques recentes. Este fenómeno resulta em implicações interessantes, particularmente sobre as propriedades dos estimadores assim como em contextos de modelação.

Para se analisarem as propriedades dos estimadores é necessário recorrer a um novo conceito designado de Teorema do Limite Central Funcional (Phillips, 1987) que permite estabelecer que, $T^{-1/2}x_t \Rightarrow \sigma^2 W(r)$, com $T \to \infty$ onde T é o tamanho da amostra e W(r) é um processo Browniano. Phillips (1987) também demonstra que o estimador de α, α_T, e correspondente teste t, obtidos pelo método dos mínimos quadrados ordinários tendo por base o modelo $x_t = \alpha x_{t-1} + e_t$ assumindo que os dados são gerados por (1) convergem para distribuições não convencionais, funções de processos Brownianos. Este novo resultado é interessante por várias razões. Em primeiro lugar a taxa de convergência de α_T é T e não a habitual $T^{1/2}$. Em segundo lugar a teoria associada à não estacionaridade veio abrir uma área de investigação muito importante, em particular relacionada com problemas de testes na qual muitos economistas, econometristas e estatísticos estão interessados.

Em particular o teste de raízes unitárias considera a hipótese nula H_0: $\alpha = 1$ (de raíz unitária) contra a alternativa H_a: $|\alpha| < 1$ (de estacionaridade); veja Dickey e Fuller (1979). O problema de testar a raíz unitária tem atraído muito interesse por várias razões: i) providencia um teste formal à determinação da ordem de integração de um ARIMA; ii) abre uma área em que os testes que são desenvolvidos dependem a) das variáveis deterministicas a incluir na regressão teste; b) da multiplicidade de raízes unitárias; e c) da presença de outros parâmetros AR e MA, parâmetros esses assimptoticamente negligenciáveis, mas que podem ter efeitos nefastos em amostras finitas.

O problema da análise e determinação de raízes unitárias foi também alargado aos processos MA (veja Davis e Dunsmuir, 1996), aos processos sazonais (Hylleberg *et al*., 1990, Osborn *et al*. (1988), Rodrigues e Taylor, 2004), a dados em Painel (Breitung e Pesaran, 2008) e a modelos não lineares do tipo SETAR (Caner e Hansen, 2001) e STAR (Smooth Transition Autoregressive), resultando num avultado número de desenvolvimentos teóricos e aplicados.

Outra área relacionada com a não estacionaridade que também viu grandes desenvolvimentos ao longo das últimas duas décadas foi a área dos processos fraccionários; veja Granger e Joyeux (1980), Crato e de Lima (1994), Robinson (1994), Breitung e Hassler (2002) e Hassler, Rodrigues e Rubia (2009).

2.3 Regressões Espúrias

Um fenómeno relacionado com as raízes unitárias é o fenómeno das relações (regressões) espúrias, conhecido dos estatísticos e econometristas desde Yule (1897) e Pearson (1897). Existem muitos exemplos de regressões espúrias na literatura. Por exemplo, Phillips (1986) ilustra este fenómeno recorrendo à relação implausível entre "the number of ordained ministers and the rate of alcoholism in Great Britain in the nineteenth century"; o próprio Yule (1926) apresenta a curiosa relação entre a "proportion of Church of England marriages to all marriages and the mortality rate over the period 1866–1911"; outro caso curioso é a relação entre o nível de preços e a quantidade cumulativa de precipitação no Reino Unido avançada por Hendry (1980).

O grande interesse neste fenómeno em econometria só surgiu no entanto com o famoso artigo de Granger e Newbold (1974) no qual com base em análise Monte Carlo, é apresentada evidência, regredindo variáveis perfeitamente independentes entre si, da significância aparente das várias relações.

A questão é que sob as condições de regularidade convencionais o método dos mínimos quadrados ordinários não apresenta evidência de relação entre duas variáveis independentes. No entanto, no estudo de Granger e Newbold (1974), as séries foram geradas como passeios aleatórios e era a não estacionaridade (resultante de raízes unitárias) das variáveis que originou estimativas de parâmetros estatisticamente diferentes de zero. Foi Phillips (1986) que veio enquadrar teoricamente o problema explicando analiticamente as causas do fenómeno das regressões espúrias.

Note-se que o problema simétrico, denominado de quebras espúrias (Nunes et al., 1996), também se pode colocar. Quando uma série é I(1), os métodos habituais de inferência válidos para séries I(0) tendem a concluir sobre a existência de uma ou mais quebras mesmo quando o processo gerador de dados não apresenta quebras.

2.4 Cointegração

O conceito de cointegração que valeu a Clive Granger em 2003 o prémio Nobel da Economia traduz um fenómeno importante para a modelação de longo prazo das séries cronológicas. De acordo com Engle e Granger (1987), um vector $\mathbf{X}_t$ de n variáveis (X_{it}, i=1, 2,….,n) diz-se cointegrado se todos os elementos de $\mathbf{X_t}$ forem integrados da mesma ordem e exista um vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ...,\alpha_n)$, tal que a combinação linear $z_t = \boldsymbol{\alpha}'\mathbf{x}_t$ é de ordem de integração inferior à de $\mathbf{x}_t$ e $\boldsymbol{\alpha} \neq 0$. $\boldsymbol{\alpha}$ é conhecido como vector de cointegração. Se existir um sistema de r α_i (i = 1,...,r) vectores, então a matriz (n×r) de vectores designa-se por matriz de cointegração. O número de vectores de cointegração é definido como a ordem de cointegração de $\mathbf{X}_t$, i.e., se $\mathbf{X}_t$ for constituído por n variáveis então poderão existir até n-1 vectores de cointegração.

O conceito de cointegração está relacionado com a noção de equilíbrio de longo prazo. Uma definição de equilíbrio entre um conjunto de variáveis $\mathbf{X}_t$ é uma relação do tipo $\boldsymbol{\alpha}'\mathbf{x}_t = 0$. Esta relação é designada por relação de cointegração ou de longo prazo entre as variáveis. No entanto, dado que é difícil de ocorrer, $Z_t = \alpha' x_t$ mede o grau de desequilíbrio do sistema, i.e., a distância do sistema à situação de equilíbrio, representando desta forma o erro de equilíbrio.

Se duas ou mais variáveis forem cointegradas, elas podem ter uma representação de curto e longo prazo através de um mecanismo corrector do erro (MCE). O equilíbrio anteriormente descrito é introduzido neste modelo como uma variável adicional sobre a forma de um termo corrector do erro.

Apesar de já existirem na literatura (na literatura estatística a ideia de combinação linear de séries com raízes unitárias ser estacionária foi analisada por Box e Tiao, 1977), foi só com Engle e Granger (1987) que o conceito de cointegração ganhou expressão. Em particular, o Teorema de Representação de Granger teve importante destaque. De acordo com este Teorema, considerando $\mathbf{X}_t$ um vector (n×1) de n variáveis cointegradas, com r vectores cointegrantes ($0 < r \leq n-1$), a seguinte representação de correcção de erro pode ser obtida,

$$A(L)(1 - L)X_t = -\gamma Z_{t-1} + u_t \qquad (2)$$

onde $Z_t = \boldsymbol{\alpha}'\mathbf{x}_t$, u_t é um vector de resíduos estacionários, A(L) é uma matriz de polinómios em L, L é o operador de desfasamento temporal convencional e γ são os coeficientes do termo corrector do erro ou coeficientes de ajustamento.

O termo $Z_{t-1} = \boldsymbol{\alpha}'\mathbf{X}_{t-1}$ é conhecido como termo corrector do erro, dado que representa o desequilíbrio entre as variáveis X_{it} (elementos de $\mathbf{X}_t$) no período t-1. Quanto maiores forem estes coeficientes, maior é a resposta de X_{it} a desvios do período anterior em relação ao equilíbrio de longo prazo. O sinal negativo indica que a próxima alteração em X_{it} vai ser de sinal oposto à de Z_{t-1}. O MCE pode ser interpretado como o mecanismo que conduz a economia para um estado de equilíbrio. Isto significa que no período t os agentes económicos corrigem parte do desequilíbrio detectado no período t-1.

Johansen e Juselius (1990) e Johansen (1988) introduziram duas estatísticas para determinar o número

de vectores cointegrantes: i) o teste do traço da matriz e ii) o teste do valor próprio máximo.

2.5 Modelos Multivariados

Os modelos VAR (*vector autoregressive*) continuam a estar entre os mais populares na análise das interrelações dinâmicas entre várias variáveis. Estes modelos permitem responder a questões como qual o impacto de um aumento da taxa de juro em variáveis macroeconómicas como o desemprego ou a inflação ao longo de vários períodos através das chamadas funções de resposta a impulsos. A razão da sua popularidade tem a ver com o facto de serem modelos lineares relativamente fáceis de estimar.

Além disso, não necessitam à partida da imposição de restrições sobre as relações entre as variáveis, ou seja, não é necessário impor uma estrutura ou modelo macroeconómico. No entanto este último ponto é também um dos maiores desafios destes modelos já que a identificação e estimação das funções de resposta a impulsos e de outros resultados destes modelos (como a decomposição da variância do erro de previsão) requerem a imposição de restrições que permitam identificar o modelo estrutural subjacente ao modelo VAR. Por exemplo considere-se o seguinte modelo VAR com apenas duas variáveis:

$$Y_{1t} = a_{12}\, Y_{2t} + b_{12}\, Y_{2,t-1} + b_{11}\, Y_{1,t-1} + e_{1t}$$
$$Y_{2t} = a_{21}\, Y_{1t} + b_{21}\, Y_{1,t-1} + b_{22}\, Y_{2,t-1} + e_{2t}$$

Em que os e_{1t} e e_{2t} são choques estruturais ortogonais com $E(e_t \mid Y_{t-1}, Y_{t-2}, \ldots)=0$. É fácil concluir que os coeficientes do modelo não são identificados. Os vários desenvolvimentos dos modelos VAR consistem precisamente em diversas formas de utilizar restrições vindas da teoria económica que permitam identificar os parâmetros do modelo. As restrições de curto prazo são as mais frequentemente utilizadas (por exemplo impor que $a_{12} = 0$, ou seja, que a variável Y_1 não reage contemporaneamente aos choques e_2). Uma revisão destes métodos aparece por exemplo em Hamilton (1994). No entanto é possível utilizar também restrições de longo prazo (Blanchard e Quah, 1989; ou King, Plosser, Stock e Watson, 1991), restrições de sinal da função de resposta a impulsos em determinados horizontes temporais (Uhlig, 2005), através da identificação de regimes em que os choques têm variâncias diferentes (Rigobon, 2003), ou através de métodos Bayesianos em que as distribuições *a priori* dos parâmetros do modelo são sugeridas por modelos estruturais macroeconómicos (DelNegro and Schorfheide, 2004).

3. Conclusão

Este artigo apresenta uma breve resenha da evolução da análise econométrica de séries cronológicas e de alguns tópicos recentes que marcaram a área devido à sua importância e impacto em termos de análise, modelação e previsão.

As diferentes características das séries cronológicas têm motivado a necessidade de desenvolvimento de métodos e técnicas de análise adequados, levando a uma dinâmica recente de investigação importante em tópicos como o Bootstrap, a não-linearidade e não-estacionaridade, a análise não paramétrica, quebras na variância, etc. A análise econométrica de séries cronológicas é uma área de investigação muito activa e que tem despertado muito interesse quer em termos aplicados como teóricos.

4. Referências

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Andrews, D.W.K., (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821-856 (Corrigendum, 71, 395-397).

Andrews, D.W.K., Ploberger, W., (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383-1414.

Bai, J., (1997) Estimation of a change point in multiple regression models. *Review of Economic and Statistics* 79, 551-563.

Bai, J., Perron, P., (2003) Computation and analysis of multiple structural change models. *Journal of Applied Econometrics* 18, 1-22.

Blanchard, O.J., e Quah, D. (1989) The Dynamic Effects of Aggregate Demand and Supply Disturbances, *American Economic Review*, 79(4), 655-73.

Box, G.E.P., e Jenkins, G.M., (1970) Time Series Analysis: Forecasting and Control.

Box, E.E.P., e Tiao, G.C., (1977) A canonical analysis of multiple time series, *Biometrika*, 64, 355–365.

Breitung, J. e Hassler, U. (2002) Inference on the cointegration rank in fractionally integrated processes, *Journal of Econometrics*, 110(2), 167-185.

Breitung, J. and Pesaran, M.H., (2008) Unit roots and cointegration in panels, Ed. Matyas, L. and Sevestre, P. *The Econometrics of Panel Data* (Third Edition), Kluwer Academic Publishers, no prelo.

Brillinger, D.R. (1975) Time Series: Data Analysis and Theory. New York: Holt, Rinchart and Winston.

Brown, R.L., Durbin, J., Evans, J.M., (1975) Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society* B 37, 149-163.

Caner, M., e Hansen, B.E., (2001) Threshold Autoregression with a Unit Root, *Econometrica*, 69(6), 1555-1596.

Crato, N. e P. J. F. de Lima, (1994) Long-range dependence in the conditional variance of stock returns, *Economics Letters*, vol. 45, no. 3, pp. 281–285.

Davidson, J., D. Hendry, F. Srba, e S. Yeo, (1978) Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom, *Economic Journal*, vol. 88, pp. 661–692.

Davies, R.B. (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247-254.

Davis, R.A., and Dunsmuir, W.T.M., (1996) Maximum-likelihood estimation for MA(1) processes with a unit root on or near the unit circle, *Econometric Theory*, 12, 1–29.

Del Negro, M. e Schorfeide, F. (2004) Priors from General Equilibrium Models for VAR's, *International Economic Review*, 45, 643-673.

Dickey, D. A. e W. A. Fuller, (1979) Distribution of the estimators for autoregressive time series with a unit root, *Journal of the American Statistical Association*, vol. 74, no. 366, part 1, pp. 427–431.

Engle, R. F. e C. W. J. Granger, (1987) Co-integration and error correction: representation, estimation, and testing, *Econometrica*, vol. 55, no. 2, pp. 251–276.

Gardner, L.A., (1969) On detecting changes in the mean of normal variates. *The Annals of Mathematical Statistics* 40, 116-126.

Granger, C. W. J. e R. Joyeux, (1980) An introduction to long-memory time series models and fractional differencing, *Journal of Time Series Analysis*, vol. 1, no. 1, pp. 15–29.

Granger, C. W. J. e P. Newbold, (1974) Spurious regressions in econometrics, *Journal of Econometrics*, vol. 2, no. 2, pp. 111–120.

Hamilton, James D, (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2), 357-84.

Hamilton, J.D., (1994) *Time Series Analysis*, Princeton, NJ: Princeton University Press.

Hannan, E.J., (1980) The estimation of the order of an ARMA process, *Annals of Statistics*, 8, 1071–1081.

Harvey, A C, (1990) *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, D. I., Leybourne, S.J. and Taylor, A.M.R., (2009) Simple, Robust, And Powerful Tests Of The Breaking Trend Hypothesis. *Econometric Theory* 25(4), 995-1029.

Hassler, U., Rodrigues, P.M.M. e A. Rubia, (2009) Testing for the General Fractional Integration Hypothesis in the Time Domain. *Econometric Theory*, no prelo.

Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo (1990), Seasonal Integration and Cointegration, *Journal of Econometrics*, 44: 215-38.

Hendry, D. (1980) Econometrics-alchemy or science? *Economica*, vol. 47, no. 188, pp. 387–406.

Hendry, D. e G. Mizon, (1978) Serial correlation as convenience simplification, not a nuisance: a comment on a study of the demand for money by the bank of England, *Economic Journal*, vol. 88, pp. 549–563.

Johansen, S. (1988), Statistical Analysis of Cointegrating Vectors, *Journal of Economic Dynamics and Control*, 12, 231-54.

Johansen, S., e Juselius, K., (1990) Maximum Likelihood Estimation and Inference on Cointegration With Applications to the Demand for Money, *Oxford Bulletin of Economics and Statistics*, 52(2), 169-210.

King, R. , C. Plosser, J. Stock, and M. Watson, (1991) Stochastic trends and economic fluctuations, *The American Economic Review*, vol. 81, no. 4, pp. 819–840.

Kim, C.J., and Nelson, C.R., (1999) *State-Space Models with Regime Switching*, Cambridge, Massachusetts, MIT Press.

Kim, D., Perron, P., (2009), Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. Journal of Econometrics, 148(1), 1-13.

Liu, J., Wu, S., Zidek, J.V., (1997) On segmented multivariate regressions. *Statistica Sinica* 7, 497-525.

MacNeill, I.B., (1974) Tests for change of parameter at unknown time and distributions on some related functionals of Brownian motion. *Annals of Statistics* 2, 950-962.

Nunes, L.C., Newbold, P., and Kuan,C.-K., (1996) Spurious number of breaks, *Economics Letters*, 50(2), 175-178.

Ohara, H.I., (1999) A unit root test with multiple trend breaks: a theory and application to US and Japanese macroeconomic time-series. *The Japanese Economic Review* 50, 266-290.

Osborn, D.R., Chui, A.P.L., Smith, J.P. and Birchenhall, C.R. (1988), Seasonality and the Order of Integration for Consumption, *Oxford Bulletin of Economics and Statistics*, 50, pp.361-377.

Pearson, K. (1897) On a form of spurious correlation which may arise when indices are used in the measurement of organs, *Proceedings of the Royal Society of London*, vol. 60, pp. 489–498.

Perron, P. (1989) The great crash, the oil price shock and the unit root hypothesis, *Econometrica*, vol. 57, pp. 1361–1401.

Perron, P., Yabu, T., (2005) Testing for shifts in trend with an integrated or stationary noise component. Manuscript in preparation, Department of Economics, Boston University.

Phillips, P. C. B. (1986) Understanding spurious regressions in econometrics, *Journal of Econometrics* 33(3), pp.311–340.

Phillips, P. C. B. (1987) Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, Vol. 74(3), 535–547.

Priestley, M.B. (1981) Spectral Analysis and Time Series (Vols. 1 & 2), London: Academic Press.

Quandt, R. E., (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* 53, 873-880.

Quandt, R.E., (1960) Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* 55, 324-330.

Rigobon, R., (2003) Identification Through Heteroskedasticity, *The Review of Economics and Statistics*, 85(4), 777-792.

Robinson, P. M. (1994) Efficient tests of nonstationary hypotheses, *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1420–1437.

Rodrigues, P.M.M. e A.M.R. Taylor (2004) Alternative Estimators and Unit Root Tests for Seasonal Autoregressive Processes. *Journal of Econometrics* 120, 35-73.

Teräsvirta, T. (2006), 'Univariate nonlinear time series models' in Patterson, Kerry and Terence C. Mills (eds.) *Palgrave Handbook of Econometrics*, Volume 1: Econometrics, Capítulo 10, pp. 396-424, Palgrave Macmillan.

Tong, H., (1990) *Non-linear time series: a dynamical system approach*, Oxford University Press, Oxford.

Tsay, R.S. (2000) Time Series and Forecasting: Brief History and Future Research, *Journal of the American Statistical Association* 95 (450), pp. 638-643.

Turner, C.M., Startz, R. and Nelson, C.R., (1989) A Markov model of heteroskedasticity, risk, and learning in the stock market. *Journal of Financial Economics* 25(1), 3-22.

Uhlig, H. (2005) What are the effects of monetary policy on output? Results from an agnostic identification procedure, *Journal of Monetary Economics*, 52(2), 381-419.

Yao, Y-C., (1988) Estimating the number of change-points via Schwarz' criterion. *Statistics and Probability Letters* 6, 181-189.

Yule, G.U., (1897) On the theory of correlation, *Journal of the Royal Statistical Society* 60(4), pp. 812–854.

Yule, G.U., (1926) Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series, *Journal of the Royal Statistical Society* 89(1), pp. 1–63.

Yule, G.U., (1927) On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. Philosophical Transactions of the Royal Society of London, Ser. A, 226, 267-298.

Zivot, E. e D. W. K. Andrews, (1992) Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis, *Journal of Business and Economic Statistics* 10, pp. 251–270.

Econometria Financeira

João Nicolau, *nicolau@iseg.utl.pt*

Instituto Superior de Economia e Gestão / Universidade Técnica de Lisboa e CEMAPRE

1. Introdução

A investigação em econometria financeira realiza-se em duas grandes áreas: uma que se preocupa fundamentalmente com o desenvolvimento de métodos econométricos adequados a dados financeiros; a outra, mais empírica, que aplica os métodos para testar hipóteses e teorias da economia financeira. Qualquer que seja a abordagem entende-se que "Financial econometrics is simply the application of econometric tools to financial data" (Robert Engle). A econometria financeira é, portanto, uma disciplina econométrica (baseada em métodos estatísticos e matemáticos) vocacionada para analisar dados financeiros. Alguns dos temas genéricos de interesse na área do desenvolvimento dos métodos econométricos são a estimação, a construção de modelos econométricos e a previsão. O escopo das aplicações econométricas à economia financeira é muito vasto. Citem-se alguns exemplos:

- Avaliação do risco (por exemplo, através do *Value at Risk*);
- Avaliação de obrigações, opções e outros activos financeiros;
- Previsão da volatilidade;
- Gestão de *portfolios*;
- Análise da previsibilidade e eficiência dos mercados.

Sendo a econometria financeira a aplicação de métodos econométricos adequados a dados financeiros, todos os métodos estatísticos que de uma forma ou outra se apliquem a dados financeiros, interessam à econometria financeira. De todo o modo, a área proeminente em econometria financeira é a das séries temporais. Estuda-se, por exemplo, a evolução temporal das cotações, taxas de câmbio, taxas de juro, etc. Por esta razão, este documento analisa essencialmente métodos econométricos para séries temporais, sobretudo os métodos que de alguma forma se adequam às características próprias das séries financeiras, como sejam, a não linearidade e a não normalidade.

É conveniente distinguir séries temporais de natureza macroeconómica e as de natureza financeira. Dados de natureza macroeconómica (consumo, produto, taxa de desemprego) podem ser observados com periodicidade mensal, trimestral ou anual; dados financeiros, como por exemplo, retornos de acções ou taxas de câmbio podem ser observados com uma frequência muito superior; nalguns casos, com intervalos de minutos ou segundos entre duas observações consecutivas. Assim, o número de observações disponíveis de dados financeiros pode situar-se na ordem das centenas de milhares, ou ainda mais. Normalmente, prefere-se trabalhar com dados diários (evitando-se os problemas de microestrutura de mercado). Com as séries macroeconómicas raramente se passam das poucas centenas de observações (quando, na melhor das hipóteses, se têm observações mensais). Os dados macroeconómicos são menos fiáveis, i.e., estão mais sujeitos a erros de medição. Com efeito, os valores apurados não resultam de valores efectivamente observados no mercado, como sucede com a generalidade das séries financeiras, mas antes de valores apurados de acordo com certa metodologia e decorrentes de inquéritos. Outra diferença assinalável decorre das propriedades estatísticas dos dois

tipos de séries. Ao contrário das séries macroeconómicas, as séries financeiras tendem a exibir habitualmente fortes efeitos não lineares e distribuições não normais.

O ponto de partida para a análise estatística é normalmente uma série de preços (por exemplo, a série das cotações de fecho do BCP num certo intervalo de tempo). De uma forma geral, o preço pode ser, por exemplo, o valor a que um intermediário financeiro informa estar disposto a pagar pela compra de um determinado activo, opção ou futuro (*bid price*), o valor a que um intermediário financeiro informa estar disposto a receber pela venda de um determinado activo, opção ou futuro (*ask price*), o valor final da transacção, o valor definido num mercado de futuros, entre outros.

2. Factos Empíricos Estilizados de Séries Temporais Financeiras

Antes de se propor um modelo estocástico para uma série financeira, é importante discutirem-se as principais regularidades empíricas da série. Em geral, há um conjunto de regularidades empíricas que são partilhadas por um grande leque de séries temporais financeiras observadas com frequência elevada (diária ou semanal). Chamam-se a essas regularidades factos empíricos estilizados, por serem comuns a muitas séries. Os principais são os seguintes:

1. *Prémio de risco positivo*. O valor esperado do retorno de um investimento no mercado de capitais deve exceder o retorno do investimento sem risco. A essa diferença designa-se prémio de risco. Este prémio deve ser positivo pois, caso contrário, não haveria motivação para aceitar um investimento com retornos incertos, quando a alternativa é um retorno garantido.
2. *Desvios padrão diferentes consoante os activos*. Os activos com maior variabilidade e, portanto, com maior risco associado, são os títulos de empresas, seguidos dos índices bolsistas e taxas de câmbio. Os bilhetes do tesouro apresentam a menor variabilidade. No âmbito dos títulos de acções, vários estudos indicam que a variabilidade dos retornos tende a diminuir à medida que a dimensão das empresas aumenta (títulos de empresas pequenas apresentam maior variabilidade).
3. *Retornos de acções e de índices tendem a apresentar assimetria negativa*. As distribuições empíricas das rendibilidades de acções e índices bolsistas tendem, em geral, a serem assimétricas negativas (normalmente ocorrem mais variações negativas fortes, i.e. *crashes*, do que variações positivas fortes).
4. *Retornos apresentam distribuições leptocúrticas*. Observa-se para a generalidade das séries financeiras que os retornos muitos altos e muito baixos ocorrem com maior frequência do que seria de esperar se os retornos seguissem uma distribuição normal. Na generalidade dos casos o coeficiente de *kurtosis* estimado vem quase sempre bastante acima de 3, o que sugere que a distribuição dos retornos (de cotações, índices, taxas de câmbio e mesmo taxas de juro) é leptocúrtica.
5. *Autocorrelações lineares baixas entre os retornos*. Em geral os coeficientes de autocorrelação dos retornos são baixos. Imagine-se uma situação hipotética em que a média dos retornos diários é zero e o coeficiente de correlação é negativo e alto em módulo. Se o retorno hoje é positivo, amanhã tenderá a ser negativo e vice-versa. Existe, portanto, uma forte possibilidade de ganho (arbitragem) com base na observação passada dos preços. Se outros participantes do mercado comprarem e venderem com base neste padrão de autocorrelação, o processo de arbitragem reduzirá rapidamente a correlação. Portanto, não é credível supor-se coeficientes de autocorrelação lineares altos.
6. *Volatility Clustering*. Já vimos que valores muitos altos e muito baixos ocorrem frequentemente (com maior frequência do que seria de esperar se as variáveis seguissem uma distribuição normal). Este valores extremos não ocorrem isoladamente: tendem a ocorrer de forma seguida, daí o termo *volatility clustering*.
7. *Forte dependência temporal da volatilidade*. Nos pontos precedentes observámos dois factos estilizados: (1) valores muitos altos e muito baixos ocorrem frequentemente e (2) estes valores extremos aparecem de forma seguida (*volatility clustering*). Neste ponto reforça-se a ideia de *volatility clustering*: não só os valores extremos tendem a aparecer de forma seguida como também há alguma persistência neste fenómeno. Isto é, se a volatilidade é alta (baixa), então é

razoável esperar que a volatilidade se mantenha alta (baixa) durante bastante tempo.

8. *Efeito assimétrico*. Tem-se observado existir alguma correlação entre a volatilidade e a ocorrência de perdas significativas nos mercados de capitais. Designa-se esta relação por efeito assimétrico. Concretamente, quando a rendibilidade de um activo é negativa espera-se, em média, um aumento de volatilidade para o período seguinte. O efeito assimétrico é, por vezes identificado como *leverage effect* depois de Black em 1976 ter notado que a volatilidade aumenta quando o mercado cai e o rácio de endividamento (*leverage ratio*) aumenta. No entanto, vários autores têm salientado que o *leverage* é muito reduzido quando comparado com o efeito assimétrico.
9. *Aumento da frequência das observações acentua a não linearidade e a não normalidade*. Pode mostrar-se que, em geral, o coeficiente de *kurtosis* tende a aumentar com o aumento da frequência amostral (por exemplo, quando se passa de observações semanais para observações diárias). Assim, a distribuição marginal dos retornos diários apresenta um maior afastamento face à distribuição normal do que a distribuição marginal dos retornos mensais (por exemplo). Também a correlação entre a magnitude dos retornos tende a acentuar-se com o aumento da frequência das observações.
10. *Efeitos de calendário*. Em certas séries a rendibilidade e/ou a volatilidade varia com o calendário.

3. Modelação Em Tempo Discreto

3.1 Introdução

Ao longo das últimas décadas os modelos ARMA têm dominado a abordagem de séries temporais (os primeiros trabalhos datam de 1927 com os modelos autoregressivos estudados por Yule). Existem razões para o sucesso dos modelos ARMA Gaussianos:

- simplicidade: as equações lineares às diferenças finitas são fáceis de tratar;
- o modelo ARMA Gaussiano é completamente caracterizado pela média, variância e pelas autocorrelações;
- são fáceis de aplicar e implementar (a maioria dos programas de estatísticas possui um módulo para tratar os modelos ARMA);
- a teoria está bastante desenvolvida: praticamente, todas as principais questões relacionadas com a estimação, inferência e previsão estão resolvidas;
- os modelos lineares apesar de simples são relativamente flexíveis e úteis na previsão.

Todavia, os modelos ARMA apresentam limitações:

- não são apropriados para dados que exibam súbitas alterações em períodos irregulares;
- não são apropriados para dados que exibam forte assimetria e achatamento e,
- obviamente, não são indicados para modelarem relações não lineares. Pode suceder que os coeficientes de autocorrelação linear sejam baixos, e existir fortes correlações não lineares entre as variáveis da sucessão (por exemplo, a autocorrelação entre os quadrados das variáveis pode ser alto). Pode suceder também que os coeficientes de autocorrelação linear dependam do nível do processo.

Tendo em conta os factos estilizados empíricos estilizados de séries temporais financeiras descritos no ponto anterior, é fácil perceber que os modelos ARMA não são em geral adequados para modelarem séries financeiras, observadas com frequência elevada. Modelos mais adequados para séries financeiras deverão ser capazes de modelarem não linearidades. Uma forma simples (mas não geral) de introduzir modelos não lineares consiste em apresentar a não linearidade através dos momentos condicionais. Considere-se o modelo

$$y_t = \mu_t + u_t, u_t = \sigma_t \varepsilon_t$$

onde $\{\varepsilon_t\}$ é um ruído branco (ou uma diferença de martingala), $\mu_t = g\left(y_{t-1}, \dots, y_{t-p}; u_{t-1}, u_{t-2}, \dots, u_{t-q}\right)$ é a média condicional de y_t e

$\sigma_t^2 = h(y_{t-1}, \dots, y_{t-p}; u_{t-1}, u_{t-1}, \dots, u_{t-q})$ é a variância condicional de y_t. O modelo é não linear na média se g é uma função não linear dos seus argumentos; o modelo é não linear na variância se σ_t^2 é não constante ao longo do tempo pois, neste caso, o processo $\{u_t\}$, definido por $u_t = \sigma_t \varepsilon_t$ é não linear, por ser um processo multiplicativo.

3.2 Modelos Não Lineares na Variância

Um dos primeiros modelos a romper com o quadro clássico da estimação ARMA, foi o modelo ARCH, proposto por Robert Engle em 1982. Embora existisse já alguma evidência de que a volatilidade não era constante, devido aos trabalhos de Benoit Mandelbrot e Eugene Fama, na década de 60, os modelos de séries temporais habituais na *empirical finance* assumiam homocedasticidade (variâncias iguais). Os modelos ARCH revolucionaram a abordagem econométrica de séries temporais: não só passam a admitir, no âmbito de um modelo paramétrico, heterocedasticidade (que era frequentemente visto apenas como um problema de dados seccionais), como também propõem um modelo dinâmico para a volatilidade.

O modelo da família ARCH (inclui, por exemplo, o GARCH, TARCH, EGARACH, etc.) pode ser representado genericamente pelas seguintes equações

$$u_t = \sigma_t \varepsilon_t$$

$$\sigma_t^2 = h(y_{t-1}, \dots, y_{t-p}; u_{t-1}, u_{t-2}, \dots, u_{t-q})$$

onde ε_t pode ser, por exemplo, um ruído branco.

Pode demonstrar-se que os modelos da família ARCH podem captar os factos estilizados 4 a 10 mencionados na secção 2. Trata-se de um enorme progresso face aos modelos ARMA. Alguns dos factos estilizados, como por exemplo, o efeito de calendário, não resultam directamente das propriedades dos modelos da família ARCH, mas é muito fácil, no quadro da estimação ARCH incorporar esses efeitos. Por outro lado, outros factos estilizados resultam directamente das propriedades dos modelos ARCH. Por exemplo, pode demonstrar-se que a distribuição marginal dos retornos de um processo ARCH possui caudas polinomiais e, portanto caudas pesadas, com *kurtosis* superior a 3, mesmo que a distribuição condicional seja Gaussiana.

3.3 Modelos Não lineares na Média

Uma classe importante de processos não lineares na média baseia-se na ideia de *regime-switching*. Podem ser usados em duas circunstâncias gerais: (a) existem alterações bruscas e inesperadas nas trajectórias dos processos (e.g., ataques especulativos, *crashes* bolsistas, anúncios públicos de medidas do governo, eventos políticos e, em geral, eventos extraordinários não antecipados); (b) existem alterações da dinâmica do processo sem alterações bruscas nas trajectórias. Por exemplo, a taxa de juro no período 1993 a 2006 exibe dois períodos com comportamento bem diferenciado: no primeiro, as taxas de juro e a volatilidade são relativamente altas e o processo evidencia uma tendência de reversão para uma média, seguindo-se, depois de 1995, um período de baixas taxas de juro, baixa volatilidade e ausência de reversão para uma média.

Para este tipo de fenómenos, os modelos com alterações (estocásticas) de regime (ou *regime-switching*) podem ser, no essencial, de dois tipos: (a) a mudança de regime é função de uma variável observável; são exemplos, modelos com variáveis impulso (*dummy*), os modelos limiares ou *threshold AR* (TAR), os modelos onde os coeficientes associados às componentes AR são funções não lineares dos valores passados do processo (STAR, *smoothed transition AR*), entre outros; (b) a mudança de regime não é observada, incluindo-se, nesta classe, os modelos onde os regimes são independentes entre si (como, por exemplo, os modelos *simple switching* ou de Bernoulli) e os modelos onde existe dependência entre os regimes (como por exemplo, os modelos MS, *Markov-Switching)*.

4. Modelação em Tempo Contínuo

4.1 Introdução

Nos últimos anos tem-se assistido a um enorme interesse na modelação em tempo contínuo. Podemos atribuir este facto ao *boom* da Matemática Financeira, que usa fundamentalmente processos em tempo contínuo para avaliar opções sobre activos, e à recente disponibilidade de séries financeiras de altíssima frequência.

O modelo base para descrever o comportamento probabilístico de uma série financeira ao longo do tempo é o processo de difusão que pode ser descrito através de uma equação diferencial estocástica (EDE)

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \quad X_0 = x \tag{1}$$

onde W_t é o processo de Wiener (padrão). Processos de difusão são processos de Markov com trajectórias contínuas quase certamente onde as probabilidades de transição $\mathrm{P(s, x, t, B)} \equiv \mathrm{P}(X_t \in B | X_s = x)$ satisfazem, para cada $s \in [t_0, T]$, $x \in \mathbb{R}$, $\varepsilon > 0$,

1. $\lim_{t \to s} \frac{1}{t-s} \int_{|y-x|<\varepsilon} \mathrm{P(s, x, t, B)} = 0$
2. existe uma função $a(s, x)$ tal que $\lim_{t \to s} \frac{1}{t-s} \int_{|y-x|<\varepsilon} (\mathrm{y - x})\ \mathrm{P(s, x, t, B)} = \mathrm{a(s, x)}$
3. existe uma função $b^2(s, x)$ tal que $\lim_{t \to s} \frac{1}{t-s} \int_{|y-x|<\varepsilon} (\mathrm{y - x})^2\ \mathrm{P(s, x, t, B)} = \mathrm{b^2(s, x)}$

De acordo com a condição 1 a ocorrência de saltos instantâneos na trajectória do processo é improvável. As condições 2 e 3 estabelecem que o processo tem uma média infinitesimal $a(s, x)$ e uma variância infinitesimal $b^2(s, x)$. A média infinitesimal (também designada por coeficiente de tendência) fornece uma medida da velocidade média do movimento descrito por X no momento s, dado que $X_s = x$ (note-se, no caso do valor esperado condicional existir, o coeficiente de tendência pode ser interpretado como $(t - s)^{-1} E(X_t - X_s | X_s = x) \approx a(s, x)$) e a variância infinitesimal (também designada por coeficiente de difusão) fornece uma medida da magnitude local das flutuações de $X_t - X_s$ dado $X_s = x$ (note-se, $(t - s)^{-1} E\left(X_t - X_s{}^2 \middle| X_s = x\right) \approx b^2(s, x)$).

Quais as vantagens em se modelar uma série financeira através de uma EDE, comparativamente à modelação em tempo discreto? Há, em primeiro lugar, uma vantagem óbvia - permite que os modelos financeiros teóricos, na grande maioria deduzidos a partir de EDEs, possam ser efectivamente aplicados ao "mundo real". Existem também vantagens (e, certamente, desvantagens) em termos puramente estatísticos. Nos modelos a tempo discreto a especificação dos dois primeiros momentos condicionais é imediata; por exemplo, na especificação do modelo a tempo discreto, $X_t = \mu_t + \sigma_t \varepsilon_t$ (para $t = 1,2, \ldots$), onde ε_t é, por exemplo, uma diferença de martingala, com variância finita igual a um, a média condicional é μ_t e a variância condicional é σ_t^2. As probabilidades de transição são fáceis de estabelecer uma vez especificada a distribuição de ε_t. Nas EDEs, os momentos e as probabilidades condicionais, associadas a observações discretas são, em geral, muito difíceis de obter. Não obstante, uma das vantagens das EDEs é a de que permitem, para um número apreciável de casos não lineares, a obtenção das distribuições estacionárias (quando existam, obviamente) que são um elemento chave para a compreensão do fenómeno. Em geral, muitas expressões de interesse, como leis de probabilidade que governam o processo de difusão, são determinadas, à parte certas condições fronteira, apenas a partir da relação dos coeficientes infinitesimais. Na generalidade dos casos, é possível estabelecer uma equação diferencial parcial (EDP) cuja solução determina a expressão de interesse. Também se obtém, em geral, uma infinidade de momentos estacionários (quando existam). No caso discreto, de equações não lineares, (por exemplo, do tipo ARCH) é geralmente difícil a obtenção de resultados limites, quer em termos de momentos estacionários, quer sobretudo em termos de distribuições estacionárias. Um exemplo destas dificuldades é mostrada por Daniel Nelson: para derivar certos resultados assintóticos dos processos ARCH, Nelson considerou processos de difusão como soluções limites de processos ARCH, quando o intervalo de tempo entre sucessivas realizações do processo tende para zero. Outra vantagem dos modelos em tempo contínuo é a de que é fácil estabelecer resultados para qualquer t pertencente a um intervalo. Nos modelos a tempo discreto, se os

dados são anuais, os resultados apenas podem referir-se a instantes múltiplos do ano. Nestes modelos, supõe-se ainda que o intervalo entre as observações é constante. A favor das EDES pode-se ainda argumentar que as variáveis económicas evoluem intrinsecamente em tempo contínuo mesmo que as trajectórias possam exibir descontinuidades, pois os processos latentes geradores das variáveis económicas são contínuos. Por exemplo, as decisões dos agentes, a informação, os gostos, a tecnologia são quase certamente processos contínuos no tempo. A economia não "pára", obviamente, entre duas observações do processo; da mesma forma, a economia não evolui de acordo com as observações disponíveis do processo. Uma questão diferente é saber se as trajectórias dos processos económicos são contínuas. Algumas séries financeiras têm provavelmente trajectórias descontínuas, isto é, evoluem através de saltos aleatórios no tempo (por exemplo, uma cotação de uma acção sujeita a reduzidas transacções não está continuamente a alterar de valor).

4.2 Extensões ao Modelo Base

Uma das extensões ao modelo (1) mais importante é aquela que permite modelar o coeficiente de difusão através de outra equação diferencial estocástica. Entende-se, neste caso, que não só os preços mas também a volatilidade dos preços admite uma representação estocástica. Estes processos designam-se por modelos de volatilidade estocástica. Para ilustrar considere-se o seguinte exemplo

$$dS_t = (r - d)S_t dt + \sigma_t S_t dW_{t1},$$

$$d\sigma_t^2 = \varphi(\delta - \sigma_t^2)dt + \omega\sigma_t dW_{t2},$$

onde W_{t1} e W_{t2} são processos de Wiener, não necessariamente independentes. Outra extensão importante são os processos de difusão com saltos de Poisson. Estes modelos adequam-se a fenómenos com alterações bruscas da trajectória, devido, por exemplo, a anúncios de política monetária, a *crashes* bolsistas, ataques especulativos súbitos, etc.

4.3 O problema da Estimação

Tal como no caso discreto, também os processos de difusão envolvem parâmetros ou funções desconhecidas que devem ser estimados a partir de observações discretas do processo. A estimação e inferência estatística são consideravelmente mais difíceis em processos de difusão do que em processos em tempo discreto. Embora o método da máxima verosimilhança para processos de difusão baseados em observações discretas apresenta as habituais boas propriedades (consistência, eficiência e distribuição assimptótica normal dos estimadores), as densidades de transição necessárias para construir a função de verosimilhança são geralmente desconhecidas. Várias abordagens de estimação têm sido propostas, como por exemplo: método dos momentos generalizados baseados no operador infinitesimal; função martingala de estimação; aproximação da verosimilhança via expansão de Hermite; aproximação da verosimilhança via aproximação numérica da equação progressiva de Kolmogorov; aproximação da verosimilhança via simulação; métodos Bayesianos; métodos baseados em modelos auxiliares (inferência indirecta e método dos momentos eficientes). Também a estimação não paramétrica tem suscitado muito interesse. A estimação mais difícil é que envolve os processos de volatilidade estocástica: não só a verosimilhança, associada a observações discretas, é geralmente desconhecida, como também o processo da volatilidade não é observado.

5. Aplicações

5.1 Opções

Uma das aplicações mais importantes da teoria dos processos estocásticos às finanças é a que respeita à determinação do preço *justo* ou prémio de uma *opção*. Uma opção *call* europeia confere ao seu detentor o direito, mas não a obrigação, de comprar um activo (por exemplo uma acção cotada na bolsa) na data de expiração do contrato T, por um preço K previamente fixado. A cotação do activo evolui estocasticamente ao longo do tempo e pode ser genericamente caracterizado como um processo estocástico $\{S_t: 0 \leq t \leq T\}$ definido num espaço de probabilidades $(\Omega, \mathcal{F}, P)$ (onde Ω é o espaço amostral, podendo ser identificado como o conjunto de todos os cenários de mercado, $\mathcal{F}$ é a álgebra-σ

dos subconjuntos de Ω e P é a medida de probabilidade). No instante T o detentor da opção pode comprar o activo pelo preço K, previamente estabelecido, e vender imediatamente por S_T, supondo obviamente que $S_T > K$. Se $S_T < K$ o detentor da opção não exerce o direito de compra. Desta forma a *receita* (*payoff*) é $max\{S_T - K, 0\}$. Nestas circunstâncias, qual o valor justo do prémio da opção no momento $t < T$? Naturalmente, o valor $max\{S_T - K, 0\}$ depende crucialmente do processo estocástico $\{S_t: 0 \leq t \leq T\}$. Fisher Black e Myron Scholes, assumindo um movimento Browniano geométrico, deduziram uma fórmula matemática para o prémio da opção. Esta fórmula, simples e extremamente útil ainda nos dias de hoje, é considerada por muitos economistas como uma das maiores realizações da teoria financeira.

5.2 Estimação da Volatilidade

A volatilidade é um tópico fundamental em finanças. O conceito de volatilidade está presente na gestão do risco, na afectação e selecção de activos, na valorização e *hedging* das opções e derivados e em muitas outras operações e estratégias financeiras (no exemplo anterior, o parâmetro mais importante que condiciona o valor $max\{S_T - K, 0\}$ é, precisamente, a volatilidade do processo $\{S_t: 0 \leq t \leq T\}$). A área da modelação e da previsão da volatilidade assenta, naturalmente, em processos estocásticos. A literatura é muita vasta nesta área, e inclui variadíssimos modelos em tempo discreto (e.g. modelos ARCH e modelos de volatilidade estocástica) e em tempo contínuo (e.g. processos de difusão univariados com coeficiente de difusão não constante e processos de difusão de segunda ordem de volatilidade estocástica).

5.3 Gestão do Risco

A gestão do risco consiste, grosso modo, em identificar as fontes de risco e em medir, controlar e gerir esse mesmo risco. Nesta área, um conceito fundamental é o *Value at Risk* ou VaR (como é usualmente conhecido na literatura). O VaR representa a perda que pode ocorrer num lapso de tempo determinado, com uma certa probabilidade α, supondo que o *portfolio* não é gerido durante o período de análise. Em termos probabilísticos, o VaR é o quantil de ordem α da distribuição teórica de ganhos e perdas. Estes ganhos e perdas evoluem ao longo do tempo e, portanto, são susceptíveis de serem modelados através de processos estocásticos.

5.4 Eficiência dos Mercado Financeiros

Uma discussão já longa na literatura debate a eficiência dos mercados financeiros. O mercado de capitais diz-se eficiente se os preços dos produtos financeiros reflectirem toda a informação disponível. Quando é libertada uma informação relevante (por exemplo, um anúncio de distribuição de dividendos de valor superior ao esperado, um anúncio de fusões ou aquisições, etc.) num mercado eficiente os agentes reagem imediatamente comprando ou vendendo de acordo com a informação e os preços ajustam-se imediatamente. Se o mercado é eficiente o preço ajusta-se rapidamente e não há oportunidades para a realização de rendibilidades anormais. Neste caso, o retorno não é previsível e, portanto, deverá ser não autocorrelacionado. Naturalmente esta discussão faz-se no âmbito de um modelo probabilístico de processos estocásticos.

5.5 Gestão de Porfolios

Um problema importante em finanças é o da selecção e constituição de *portfolios* de acordo com o princípio geral da obtenção da máxima rendibilidade com a menor volatilidade (risco) possível. Existem várias abordagens para obter a rendibilidade e a volatilidade mas a mais conveniente e adequada diz respeito às previsões (temporais) de rendibilidade e volatilidades associadas aos activos que constituem o *portfolio*. Com efeito, a decisão sobre constituição de *portfolio* dependerá da rendibilidade e da volatilidade futura dos activos financeiros que constituem o portfolio. Trata-se, portanto, de um problema de previsão que deve ser tratado, naturalmente, no âmbito dos processos estocásticos.

6. Referências Bibliográficas[1]

Aït-Sahalia Y., (1996), "Nonparametric Princing of Interest Rate Derivative Securities", *Econometrica*, 64, 527-560.

Aït-Sahalia, Y. (2002), *"Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach." Econometrica, 70(1), 223-262.*

Amin and Ng, (1993), "Option Valuation with Systematic Stochastic Volatility", *Journal of Finance,* 48(3), 881-910.

Andersen, T.G., T. Bollerslev, F.X. Diebold and H. Ebens, (2001), "The distribution of realized stock return volatility", *Journal of Financial Economics*, 61, 43-76.

Bachelier L. (1900), *Théorie de la Spéculation*, thèse de Mathématique, Paris.

Barndorff-Nielsen, O.E. and N. Shephard (2006), "Econometrics of testing for jumps in financial economics using bipower variation", *Journal of Financial Econometrics*, 4, 1-30.

Bibby, B., and M. Sorensen (1995), "Martingale Estimation Function for Discretely Observed Diffusion Process", *Bernoulli, 1, 17-39*.

Black, F. and M. Scholes, (1973), "The Pricing of Options and Corporate Liabilities, *Journal of Political Economy*", *81, pp. 637-654*.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroscedasticity", *Journal of Econometrics* 31, 307-327.

Bollerslev, T., R.Y. Chou and K.F. Kroner (1992) "ARCH modeling in finance: a review of the theory and empirical evidence", *Journal of Econometrics,* 52, 5-59.

Cox J.C., Ingersoll J.E., Ross S. A., (1985), "A Theory of the Term Structure of Interest Rates", *Econometrica*, 53, 385-407.

Danielsson J., (1994), "Stochastic Volatility in Asset Prices - Estimation With Simulated Maximum Likelihood", *Journal of Econometrics*, 64, 375-400.

Duffie D., (1988), *Security Markets : Stochastic Models*, Academic Press.

Engle R., (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica*, 50, 987-1008.

Engle R., (2001), "GARCH101: The Use of ARCH/GARCH Models in Applied Econometrics", *Journal of Economic Perspectives,* 15, 157-168

Engle, R. and Jeff Russell, (1998), " Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data", *Econometrica* 66.

Eraker, B. (2001), "MCMC Analysis of Diffusion Models With Application to Finance", *Journal of Business & Economic Statistics*, 19, 177-191.

Fama, E. (1976), "Forward Rates as Predictors of Future Spot Rates", *Journal of Financial Economics*, 361-77.

Gallant A., and G. Tauchen (1996), "Which moments to match?", *Econometric Theory*, 12, 657-681.

[1] Lista de alguns artigos relevantes na área dos processos estocásticos em finanças. Inclui também algumas publicações do autor na área da Econometria Financeira.

Hansen L.P. (1982). "Large Sample Properties of Generalized Methods of Moments", *Econometrica*, 50.

Hansen, L., and J. Scheinkman (1995): "Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes", Econometrica, 63, 767-804.

Hull J, and A. White (1987), "The Pricing of Options on Assets with Stochastic Volatilities", *Journal of Finance*, 42, 281-300.

Jacquier, Polson, and Rossi (1994), "Bayesian Analysis of Stochastic Volatility Models", *Journal of Business and Economic Statistics,* 12,371-389

Kessler, M. (1997): "Estimation of an Ergodic Diffusion from Discrete Observations", *Scandinavian Journal of Statistics*, 24, 211- 229.

Lintner J. (1965), "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets", *Review of Economics and Statistics*, 47, 13-37.

Lo, A. (1988): "Maximum likelihood estimation of generalized Ito processes with discretely sampled data", *Econometric Theory*, 4, 231--247.

Marakowitz H. (1952), "Portfolio selection", *Journal of Finance,* 7, 77-91.

Merton R.C., 1990, *Continuous Time Finance,* Cambridge, M.A. Blackwell.

Merton, R.C., 1973, "Theory of Rational Option Pricing", Bell Journal of Economics and Management Science, 4, 141-183.

Nelson D.B., (1990a), "ARCH Models as Diffusion Approximations", *Journal of Econometrics*, 45, 7-38.

Nelson D.B., (1991), "Conditional Heteroskedasticity in Asset Returns: A New Approach", *Econometrica*, 59.

Nicolau, J. (2002) "New Technique for Simulating the Likelihood of Stochastic Differential Equations", *The Econometrics Journal*, 5, 2002.

Nicolau, J. (2002) "Stationary Processes that Look Like Random Walks -- the Bounded Random Walk Process in Discrete and Continuous Time", *Econometric Theory,* 18.

Nicolau, J. (2003) "Bias Reduction in Nonparametric Diffusion Coefficient Estimation", *Econometric Theory,* 19.

Nicolau, J. (2005), "Processes with Volatility-Induced Stationarity. An Application for Interest Rates", *Statistica Neerlandica*, 59, 376-396.

Nicolau, J. (2005). "A Method for Simulating Non-Linear Stochastic Differential Equations in R1". *Journal of Statistical Computation and Simulation*, 75, 595-609.

Nicolau, J. (2007), "A Discrete and a Continuous-Time Model Based on a Technical Trading Rule", *Journal of Financial Econometrics,* 5, 266-284.

Nicolau, J. (2007), "Non-Parametric Estimation of Second Order Stochastic Difference Equations", *Econometric Theory,* 23.

Nicolau, J. (2008), "Modeling Financial Time Series Through Second Order Stochastic Differential Equations", *Statistics and Probability Letters*, 75, 595-609.

Pedersen, A. (1995), "A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations", *Scandinavian Journal of Statistics,* 22, 55-71.

Sharpe W. (1963), “A simplified model for portfolio analysis”, *Management Science,* 9, 277-93.

Sharpe W. (1964), “Capital asset prices: a theory of market equilibrium under conditions of risk”, *Journal of Finance,* 19.

Sørensen M., (1995), “Martingale Estimation Function for Discretely Observed Diffusion Process”, *Bernoulli*, 1.

Taylor S.J. (2008), *Modelling Financial Time Series*, Second Edition, John Wiley & Sons.

Yoshida, N. (1992), "Estimation for Diffusion Processes from Discrete Observations", *Journal of Multivariate Analysis* 41, 220-242.

O Bootstrap para Estatísticas HAC e os seus Competidores

Sílvia Gonçalves, *silvia.goncalves@umontreal.ca*

Université de Montréal, Canada

1. Introdução

O bootstrap é um método de inferência que pode ser utilizado para estimar a função de distribuição (ou funcionais dela, tais como a média ou a variância) de um determinado estimador ou estatística de teste. A ideia subjacente ao bootstrap é muito simples: tratam-se os dados disponíveis como sendo a população para realizar a inferência.

Desde a sua introdução por Efron em 1979, o bootstrap tornou-se muito popular em econometria. Uma das razões da sua popularidade é a sua simplicidade. Por exemplo, o bootstrap tornou-se num dos métodos padrão para a obtenção de erros padrões de estimadores complicados quando as suas variâncias assimptóticas são desconhecidas ou difíceis de derivar. Dado que o poder computacional melhorou substancialmente ao longo do tempo, o bootstrap tornou-se numa alternativa relativamente barata a métodos de inferência mais complicados baseados em derivações assimptóticas. A outra razão pela qual o bootstrap é um método de inferência popular prende-se com o facto deste ter um desempenho melhor em amostras finitas do que outros métodos alternativos baseados em aproximações assimptóticas de primeira ordem, reduzindo substancialmente as distorções de amostras finitas associadas a estas aproximações.

O objectivo deste artigo é discutir o bootstrap para dados dependentes no contexto de estatísticas *t* baseadas em estimadores de variância consistentes à heteroscedasticidade e autocorrelação (designados na literatura por HAC). Este é um exemplo bem conhecido em que a distribuição normal assimptótica de primeira ordem convencional nos dá uma fraca aproximação à distribuição de amostras finitas da estatística de interesse. Estimadores HAC dependem de dois importantes parâmetros de afinação: a janela do ponderador (*kernel)* e o parâmetro relativo à largura de banda (*bandwidth)*. A escolha destes parâmetros é importante em amostras finitas, mas não é captada pelas aproximações da normal padrão. Uma aproximação assimptótica alternativa foi recentemente proposta por Kiefer e Vogelsang (2005), onde o parâmetro relativo à largura de banda é modelizado como uma proporção fixa do tamanho da amostra. Esta nova teoria assimptótica capta a escolha da largura de banda e da função ponderadora e consequentemente tem um melhor desempenho em amostras finitas. O *bootstrap* em bloco oferece outra aproximação. Neste artigo, revêem-se estas aproximações e discute-se o seu desempenho em amostras finitas, tendo por base o trabalho de Gonçalves e Vogelsang (2009). Para simplificar a exposição, iremos focar-nos na média amostral. Primeiro, na Secção 2, revêem-se algumas propriedades do bootstrap i.i.d. quando aplicado a dados i.i.d. Na Secção 3, discutem-se as razões que justificam a falha do bootstrap i.i.d. quando os dados são dependentes. A Secção 4 revê o método do bootstrap em bloco e os seus competidores no contexto de séries cronológicas. A Secção 5 conclui.

2. O bootstrap i.i.d. para a média amostral de dados i.i.d.

Suponha que $\{X_t: t=1,\ldots,n\}$ representa uma amostra identicamente e independentemente distribuída (i.i.d.) de uma população F, de média μ e variância $\gamma(0)$. O estimador de μ é a média amostral

$$\hat{\mu}_n = \bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t = \mathrm{f}(\mathcal{X}_\mathrm{n})$$

onde $\chi_n = (X_1, \ldots, X_n)$. Suponha que o objectivo é estimar a variância de $\hat{\mu}$. Neste contexto i.i.d. simples,

$$\sigma^2 \equiv Var\left(\sqrt{n}\hat{\mu}\right) = \gamma(0). \quad (1)$$

Um estimador padrão é

$$\hat{\sigma}^2 \equiv \widehat{Var}\left(\sqrt{n}\hat{\mu}\right) = \tilde{\gamma}(0) \; onde \; \tilde{\gamma}(0) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad (2)$$

que corresponde à variância amostral de χ_n.

Se F fosse conhecido, poderíamos aproximar a variância de $\hat{\mu}$ arbitrariamente bem através da aleatorização de Monte Carlo. Poderíamos gerar muitas amostras aleatórias de F e calcular a variância amostral sobre as replicações de Monte Carlo de $\hat{\mu}$ como aproximação da verdadeira variância da população dada por (1). O problema é que não se conhece F. O bootstrap simplesmente substitui F por $\hat{F}$, uma estimativa de F, e depois gera amostras aleatórias de $\hat{F}$. Em particular, o bootstrap não paramétrico proposto por Efron (1979) consiste em gerar amostras i.i.d. dos dados originais χ_n, o que corresponde a considerar $\hat{F}$ a função de distribuição empírica.

Considere $\chi^*_n = (X^*_1, \ldots, X^*_n)$ uma amostra bootstrap i.i.d de χ_n. Uma forma conveniente de escrever as observações bootstrap é $X^*_t = X\tau_t$, onde τ_t é um valor i.i.d. de uma distribuição uniforme sobre $\{1, \ldots, n\}$. Considere P^* (E^* e Var^*) a medida de probabilidade induzida pelo bootstrap (valor esperado bootstrap e variância bootstrap), condicional aos dados. Podemos avaliar a estatística de interesse sobre χ^*_n e obter $\hat{\mu}^* = n^{-1}\sum_{t=1}^{n} X_t^* = f(\mathcal{X}_n^*)$ que é o análogo bootstrap de $\hat{\mu}$. Obtém-se que

$$E^*(\hat{\mu}^*) = \frac{1}{n}\sum_{t=1}^{n} E^*(X_t^*) = \frac{1}{n}\sum_{t=1}^{n}(X_t) \equiv \hat{\mu}. \quad (3)$$

Em particular, $E^*(X_t^*) = E^*(X_1^*) = \frac{1}{n}\sum_{t=1}^{n} X_t$, onde a primeira igualdade é verificada porque as observações bootstrap são identicamente distribuídas (logo, os seus momentos não se alteram com t) e a segunda igualdade verifica-se dado que cada observação em χ_n tem a probabilidade $\frac{1}{n}$ de ser escolhida para a amostra bootstrap. De igual modo, podemos mostrar que

$$\sigma^{*2} \equiv Var^*\left(\sqrt{n}\hat{\mu}^*\right) = \hat{\gamma}(0) \; onde \; \hat{\gamma}(0) = \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X}_n)^2. \quad (4)$$

Se compararmos (4) com (2), podemos observar que ambos são muito próximos. A diferença é que a variância do bootstrap se baseia em $\hat{\gamma}(0)$, enquanto que a variância habitual se baseia em $\tilde{\gamma}(0)$, que utiliza um factor de ajustamento dos graus de liberdade.

Para o caso especial da media amostral, a variância bootstrap é uma expressão fechada conhecida dos dados originais dada por (4). Logo, não necessitamos de recorrer a métodos de simulação de Monte Carlo para a aproximar. No entanto, nem sempre é assim. Para estatísticas mais complicadas (por exemplo, qualquer função não linear de $\hat{\mu}$), a forma fechada do estimador da variância bootstrap não é conhecida, e nestes casos deverá ser aproximado através de simulações de Monte Carlo.

Dado (2) e (4), a consistência do estimador da variância *bootstrap* obtêm-se sob os pressupostos habituais que garantem a consistência da variância amostral para a variância da população. Isto justifica a utilização do bootstrap para estimar a variância de $\hat{\mu}$. No entanto, o bootstrap é uma ferramenta muito mais poderosa: permite-nos aproximar toda a distribuição de $\hat{\mu}$. De facto, a teoria das expansões Edgeworth (e.g. Hall, 1992) sugere que se o objectivo for realizar um teste de hipóteses ou construir um intervalo de confiança para μ, devemos utilizar o bootstrap para estimar os quantis de uma estatística devidamente *studentized* (cuja distribuição limite é livre de parâmetros perturbadores) de forma a obter-se um refinamento assimptótico sobre a distribuição assimptótica padrão de primeira ordem. De seguida discute-se a aplicação do bootstrap para estimar a função de distribuição de uma estatística t.

Considere

$$t \equiv \frac{\sqrt{n}\,(\hat{\mu} - \mu)}{\hat{\sigma}}.$$

A distribuição em amostras finitas de t só é conhecida se considerarmos pressupostos distribuicionais fortes. A abordagem padrão baseia-se na sua aproximação recorrendo à teoria assimptótica de primeira ordem. Em particular, sob condições mais ligeiras, o teorema do limite central implica que $t \xrightarrow{d} N(0,1)$, o

que justifica a utilização dos quantis da distribuição *N(0,1)* para efeitos de inferência.

Uma alternativa à distribuição normal padrão é utilizar-se o bootstrap para estimar os quantis de *t*. Considere t^* o análogo bootstrap de *t*:

$$t^* \equiv \frac{\sqrt{n}(\hat{\mu}^* - \hat{\mu})}{\hat{\sigma}^*},$$

onde $\hat{\mu}^*$, $\hat{\mu}$ e $\hat{\sigma}^*$, são os análogos bootstrap de $\hat{\mu}$, μ e $\hat{\sigma}$, respectivamente. Em particular, $\hat{\sigma}^{*2}$ é dado pela expressão (2) mas avaliado com dados bootstrap $\mathcal{X}_n^*$ em vez de $\mathcal{X}_n$. O análogo bootstrap de μ é $\hat{\mu}$ dado que a media amostral bootstrap $\hat{\mu}^*$ é centrada em $\hat{\mu}$, i.e. $E^*(\hat{\mu}^*) = \hat{\mu}$, como demonstrado *em* (3).

Sob condições de regularidade fracas (veja e.g. Bickel e Freedman, 1981) podemos mostrar que o boostrap é assimptoticamente correcto de primeira ordem, i.e.

$$\sup_{x\in\mathbb{R}}|P^*(t^* \le x) - P(t \le x)| \xrightarrow{p} 0, \ com \ n \to \infty. \qquad (5)$$

Logo, podemos utilizar os quantis empíricos da distribuição bootstrap de t^* para aproximar os quantis da distribuição de *t*. Neste contexto i.i.d., podemos fortalecer (5) mostrando que o erro implícito na aproximação bootstrap converge para zero a uma taxa mais rápida do que o erro implícito na aproximação normal (veja Hall, 1992, Capítulo 3). Isto é conhecido na literatura sobre bootstrap como refinamento assimptótico do bootstrap.

3. A falha do bootstrap em dados dependentes

Suponha agora que $\{X_t: t=1,\ldots,n\}$ é uma amostra de *n* observações obtidas de um processo de séries cronológicas estritamente estacionário $\{X_t: t \in \mathbb{Z}\}$. Considere também que $\{X_t: t \in \mathbb{Z}\}$ é fracamente dependente no sentido em que $\{X_t: t \le 0\}$ e $\{X_t: t \ge k\}$ se tornam assimptoticamente independentes com $k \to \infty$. A média amostral $\hat{\mu}$ continua a ser um estimador consistente de $\mu = E(X_t)$ mas a sua variância já não é dada por (1). De facto, neste caso,

$$\sigma^2 \equiv Var\left(\sqrt{n}\hat{\mu}\right) = \gamma(0) + 2\sum_{\tau=1}^{n}\left(1 - \frac{\tau}{n}\right)\gamma(\tau), \qquad (6)$$

onde $\gamma(\tau) = Cov(X_t X_{t+\tau})$, para qualquer $\tau = 0,1,\ldots$. Note que $\gamma(0) = Var(X_t)$.

A variância dada em (6) é conhecida na literatura econométrica como a variância de longo prazo. A presença em (6) das autocovariâncias $\gamma(\tau)$ complica substancialmente o problema da estimação das variâncias. Também origina problemas para o bootstrap i.i.d., como foi observado por Singh (1981).

Considere uma amostra bootstrap $\mathcal{X}_n^* = \{X_t^*: t = 1, \ldots, n\}$ de $\mathcal{X}_n = \{X_t: t = 1, \ldots, n\}$ utilizando o bootstrap i.i.d. de Efron (1979). Condicional sobre $\mathcal{X}_n$, X_t^* é i.i.d. para qualquer *t=1, …, n*, implicando que a estrutura de dependência do conjunto dos dados originais é perdida. Em particular, σ^{*2}, a variância bootstrap de $\sqrt{n}\hat{\mu}^*$, é ainda dada por (4). Dado que $\sigma^{*2} \xrightarrow{p} \gamma(0)$, esta não considera as autocovariâncias $\gamma(\tau)$ em (6). Isto implica que o bootstrap i.i.d. não pode ser utilizado para estimar a variância nem os quantis da distribuição de $\sqrt{n}(\hat{\mu} - \mu)$. Como iremos ver abaixo, o bootstrap i.i.d. ainda pode ser utilizado para estimar os quantis da estatística *t* "studentized".

4. O bootstrap em bloco e aproximações assimptóticas relacionadas

A falha do bootstrap i.i.d. no contexto de series cronológicas motivou o desenvolvimento de métodos bootstrap alternativos para dados dependentes. Nesta secção, analisarei um desses métodos, o bootstrap de blocos móveis (MBB) (cf., Götze e Künsch, 1989 e Liu e Singh, 1992), e discutirei como é que ele se relaciona com métodos de inferência alternativos que se baseiam em teoria assimptótica de primeira ordem.

Estatísticas "Studentized" baseadas em erros padrão HAC

Sob condições de regularidade bem conhecidas (veja e.g. Newey e West, 1987 e Andrews, 1991), um estimador consistente da variância de longo-prazo, σ^2, em (6) é o estimador da variância HAC. Este tem a seguinte forma geral

$$\sigma^2_{HAC} = \hat{\gamma}(0) + 2\sum_{\tau=1}^{n} k\left(\frac{\tau}{n}\right)\hat{\gamma}(\tau), \qquad (7)$$

onde *k(x)* é uma função ponderadora tal que $k(x) = k(-x)$, $k(0) = 1$, $|k(x)| \leq 1$, $k(x)$ é contínuo em $x = 0$, e $\int_{-\infty}^{\infty} k^2(x) < \infty$. Aqui, $\hat{\gamma}(\tau) = n^{-1}\sum_{t=\tau+1}^{n}(X_t - \bar{X}_n)(X_{t-\tau} - \bar{X}_n)$ são as autocovariâncias amostrais relativas ao desfasamento τ de $\{X_t\}$. *M* é o parâmetro relativo à largura de banda, que pode funcionar como o desfasamento de truncagem para os ponderadores de modo que $k(x) = 0$ para $|x| > 1$. O ponderador de Bartlett utilizado no popular estimador HAC proposto por Newey-West (1987) é um exemplo.

Uma estatística "studentized" baseada em erros padrão HAC é dada por

$$t_{HAC} \equiv \frac{\sqrt{n}(\hat{\mu} - \mu)}{\hat{\sigma}_{HAC}}.$$

Tal como no contexto i.i.d. puro, a distribuição desta estatística t não é conhecida em amostras finitas. Logo, para efeitos de inferência temos que a aproximar.

A aproximação assimptótica da normal padrão

Suponha que $\hat{\sigma}^2_{HAC}$ é um estimador consistente de σ^2, que requer que $M \to \infty$ à medida que $n \to \infty$, mas $M/n \to 0$. Sob estas condições, $t_{HAC} \xrightarrow{d} N(0,1)$. Dado que a aproximação normal se baseia na consistência de $\hat{\sigma}^2_{HAC}$ para σ^2_{HAC} obtém-se a mesma distribuição limite normal padrão independentemente do ponderador ou do parâmetro relativo à largura de banda utilizado. Como estas escolhas têm impacto em amostras finitas, a aproximação normal padrão tem um mau desempenho em amostras finitas.

A aproximação assimptótica com *b*-fixo

Uma aproximação alternativa para $\hat{\sigma}^2_{HAC}$ foi proposta por Kiefer e Vogelsang (2005). Suponha que a largura de banda é modelada da seguinte forma $M = bT$, com *b* uma constante fixa em (0, 1]. Dado que b é mantido fixo, esta abordagem tem sido designada por abordagem assimptótica com *b*-fixo. Sob a abordagem assimptótica com b-fixo, $\hat{\sigma}^2_{HAC}$ converge para uma variável aleatória (e não para uma constante) que depende do ponderador e da largura de banda. Como consequência, t_{HAC} tem uma distribuição não convencional. Esta distribuição limite é útil para inferência dado que reflecte a escolha da largura de banda e do ponderador e é assimptoticamente *pivotal* (i.e. independente de parâmetros perturbadores) e os valores críticos podem ser tabulados. Por exemplo, sob condições de regularidade adequadas, Kiefer e Vogelsang (2005) mostraram que

$$t_{HAC} \xrightarrow{d} \frac{W(1)}{\sqrt{Q(b)}},$$

onde *W(r)* é um processo padrão de Wiener e *Q(b)* é uma variável aleatória que depende do ponderador utilizado.

Dado que a aproximação com b-fixo reflecte a escolha de *b* e a função ponderadora (através da forma de *Q(b)*), é esperado que tenha melhor desempenho em amostras finitas do que a aproximação *N(0,1)*. Isto foi confirmado teoricamente e por simulação.

A aproximação bootstrap em bloco

A ideia subjacente ao bootstrap em bloco é a de reamostragem dos blocos de observações consecutivas em vez de observações individuais. Desta forma, preservamos a estrutura de dependência dos dados originais em cada bloco e, desde que as observações sejam fracamente dependentes, a independência entre blocos é em termos assimptóticos de primeira ordem irrelevante. Para descrever o MBB, considere que l denota o tamanho do bloco e k o número de blocos. Suponha para simplificar que $k = n/$ l. Considere $B_{t,\ell} = \{X_t, X_{t+1}, \dots, X_{t+\ell-1}\}$ o bloco de l observações consecutivas com início em t (note que l = 1 corresponde ao bootstrap i.i.d. de Efron). O MBB procede à reamostragem de k blocos aleatoriamente com reposição do conjunto de n – l +1 blocos que se sobrepõem $\{B_{1,\ell}, B_{2,\ell}, \dots, B_{n-\ell+1,\ell}\}$. Assumindo que $I_1, \dots, I_k$ são variáveis aleatórias i.i.d distribuídas uniformemente em $\{1, \dots, n - l - 1\}$, temos que $\{X_t^* = X_{\tau_t} : t = 1, \dots, n\}$ onde τ_t define um índice aleatório dado por $\{\tau_t\} = \{I_1, \dots, I_{1+\ell-1}, \dots, I_k, \dots, I_{k+\ell-1}\}$.

Como no caso i.i.d., para o MBB também existe uma expressão fechada para $\sigma^{*2} = Var^*(\sqrt{n}\hat{\mu}^*)$. No entanto, e contrariamente a (4) para o bootstrap i.i.d., a expressão para a variância de MBB contem termos que dependem da função de autocovariâncias da amostra. De facto, podemos demonstrar que a variância MBB é assimptoticamente equivalente à de um estimador HAC baseado num ponderador de Bartlett. Logo, a variância de MBB é um estimador consistente da variância de longo-prazo.

Para obter uma aproximação à distribuição de t_{HAC} através do bootstrap em bloco, temos que construir uma estatística bootstrap "studentized". Várias escolhas existem na literatura, dependendo da escolha do estimador da variância bootstrap σ^{*2}.

Um abordagem natural consiste simplesmente em substituir os dados bootstrap pelos dados originais nas formulas utilizadas para construir a estatística t original. Em particular, constrói-se,

$$t^* = \frac{\sqrt{n}(\hat{\mu}^* - \hat{\mu})}{\hat{\sigma}^*_{HAC}}, \qquad (8)$$

onde $\hat{\sigma}^*_{HAC}$ é dado por (7), mas avaliado sobre os dados bootstrap $\{X_t^*\}$.[1]

Uma comparação das diferentes aproximações

De seguida ilustramos o desempenho em amostras pequenas das diferentes aproximações com base em alguns resultados de Monte Carlo obtidos por Gonçalves e Vogelsang (2009). Suponha que $X_t = \mu + \varepsilon_t$, onde $\mu = 0$ e $\varepsilon_t = \rho\varepsilon_{t-1} + (1-\rho^2)^{1/2}u_t$, com $\{u_t\}$~i.i.d.N(0,1), $\varepsilon_1 = 0$ e $\rho \in \{0, 0.5, 0.9\}$. O objectivo é testar H_0: $\mu \leq 0$ contra H_1: $\mu > 0$ a um nível de significância de 5% utilizando t_{HAC}. São geradas 10000 amostras aleatórias de tamanho $n = 50$. Rejeitamos a hipótese nula sempre que $t_{HAC} > t_c$, onde t_c é o valor crítico obtido com base em cada um dos três métodos discutidos anteriormente. O MBB baseia-se em 999 replicações bootstrap e consideramos l = 1 e l = 5. A Figura 1 apresenta as verdadeiras taxas de rejeição para 25 valores de M. Podemos sumariar os resultados da seguinte forma. Em primeiro lugar, a aproximação com b-fixo domina a aproximação N(0,1). Em segundo lugar, o bootstrap i.i.d. aplicado a t* segue quase exactamente a aproximação assimptótica com b-fixo. Logo, é assimptoticamente válido mesmo quando os dados são dependentes. Isto deve-se ao facto da distribuição assimptótica de t_{HAC} não depender da estrutura de dependência dos dados. Por fim, o MBB com tamanhos de blocos maiores do que um têm um desempenho superior à aproximação assimptótica com b-fixo (e normal) quando a dependência é forte.

[1] Esta abordagem foi designada como ``naive" por Davison e Hall (1993) e Götze e Künsch (1996), e os autores avisaram que esta não prometia refinamento assintoticos em relação à aproximação normal padrão. Em vez disso, eles sugeriram uma forma especial de recentrar e de "studentization" da estatística *t* do bootstrap.

5. Conclusão

Neste artigo foram revistas três aproximações diferentes da distribuição em amostras finitas dum teste t robusto baseado em estimadores HAC: a aproximação *N(0,1)*, a recentemente desenvolvida aproximação assimptótica com b-fixo e o *naive block bootstrap*, conforme analisado em Gonçalves e Vogelsang (2009). Um dos maiores desafios na aplicação destes métodos é a escolha da largura de banda/tamanho dos blocos, que para efeitos de brevidade não foram considerados neste artigo.

6. Referências

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation, Econometrica, 59, 817-858.

Bickel, P. e D. Freedman, 1981. Some asymptotic theory for the bootstrap, Annals of Statistics, 9, 1196-1217.

Davison, A.C. e P. Hall, 1993. On studentizing and blocking methods for implementing the bootstrap with dependent data, Australian Journal of Statistics, 35, 215-224.

Efron, B., 1979. Bootstrap methods: another look at the jackknife, Annals of Statistics, 7, 1-26.

Gonçalves, S., e T. Vogelsang, 2009. Block bootstrap HAC robust tests: the sophistication of the naive bootstrap, mimeo, Université de Montréal.

Götze, F., e H.R. Künsch, 1996. Second-order correctness of the blockwise bootstrap for stationary observations, Annals of Statistics, 24, 1914-1933.

Hall, P., 1992. The bootstrap and Edgeworth expansion. Springer, New York.

Kiefer, N.M. e T. J. Vogelsang, 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. Econometric Theory, 21, 1130-1164.

Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations, Annals of Statistics, 17, 1217-1241.

Liu, R.Y., e K. Singh, 1992. Moving blocks jackknife and bootstrap capture weak dependence, in Exploring the Limits of the Bootstrap, ed. by R. LePage and L. Billiard. New York: Wiley.

Newey, W. e K.D. West, 1987. A simple positive semi-definite, heteroskedastic and autocorrelation consistent covariance matrix, Econometrica, 55, 703-708.

Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. Annals of Statistics, 9, 1187-1195.

O Método Generalizado dos Momentos

Joaquim J. S. Ramalho, *jsr@uevora.pt*

Departamento de Economia e CEFAGE-UE
Universidade de Évora

1. Introdução

Durante muitos anos, foi prática corrente em Econometria a utilização quase exclusiva de modelos e métodos de estimação que requerem fortes pressupostos distribucionais. A veracidade desses pressupostos era raramente colocada em causa, justificando-se a sua adopção pelas propriedades óptimas que os estimadores assim obtidos potencialmente poderiam ter. Esta prática foi sendo abandonada a pouco e pouco, sendo hoje comum quer a aplicação de testes que permitem avaliar a adequabilidade dos pressupostos assumidos quer a utilização de métodos de estimação menos exigentes em termos de pressupostos.

Um dos métodos de estimação que é actualmente bastante popular em Econometria é o Método Generalizado dos Momentos (abreviado, de ora em diante, por GMM, nome pelo qual é conhecido na literatura em língua inglesa). Para poder ser aplicado, este método requer simplesmente a especificação de um certo número de condições de momentos, as quais são função das variáveis e dos parâmetros de interesse do modelo. Embora Karl Pearson tenha sido o primeiro investigador a reconhecer a possibilidade de utilizar condições de momentos como base para a estimação de parâmetros há mais de cem anos atrás, foi apenas após a publicação do artigo pioneiro de Hansen (1982) sobre o GMM que essa forma de estimação alternativa se popularizou de um modo extraordinário na literatura econométrica.

Na base deste desenvolvimento, para além da óbvia vantagem a nível dos pressupostos que é necessário assumir, estão dois factos principais. Por um lado, o GMM inclui vários estimadores igualmente populares como casos particulares (mínimos quadrados, variáveis instrumentais, máxima verosimilhança, etc.), o que permite estudá-los de uma forma integrada. Por outro lado, devido à sua flexibilidade e generalidade, o GMM pode ser facilmente aplicado à estimação de modelos não lineares que de outra forma exigiriam pressupostos adicionais e, mesmo assim, seriam muito complicados de estimar.

Neste artigo, descreve-se de forma sumária em que consiste o GMM e quais as suas principais aplicações, limitações e alternativas. Para uma descrição mais pormenorizada, deve-se consultar, por exemplo, Newey e McFadden (1994), Mátyás (1999) e Hall (2005).

2. Do Método dos Momentos ao Método Generalizado dos Momentos

Em Estatística, o termo *momento* é habitualmente usado para designar o valor esperado de uma determinada potência de uma variável aleatória. Por exemplo, o momento de ordem r da variável aleatória discreta y com função de probabilidade $f(y)$ definida no espaço amostral S é calculado como:

$$E\left(y^r\right)=\mu_r=\sum_S f\left(y\right)y^r .$$

Na ausência de conhecimento sobre *f(y)*, o Método dos Momentos (MM) sugerido por Pearson permite estimar μ_r através da resolução da *condição de momentos amostral*:

$$\frac{1}{N}\sum y_i^r-\hat{\mu}_r=0,$$

a qual representa a contrapartida na amostra de $E\left(y^r-\mu_r\right)=0$.

A aplicação do MM pode ter por base mais do que uma condição de momentos amostral mas é necessário que o número de parâmetros a estimar seja igual ao número de condições de momentos. Por vezes, o número de condições de momentos que é possível e faz sentido ter em conta pode ser superior ao número de parâmetros do modelo, o que implica a necessidade de seleccionar previamente quais as condições que devem ser usadas na estimação e, por consequência, quais as que devem ser excluídas. A impossibilidade de usar toda a informação disponível sobre o modelo de interesse é uma grande desvantagem do MM, a qual é evitada pelo GMM da forma que se descreve de seguida.

Vamos supor é possível definir *s* condições de momentos,

$$E\left[\,g\left(z,\beta\right)\right]=0, \qquad (1)$$

onde *g(z,β)* representa uma determinada função das variáveis *z* e dos *k* parâmetros *β* do modelo de interesse, com $s \geq k$. Um estimador para *E[g(z,β)]* é naturalmente dado por

$$g_n\left(z,\hat{\beta}\right)=\frac{1}{N}\sum_{i=1}^{N} g\left(z_i,\hat{\beta}\right),$$

onde $\hat{\beta}$ representa um estimador consistente de *β*. A questão que se coloca é como obter $\hat{\beta}$ usando a informação contida em (1).

A ideia subjacente ao GMM é muito simples. O vector de parâmetros de interesse é estimado tendo por objectivo obter para $g_n\left(z,\hat{\beta}\right)$ um valor tão próximo de zero quanto possível de modo a reflectir aquilo que acontece na população. Se o número de condições de momentos e de parâmetros for idêntico ($s = k$), então o GMM corresponde a uma aplicação trivial do MM, sendo $\hat{\beta}$ obtido através da resolução do sistema $g_n\left(z,\hat{\beta}\right)=0$. O mesmo procedimento não pode ser aplicado quando o modelo está sobre-identificado ($s > k$) pois em geral não existe nenhum valor de $\hat{\beta}$ que permita obter aquela igualdade. Neste caso, o número de equações tem de alguma forma de ser reduzido para *k*, tendo Hansen (1982) proposto usar *k* combinações lineares das *s* condições de momentos iniciais.

Em qualquer dos casos, o estimador GMM para *β* corresponde, por definição, ao valor de $\hat{\beta}$ que minimiza a seguinte função quadrática das condições de momentos amostrais:

$$\left[\frac{1}{N}\sum_{i=1}^{N} g\left(z_i,\hat{\beta}\right)'\right]W_n\left[\frac{1}{N}\sum_{i=1}^{N} g\left(z_i,\hat{\beta}\right)\right], \qquad (2)$$

onde W_n é uma matriz simétrica *s* x *s* de ponderadores que pode depender das observações e converge para a matriz positiva definida *W*. Daqui resulta que as *k* condições de primeira ordem que caracterizam o estimador GMM são dadas por

$$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial g\left(z_i,\hat{\beta}\right)'}{\partial\beta}\right]W_n\left[\frac{1}{N}\sum_{i=1}^{N}g\left(z_i,\hat{\beta}\right)\right]=0\,,$$

as quais implicitamente definem as combinações lineares propostas por Hansen (1982).

Tal como demonstrado por Hansen (1982), qualquer que seja a escolha efectuada para a matriz *W*, o estimador GMM é consistente e assimptoticamente normal. Pelo contrário, diferentes escolhas para *W* conduzem a estimadores GMM com diferentes níveis de eficiência. Hansen (1982) demonstrou que o nível máximo de eficiência é atingido quando *W* corresponde ao inverso da matriz de covariâncias das condições de momentos, definida por $V \equiv E\left[g\left(z,\beta\right)g\left(z,\beta\right)'\right]$.

A matriz óptima de ponderadores, V^{-1}, depende de β, o que significa que, em termos práticos, é necessário dispor de uma estimativa inicial desse parâmetro. Desde que essa estimativa seja consistente, as propriedades assimptóticas do GMM não são afectadas (Hansen, 1982). Normalmente, essa estimativa preliminar de β é obtida aplicando numa primeira fase o GMM usando como ponderadora a matriz identidade. O estimador GMM eficiente pode assim ser interpretado como um estimador GMM a dois passos. Para estimadores GMM alternativos, pode-se consultar Hansen, Heaton and Yaron (1996).

Como referido anteriormente, o GMM inclui como casos particulares muitos estimadores amplamente conhecidos. No âmbito do GMM, esses estimadores diferem entre eles apenas pelas diferentes funções $g(z,\beta)$ que os caracterizam. Por exemplo, o estimador dos mínimos quadrados é obtido quando se assume que $E(X'u) = 0$, onde *X* representa a matriz de regressores e *u* é o termo erro da regressão. Similarmente, $E(T'u) = 0$, onde *T* representa uma matriz de instrumentos, produz o estimador das variáveis instrumentais (pressupondo que se assume a existência de homocedasticidade), enquanto que quando $g(.)$ representa a função score se obtém o estimador da máxima verosimilhança.

3. Testes de especificação

No âmbito do GMM, o teste de especificação mais conhecido é o chamado teste de sobre-identificação, ou teste *J*, proposto por Hansen (1982). A ideia por trás deste teste é também muito simples e intuitiva, baseando-se no facto de se utilizarem *s* condições de momentos quando apenas *k* seriam necessárias para estimar os parâmetros de interesse, isto é, existem $s - k$ condições de sobre-identificação. O modo mais evidente de testar a especificação de um modelo definido por (1) consiste em verificar se o valor de todas as condições de momentos amostrais é aproximadamente zero ou não, já que o GMM apenas impõe que *k* combinações lineares delas o sejam. Assim, a estatística *J* é dado simplesmente pelo produto de *N* pelo valor da função objectivo (2), tendo uma distribuição assimptótica de qui-quadrado com $s - k$ graus de liberdade.

A avaliação de restrições paramétricas no contexto do GMM pode ser feita de modo similar ao que acontece com outros estimadores. Para este fim, Newey e West (1987) derivaram testes Wald, score e do tipo LR, enquanto Newey e McFadden (1994) desenvolveram testes de Hausman. Testes para a validade de sub-conjuntos de condições de momentos foram propostos por Newey (1985) e Eichenbaum, Hansen e Singleton (1988). Finalmente, testes para hipóteses não encaixadas foram desenvolvidos por Singleton (1985) e Smith (1992).

4. Aplicações

O GMM começou por ganhar maior notoriedade na área das séries temporais mas tem sido também bastante utilizado em aplicações com dados seccionais ou de painel. Para exemplos de aplicações possíveis do GMM, consultar Ogaki (1993) e Hall (2005).

Com dados seccionais, o GMM tem sido usado essencialmente como uma alternativa ao método dos mínimos quadrados a dois passos quando se suspeita da existência de heterocedasticidade em modelos de regressão linear. Outras aplicações incluem certos modelos de regressão exponencial com variáveis explicativas endógenas (Mullahy, 1997), modelos para amostras sujeitas a estratificação endógena (Imbens, 1992) e modelos microeconométricos corrigidos para a não resposta (Ramalho e Smith, 2009).

Com dados de natureza temporal, uma das grandes vantagens do GMM é a de permitir lidar com problemas de autocorrelação de modo relativamente simples, através da utilização de uma matriz ponderadora que reflicta essa situação. Outra vantagem é a possibilidade de se trabalhar apenas com as condições de momentos que são efectivamente implicadas pela teoria económica, sem necessidade de adicionar pressupostos distribucionais, como se fazia anteriormente em modelos não lineares de expectativas racionais (Hansen e Singleton, 1982). Exemplos de aplicações do GMM no contexto temporal incluem modelos de ciclos de negócios, modelos de volatilidade estocástica e modelos de avaliação de activos.

Algumas das mais recentes e interessantes aplicações do GMM ocorreram na estimação de modelos para dados de painel, nomeadamente em casos em que existe heterogeneidade não observada. Sob o pressuposto de que as variáveis explicativas não estão correlacionadas com o termo erro em nenhum período de tempo (excluindo o chamado efeito fixo), é possível construir uma multiplicidade de estimadores GMM através da adição de condições de ortogonalidade representando essa ausência de relação, os quais são naturalmente mais eficientes que o estimador de efeitos fixos tradicional. Quando o modelo contém ainda a variável dependente desfasada entre os regressores, então o GMM é já o método de eleição para obter estimadores consistentes para os parâmetros de interesse; ver, por exemplo, Arellano e Bond (1991) e Blundell e Bond (1998).

5. Limitações e métodos alternativos

Apesar da sua corrente popularidade, o GMM possui algumas características indesejadas. Acima de tudo, a distribuição assimptótica dos estimadores GMM parece constituir uma aproximação de fraca qualidade à sua distribuição em amostras finitas. De facto, existe uma grande variedade de estudos de simulação de Monte Carlo que mostram claramente que os estimadores GMM para os parâmetros de interesse são por norma claramente enviesados em pequenas amostras, enquanto que a dimensão estimada dos testes de especificação associados a esses estimadores é frequentemente substancialmente diferente da esperada de acordo com a teoria assimptótica.

O nível de preocupação acerca das propriedades dos estimadores GMM em amostras finitas tem sido tal que em 1996 uma edição da prestigiada revista *Journal of Business Economics & Statistics* foi integralmente dedicada a esta questão. Entre outros aspectos, Andersen e Sorensen (1996) confirmaram que o comportamento do GMM em modelos de volatilidade estocástica decai substancialmente à medida que o número de instrumentos (isto é, condições de momentos) aumenta e Altonji e Segal (1996) concluíram que o estimador GMM eficiente a dois passos pode sofrer de níveis de enviesamento muito superiores ao do estimador GMM não eficiente baseado na matriz de identidade.

Naturalmente, dado o comportamento inadequado do GMM em amostras de reduzida dimensão, têm vindo a ser sugeridos métodos de estimação alternativos para lidar com modelos definidos por condições de momentos. Na mesma edição do *Journal of Business Economics & Statistics*, Hansen, Yeaton and Yaron (1996) propuseram o *continuous-updating* GMM, que difere do GMM a dois passos pelo facto da matriz ponderadora V^{-1}, que depende dos parâmetros de interesse β, não ser estimada num primeiro passo mas sim em simultâneo com β. Estes autores demonstraram que os dois estimadores são assimptoticamente equivalentes e que em pequenas amostras o novo estimador é aproximadamente centrado em termos medianos. Contudo, a sua computação é muito mais complicada e tende a exibir níveis de dispersão muito mais elevados, pelo que a utilização do *continuous-updating* GMM em aplicações econométricas tem sido praticamente nula.

Existe ainda uma classe de estimadores alternativos ao GMM que tem sido largamente estudada a nível teórico mas, dada a sua difícil computação, raramente tem sido usada a nível prático. Essa classe, chamada de verosimilhança empírica generalizada (GEL) por Newey e Smith (2004), tem três características muito atractivas: (i) ao contrário do GMM, não é necessário definir nenhuma matriz de ponderadores; (ii) também em oposição ao GMM, uma versão ponderada de *todas* as condições de momentos são satisfeitas na amostra, em vez de apenas *uma combinação linear* delas; e (iii) embora equivalente ao GMM em termos de teoria assimptótica de primeira ordem, as propriedades assimptóticas de ordem superior do GEL parecem ser muito mais apelativas.

Na literatura estatística, o método da verosimilhança empírica foi introduzido por Owen (1988). A sua introdução na literatura econométrica e aplicação a modelos de condições de momentos foi feita por Qin and Lawless (1994) e Imbens (1997). Outros métodos semelhantes, que também pertencem à classe GEL, nomeadamente o método de *exponential tilting*, foram posteriormente desenvolvidos por Kitamura e Stutzer (1997) e Imbens, Spady and Johnson (1998).

Todos os estimadores GEL partilham uma característica comum: utilizam um estimador da função de densidade dos dados que é mais eficiente do que o usado pelo GMM pois, ao contrário deste, na sua construção é também explorada a informação contida nas condições de momentos. Assim, enquanto que o GMM se baseia na função de densidade empírica, que atribui o mesmo peso a cada observação, a função de densidade GEL atribui um peso diferente a cada observação, sendo esse peso estimado (em simultâneo com os parâmetros de interesse) de forma a impor na amostra todas as condições de momentos. Para detalhes sobre os estimadores GEL, consultar os artigos citados.

Finalmente, outras alternativas ao GMM incluem o Método Simulado dos Momentos e o Método Eficiente dos Momentos. Dada a sua complexidade, também estes métodos têm sido pouco utilizados em trabalho aplicado. Para *surveys* sobre estes métodos, consultar Carrasco e Florens (2002) e Gourieroux e Monfort (1996). A possibilidade de usar técnicas de bootstrap na correcção do enviesamento do GMM em pequenas amostras tem também sido alvo de alguns estudos. Hall e Horowitz (1996), Brown e Newey (2002) e Ramalho (2006) propuseram métodos bootstrap alternativos que, de acordo com a evidência obtida até ao momento através de estudos de Monte Carlo, parecem funcionar razoavelmente bem na atenuação das distorções geralmente apresentadas pelo GMM em pequenas amostras.

6. Conclusão

O GMM tem por finalidade obter estimadores para os parâmetros de modelos que são apenas definidos por condições de momentos. A maior parte dos modelos econométricos pode ser expressa desta forma, o que realça o importante papel que o GMM desempenha na Econometria. A sua aplicação torna-se mesmo essencial em certos modelos, como modelos não lineares de expectativas racionais e modelos dinâmicos para dados de painel com efeitos fixos. Pese embora o risco que representa a sua utilização em amostras de pequena dimensão, estamos convencidos que o GMM continuará a ser o método mais usado em trabalho aplicado na estimação de modelos definidos por condições de momentos, dada a complexidade das alternativas existentes e a possibilidade de aplicar correcções usando o bootstrap ou métodos similares.

Bibliografia

Altonji, J.G. and Segal, L.M. (1996), "Small-sample bias in GMM estimation of covariance structures", *Journal of Business & Economic Statistics*, 14(3), 353-365.

Andersen, T.G. and Sorensen, B.E. (1996), "GMM estimation of a stochastic volatility model: a Monte Carlo study", *Journal of Business & Economic Statistics*, 14(3), 328-352.

Arellano, M. and Bond, S. (1991), “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations”, *Review of Economic Studies*, 58, 277-297.

Blundell R. and Bond, S. (1998), “Initial conditions and moment restrictions in dynamic panel data models”, *Journal of Econometrics*, 87, 115-143.

Brown, B.W. and Newey, W.K. (2002), “Generalised method of moments, efficient bootstrapping, and improved inference”, *Journal of Business and Economic Statistics,* 20, 507-517.

Carrasco, M. and Florens, J.P. (2000), “Generalization of GMM to a continuum of moment conditions”, *Econometric Theory*, 16, 797-834.

Eichenbaum, M.S., Hansen, L.P. and Singleton, K.J. (1988), "A time series analysis of representative agent models of consumption and leisure choice under uncertainty", *Quarterly Journal of Economics*, 103, 51-78.

Gourieroux, C. and Monfort, A. (1996), *Simulation-Based Econometric Methods*, Oxford University Press.

Hall, A. (2005), *Generalized Method of Moments*, Oxford University Press.

Hall, P. and Horowitz, J.L. (1996), "Bootstrap critical values for tests based on generalised-method-of-moments estimators", *Econometrica*, 64(4), 891-916.

Hansen, L.P. (1982), "Large sample properties of generalised method of moments estimators", *Econometrica*, 50(4), 1029-1054.

Hansen, L.P., Heaton, J. and Yaron, A. (1996), "Finite-sample properties of some alternative GMM estimators", *Journal of Business & Economic Statistics*, 14(3), 262-280.

Hansen, L.P. and Singleton, K.J. (1982), “Generalized instrumental variables estimation of nonlinear rational expectations models”, *Econometrica*, 50(59, 1269-1286.

Imbens, G.W. (2002), "Generalized method of moments and empirical likelihood", *Journal of Business & Economic Statistics*, 20(4), 493-506.

Imbens, G.W. (1997), "One-step estimators for over-identified generalised method of moments models", *Review of Economic Studies*, 64, 359-383.

Imbens, G.W., Spady, R.H. and Johnson, P. (1998), "Information theoretic approaches to inference in moment condition models", *Econometrica*, 66(2), 333-357.

Kitamura, Y. and Stutzer, M. (1997), "An information-theoretic alternative to generalised method of moments estimation", *Econometrica*, 65(4), 861-874.

Mátyás, L. (1999), *Generalized Method of Moments Estimation*, Cambridge University Press.

Mullahy, J. (1997), "Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior", *Review of Economics and Statistics*, 79(4), 586-593.

Newey, W.K. (1985b), "Maximum likelihood specification testing and conditional moment tests", *Econometrica*, 53(5), 1047-1070.

Newey, W.K. and McFadden, D. (1994), "Large sample estimation and hypothesis testing", in Engle, R. F. and McFadden, D. L. (eds.), *Handbook of Econometrics*, Vol. 4, Elsevier Science Publishers, pp. 2111-2245

Newey, W.K. and Smith, R.J. (2004), "Higher order properties of GMM and generalized empirical likelihood estimators", *Econometrica*, 72(1), 219-255.

Newey, W.K. and West, K.D. (1987), "Hypothesis testing with efficient method of moments estimation", *International Economic Review*, 28, 777-787.

Ogaki, M. (1993), "Generalized method of moments: econometric applications", in Maddala, G. S., Rao, C. R. and Vinod, H. D. (eds.), *Handbook of Statistics*, Vol. 11, Elsevier Science Publishers, pp. 455-488.

Owen, A.B. (1988), "Empirical likelihood ratio confidence intervals for a single functional", *Biometrika*, 75(2), 237-249.

Qin, J. and Lawless, J. (1994), "Empirical likelihood and general estimating equations", *Annals of Statistics*, 22(1), 300-325.

Ramalho, J.J.S. (2006), "Bootstrap bias-adjusted GMM estimators", *Economics Letters*, 92(1), 149-155.

Ramalho, E.A. and Smith, R.J. (2009), "Discrete choice nonresponse", mimeo.

Singleton, K.J. (1985), "Testing specifications of economic agents' intertemporal optimum problems in the presence of alternative models", *Journal of Econometrics*, 30, 391-413.

Smith, R.J. (1992), "Non-nested tests for competing models estimated by generalised method of moments", *Econometrica*, 60(4), 973-980.

Dados de Painel

Paulo Guimarães, *pguimaraes2001@gmail.com*

Universidade da Carolina do Sul, EUA

1. Introdução

O termo "dados em painel" ou "dados longitudinais" é vulgarmente utilizado para designar bases de dados constituídas por variáveis que integram observações com uma dimensão seccional e temporal. Exemplos deste tipo de dados são observações para um conjunto de indivíduos, firmas ou países ao longo do tempo. Contudo os dados em painel podem integrar mais de duas dimensões ou ter outras dimensões que não espaço e tempo. Para assentar ideias iremos considerar como referência a situação mais comum em que os dados integram apenas duas dimensões, seccional e temporal.

O aumento do número de bases de dados em painel disponíveis para investigação encontra paralelo na crescente procura por este tipo de dados. A razão para este interesse nos dados em painel tem a ver com as reconhecidas vantagens que se lhes atribuem. Em primeiro lugar, pela sua própria natureza, os dados em painel permitem uma inferência mais precisa, pois lidam com um muito maior número de observações (e graus de liberdade) do que os dados puramente seccionais ou temporais. Visto que tratam com múltiplas observações para a mesma unidade, estes dados permitem controlar para características não-observadas dessas mesmas unidades. Também, porque misturam as diferenças inter-unidades com a dinâmica intra-unidades, permitem-nos estudar a importância do desfasamento temporal no comportamento das unidades. Outras vantagens dos dados em painel são a possibilidade de permitirem o teste de hipótese de comportamento mais sofisticadas assim como o estudo das fundações micro na análise de dados agregados. Recordemos que a análise de dados agregados é tipicamente baseada na premissa do "agente representativo" mas a existência de heterogeneidade individual pode por em causa a validade da análise agregada.

Trataremos aqui do caso em que pretendemos modelizar o comportamento de uma variável y_{it} composta por N unidades seccionais observadas ao longo de T períodos de tempo, sendo que $i=1,\ldots, N$ e $t = 1,\ldots, T$. Designaremos genericamente por x_{it} as variáveis explicativas destes modelos. Um painel de dados é considerado balanceado se existem observações para todas as variáveis para todas as unidades seccionais em todos os períodos de tempo. Se tal não acontece então trata-se de um painel não-balanceado. A existência de dados não-balanceados não é por si um problema, desde que o mecanismo gerador dos dados em falta não seja endógeno ao modelo.

2. Modelos Lineares de Dados em Painel

Um dos modelos mais comuns para dados em painel modeliza a heterogeneidade não observada utilizando uma regressão linear simples mas permitindo um intercepto diferente para cada unidade do painel. Neste caso,

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}$$

onde β é o vector de coeficientes associado às variáveis explicativas, α_i é uma variável aleatória que

captura a heterogeneidade não observada e ε_{it} é o termo de perturbação usual. Neste caso α_i captura todas as características da unidade que se mantêm constantes ao longo do tempo quer sejam observadas ou não. Por exemplo, se a unidade forem indivíduos então características como o sexo ou a naturalidade são capturadas por α_i.

O modo de tratamento dos α_i determina o tipo de modelo a usar. Se assumirmos que α_i não está correlacionado com x_{it} então os α_i poderão ser tratados como um termo de perturbação adicional. Este tipo de modelos são designados como modelos de "efeitos aleatórios". Se assumirmos que os α_i estão correlacionados com x_{it} então passamos a lidar com um modelo de "efeitos fixos" e a solução agora consiste em estimar os α_i (os "efeitos fixos") ou efectuar uma transformação do modelo que remova os α_i mas permita ainda a estimação dos coeficientes associados às variáveis de interesse. Note-se que neste contexto a designação "efeito fixo" tem um significado diferente daquele comummente utilizado na Estatística.

No caso do modelo linear, o estimador de "efeitos aleatórios" é implementado assumindo que α_i e ε_{it} são variáveis aleatórias i.i.d. não correlacionadas, homocedásticas e de média zero. Baseado nestas hipóteses é fácil calcular a matriz de variâncias e covariâncias de $\alpha_i + \varepsilon_{it}$ e a partir daí aplicar o estimador de "mínimos quadrados generalizados exequíveis" que, como é sabido, é consistente e assimptoticamente eficiente. Note-se que se aplicarmos directamente mínimos quadrados, ignorando a estrutura de correlação dos erros, obteremos na mesma estimadores consistentes mas não eficientes para β (chama-se a este modelo, o modelo "pooled" para dados em painel). Por contrapartida, as estimativas dos desvios-padrão dos estimadores de mínimos quadrados virão incorrectamente calculadas pois ignoram a existência de correlação (temporal) entre observações para a mesma unidade.

Quando assumimos a existência de "efeitos fixos" estamos a admitir a possibilidade de existência de correlação entre os α_i e ε_{it}. Neste caso existem vários estimadores que permitem lidar com os "efeitos fixos". O mais comum é o estimador "within" usualmente obtido a partir de uma regressão que transforma todas as variáveis por subtracção das médias temporais, isto é, efectuando uma regressão do tipo

$$y_{it} - \bar{y}_i = \alpha_i + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$$

em que todas as variáveis são calculadas como desvios de cada unidade do painel em relação à sua média temporal. Esta transformação elimina todas as variáveis que não exibem variação temporal incluindo obviamente os α_i. Se a dimensão seccional for pequena então o estimador "within" poderá ser obtido de outra forma sem necessidade de transformar o modelo. Bastará neste caso estimar um modelo pelo método dos mínimos quadrados que para além de x_{it} inclui ainda *N* variáveis "dummy" (mas exclui o intercepto da regressão) que identificam se a observação pertence ou não à unidade do painel. Um outro estimador para modelos com efeitos fixos, é o estimador de primeiras-diferenças. Este estimador é obtido aplicando mínimos quadrados às primeiras diferenças temporais dos dados

$$y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{it-1})$$

Admitindo que ε_{it} segue as hipótese habituais então para $T > 2$ o estimador "within" é mais eficiente do que o estimador a primeiras-diferenças. Deverá ser realçado que se de facto se verificar a hipótese de correlação entre os α_i e x_{it} então os estimadores de mínimos quadrados ou os estimadores de "efeitos aleatórios" são inconsistentes. Esta é uma das razões porque os economistas tendem a preferir os estimadores de efeitos fixos, pois estes estimadores mantêm-se consistentes independentemente da existência ou não de correlação entre α_i e x_{it} embora sejam menos eficientes que o estimador de efeitos aleatórios se de facto essa correlação for nula.

Do ponto de vista prático existe uma outra distinção importante entre "efeitos fixos" e "efeitos aleatórios". No modelo com "efeitos aleatórios" é possível identificar tanto o efeito marginal $\partial E[y_{it}|\alpha_i, \mathbf{x}_{it}]/\partial \mathbf{x}_{it}$ assim como $E[y_{it}|\mathbf{x}_{it}]$ enquanto que no modelo com "efeitos fixos" apenas é possível identificar os efeitos marginais para variáveis que tenham variação temporal (isto acontece porque os "efeitos fixos" absorvem todos os factores que são constantes na dimensão temporal como por exemplo o sexo do indivíduo). Mesmo assim as estimativas obtidas podem ser muito imprecisas se a maior parte da variabilidade for do tipo seccional.

Como vimos anteriormente a opção entre um estimador de efeitos aleatórios ou fixos não é inconsequente. Se o investigador estiver a lidar com dados obtidos a partir de um experimento controlado então fará todo o sentido a utilização do estimador de efeitos aleatórios pois não existe razão para suspeitar da existência de correlação entre α_i e x_{it}. Noutras circunstâncias poderá ser utilizado um teste de Hausman para ajudar a decidir qual o modelo apropriado. Se o teste de Hausman mostra uma diferença estatisticamente significativa entre os dois estimadores então isto deverá ser interpretado como evidência a favor do modelo de efeitos fixos.

A aplicação do estimadores de "efeitos fixos" ou "aleatórios" poderá não remover completamente a correlação entre os membros da mesma unidade, pelo que uma atitude mais conservadora consiste em utilizar um estimador robusto (tipo White/Huber) para a matriz de variâncias e covariâncias ou "bootstrapping" (note-se que estas técnicas têm de ser adaptadas à estrutura de painel dos dados). Ao fazer isto estamos também a prevenir contra a possível existência de heterogeneidade no termo de perturbação habitual.

A verificação das propriedade assimptóticas dos estimadores requer que se especifique se o aumento da dimensão amostral é efectuado por via de *N* ou *T* (ou ambos). Em grande parte das aplicações faz mais sentido admitir que seja *N* a tender para infinito pois a dimensão temporal dos painéis é geralmente pequena enquanto a dimensão seccional é elevada, centrando-se o interesse na modelização da heterogeneidade não-observada. Neste caso, e para o modelo com efeitos fixos, as estimativas dos α_i são inconsistentes porque o número de parâmetros α_i aumenta com *N*, embora as estimativas de β não venham afectadas. Se por outro lado tivermos um painel com uma dimensão temporal elevada poderá fazer sentido modelizar também a autocorrelação temporal.

Os modelo lineares discutidos anteriormente assumem a inexistência de correlação entre o termo de perturbação e os restantes termos do modelo. Mas esta hipótese é inconsistente com modelos que incluem entre os regressores variáveis endógenas ou variáveis dependentes desfasadas. Para lidar com este problema torna-se necessário utilizar variáveis instrumentais. Esta tarefa é facilitada pela natureza dos dados, pois podemos utilizar os desfasamentos temporais de uma variável como seu próprio instrumento. O método de estimação habitual destes modelos é o método dos momentos generalizados (GMM) tipicamente aplicado a um modelo transformado por forma a eliminar os efeitos individuais. O conhecido estimador de Arellano-Bond, utilizado para lidar com modelos de painel dinâmicos, ou seja, modelos do tipo,

$$y_{it} = \gamma y_{it-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

é um exemplo dum estimador de painel GMM.

3. Modelos não lineares de Dados em Painel

A modelização utilizada para lidar com dados em painel nos modelos lineares pode ser estendida a modelos não lineares. No entanto os resultados conhecidos para o modelo linear não são generalizáveis aos modelos não-lineares. Por exemplo, em painéis curtos, a introdução de efeitos fixos em vários modelos não-lineares torna os estimadores dos parâmetros associados aos x_{it} inconsistentes. Este problema é conhecido na literatura como o problema dos parâmetros incidentais. Note-se que isto não acontece no modelo linear, onde para painéis curtos os estimadores de β são consistentes embora os estimadores dos α_i sejam inconsistentes. Para alguns modelos não-lineares como por exemplo o modelo logit para dados binários e o modelo de regressão de Poisson é possível eliminar os α_i calculando a função de máxima verosimilhança condicionada a uma estatística suficiente dos α_i. A maximização da função de máxima verosimilhança condicionada produz estimadores consistentes para β embora (com a excepção do modelo linear e modelo de Poisson) esses estimadores sejam menos eficientes. A utilização da máxima verosimilhança condicionada apenas é possível para alguns modelos não-lineares. Uma outra opção para a estimação de modelos com efeitos fixos consiste na inclusão explícita de variáveis dummy para os efeitos individuais. Esta alternativa funcionará se o número de variáveis dummy for relativamente pequeno, caso contrário poderá ser impraticável estimar um modelo com um elevado número de regressores.

Quanto aos modelos de efeitos aleatórios convém notar que, tal como para o modelo linear, a existência de correlação entre os efeitos individuais e os outros regressores torna os estimadores inconsistentes. Desde que seja possível estimar de forma consistente o modelo com efeitos fixos então poderá ser implementado um teste de Hausman por forma a decidir qual a modelização apropriada. Nos modelos com efeitos aleatórios admite-se que os efeitos individuais seguem uma distribuição conhecida. Dependendo do modelo em causa pode ser conveniente assumir uma distribuição particular para os efeitos aleatórios que permita obter de forma explícita a distribuição incondicional dos dados. Por exemplo, se num modelo de Poisson para dados em painel se assumir que os efeitos aleatórios seguem uma distribuição gama então a distribuição incondicional dos dados será uma distribuição binomial negativa. Na maior parte dos casos não é possível obter de forma explícita a distribuição incondicional mas mesmo assim é possível maximizar a função de máxima verosimilhança utilizando métodos numéricos.

Hoje em dia vários "packages" estatísticos integram estimadores próprios para dados em painel. Os modelos de efeitos aleatórios que temos vindo a discutir são um caso particular dos modelos mistos ("mixed models") e portanto podem ser estimados com qualquer package estatístico que suporte modelos mistos como por exemplo o SAS, o R e o S-plus. No entanto, "packages" como o Stata e o LIMDEP são especializados neste tipo de dados e oferecem uma enorme variedade de estimadores.

4. Referências Bibliográficas

Cameron, C. and P. Trivedi (2005), Microeconometrics: Methods and Applications. Cambridge: Cambridge University Press.

Davidson, R. and J. MacKinnon (1993), Estimation and Inference in Econometrics. New York: Oxford University Press.

Greene, W. (2004), The Behavior of the Fixed effects Estimator in Nonlinear Models,, *Econometrics Journal*, Vol. 7, pp. 98-119.

Greene, W. H. (2003). Econometric Analysis. 5th ed. Upper Saddle River: Prentice Hall.

Woolridge, J. (2002). Econometric Analysis of Cross-Section and Panel Data. MIT Press.

Loxodromia da vida humana: Uma introdução à análise estatística da duração*

Carlota Louro, *carlouro.spub@fcm.unl.pt*
Faculdade de Ciências Médicas, Universidade Nova de Lisboa

Pedro Portugal, *pportugal@bportugal.pt*
Banco de Portugal e Faculdade de Economia, Universidade Nova de Lisboa

"Assi se estima a cousa: como se sabe julgar."
D. Francisco de Portugal, primeiro conde de Vimioso

1. Introdução

A análise estatística de fenómenos em que o interesse central reside na contagem do tempo decorrido até à verificação de determinado acontecimento ou, numa terminologia mais rigorosa, de um evento terminal, é o objecto da *análise da duração*. Dito de outro modo, a *análise da duração* trata da modelação estatística da ocorrência de transições entre diferentes estados. Os fundamentos da análise de duração são tributários da teoria dos processos estocásticos de renovação, dos processos pontuais e dos processos de contagem.

2. Loxodromia da vida humana

Comece-se por definir T como uma variável aleatória contínua não-negativa que representa a duração num dado estado. Seja $f(t)$ a correspondente função densidade de probabilidade e $F(t)$ a função de distribuição cumulativa, que será dada por

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

A probabilidade de um indivíduo se manter nesse estado até t é, então, dada pela *Função de Sobrevivência*

$$S(t) = P(T > t) = 1 - F(t).$$

O conceito fundamental na análise de duração é o de taxa de quebras ou *função "hazard"*, que é definida como

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

e que mede a taxa instantânea de saída no momento t, dado que o indivíduo sobreviveu no estado até t.

De acordo com Tiago de Oliveira (1990), "o conceito de taxa de quebras, sob o aspecto de força de mortalidade (Demografia) foi criado em 1757 por Soares de Barros e Vasconcelos, um estrangeirado... que o publicou no artigo "Loxodromia da Vida Humana", Mem. Real Academia de Sciencias de Lisboa, 1ª

* Os autores agradecem os comentários e sugestões de António Antunes, José António Ferreira Machado, Carlos Robalo Marques e Paulo Rodrigues.

série, I, 1759. Soares de Barros e Vasconcelos usa *1/h(t)* (chamada força da vida), que interpreta correctamente; note-se que *1/h(t)* é hoje um instrumento importante na Estatística dos Extremos. Só mais tarde Gompertz (1825) e Makeham (1860) redescobrem o conceito e o utilizam em Demografia e Actuariado."

Associada com a função "hazard", define-se a *função "hazard" cumulativa*

$$\Lambda(t) = \int_0^t h(u)du$$

que é especialmente utilizada na análise de especificação. A evolução temporal da taxa "hazard" é caracterizada por *dh(t)/dt*, define a importante noção de *dependência da duração*. Diz-se que a dependência da duração é negativa (positiva) quando a taxa "hazard" diminui (aumenta) com a passagem do tempo.

Uma outra função com interesse na análise da duração é a do *valor esperado da duração condicionado à sobrevivência no estado até s*

$$e(s) = \int_s^\infty \frac{tf(t)dt}{S(s)} = s + \int_s^\infty \frac{S(t)dt}{S(s)}$$

que permite deduzir a expressão da duração média como o integral da função de sobrevivência

$$e(0) = E(T) = \int_0^\infty S(t)dt.$$

3. Durações incompletas

Muitas vezes as observações sobre a duração de um dado episódio não são exactas. Frequentemente são incompletas, seja porque somente é conhecido que a duração excede um dado valor (neste caso dizem-se *censuradas à direita*), ou porque é sabido que a contagem exacta da duração foi iniciada após ter decorrido já algum tempo (neste caso dizem-se *censuradas à esquerda*). É também possível a simultaneidade destas duas situações gerando observações *censuradas por intervalo*.

Seja T^* uma variável aleatória representando a duração na ausência de qualquer censura e seja C o tempo de censura. A duração observada será então $T=min(T^*,C)$. É conveniente, nestas circunstâncias, introduzir o indicador de censura para o indivíduo *i*

$$\delta_i = \begin{cases} 0, se\ T_i^* > C_i; \\ 1, se\ T_i^* \leq C_i. \end{cases}$$

A função de verosimilhança para uma amostra do par (t_i, δ_i) de dimensão n, no caso em que o tempo de censura não é informativo sobre os parâmetros da distribuição da duração, simplifica para

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

A abordagem paramétrica à estimação da função de sobrevivência pressupõe a especificação da distribuição da duração definida por um número finito de parâmetros. Deve ser exercido especial cuidado na escolha da função de distribuição, já que a utilização de funções de distribuição inadequadas conduz a estimativas inconsistentes dos parâmetros de interesse, e em particular, a resultados erróneos sobre a dependência da duração. Importa ainda ter presente que a duração somente admite valores não-negativos. As distribuições exponencial, Weibull, log-normal, log-logística, Gompertz, Pareto, gama e gama generalizada têm sido frequentemente utilizadas.[2] Sublinhe-se ainda que estas distribuições implicam diferentes comportamentos da função "hazard" no que diz respeito à duração da dependência.

[2] Ver Addison e Portugal (1987) para uma discussão sobre a escolha da distribuição da duração.

4. Estimação não-paramétrica da função de sobrevivência

Uma forma conveniente de descrever a função de sobrevivência é dada pela exibição do seu gráfico. Para uma amostra aleatória de dimensão n de durações completas a função de sobrevivência empírica pode ser simplesmente obtida por

$$\widehat{S(t)} = \frac{n^{\underline{o}}\ de\ T \geq t}{n}.$$

Com dados censurados à direita a função de sobrevivência pode ser obtida através do celebrado estimador de Kaplan-Meier,

$$\hat{S}(t) = \prod_{t_j > t} (1 - \hat{\theta}_j)$$

em que $\hat{\theta}_j = n_j/r_j$ com n_j sendo igual ao número de indivíduos que falham no momento t_j e r_j correspondendo ao número de indivíduos em risco de saírem no momento imediatamente anterior a t_j.

5. Análise de regressão

Na análise econométrica da duração importa considerar os efeitos dos regressores *x* (em que *x* é um vector de variáveis explicativas) no comportamento de *f(t | x), h(t | x)* e *S(t | x)*. Na formulação do modelo de efeitos proporcionais (Cox, 1972) assume-se que os regressores influenciam de forma proporcional a taxa "hazard"

$$h(t \mid x) = h_0(t) exp(x'\beta)$$

em que $h_0(t)$ representa *h(t)* quando $x = 0$. Dito de outra forma, *x* afecta linearmente o comportamento de *ln[h(t)]*. Assim, nesta especificação, o rácio entre a taxa "hazard" de duas sub-populações permanece constante ao longo do tempo (admitindo *x* constante ao longo do tempo).

6. O plano amostral: o paradoxo do autocarro

O processo gerador da amostra de durações observadas pelo investigador tem consequências decisivas sobre a forma de interpretar a informação recolhida (por exemplo, o significado da duração média) e, portanto, sobre a especificação da função de verosimilhança. É absolutamente crucial na análise da duração caracterizar o plano amostral e avaliar as condições de identificação dos parâmetros de interesse. No contexto da investigação da experiência dos desempregados, importará distinguir três situações: amostragem sobre o fluxo; amostragem sobre o *stock*; e a amostragem sobre um intervalo fixo.

6.1 Amostragem sobre o fluxo

A primeira ideia a reter será a de que a amostragem sobre fluxos se relaciona de forma directa com os parâmetros da distribuição de duração da população. Isto é, uma amostra aleatória que acompanhe os indivíduos desde o início do episódio será representativa da população. No caso da amostragem sobre fluxos a função de verosimilhança apropriada será:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.$$

6.2 Amostragem sobre o *stock*

Já no caso de amostragens sobre o *stock* - isto é, sobre a duração decorrida num estado pelos indivíduos que num dado instante se encontram nesse estado - é necessário ter presente, por um lado, que todas as durações são incompletas, e, por outro lado (e mais importante), que amostras recolhidas de acordo com este plano amostral tendem a sobre-representar sistematicamente as durações mais longas. Este último aspecto é conhecido na literatura como "*length bias sampling*". Um exemplo esclarecedor de "*length bias sampling*" é o chamado paradoxo do autocarro (Feller, 1966). Admita-se que numa cidade os autocarros circulam exactamente à tabela com um intervalo de 60 minutos. Um passageiro que chegue aleatoriamente a uma

paragem, esperará, em média, 30 minutos pelo autocarro. Suponha-se agora que os autocarros chegam aleatoriamente, de acordo com uma distribuição Poisson, passando, em média, um autocarro em cada 60 minutos. O resultado paradoxal é que o mesmo passageiro esperará agora, em média, 60 minutos, o dobro da situação inicial.

No caso da amostragem sobre *stocks* a função de verosimilhança é ponderada pela probabilidade de um dado indivíduo ser observado, a qual é proporcional à duração média. Admitindo um fluxo de entrada constante, a expressão da função de verosimilhança será

$$L = \prod_{i=1}^{n} \frac{S(t_i)}{\mu}$$

em que μ é a duração média.

6.3 Amostragem sobre um intervalo fixo

Ainda que utilizando uma amostragem sobre o *stock*, por vezes a informação sobre a duração é recolhida em dois ou mais momentos. No primeiro momento é obtida informação sobre a duração decorrida e nos períodos subsequentes obtém-se informação sobre a realização ou não de uma transição. Neste caso é possível condicionar a probabilidade de transição na duração decorrida e obter os parâmetros da função distribuição da duração pela maximização da função:

$$L = \prod_{i=1}^{n} \pi^{\delta_i} (1-\pi)^{\delta_i} \text{ em que } \pi = \frac{S(t) - S(t+h)}{S(t)}$$

e em que *h* denota o período de seguimento.

7. Heterogeneidade individual não-observada: a lebre e a tartaruga

O problema da heterogeneidade não observada na análise da duração pode ser visto, à semelhança da discussão dos modelos de efeitos aleatórios, como um problema de especificação incompleta. A presença de heterogeneidade individual não observada pode ter como fonte erros de medida na duração ou nos regressores, omissão de variáveis relevantes, inadequada formulação da forma funcional ou da função "hazard". A heterogeneidade individual não observada acarreta como consequência, na generalidade das situações, a inconsistência dos estimadores.

Em particular, é sabido que enviesa de forma sistemática a estimação no sentido de favorecer a dependência da duração negativa. A razão deste enviesamento radica na alteração da composição da amostra ao longo do tempo. Isto acontece porque à medida que o tempo passa, a amostra é crescentemente constituída por indivíduos que têm atributos não observados que lhe dificultam a transição para outro estado. Suponha-se que numa dada amostra existem lebres e tartarugas. O econometrista não distingue umas das outras. Iniciada a corrida, as lebres (com "hazards" mais elevadas) tenderão a completar mais rapidamente o trajecto, fazendo com que, à medida que o tempo passa, a amostra seja composta por cada vez mais tartarugas (com "hazards" menores), projectando assim uma ilusão de dependência da duração negativa.

Uma forma directa de formalizar este problema é através da incorporação de um termo de perturbação aleatória *v* que representa um efeito individual específico, com função densidade de probabilidade *g(v)*. Sendo assim, o problema pode ser abordado no âmbito da temática da *mistura de distribuições*. A função de sobrevivência incondicional resulta então

$$S(t) = E_v[S(t \mid v)] = \int_v g(v) S(t \mid v) dv.$$

Duas abordagens alternativas têm sido propostas para incorporar a heterogeneidade individual não observada nos modelos de duração. Num caso a função paramétrica *g(v)* é explicitada (assumida) permitindo derivar a função de sobrevivência incondicional (abordagem paramétrica). A distribuição *gama*

com média unitária é frequentemente utilizada para este efeito porque permite (à semelhança das distribuições da família exponencial) uma solução fechada para $S(t)$.[3] Alternativamente, têm sido propostas abordagens não paramétricas. Nestes casos a função densidade de probabilidade da heterogeneidade não observada é aproximada através da estimação de uma função empírica discreta com um número pré-determinado (ou não) de pontos de suporte. Neste último caso será possível estimar a proporção de indivíduos associada a cada grupo (por exemplo, a fracção de "lebres" e de "tartarugas").

8. Riscos defectivos: o ovo da tartaruga

A presença de sobreviventes de longa duração, isto é, de indivíduos cuja probabilidade de transição para outro estado é zero, obriga a especificar a distribuição da duração como uma distribuição degenerada. Na epidemiologia, situações de sobrevivência de longa duração poderão ser geradas pela presença de indivíduos imunes ou curados. Uma forma de incorporar a presença de sobreviventes de longa duração passa pela consideração de uma probabilidade p do indivíduo poder vir a transitar para outro estado e uma probabilidade (1 – p) ficar para sempre estado inicial. A função de sobrevivência incondicional poderá então ser expressa como:

$$S(t) = (1 - p) + pS_2(t)$$

em que $S_2(t)$ identifica a função de sobrevivência condicionada à possibilidade de transição. Uma característica interessante deste modelo radica precisamente na possibilidade de estimar a proporção de sobreviventes de longa duração ($1 - p$). Uma vez especificada a estrutura da função de sobrevivência como um modelo de mistura de distribuições, a função "hazard" incondicional é definida como

$$h(t) = \frac{pf(t)}{(1 - p) + pS_2(t)}.$$

Admita-se que se pretende analisar, com base numa amostra de tartarugas, o tempo que decorre até à postura de ovos. Desafortunadamente, a amostra é constituída por tartarugas macho e tartarugas fêmea, e o econometrista não é capaz de distinguir o género das tartarugas. O modelo de sobrevivência de longa duração poderá estimar a proporção de tartarugas macho e a distribuição do tempo até à postura das tartarugas fêmea.

Referências

Addison, J. T. e P. Portugal (1987) "On the Distributional Shape of Unemployment Duration," *Review of Economics and Statistics*, 69.

Addison, J. T. e P. Portugal (1997) "Some Specification Issues on Unemployment Duration," *Labour Economics*, 5.

Cox, D. (1972) "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 34.

Feller (1966) *An Introduction to Probability Theory and Its Applications* Vol 2, Wiley, New York.

Kaplan EL e P. Meier (1958) "Non parametric estimation from incomplete observations." *Journal of the American Statistical Association* 53.

[3] Ver Addison e Portugal (1997) para uma discussão sobre a escolha da distribuição de mistura.

SPE e a Comunidade

Estatística multivariada no R

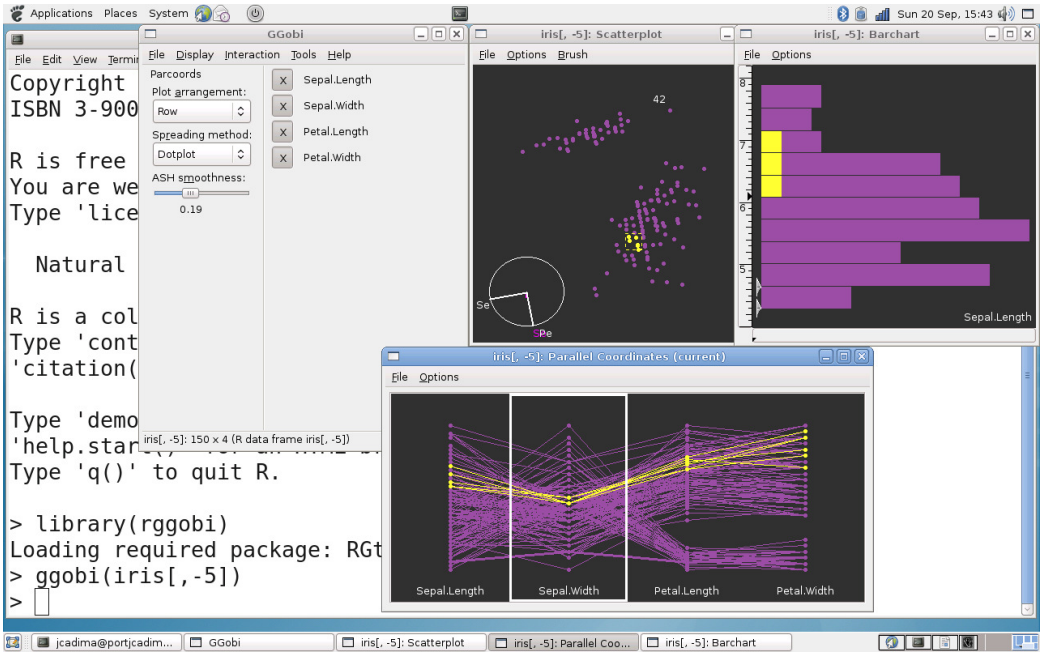
Jorge Cadima, *jcadima@isa.utl.pt*

Instituto Superior de Agronomia, Universidade Técnica de Lisboa

1. Introdução

Os principais métodos de estatística multivariada encontram-se disponíveis no R, quer integrando a distribuição base, quer espalhados por alguns dos mais de 2000 módulos adicionais (*Packages*, na linguagem do R) contribuídos por utilizadores do R. A utilização destes métodos no ambiente de trabalho do programa R traz ao utilizador as seguintes vantagens:

- fiabilidade do código (pelo menos para as funções da distribuição base);
- ambiente de trabalho flexível, que permite com facilidade pre-processar os dados e manipular e submeter a ulterior análise os resultados;
- ambiente de trabalho poderoso, que permite escrever código, incluindo através da programação de funções (subrotinas) para efectuar análises mais complexas ou que envolvam vários procedimentos.



No R estão também disponíveis ferramentas para a visualização gráfica bi- ou tri-dimensional de dados. Uma das mais poderosas baseia-se no software de visualização gráfica em código aberto Ggobi (http://www.ggobi.org), que (estando instalada no nosso sistema) pode ser invocado a partir duma sessão de trabalho do R por intermédio do *package* **rggobi**. Esta ferramenta gráfica permite, entre outras coisas, rodar nuvens de pontos dos dados (de forma automática ou manualmente, controlando com o rato), identificar observações individuais (*identify*) ou em grupo (*brush*) e criar histogramas ou gráficos em coordenadas paralelas onde os indivíduos seleccionados sejam destacados através de cores diferentes, como ilustrado na imagem.

Segue-se uma breve introdução à utilização no R de algumas das mais usadas técnicas multivariadas. Para exemplificar, será usado o famoso conjunto de dados de Fisher com medições morfométricas em *n*=150 lírios, 50 de cada uma de três espécies. Estes dados encontram-se disponíveis na distribuição base do R, no objecto **iris**. Trata-se duma *data frame*, com cinco colunas, de 150 valores cada. As quatro primeiras colunas são variáveis numéricas, com comprimentos e larguras das sépalas e pétalas de cada lírio. A quinta coluna, é um factor, com três níveis (*setosa, versicolor* e *virginica*) correspondentes às espécies observadas.

2. Análise em Componentes Principais e *biplots*

Uma Análise em Componentes Principais (ACP) [1] pode ser efectuada através de duas funções, ambas incluídas no *package* **stats** (que integra a distribuição base do R, estando automaticamente disponível em qualquer versão do R instalada da forma usual). Trata-se das funções **prcomp** e **princomp**. A primeira tem duas vantagens, que aconselham a sua utilização:

1. obtém a informação necessária a uma ACP através da decomposição em valores singulares (DVS) da matriz de dados, enquanto que a segunda função se baseia numa decomposição espectral (em valores e vectores próprios) da matriz de covariâncias (ou de correlações) dos dados. A maior estabilidade numérica da DVS aconselha o uso da função **prcomp**;
2. a função **princomp** exige que nos dados analisados haja um número de observações (que, por omissão, são associadas pelo R às linhas da matriz de dados) superior ao número de variáveis observadas (colunas da matriz). Esta restrição não existe na função **prcomp**.

Aqui aborda-se apenas a função **prcomp** (podendo as particularidades da outra função ser consultadas da forma usual, com recurso à função **help**). A função **princomp** é mantida sobretudo por razões de compatibilidade com o *S-Plus*, o software estatístico comercial que, tal como o R, é baseado na linguagem de programação *S*.

A passagem dos dados para a função **prcomp** faz-se através duma matriz ou *data frame* (naturalmente com dados numéricos). Trata-se do único argumento obrigatório desta função. Assim, para efectuar uma ACP aos dados dos lírios, basta escrever **prcomp(iris[,-5]).** Note-se que se torna necessário excluir a quinta coluna do objecto **iris** da análise, uma vez que se trata de um factor, cuja inclusão na ACP não faz sentido.

Caso se deseje efectuar uma ACP sobre a matriz de correlações, isto é, sobre os dados normalizados, deve utilizar-se o argumento lógico **scale**, associando-lhe o valor "verdade":

> prcomp(iris[,-5], scale=TRUE)

Alternativamente, seria possível efectuar primeiro a normalização dos dados (através da função **scale**) e depois invocar a função **prcomp** sem ulteriores argumentos: **prcomp(scale(iris[,-5]))**.

O comando **prcomp** produz como resultado uma lista com vários objectos, dois dos quais são mostrados, por omissão:

- **sdev** - um vector com os *p* desvios padrão das componentes principais (ou seja, com os valores singulares da matriz de dados - centrada e a dividir por n+1 - que são também as raízes quadradas dos valores próprios da matriz de covariâncias dos dados analisados);
- **rotation** – uma matriz *p*x*p*, cujas coluna *k* contém os coeficientes (*loadings*) na combinação linear das variáveis que definem a componente principal *k*.

A lista devolvida pelo comando produz ainda um terceiro objecto que apenas é mostrado caso seja explicitamente pedido. Trata-se duma matriz *n*x*p*, cujas colunas contêm os valores (*scores*) de cada indivíduo em cada componente principal. Os autores da função **prcomp** deram um nome curioso a este objecto: apenas **x**. Para visualizar os *scores* nas componentes principais dos dados dos lírios, basta

solicitar, da forma usual, a componente da lista através do seu nome: **prcomp(iris[,-5])$x**

Como é sabido, há uma indeterminação nos sinais em cada coluna das matrizes **rotation** e **x**, no sentido que as colunas podem ser multiplicadas por -1 de forma arbitrária. É possível que as funções **prcomp** e **princomp**, ou até a mesma função em diferentes binários executáveis do R, produzam variações nestes sinais, como adverte a *help page* da função **prcomp**.

Como foi referido, o *output* do comando **prcomp** é uma lista. Mais formalmente, trata-se dum objecto da classe *prcomp*, que herda da classe *list*, o que significa que é uma lista de natureza especial, para a qual podem ser escritos métodos específicos, e que na ausência de métodos específicos para um objecto de tipo *prcomp*, aplicar-se-ão os métodos para a classe mais geral *list*.

Uma das funções mais úteis para aplicar aos resultados duma ACP no R é a função **summary**, que constrói uma tabela com os habituais indicadores da qualidade de cada componente principal:

> summary(prcomp(iris[,-5]))

Se o resultado duma ACP fôr passado para a função *genérica* **plot**, será produzido (numa janela gráfica separada) um *screeplot*, na forma dum histograma das variâncias associadas a cada componente principal.

O *biplot* [2] é, tal como a ACP, uma técnica baseada na decomposição em valores singulares duma matriz de dados. A distribuição base do R inclui a função **biplot**, que produz um gráfico bidimensional com marcadores de indivíduos (pontos) e marcadores de variáveis (vectores) a partir dos quais se reconstrói, aproximadamente, boa parte da informação associada à matriz de dados original. A função **biplot** admite, como argumento de entrada, o resultado duma ACP: **biplot(prcomp(iris[,-5]))**

3. Análises Classificatórias

O vasto conjunto de técnicas de classificação (*clustering*) é disponibilizado no R através de várias funções e *packages*.

Uma *análise classificatória hierarquizada* (*hierarchical clustering*) [3] tem como ponto de partida uma matriz de distâncias ou dissemelhanças entre os *n* indivíduos que se pretende classificar. Essas dissemelhanças podem ser determinadas de diferentes formas (incluíndo avaliações subjectivas). As formas mais objectivas têm por base um conjunto de dados observados. O R disponibiliza a função **dist** para construir uma matriz de distâncias/dissemelhanças a partir duma matriz ou *data frame* de dados. Por omissão, são calculadas as distâncias euclideanas usuais entre os indivíduos (linhas da matriz ou *data frame*). As distâncias ℓ_2 entre os 150 lírios de Fisher calculam-se com **dist(iris[,-5])**.
A função **dist** admite algumas outras opções para os critérios de dissemelhança, que podem ser invocadas através do argumento **method**. As opções actualmente disponíveis (versão 2.9.2 do R) são, além da "*euclidean*", "*maximum*" (norma do supremo), "*manhattan*" (norma ℓ_1), "*canberra*", "*binary*" (para dados de tipo 0/1) ou "*minkowski*" (com potência *p* controlável através do argumento com esse nome). Assim, as distâncias ℓ_1 entre os lírios serão obtidas com o comando:

> dist(iris[,-5], method="manhattan")

Uma análise classificatória hierarquizada no R pode efectuar-se através da função **hclust**, automaticamente disponível na distribuição base do R (package *stats*). Esta função tem um único argumento obrigatório, que é uma estrutura de dissemelhanças, como as que são criadas pela função **dist**. A classificação hierárquica será, por omissão, baseada no critério de agregação do vizinho mais distante, que considera a distância entre dois grupos de observações como sendo a maior das dissemelhanças entre um indivíduo de cada grupo (*complete link*). Outros critérios de agregação de grupos estão disponíveis através do argumento **method** da função **hclust**, que actualmente tem como

valores possíveis: "*ward*", "*single*", "*complete*", "*average*", "*mcquitty*", "*median*" e "*centroid*".

Os resultados são de difícil leitura sem auxiliares ulteriores, como a função **plot** (que tem um método específico para objectos de classe *hclust*, produzidos pela função homónima, e que também pode ser invocado explícitamente com o nome **plclust**). Esta função produz a tradicional representação gráfica das classificações hierárquicas: um dendrograma.

A função **cutree** ajuda a interpretar dendrogramas com numerosas "folhas" (indivíduos), que podem não ser de fácil leitura. Esta função exige como argumento um objecto de classe *hclust*, e ainda um de dois parâmetros: ou um número *k* de classes em que se deseja agrupar os indivíduos, ou uma altura *h* à qual se deseja cortar o dendrograma, para identificar as classes que um tal corte produziria. Em ambos os casos, o *output* da função será uma listagem do grupo a que pertencerá cada individuo.

Um conjunto alternativo de funções para análises classificatórias encontra-se num *package* mais recente, chamado **cluster**, baseado em [4]. Trata-se de um *package* incluído nas distribuições base do R, mas que é necessário carregar explicitamente para uma sessão de trabalho:

> library(cluster)

Neste *package* encontram-se duas funções para análises classificatórias hierárquicas: a função **agnes**, que efectua uma análise hierárquica agregadora, isto é, "de baixo para cima", começando por considerar (tal como a função **hclust**) que cada individuo constitui um grupo e procurando depois agregar indivíduos em grupos o mais possível internamente homogéneos e heterogéneos entre si; e a função **diana**, que efectua uma análise desagregadora, isto é, "de cima para baixo", iniciando num único grande grupo e procurando em seguida as melhores desagregações possível. O *input* destas funções é, tal como para a função **hclust**, uma estrutura de dissemelhanças que pode ser produzida pela função **dist**. Mas o *package* disponibiliza uma função própria para calcular matrizes de dissemelhança, a função **daisy**, que tem a particularidade de poder lidar com dados nominais ou ordinais, através do coeficiente de dissemelhança generalizado de Gower. O *output* pode, também aqui, ser passado para a função **plot**, produzindo, além do dendrograma correspondente, uma outra visualização gráfica (a *banner*). As funções **agnes** e **diana** também calculam indicadores numéricos da qualidade da classificação produzida (veja-se a documentação de cada função para ulteriores pormenores).

Métodos de *classificação não-hierárquica* estão também disponíveis no R, quer através do *package* **cluster**, quer em algumas outras funções. Entre estas últimas, encontra-se a função **kmeans**, que agrupa indivíduos em *k* classes, a partir duma matriz de dados numéricos. O número de classes desejado é indicado através do argumento **centers**, que pode ter por valor uma matriz *k*x*p* com *k* conjuntos de valores nas *p* variáveis analisadas e que constituirão os centróides dos grupos em torno dos quais agrupar os indivíduos observados, ou apenas o número *k* de grupos que se deseja constituir (em cujo caso serão aleatoriamente seleccionados *k* de entre os *n* indivíduos na matriz de dados para constituir os centros dos grupos). O algoritmo das *k* médias procede a associar os indivíduos a grupos, tendo por critério a minimização da soma de quadrados das distâncias de cada indivíduo aos centros do respectivo grupo. Eis um exemplo de aplicação desta função, escolhendo o primeiro lírio de cada espécie como centróides dos futuros grupos:

- **kmeans(iris[,-5],centers=iris[c(1,51,101),-5])**

4. Análises Discriminantes

As tradicionais Análises Discriminantes linear (de Fisher) e quadrática podem ser efectuadas com as funções **lda** e **qda**, respectivamente, ambas disponíveis no pacote **MASS** [5], que tem de ser explicitamente carregado para estar disponível numa sessão de trabalho, através do comando **library(MASS)**.

A função **lda** pode ser invocada passando como argumentos de entrada uma matriz ou *data frame* com

as variáveis numéricas cujas combinações lineares definem as funções discriminantes, e ainda com o argumento **group**, que deve ser um factor indicado as classes cuja discriminação se deseja efectuar. Assim, a discriminação linear das espécies de lírios, com base nas quatro variáveis numéricas observadas pode ser pedida com o comando **lda(iris[,-5] , group=iris[,5])**. Alternativamente, o argumento de entrada pode ser uma fórmula, semelhante às utilizadas nas funções de ajustamentos de regressões lineares ou ANOVAs, em que a variável resposta (à esquerda do símbolo ~) seja o factor dos grupos que se pretende discriminar, e as parcelas do lado direito sejam as variáveis numéricas em que essa discriminação se baseia. De forma totalmente análoga se pode invocar a função **qda**. É também possível visualizar a discriminação produzida pelas duas primeiras funções discriminantes lineares passando o resultado da função **lda** para o comando **plot**.

O pacote **mda** (de *mixture discriminant analysis*) contém também funções úteis para análises discriminantes.

5. Métodos de *scaling*

Entre os métodos que visam a representação gráfica de indivíduos, a partir duma matriz de dissemelhanças, a abordagem mais clássica, de Gower, é também conhecida por Análise em Coordenadas Principais. Essa técnica está disponível na distribuição base do R através da função **cmdscale** (***c****lassical* ***m****ulti****d****imensional* ***scal****ing*). Esta função aceita como argumento uma estrutura de dissemelhanças, como as que são produzidas pela já referida função **dist** (e ainda, a indicação do número de dimensões nas quais se pretende fazer a representação métrica das dissemelhanças, através do parâmetro *k* que, por omissão, toma o valor 2). Uma Análise em Coordenadas Principais da matriz de distâncias ℓ_1 entre os 150 lírios obtém-se pelo comando **cmdscale(dist(iris[,-5],method="manhattan"))**. O comando **plot** lida bem com o resultado produzido pelo comando **cmdscale**, gerando um gráfico dos indivíduos em duas dimensões.

De entre as muitas abordagens não-métricas de *scaling*, duas estão disponíveis através das funções **isoMDS** e **sammon**, ambas incluídas no já referido *package* **MASS**. A utilização destas funções faz-se de forma semelhante ao que já foi descrito para a função **cmdscale**.

6. Distribuições multivariadas

Algumas funções do R permitem trabalhar com distribuições multivariadas. A função **mvrnorm** (do pacote **MASS**) gera observações duma multinormal com vector médio (argumento **mu**) e matriz de variâncias (argumento **Sigma**) especificados como argumentos da função. Nos *packages* **mvtnorm** e **mnormt** (que é necessário descarregar a partir do repositório CRAN – cran.r-project.org – instalar e carregar para a sessão de trabalho, antes de estarem disponíveis) encontram-se várias funções para trabalhar com a multinormal e a distribuição t multivariada. Em particular, o primeiro destes packages disponibiliza funções da família d-p-q-r (disponíveis para as distribuições univariadas) que geram valores de densidades, funções distribuição cumulativas, quantis e amostras simuladas, respectivamente, para a multinormal e a t multivariada. O package **sn** disponibiliza funções para as distribuições multinormal e t multivariada assimétricas.

Além das funções e métodos aqui referidos, estão disponíveis numerosas outras funções e métodos na distribuição base ou nos *packages* adicionais do R. Informações mais completas sobre ferramentas disponíveis no R para dados multivariados ou técnicas de estatística multivariada estão disponíveis no documento *cran.r-project.org/web/views/Multivariate.html*

Bibliografia

[1] Jolliffe, I.T. (2002). *Principal Component Analysis*, 2d ed., Springer-Verlag.
[2] Gabriel, K. R. (1971). The biplot graphical display of matrices with application to principal component analysis. *Biometrika,* 58, 453-467.
[3] Everitt, B. (1974). *Cluster Analysis.* London: Heinemann Educ. Books.
[4] Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
[5] Venables, W.N. e Ripley, B.D. (2002). *Modern Applied Statistics with S (fourth edition)*, Springer-Verlag.

A utilização do R na produção de informação estatística

Rita Sousa, *rita.sousa@ine.pt*
Instituto Nacional de Estatística - DME/ME

Pedro Campos, *pedro.campos@ine.pt*
Instituto Nacional de Estatística - DME/ME

Ana Patrícia Martins, *apsc.martins@gmail.com*

Ana Luísa Quitério, *anaquiterio@gmail.com*

1. Nota Introdutória

O R é uma linguagem e um ambiente de desenvolvimento integrado para análise estatística de dados e construção de gráficos, com a vantagem de constituir uma ferramenta *open source* gratuita. Uma das grandes virtudes do R é a semelhança entre a linguagem de construção do *software* e a desenvolvida pelos utilizadores (Everitt, B. S. e Hothorn, T., 2006). Para além disso, como é um *software open source*, possibilita ao utilizador o acesso aos procedimentos existentes, facilitando assim a criação de novas funcionalidades. Este *software* é particularmente útil para profissionais e investigadores das áreas da Estatística e da Matemática, pela diversidade de mecanismos incorporados que lhes permite conceber toda a análise, desde a organização dos dados à execução de cálculos, gráficos e à implementação de metodologias de análise de dados. Apesar de ter uma forte utilização e influência das comunidades de Matemáticos e Estatísticos, o R tem vindo a conquistar profissionais e investigadores de muitas outras áreas de saber. Neste momento o R é uma linguagem muito utilizada não só na comunidade académica e científica mas também no meio empresarial.

A utilização do R nos inquéritos por amostragem trouxe muitas vantagens aos produtores de informação estatística, nomeadamente ao Instituto Nacional de Estatística (INE). Neste artigo apresenta-se a génese da utilização do R no INE e descrevem-se muito sucintamente os *packages* do R que são utilizados para a estimação de variáveis em inquéritos por amostragem, tais como o *package survey*, o *SAE* (para estimação em Pequenos Domínios) e os *packages Hmisc* e *mice*.

Este trabalho foi elaborado por quatro colaboradores que iniciaram desenvolvimentos em R no INE, dois dos quais já não se encontram no Instituto.

2. O INE e o R

No INE, a utilização do R teve início em 2005 com o objectivo de replicar processos já existentes, até então exclusivamente realizados noutras linguagens. A opção pelo R mostrou-se uma aposta ganha, uma vez que permitiu um maior controlo da programação e um ganho significativo nos tempos de execução dos procedimentos. Para além de todas as potencialidades computacionais conhecidas do R, o facto de estar associado a uma comunidade muito activa geradora de novos contributos e actualizações permanentes aos *packages* disponíveis constitui um ponto fundamental na escolha desta linguagem. O *package* mais utilizado pelo INE é o *survey,* desenvolvido por Thomas Lumley, professor da Universidade de Washington. O contacto com o autor e os comentários dos utilizadores têm permitido uma actualização deste *package*, muitas vezes adaptado à própria realidade do INE. A primeira grande aplicação desenvolvida em R no INE, destinava-se à execução de tarefas inerentes ao Inquérito ao Emprego (IE) e dado que os resultados obtidos foram bons, o processo tem vindo a ser

utilizado em diversos outros inquéritos, tais como o Inquérito Nacional de Saúde (INS), o Inquérito à Utilização de Tecnologias da Informação e da Comunicação (IUTIC), entre outros.

Após o desenvolvimento de diversos programas, já de considerável complexidade, tornou-se notória a necessidade de se criar uma aplicação que permitisse a sua utilização tanto por utilizadores mais especializados como por utilizadores comuns, não familiarizados com as questões metodológicas. Assim, a criação de uma interface gráfica (com recurso ao *package tcltk*) tem sido um dos projectos em desenvolvimento no INE (IE-R), para além de toda a utilização paralela que se faz do R no âmbito da investigação e implementação das metodologias estatísticas.

3. Da amostragem à estimação

O R é um dos *softwares* mais utilizados em investigação com recurso a metodologias estatísticas, sendo a amostragem e a estimação duas das áreas de maior utilização.
De seguida descrevem-se alguns dos *packages* utilizados na produção de informação estatística através dos inquéritos por amostragem do INE.

A amostragem e o *package survey*

No contexto da amostragem existem diversos *packages* em R que permitem aplicar os métodos probabilísticos mais usuais. No INE recorre-se mais frequentemente ao *package survey* (Lumley, T., 2009) pelo facto deste permitir a análise de dados provenientes de planos complexos de amostragem (Cochran, W.G., 1977).

A produção de estatísticas oficiais passa muitas vezes pela elaboração de inquéritos com diversas especificidades que tornam o processo de análise e de estimação mais complexo, como é o caso do desenho amostral muti-etápico, da pós-estratificação, da calibração de ponderadores, da imputação múltipla, etc. Assim, o *package survey* disponibiliza diferentes funções pensadas para responder a algumas destas situações de maior complexidade, como por exemplo os métodos de reamostragem: *Bootstrap* e *Jackknife* (Särndal et al., 1992) para a estimação da variância dos estimadores, quando estes não são directamente deduzíveis de forma analítica. Isto acontece com importantes operações estatísticas, como o IE, o INS, o IUTIC, entre outros.

Desde 2005 que, no INE, o projecto IE-R tem permitido a automatização dos principais apuramentos deste inquérito, até então produzidos em SAS. Grande parte da programação passa a ser substituída por funções existentes no *package survey*, tendo-se verificado ganhos significativos nos tempos de execução e uma maior flexibilidade de alteração dos procedimentos. Para além disso, a utilização do R neste inquérito tem despertado o interesse para a investigação de novas metodologias estatísticas no sentido de explorar e inferir sobre a qualidade da informação produzida. Por exemplo, a difusão periódica e sequencial dos resultados ao longo do tempo conduz-nos à análise de séries temporais, que muitas vezes exigem um estudo mais aprofundado, como o que se faz no IE com recurso ao R (Quitério, A., 2008). A abertura e transparência deste *software* possibilitam uma partilha e evolução constantes, tanto por parte dos autores como dos próprios utilizadores. O autor deste *package*, Thomas Lumley, tem feito alterações frequentes de código quando, por algum motivo, as funções existentes não se ajustam à realidade dos inquéritos do INE.

Estimação em Pequenos Domínios

Quando nos inquéritos por amostragem as áreas a estudar fornecem amostras com dimensão suficientemente grande, conseguem-se obter estimativas com precisão adequada. No entanto, nos casos em que as amostras são de dimensão insuficiente, as estimativas poderão apresentar uma variabilidade inaceitável. Para obter estimativas para as variáveis de interesse em áreas com pouca ou nenhuma informação disponível, recorre-se a informação de outras áreas e/ou a informação de variáveis auxiliares (Tukey, J.W., 1958). A Estimação em Pequenos Domínios (EPD) é frequentemente utilizada nestes casos, especialmente na estimação da taxa de desemprego do IE em áreas geográficas muito desagregadas. A EPD é também utilizada na estimação do rendimento das famílias nos inquéritos às condições de vida e rendimento em vários países. Os estimadores utilizados neste âmbito chamam-se "directos", quando utilizam informação apenas do domínio de estudo,

podendo recorrer a informação auxiliar dentro do próprio domínio. Por outro lado, os estimadores "indirectos" são aqueles que recorrem a informação das variáveis de estudo ou de variáveis auxiliares fora do domínio de estudo.

O *package SAE* (*Small Area Estimation*), com código desenvolvido em R por Virgilio Gómez Rubio e Nicola Salvati no âmbito do BIAS Project (http://www.bias-project.org.uk), permite facilitar o trabalho de Estimação em Pequenos Domínios (Rubio, V.G. e Salvati, N., 2007).

Imputação de não respostas

No âmbito da estimação, a qualidade não se afere apenas pela avaliação do erro amostral mas também pela correcção de factores alheios às técnicas de amostragem, como por exemplo as não respostas. A imputação é um método estatístico para o tratamento de não respostas, em que os valores omissos são substituídos por estimativas, recorrendo a métodos dedutivos ou determinísticos (Särndal et al., 2004). O R disponibiliza um conjunto significativo de *packages*, dos quais destacamos o *Hmisc* na imputação *Bayesiana*, na imputação pela média ou pelo 3º quartil e o *package mice* na imputação pelo método Hot Deck e por modelos lineares generalizados. Nesta temática, foram feitos importantes desenvolvimentos no tratamento de não respostas do IUTIC e do IE (Martins, Ana Patrícia, 2007), também com recurso a funções disponíveis no R.

Tratamento da Confidencialidade Estatística

Para proteger a informação confidencial existem métodos de controlo de divulgação (*Statistical Disclosure Control*). Estes métodos têm como objectivo, quer a protecção na difusão dos dados agregados, como por exemplo os quadros de apuramento dos inquéritos, quer na difusão dos microdados (dados individualizados). Os métodos de controlo de divulgação que são utilizados seguem critérios predefinidos que procuram por um lado minimizar o risco de identificação das unidades estatísticas a que se referem e, por outro lado, preservar algumas das propriedades estatísticas dos microdados (como por exemplo, os totais). Existem, no R, dois *packages* vocacionados para o tratamento do segredo estatístico de dados individualizados e de dados agregados: *sdcMicro* e *sdcTable.* Ambos se baseiam em métodos implementados no software μ-Argus e τ-Argus. Os métodos mais utilizados para tratamento do segredo em microdados são a Recodificação global, a Supressão local, PostRAndomisation Method (PRAM) e Micro-Agregações Numéricas. Os métodos mais utilizados para tratamento do segredo em macrodados são o método Hipercubo e o método Optimal (Willenborg, L., Waal, T. de, 1996, 2001).

Nos parágrafos anteriores foram referidos alguns dos muitos exemplos em que o R contribui com aplicações que permitem resolver problemas concretos através de soluções que se encontram disponíveis e de acesso livre. De seguida, descrevem-se algumas aplicações associadas aos *packages* descritos no presente capítulo.

4. Aplicações

A produção de informação estatística passa por diferentes fases, que vão desde o planeamento, até à recolha, ao tratamento e análise dos dados. Quer na fase da concepção quer na fase da operação propriamente dita, a definição das metodologias estatísticas assume um papel de extrema importância. Nos inquéritos por amostragem, a produção de estatísticas e dos respectivos indicadores de precisão pode ser uma tarefa difícil, principalmente quando se baseiam em planos de amostragem complexos. Nesses casos, a utilização de informação auxiliar e o recurso a técnicas como a calibração permitem minorar o erro resultante do carácter aleatório das amostras e de outros factores não amostrais (como é o caso das não respostas). A utilização do R, em particular do *package survey*, facilita bastante a implementação destas técnicas. Neste contexto, podemos destacar algumas das funções mais utilizadas:

- *svydesign* – na definição do desenho amostral;
- *as.svrepdesign* – na criação de réplicas com recurso a métodos de reamostragem, como o método de *Jackknife*;

- *calibrate* – na calibração dos ponderadores, fazendo um ajustamento com recurso a informação auxiliar mais recente;
- *svyby* – na obtenção de estimativas e respectivos indicadores de precisão.

Para o tratamento das não respostas, podemos destacar as funções *impute* e *impute.transcan* do *package Hmisc* e as funções *complete* e *glm.mids* do *package mice*:

- *impute* – na imputação por processos aleatórios, pela média ou pelo 3º quartil;
- *impute.transcan* – na imputação Bayesiana;
- *complete* – na imputação pelo método Hot Deck;
- *glm.mids* – na imputação por modelos lineares generalizados.

Quanto ao *package SAE*, os modelos de regressão constituem uma importante ferramenta na conjugação de informação auxiliar de diferentes domínios. Assim, as aplicações do *SAE* focam-se nos modelos lineares, com diferentes tipos de estimadores, como por exemplo os sintéticos e os compósitos. Neste tema destacamos a aplicação da seguinte função:

- *lm* – para aplicar um modelo linear na estimação de uma dada variável, num determinado domínio de interesse.

Quanto aos *packages sdcMicro* e *sdcTable* para o tratamento do segredo estatístico, destacam-se as funções:

- *Microaggregation* (*sdcMicro*) – Os registos são agrupados com base numa medida de proximidade das variáveis de interesse. Os pequenos grupos de registos são usados para calcular agregados para essas variáveis, que são divulgados em substituição do registo individual de valores.
- *protectTable* (*sdcTable*) – Para supressão de dados tabulares, o processo desenvolve-se em duas etapas: a supressão primária (identificação e supressão das células confidenciais) e a supressão secundária (identificação e supressão das células não confidenciais, de modo a proteger as células confidenciais).

Relativamente ao *package tctlk*, existe um elevado número de funções que permitem construir uma interface gráfica, à semelhança da que está na base do *package R Commander – Rcmdr* (que disponibiliza botões e menus de acesso às principais funções de análise de dados em R). Como exemplo apresenta-se na figura 1 uma imagem de algumas das janelas que compõem a aplicação do projecto IE-R, ainda em desenvolvimento.

Figura 1 – *Aplicação IE-R – janela principal e janela dos apuramentos trimestrais do IE*

5. Notas Finais

Ao longo do texto foram apresentadas algumas aplicações do R na produção de informação estatística realizada pelo INE. A introdução do R na produção de informação a partir de inquéritos por amostragem teve como objectivo principal o de efectuar uma optimização e enriquecimento dos procedimentos utilizados na produção de informação estatística, que se traduzem na versatilidade e na velocidade (já confirmada) nos apuramentos de alguns inquéritos, tais como o Inquérito ao Emprego (IE). O *package survey*, desenvolvido em R por Thomas Lumley, professor da Universidade de Washington tem sido muito utilizado na estimação de indicadores do IE. O contacto com o autor e os comentários dos utilizadores têm permitido uma actualização permanente deste *package*, o que permite aos técnicos do INE uma ligação estreita com o meio científico, permitindo desta forma o desenvolvimento e aplicação de métodos de estimação de forma muito mais eficaz. Para além dos métodos desenvolvidos, o R permite também o desenvolvimento de interfaces gráficas. Está a ser desenvolvida em R uma aplicação informática (IE-R) que será de fácil utilização para qualquer utilizador, não necessariamente familiarizado com a metodologia do IE.

A estas aplicações do R associam-se outras, tais como o tratamento de não respostas, a estimação em pequenos domínios e o tratamento do segredo estatístico.

Agradecimentos

Agradece-se aos colegas do Serviço de Métodos, do Departamento de Metodologia e Sistemas de Informação do INE, em particular ao Daniel Fernandes, Mafalda Cabral, Paula Marques e Sílvia Mina, pelos contributos na componente do projecto IE-R e à colega Daniela Oliveira, pelos contributos no tratamento da confidencialidade estatística em R.

Referências Bibliográficas

Cochran, W.G. (1977). *Sampling techniques*. John Wiley & Sons.

Everitt, B. S. and Hothorn, T. (2006). *A Handbook of Statistical Analyses using R*. Chapman & Hall/CRC.

Lumley, T. (2009). *Survey: analysis of complex survey samples*. R package version 3.16.

Martins, Ana Patrícia (2007). *Imputação Múltipla – Aplicação Prática aos dados do Inquérito ao Emprego.* Tese de Mestrado. Faculdade de Ciências da Universidade de Lisboa.

Quitério, Ana (2008). *Modelos de Regressão Dinâmica na Revisão das Séries do Inquérito ao Emprego, Tese de Mestrado*. Faculdade de Ciências da Universidade de Lisboa.

Quitério, A., Martins, A. P., Campos, P. e Sousa, R. (2008). *Inquérito ao Emprego – Aplicação no Software R*. Actas do Congresso da Sociedade Portuguesa de Estatística, Lisboa.

Rubio, Virgilio Gómez e Salvati, Nicola (2007). *Introduction to Small Area Estimation*, disponível em: http://www.bias-project.org.uk/software/

Särdnal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, Springer.

Särndal, C. and Lundstrom, S. (2004). *Estimation in Surveys with Nonresponse*. John Wiley & Sons.

Tukey, J.W. (1958). *Bias and Confidence in not-quite large samples*. Annals of Mathematical Statistics, 29-614.

Willenborg, L., Waal, T. de (1996). *Statistical Disclosure Control in Practice*. Springer Series: Lecture Notes in Statistics, Vol. 111.

Willenborg, L., Waal, T. de (2001). *Elements of Statistical Disclosure Control*. Springer Series: Lecture Notes in Statistics, Vol. 155.

Links de Interesse

Análise de Inquéritos com R: http://faculty.washington.edu/tlumley/survey/
BIAS Project: http://www.bias-project.org.uk
Conteúdos do R: http://neacm.fe.up.pt/CRAN/
Imputação Múltipla: http://www.multiple-imputation.com/
Mailing Lists do R: http://tolstoy.newcastle.edu.au/R/
Página Principal do R: http://www.r-project.org

R "casa" com Regressão Logística

João Gomes, *jjgomes@fc.ul.pt*
CMAF, Faculdade de Ciências Universidade de Lisboa

Sónia Nobre, *soninobre@hotmail.com*
Unidade de Gastrenterologia Centro Hospitalar de Cascais

"Os padrinhos são o estatístico e a médica"

1. Encontro

A peritonite bacteriana espontânea (PBE) é uma infecção que surge frequentemente em doentes com cirrose hepática (doença do fígado avançada). Pode ser muito grave, alcançando uma taxa de mortalidade intra-hospitalar de 20 a 40%, (Song, et al., 2006 and Thuluvath, Morss, & Thompson, 2001).

A identificação precoce de doentes de risco é crucial para a melhoria do prognóstico desta doença.

Existe um sistema de classificação amplamente usado para estratificar o risco dos doentes com cirrose hepática (classificação de Child-Turcotte-Pugh), mas apresenta deficiências relativas, designadamente, à subjectividade de alguns parâmetros avaliados.

Mais recentemente, surgiu o score MELD (Model for End-Stage Liver Disease), que se calcula através de uma fórmula matemática computorizada, com recurso a 3 variáveis objectivas, fornecidas por análises sanguíneas (1. creatinina – parâmetro de avaliação da função dos rins; 2. INR – parâmetro de coagulação do sangue; 3. bilirrubina – parâmetro de avaliação da função do fígado). Foi demonstrado que este score prediz a sobrevida a curto prazo de doentes com cirrose hepática, mas não existem dados relativos aos doentes com PBE (Kamath, et al., 2001).

Para além disso, outras variáveis poderão estar implicadas na sobrevida destes doentes, nomeadamente a idade mais avançada.

Portanto, do ponto de vista médico, existe a necessidade de estabelecer um modelo simples, célere e fidedigno, para aferir a probabilidade de sobrevida de um doente que entra no hospital com PBE, de modo a adoptar as medidas e instituir os tratamentos adequados e assim melhorar o prognóstico destes doentes. (Nobre, Cabral, Gomes, & Leitão, 2008).

2. O namoro

Numa primeira fase avalia-se a capacidade dos preditores da variável resposta "morte" (variável com dois níveis: "morte" e "sobrevida"): a idade, o sexo, valores laboratoriais, meld, etc. Este trabalho, por não ter sido efectuado no R, não será apresentado aqui.

O **teste T** para os preditores contínuos e o **teste do Qui-Quadrado em tabelas de contingência** para os preditores definidos em categorias, permitem afirmar que as únicas co-variáveis que mostram diferenças significativas entre "morte" e "sobrevida" são "meld" e "idade", ambas variáveis contínuas.

Com o objectivo de analisar cada uma delas comecemos por introduzir o nosso conjunto de dados (vamos chamar-lhe ***td***):

```
################################
td=read.table("todos.txt",h=T)        # Leitura de um ficheiro tipo "txt" chamado "todos"
head(td)          # Das 92 observações serão apresentadas 3
```

	idade	meld	morte
1	43	22	0
2	47	24	0
3	68	21	0

Tabela 1: Três das observações de um total de 92

```
td$morte=factor(td$morte)        #considerar a variável "morte" categórica
summary(td)      #Resumo descritivo dos dados
```

```
     idade            meld        morte
 Min.   :35.00   Min.   : 9.00   0:63
 1st Qu.:49.75   1st Qu.:16.00   1:33
 Median :63.00   Median :21.00
 Mean   :60.69   Mean   :22.32
 3rd Qu.:71.00   3rd Qu.:28.00
 Max.   :90.00   Max.   :48.00
```

Tabela 2: Algumas medidas de localização de "idade" e "meld" e contagem de "morte"

Analisemos a "importância" da variável independente *"meld"* relativamente à variável resposta:

```
                                        #MELD
#############################################################################################
#Resumo descritivo de "meld" mas separada por nível de resposta
v=summary(td$meld[td$morte==0])       #  "morte=0" significa "sobrevida"
m=summary(td$meld[td$morte==1])       # "morte=1" significa "morte"
list(v=v,m=m)    # O comando "list" cria um "output" conjunto, dos dois níveis de resposta, da variável "meld"
```

```
$v
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   14.50   18.00   19.08   24.00   34.00
$m
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   23.00   28.00   28.52   33.00   48.00
```

Tabela 3: As observações correspondentes a sobreviventes têm menor valor de "meld"

Façamos agora uma análise gráfica:

```
########################
par(mfrow=c(1,1),font.axis=2,font.lab=2,cex.lab=1.2)       #parâmetros a definir para o gráfico
boxplot(meld~morte,ylab="meld",xlab="morte=1,sobrevida=0",td)          #box-plot
```

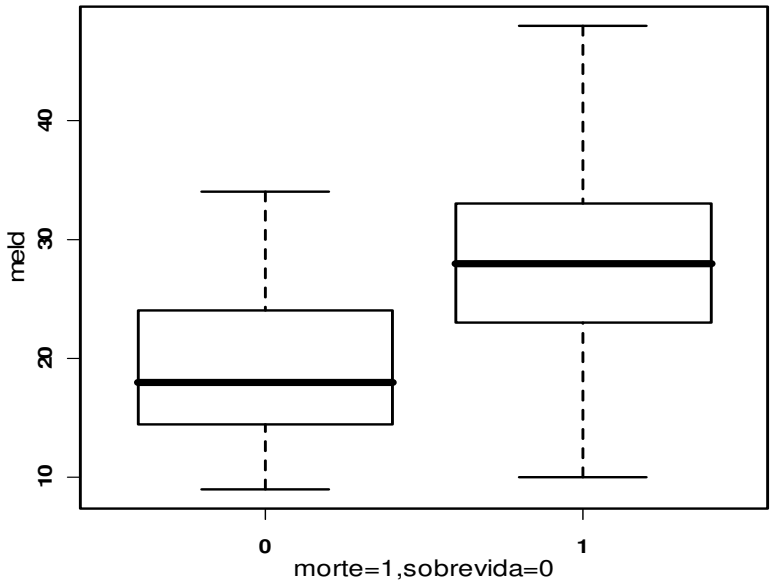


Figura 1: A box-plot diferencia de forma clara os valores de "meld" para "morte" e "sobrevida"

A proporção de "mortes" por decil indicará se o modelo logístico é um bom candidato a "noivo":

```
################################
q=quantile(td$meld,seq(0,1,0.1))              #quantis da variável meld: "Mín", 0.1, 0.2, …., 0.9 e "Max"
meldcat=rep(0,dim(td)[1])    #cria uma nova variável chamada "meldcat" com valor 0 para todas as observações
for (i in 1:10)     meldcat[td$meld>=q[i] & td$meld<q[i+1]]=mean(c(q[i],q[i+1])) # "meldcat" dá a média por quantil
meldcat[td$meld >= q[10] ] <- mean(c(q[10],q[11]))
Morte=as.numeric(td$morte)-1   # "morte" passa a ser numérica → 0 ou 1
tab=by(Morte, meldcat, mean)    # Cria uma tabela com a proporção de "morte=1" por classe
meldclas=as.numeric(names(tab))
tab=as.vector(tab)
mod1=glm(morte~meld,binomial(logit),td)          # Criação de um modelo logístico, digamos, o primeiro "beijo"
plot(Morte~meld,td)       # Dados 0 ou 1 como resposta ao valor de "meld"
points(fitted(mod1)~td$meld, pch=19,cex=1.1,col="blue") # A "probabilidade" de "morte" atribuída pelo modelo
points(meldclas,tab,col="red",pch=21,cex=1.2,ylim = c(0, 1))  #A proporção de "morte=1" por cada classe criada
```

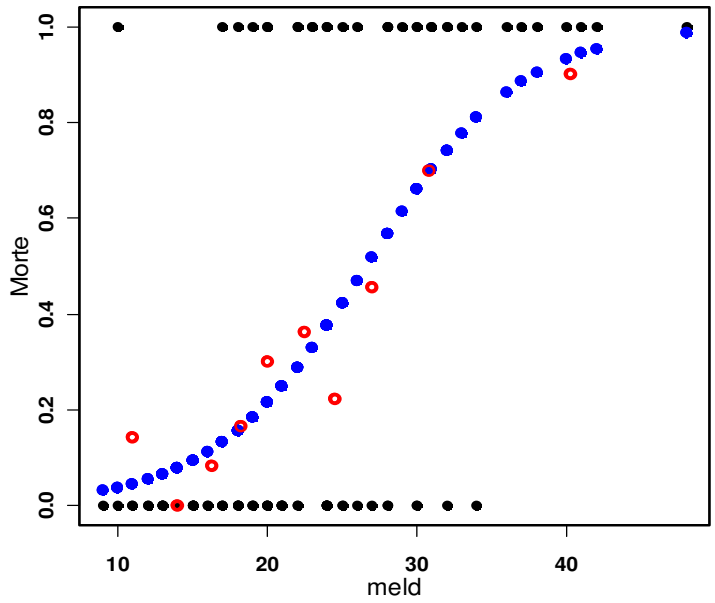


Figura 2: A adaptação dos decis à curva indica que a Regressão Logística será um "bom partido"

A Figura 2 indica-nos que a proporção de "morte=1" por decil de "meld" tem um comportamento semelhante à curva logística resultante do modelo ajustado; deste modo o modelo logístico perfila-se como primeira opção para candidato a modelo.
Da mesma forma, a "importância" da variável independente *"idade"* relativamente à variável resposta "morte" pode ser revelada através de um modelo de regressão logístico simples (teste de WALD (Hosmer & Lemeshow, 2000) , p-value<0.1) e ainda numa perspectiva gráfica semelhante à anterior

```
#IDADE
#############################################################################################
mod2=glm(morte~idade,binomial,td)  #O segundo "beijo"
summary(mod2)$coef  # Output do R para um Modelo Linear Generalizado (Tabela de coeficientes e teste de Wald)
##############
```

```
                Estimate Std. Error   z value   Pr(>|z|)
(Intercept) -2.33938026 1.06224653 -2.202295 0.02764447
idade        0.02755013 0.01671938  1.647796 0.09939452
```

Tabela 4: A variável "idade" revela alguma importância, independentemente do valor de "meld"

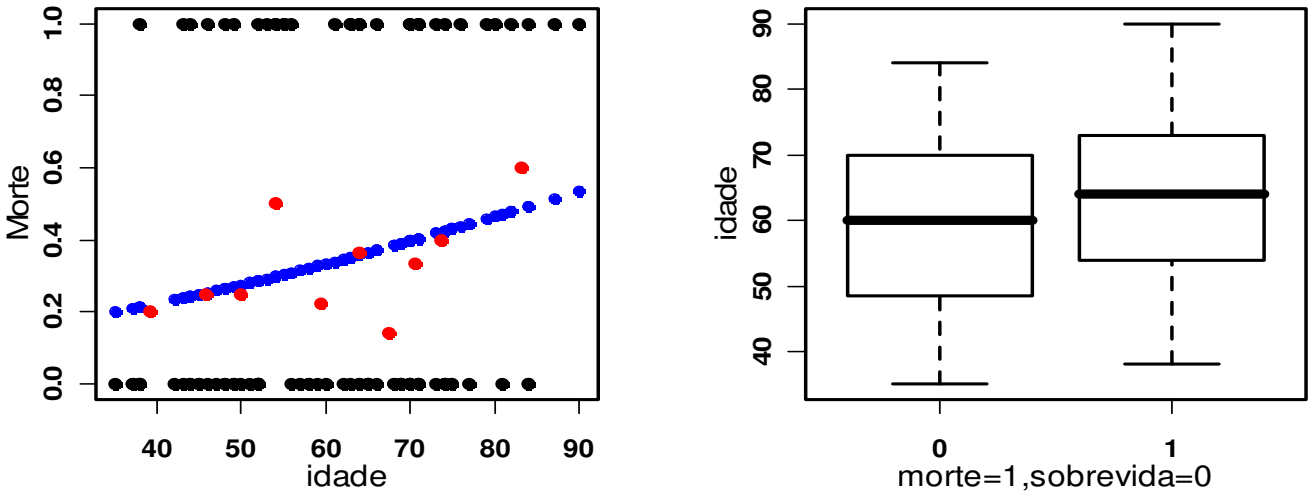


Figura 3: A proporção de "morte" cresce com os decis. A idade é mais elevada para "morte=1"

3. Casamento

Estamos então na fase de considerar que o "casamento", não só é possível, como terá tudo para dar certo. Os preditores da variável "morte", a considerar, serão "idade" e "meld". Construamos o modelo:

```
#MELD+IDADE
#############################################################################################
mod3=glm(morte~idade+meld,binomial,td) # O modelo
summary(mod3) #Output simplificado do modelo
```

```
glm(formula = morte ~ idade + meld, family = binomial, data = td)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7476  -0.6418  -0.4077   0.4760   2.6562
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.12405    2.11291  -4.318 1.57e-05 ***
idade        0.05379    0.02233   2.409    0.016 *
meld         0.22372    0.04903   4.563 5.04e-06 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 (Dispersion parameter for binomial family taken to be 1)
    Null deviance: 123.550  on 95  degrees of freedom
Residual deviance:  83.322  on 93  degrees of freedom
AIC: 89.322
```

Tabela 5:As variáveis "idade" e "meld" explicam de forma significativa a variável "morte"

Para analisar algumas características da "população", fornecidas pelo modelo, precisamos de introduzir dois conceitos: $ODDS_{x,z}$ para {meld=x,idade=z} e $ODDS\_RATIO_{k,m}$(vamos supor k>0 e z>0).

- $ODDS_{x,z}(O_{x,z})$ - Razão entre a proporção de indivíduos na população com {morte=1|x,z} e {morte=0|x,z} ($O_{x,z} = \frac{P(Y=1|x,z)}{P(Y=0|x,z)}$) (Por exemplo, para 2/3 vs 1/3 o ODDS é 2).
- $ODDS\_RATIO_{k,m}(OR_{k,m})$ - Razão entre $O_{x+k,z+m}$ e $O_{x,z}$

No caso do modelo logístico $O_{x,z} = \exp(\beta_0 + \beta_{meld}x + \beta_{idade}z)$ e $OR_{k,m} = \frac{O_{k+x,m+x}}{O_{k,x}} = Exp(k\beta_{meld} + m\beta_{idade})$, o que implica, assim, características especiais na população que obrigam a um $OR_{k,m}$ não dependente de x e z. A principal restrição do modelo também é a sua maior força pela simplicidade que implica.
Por exemplo, dois indivíduos com o mesmo valor de "meld" mas com 10 anos de diferença, têm $OR_{0,10}$ expresso pela seguinte expressão

$$OR_{0,10} = \frac{O_{0+x,7+x}}{O_{k,x}} = Exp(0\beta_{meld} + 10\beta_{idade}) = Exp(10\beta_{idade}),$$

enquanto se ambos tiverem a mesma idade mas diferença de 1 unidade no valor de "meld" o seu $OR_{1,0}$ será

$$OR_{1,0} = \frac{O_{1+x,0+x}}{O_{k,x}} = Exp(\beta_{meld} + 0\beta_{idade}) = Exp(\beta_{meld}).$$

O R, como bom "companheiro", obviamente que estimará estes valores:

```
##################################
id10=exp(10*coef(mod3)[2])  #ODDS_RATIO para 10 anos de diferença e o mesmo valor de "meld"
meld1=exp(coef(mod3)[3]) #ODDS_RATIO por unidade de "meld" para indivíduos com a mesma idade
c("10anos"=id10,"   1 unidade"=meld1)
```

```
10anos.idade      1 unidade.meld
    1.712439           1.250722
```

Tabela 6: ODD_RATIO de 10 anos e ODDS_RATIO de 1 unidade no valor de meld

O modelo estima,assim, um aumento de cerca de 71% no valor do ODDS em cada 10 anos e um aumento de 25% no valor do ODDS por cada unidade a mais no valor de "meld".
Já agora podemos estimar o valor de $O_{x,z}$ ou mesmo P(morte=1|meld=x,idade=z) com x e z nas respectivas médias amostrais:

```
################################################
#ODDSx,z para x e z nas respectivas médias amostrais
id1=mean(td$idade);meld1=mean(td$meld) # Calcula as médias
x1z1=data.frame(idade=id1,meld=meld1)    #Cria um "novo" individuo com características médias
Oxz=exp(predict(mod3,x1z1))    #Odds para o índividuo médio
Pmorte=1/(1+exp(-predict(mod3,x1z1))) # Probabilidade de "morte" para o índividuo médio
c(med_idade=id1,med_meld=meld1,ODDS=Oxz,Pmorte=Pmorte)
```

```
med_idade   med_meld       ODDS       Pmorte
60.6875000 22.3229167  0.4208584  0.2962001
```

Tabela 7: Com base no modelo, indivíduos com 60 anos e meld=22 terão, em média, um ODDS de 0.42 e probabilidade de "morte" de 0.29

4. Avaliação do Desempenho

Vamos admitir o modelo à fase de avaliação. A curva ROC (Hosmer & Lemeshow, 2000) é uma das técnicas mais eficazes e por isso uma das mais utilizadas.

Na curva ROC o primeiro conceito a reter é o de **"cut_off"**. O modelo permite estimar P(morte=1) para cada uma das observações. Se fixarmos um valor **c** $\{c : 0 \le c \le 1\}$ e se a cada observação atribuirmos o valor 1 quando a sua estimativa é superior ou igual a **c** ($\hat{y} = 1$) e o valor 0 caso contrário ($\hat{y} = 0$), estamos a definir um {"cut_off"=**c**}.

Com base neste conceito, vamos ter para cada "cut_off" um vector bidimensional $(\hat{\mathbf{y}} = atributos, \mathbf{y} = observados)$ constituído por pares (0,0), (0,1), (1,0) ou (1,1). Obviamente que muitos pares (0,0) e (1,1) são indicadores de uma boa capacidade preditiva do modelo (número de "acertos") para um $"cut\_off" = c$.

Sem nos alongarmos mais e partindo do principio que o leitor, ou já está familiarizado com o conceito, ou, antes de continuar, irá fazer uma pesquisa mais aprofundada sobre o tema (Ver, por exemplo, Hosmer & Lemeshow, 2000), calculemos alguns valores interessantes:

```
                                                    #ROC
#################################################################################################
library(ROCR) # "Package" disponível no R para avaliar a curva ROC e afins
pred=prediction( fitted(mod3), td$morte)  #Comando "chave" associado a ROCR
perf=performance(pred,"acc")   # Calcula o número de "acertos" (Pares (0,0) e (1,1)
par(mfrow=c(1,1),mgp=c(2,1,0),font.axis=2,font.lab=1.2,lwd=1,cex.lab=1.2)
par(mai=c(1,1,1,1),col=2)
cut_off=.Call("R_get_slot", perf, "x.values")       #"Slot" que nos permite conhecer os "cut_off" utilizados
(cf= as.vector (cut_off[[1]]))
```

```
  Inf 0.993 0.990 0.969 0.948 0.936 0.934 0.908 0.904 0.893 0.892 0.882 0.803 0.783 0.747
                                                  .....
    0.067 0.061 0.060 0.050 0.048 0.040 0.039 0.029 0.025 0.023 0.021 0.018 0.011
```

Tabela 8: Alguns do valores criados pelo "package" para o "cut-off"

```
#SENSIBILIDADE VS ESPECIFICIDADE
###################################
sens=performance(pred,"sens") #'tpr"->verd. positivos(sensibilidade), P(y^=1|y=1)
espec=performance(pred,"spec") #'tnr"->verd. negativos (especificidade),P(y^=0|y=0)
sensibilidade=.Call("R_get_slot", sens, "y.values")[[1]]    #Retira o "slot" com a "sensibilidade" por "cut-off"
especificidade=.Call("R_get_slot", espec, "y.values")[[1]] #Retira o "slot" com a "especificidade" por "cut-off"
plot(sensibilidade~cf,ylab="Sens. vs Espec.",xlab="Cut_off",cex=0.7,col=2)
points(especificidade~cf,t="l",lwd=2,col=3)
optimo=cf[which.min(abs(sensibilidade-especificidade))]   # "optimo" será o ponto de intersecção das curvas
points(-0.1~optimo,t="h",lwd=3,col="black")
```

```
text(0.53,0.03,round(optimo,2),col=4,font=2)
text(0.33,0.03,"cut_off optimo=",col=4,font=2)
legend(0.6,0.8,c("Sensibilidade","Especificidade"),text.col=c(2,3),bty="n",lwd=c(3,2),lty=c(3,1),col=c(2,3))
```

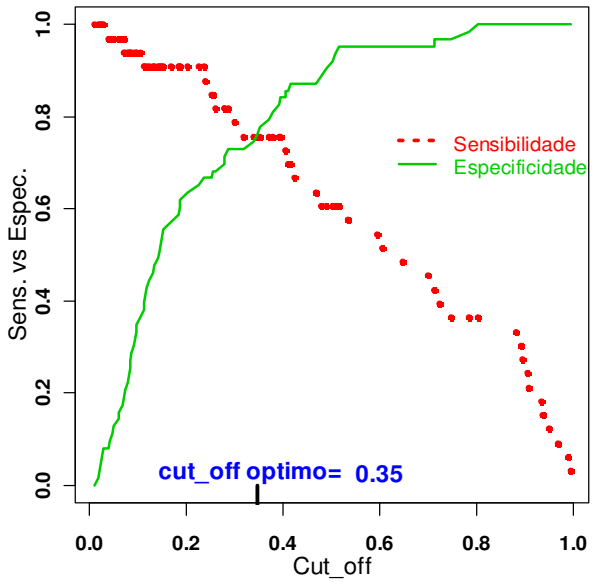


Figura 4: Nesta figura vê-se o "cut_off" óptimo, a sensibilidade e especifidade

A Figura 4 indica que a sensibilidade é relativamente baixa. Assim o "cut-off" que torna o modelo com melhor capacidade preditiva está abaixo de 0.5. O "cut_off" óptimo é o que maximiza, simultaneamente, a sensibilidade e a especificidade do modelo.
Outra "informação" fornecida pelo "package" é a área sob a curva "sensibilidade vs 1-especificidade". A esta área é usual chamar "**A**rea **U**nder **C**urve" (AUC).

```
# AREA SOB A CURVA ROC
#########################################################################
matplot(1-especificidade,sensibilidade,t="l",lwd=2,col=2,main="Curva ROC")
area= performance(pred, "auc")
area=.Call("R_get_slot", area, "y.values")[[1]]    "Slot" que fornece o valor de AUC
legend(0.5,0.56, "AUC=",bty = "n", cex=1.2,text.col ="black")
text(0.78,0.5,round(area,2),cex=1.2,col="black")
```

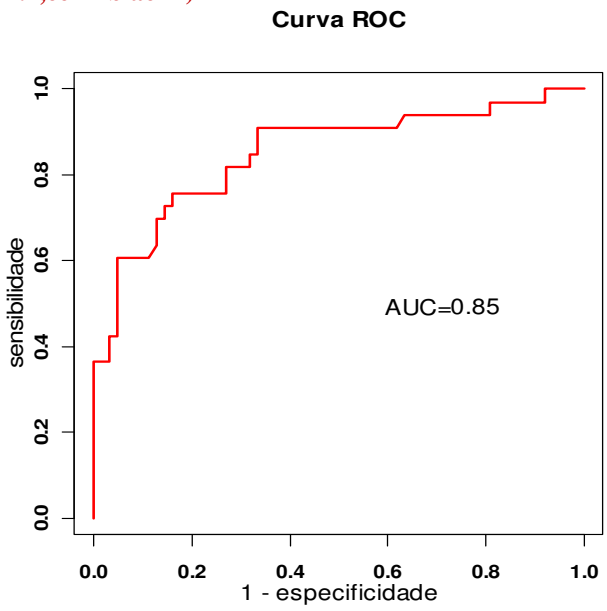


Figura 5: Esta curva ROC permite afirmar que o modelo tem uma elevada capacidade predictiva

Outras técnicas de diagnóstico estão ainda disponíveis mas não serão aqui analisadas.

5. Descendência

Para terminar, é importante perceber como este modelo poderá ser útil:
Do ponto de vista prático, é muito importante para o médico dispor de uma "ferramenta" que lhe permita avaliar precocemente se o doente com PBE que está a observar é de elevado risco, porque deste modo adoptará um conjunto de medidas e instituirá tratamentos mais específicos, que em última análise poderão melhorar o prognóstico desse doente.

```
# ATRIBUIÇÃO ATRAVÉS DO PONTO DE CORTE
#####################################################
require(grDevices)
beta=coef(mod3)
f=function(idade,meld)
{
eta=beta[1]+beta[2]*idade+beta[3]*meld
1/(1+exp(-eta))
}
idade=seq(30,95,0.5)
meld=seq(8,50,0.2)
z=matrix(0,nrow=length(idade),ncol=length(meld))
for(i in 1:length(idade))
for (j in 1:length(meld))
z[i,j]=f(idade[i],meld[j])
image(idade,meld,z>optimo,col=c("azure3","azure1"),xlab="AGE",ylab="MELD")
legend(70,50,c("atrib=morte","atrib=sobrevida"),pch=c(19,21),bg="white")
y=td$morte
points(td$idade[which(y==0)],td$meld[which(y==0)],pch=21)        # Observações
points(td$idade[which(y==1)],td$meld[which(y==1)],pch=19)
contour(idade,meld,z,labcex=1,add=T,col=1)
```

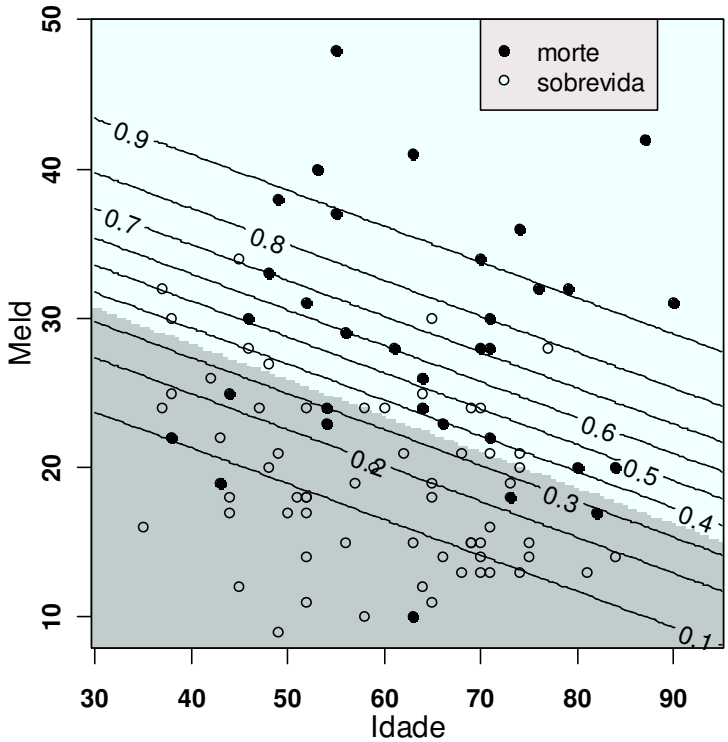


Figura 6: É possível visualizar de forma simples a "qualidade" do modelo, comparando a "atribuição" com o "real"

As duas cores da Figura 6 separam as observações por estimativa da probabilidade de morte, acima (cor clara e "atribuição=morte") ou abaixo do "cut-off" óptimo ("atribuição=sobrevida"). As curvas de nível permitem ainda visualizar o que aconteceria para outros "cut-off". Assim, a Figura 6 "promete" ajudar o médico a atribuir um protocolo a cada novo doente, após validação futura do modelo.

6. Bibliografia

Faraway, J. J. (2006). *Extending the Linear Model with R.* Chapman & Hall/CRC.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons .

Kamath, P. S., Wiesner, R., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., et al. (2001). A model to predict survival in patients with end-stage liver disease. *Hepatology , 33*, pp. 464-470.

McCullagh, P., & Nelder, J. (1983). *Generalized Linear Models.* Chapman & Hall/CRC.

Nobre, S. R., Cabral, J. E., Gomes, J. J., & Leitão, M. C. (2008, Dec). In-hospital mortality in spontaneous bacterial peritonitis: a new predictive model. *Eur J Gastroenterol Hepatol. , 20* (12), pp. 1176-1181.

Song, J. Y., Jung, S. J., Park, C. W., Sohn, J. W., Kim, W. J., Kim, M. J., et al. (2006). Prognostic significance of infection acquisition sites in spontaneous bacterial peritonitis: nosocomial versus community acquired. *J Korean Med Sci , 21*, pp. 666-671.

Thuluvath, P. J., Morss, S., & Thompson, R. (2001). Spontaneous bacterial peritonitis – in-hospital mortality,predictors of survival, and health care costs from 1988 to 1998. *Am J Gastroenterol , 96*, pp. 1232-1236.

As Estáveis no R

Helena Iglésias Pereira, *hmpereira@fc.ul.pt*

CEAUL e Departamento de Estatística e Investigação Operacional, FCUL

1. INTRODUÇÃO

É comum em estatística aplicada, assumir que os fenómenos aleatórios observados são o efeito de um grande número de causas independentes e não observáveis que se adicionam resultando no fenómeno em estudo. Pelo teorema Limite Central a soma de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) convenientemente centrada e reduzida, tem distribuição assintótica normal. Mais geralmente, a soma de v.a.'s i.i.d., convenientemente normada tem ainda distribuição normal, desde que se imponham algumas condições no comportamento assintótico do segundo momento truncado das parcelas.
Pelo Teorema Limite Central Generalizado, se a soma de v.a.'s i.i.d. tem distribuição limite não degenerada esta distribuição tem de ser um elemento da classe das leis *estáveis*, de que a normal é o único elemento com variância finita.

A classe das distribuições estáveis é caracterizada por quatro parâmetros, usualmente designados por α, β, a e c, respectivamente o expoente característico, parâmetro de assimetria, localização e escala.

As variáveis aleatórias estáveis têm propriedades aditivas interessantes e são absolutamente contínuas, mas somente se conhecem expressões analíticas das funções densidade de probabilidade (f.d.p.) correspondentes aos casos: $\alpha=2$ (normal), $\alpha=1$, $\beta=0$ (Cauchy) e $\alpha=1/2$, $\beta=1$ (Lévy).

Este facto aliado à não existência de alguns momentos destas distribuições, dificulta muito o problema da inferência estatística em modelos estáveis. No entanto, estas distribuições são usadas numa grande variedade de problemas de economia, finanças, engenharia e estatística [1].

Mais recentemente, a existência de programas de computador com suficiente precisão permite calcular as funções densidade, as funções de distribuição, os quantis e gerar amostras aleatórias destas distribuições, nomeadamente o software **R** 2.**6.2**, como iremos ilustrar ao longo deste artigo.

2. MODELOS ESTÁVEIS: DEFINIÇÕES E PROPRIEDADES

Seja $\{Y_i\}_{n\in \mathbb{N}}$ uma sucessão de v.a.'s i.i.d., a v.a. X diz-se estável sse para todo o $n \in \mathbb{N}$ existem constantes $a_n > 0 \;\; e \;\; b_n \in \mathbb{R}$ tais que

$$\frac{\sum_{i=1}^{n} Y_i - b_n}{a_n} \xrightarrow{d} X, \quad n \to +\infty \qquad (2.1)$$

A classe das distribuições estáveis tem função característica (f.c.) da forma

$$\phi(t) = exp\left\{ iat - c|t|^{\alpha}\left[1 - i\beta\frac{t}{|t|}\omega(|t|,\alpha)\right]\right\} \tag{2.2}$$

$$\omega(|t|,\alpha) = \begin{cases} tan(\pi\alpha/2), & \alpha \neq 1 \\ -\frac{2}{\pi}ln|t|, & \alpha = 1 \end{cases} \qquad a \in \Re,\ c > 0,\ \alpha \in (0,2],\ \beta \in [-1,1] \tag{2.3}$$

Existem outras representações da f.c. de uma distribuição estável [5], mas esta é a mais usual. As distribuições estáveis têm propriedades interessantes:

I) As caudas da função de distribuição (f.d.) *F* de uma v.a. X estável satisfazem

$$\begin{aligned} &x^{\alpha}[1 - F(x)] \to C\frac{k_2}{k_2 + k_1}\frac{2-\alpha}{\alpha} \\ &x^{\alpha}F(-x) \to C\frac{k_1}{k_2 + k_1}\frac{2-\alpha}{\alpha} \\ &quando\ x \to +\infty \quad e \quad (C > 0,\ k_1, k_2 \geq 0\ \ e\ \ k_1 + k_2 > 0) \end{aligned} \tag{2.4}$$

Donde se conclui que o expoente característico (e.c.) α está intimamente relacionado com o comportamento das caudas da f.d., sendo o peso destas tanto menor quanto maior o e.c. (a normal é a estável com caudas mais leves).

II) Toda a distribuição estável de e.c. α ($0 < \alpha < 2$) tem momentos absolutos finitos de ordem $\gamma \in (0, \alpha)$ (esta propriedade é consequência da anterior).

III) Dado que a função característica $\phi(t)$ é absolutamente integrável, todas as distribuições estáveis são absolutamente contínuas.

Por outro lado, o parâmetro de assimetria β "compara" o peso da cauda direita com o peso da cauda esquerda

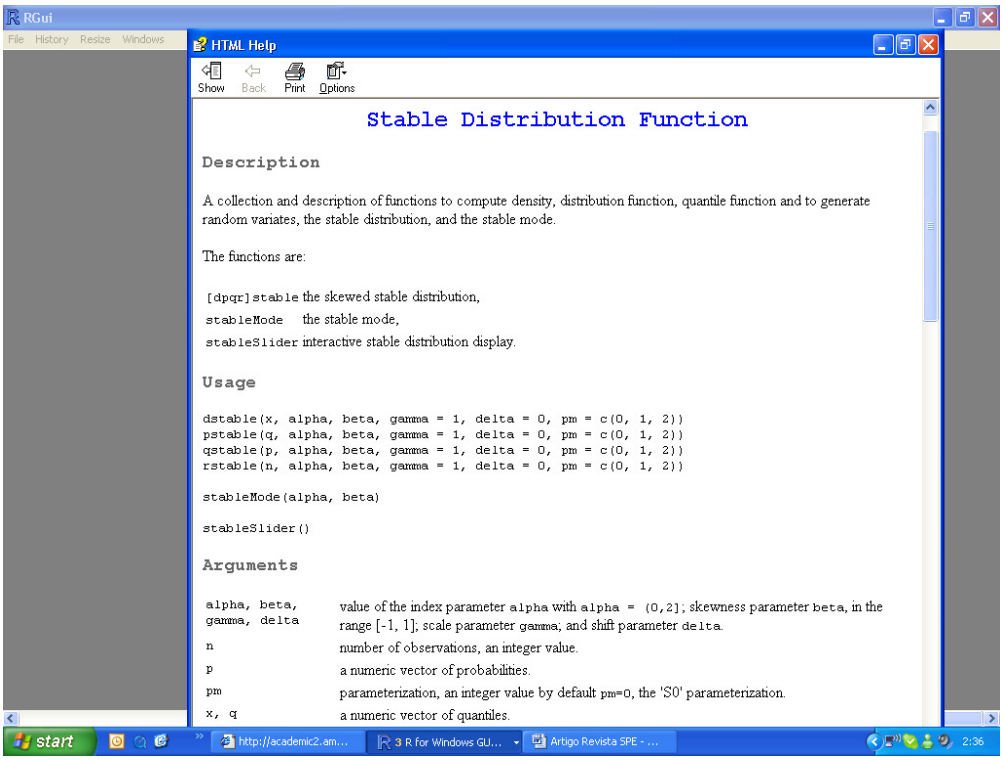
$$\begin{aligned} &\lim_{x \to +\infty} \frac{1 - F(x)}{1 - F(x) + F(-x)} = \frac{k_2}{k_1 + k_2} \\ &\lim_{x \to +\infty} \frac{F(-x)}{1 - F(x) + F(-x)} = \frac{k_1}{k_1 + k_2} \\ &tendo - se \quad \beta = \frac{k_2 - k_1}{k_1 + k_2} \end{aligned} \tag{2.5}$$

E quando $\alpha \neq 1$, tem-se ainda a seguinte relação [2]

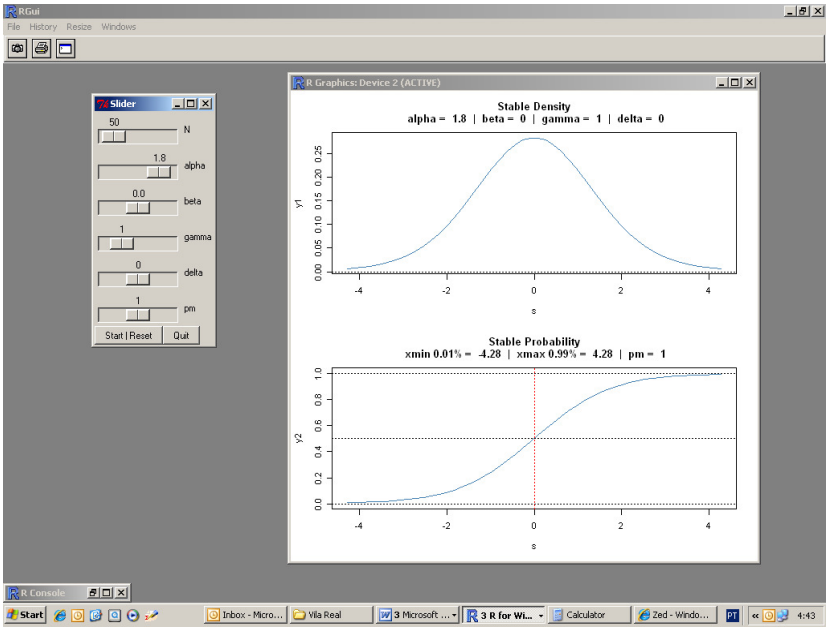
$$F(0;\alpha,\beta) = \frac{1}{2} - \frac{arctan(\beta\, tan(\pi\alpha/2))}{\pi\alpha} \tag{2.6}$$

Também para α=1 se pode verificar que $F(0;\alpha,\beta)$ decresce com β [4]. Podemos pois concluir que o parâmetro de assimetria está relacionado com as caudas da f.d. e com o valor desta no ponto x=0.

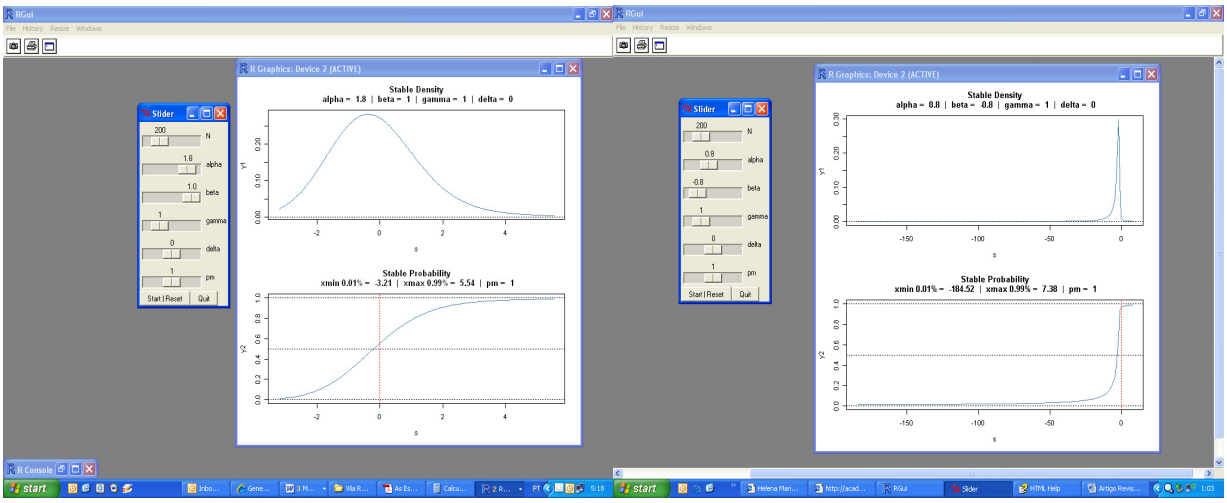
A função **stableSlider** da biblioteca *Rmetrics* do software **R** versão **2.6.2**, permite visualizar a f.d. e a f.d.p. de uma v.a. estável de parâmetros (α,β,a,c) e que passaremos a designar por $S(\alpha,\beta,a,c)$.



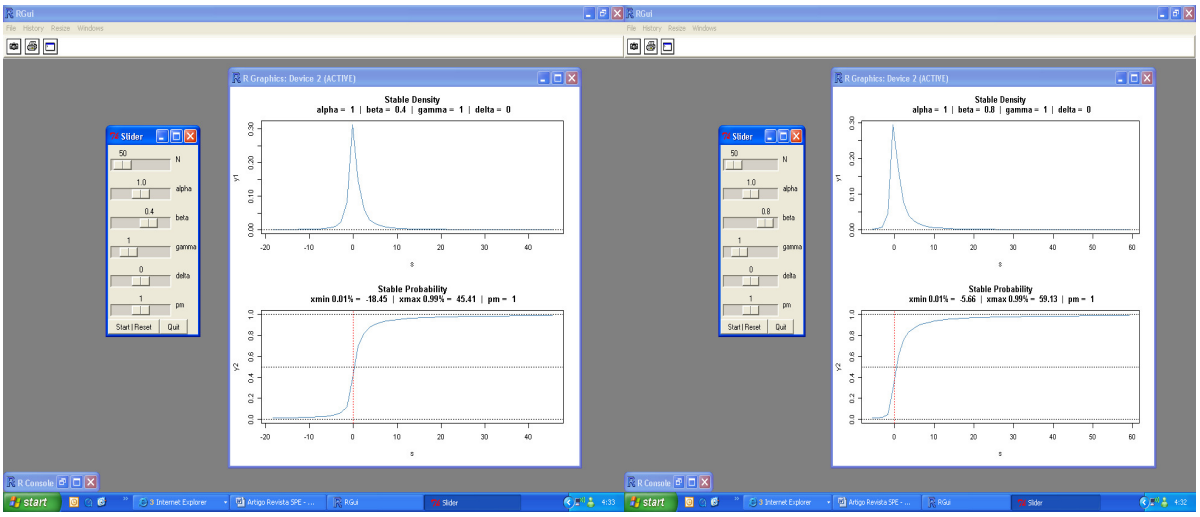
Nesta função a parametrização 1 é a que corresponde à representação (2.2), para a qual se verifica a relação (2.6). Quando β=0 a distribuição é simétrica e portanto $F(0;\alpha,\beta)=0.5$, $\alpha\neq 1$.



Para outros valores de β tem-se por exemplo, $F(0;1.8,1)=0.5556$ e $F(0;0.8,-0.8)=0.9715$.



Quando α=1 a relação (2.6) não é válida, mas podemos verificar que $F(0;\alpha,\beta)$ decresce com β.



3. GERAÇÃO DE ESTÁVEIS: UMA APLICAÇÃO

Seja X uma variável aleatória estável de parâmetros reais $\alpha\in(0,2]$, $\beta\in[-1,1]$, $a\in R$, $c>0$. Suponhamos que todos os parâmetros são conhecidos excepto o parâmetro de assimetria β, e que a=0 e c=1 por comodidade de cálculo. Como já foi referido anteriormente tem-se

$$F(0;\alpha,\beta)=\frac{1}{2}-\frac{arctan\left(\beta\tan(\pi\alpha/2)\right)}{\pi\alpha} \qquad (3.1)$$

E resolvendo em ordem a β obtém-se

$$\beta=\frac{tan\left(\pi\alpha\left(\frac{1}{2}-F(0;\alpha,\beta)\right)\right)}{tan(\pi\alpha/2)} \qquad (3.2)$$

A relação anterior permite obter um estimador do parâmetro β quando α conhecido. Seja $(X_1,...X_n)$ uma amostra aleatória de uma população estável padrão $S(\alpha,\beta,0,1)$ que designaremos abreviadamente por $S(\alpha,\beta)$, então

$$\beta_n^* = \frac{tan\left(\pi\alpha\left(\frac{1}{2} - F_n^*(0;\alpha,\beta)\right)\right)}{tan(\pi\alpha/2)} \qquad (3.3)$$

onde $F_n^*(x)$ é a função de distribuição empírica.

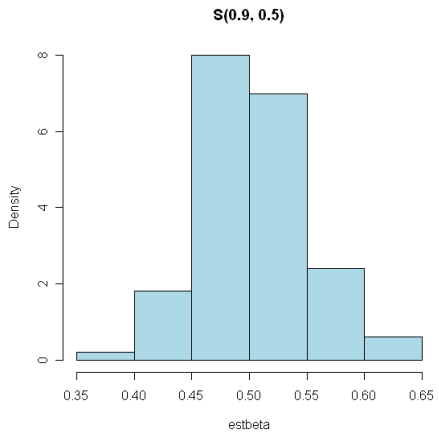
A função **rstable** (n, alpha, beta, gamma, delta, pm=c (0, 1, 2)), onde gamma é o parâmetro de escala *c*, delta o parâmetro de localização *a* e o parâmetro pm indica a parametrização utilizada, permite gerar amostras de uma distribuição $S(\alpha,\beta,a,c)$. A parametrização que nos interessa é a 1 como já referimos.

A título de exemplo, incluímos o programa para a obtenção de estimativas de β a partir da geração de *m* amostras de dimensão *n* de uma estável padrão *S (0.9,0.5).*

```
n<-500
m<-100
alpha<-0.9
beta<-0.5
s<-matrix(0,nrow=n,ncol=m)
freq<-matrix(0,nrow=1,ncol=m)
F0_emp<-matrix(0,nrow=1,ncol=m)
estbeta<-matrix(0,nrow=1,ncol=m)
set.seed(011)
for(j in 1:m){s[,j]<-rstable(n,alpha,beta,1,0,1)
freq[,j]<-length(subset(s[,j],s[,j]<=0))
F0_emp[,j]<-freq[,j]/n  # f.d. empírica no pto x=0
estbeta[,j]<-tan(alpha*pi*(0.5- F0_emp[,j]))/tan(pi*alpha/2)}
F0_emp
estbeta
round(mean(estbeta),4)
[1] 0.5046

var<-matrix(0,nrow=1,ncol=m)
for(j in 1:m){ var[,j]<-((estbeta[,j]-mean(estbeta)))^2/(m-1)}
round(sqrt(sum(var)),4)
[1] 0.0472

hist(estbeta,  freq=FALSE,  right=FALSE,  breaks="Sturges",  xlab="estbeta",
main="S(0.9, 0.5)", col="lightblue")
```



Ao apresentarmos este estimador do parâmetro β obtido em situações muito particulares, todos os restantes parâmetros são conhecidos, pretendemos simplesmente dar um exemplo de aplicação da função rstable() do *package Rmetrics* do **R** versão **2.6.2**.

4. AGRADECIMENTOS

A autora agradece ao Professor John P. Nolan os seus valiosos esclarecimentos relativamente à parametrização a utilizar na função *rstable*.

5. REFERÊNCIAS

[1] Adler, Robert J., Feldman, Raisa E., Taqqu, Murad S. (1998). *A Practical Guide to Heavy Tails,* Birkhauser.
[2] Iglésias Pereira, H. (2009). As estáveis no R. *Notas e Comunicações,* 9/2009 .
[3] Nolan, John P. (1998). Parameterizations and models of stable distributions. *Statistics and Probability Letters*, 38, p. 187-195.
[4] Nolan, John P. (2009). *Stable Distributions- Models for Heavy Tailed Data*, Capítulo 1, http://academic2.american.edu/~jpnolan/stable/stable.html.
[5] Zolotarev, V.M. (1986). One-dimensional Stable Distributions. *Am. Math. Society*. Providence, R.I.

Detecção de Fraude usando o R: um caso de estudo

Luís Torgo, *ltorgo@liaad.up.pt*

LIAAD / Inesc Porto, LA / Faculdade de Ciências / Universidade do Porto

1 Introdução

Este pequeno artigo pretende ilustrar a utilização do R [11] num problema concreto que é enquadrável no cenário mais lato da análise de dados para apoiar a tarefa de detecção de fraude em ambientes com recursos de inspecção limitados. A motivação para o uso desta ferramenta está relacionada fortemente com as suas características de código aberto e também a sua disponibilização gratuita. Estas características permitem uma fácil adaptação de ferramentas existentes a novos problemas. Para além disso o poder da linguagem de programação que lhe está associada permite o fácil e rápido desenvolvimento de protótipos que podem ser usados para testar ideias. Finalmente, ao disponibilizar uma quantidade impressionante, e em constante crescimento, de ferramentas e métodos de análise de dados, o R facilita grandemente o teste, avaliação e comparação de diferentes metodologias para abordar um problema.

A detecção de fraude é uma tarefa com forte impacto económico e social em inúmeras áreas de actividade humana. As ferramentas informáticas têm vindo a ser cada vez mais usadas para a recolha de dados sobre uma grande parte das actividades humanas, levando muitas vezes à criação de bases de dados de tamanho demasiado grande para a inspecção humana. Pela sua sensibilidade, as actividades onde a detecção de fraude tem lugar, também se encontram fortemente monitorizadas. Neste contexto, o uso de ferramentas de análise de dados para suportar e apoiar a tarefa de detecção de fraude trás vantagens óbvias.

O comportamento fraudulento é normalmente um desvio à norma na actividade em causa. Assim, não é surpreendente que a detecção destes desvios esteja relacionada com a detecção de outliers nas bases de dados que descrevem estas actividades. A detecção de outliers é uma tarefa de análise de dados por demais estudada e explorada em inúmeras áreas do saber (e.g. [5, 2, 4]). O papel das ferramentas de análise de dados no contexto da detecção de fraude é o de fornecer pistas que possam ajudar na tarefa de inspecção dos casos suspeitos. Estas tarefas de inspecção estão normalmente sujeitas a fortes restrições de recursos tanto humanos como financeiros. De facto, em muitas organizações não existem recursos que possibilitem a inspecção de todos os casos minimamente suspeitos. Assim, é importante para estas organizações direccionar os seus recursos para os casos mais relevantes. Esta noção de relevância é obviamente dependente do domínio de aplicação em causa. Poderá ser relevância económica, por exemplo interessando detectar as situações com maior retorno financeiro, mas também poderá estar ligada a outros factores, por exemplo sociais, número de detecções, etc..

Uma grande parte dos métodos de detecção de outliers existentes limitam-se a fornecer uma classificação das observações em normais ou desviantes, de acordo com uma qualquer metodologia. Este tipo de abordagens é pouco adequado a cenários de detecção de fraude com recursos limitados. De facto, é fácil deparar com situações em que são sinalizadas mais observações desviantes do que aquelas que os recursos actuais permitem inspeccionar. Neste contexto, são de muito maior utilidade ferramentas de detecção de outliers que forneçam como resultado um ranking de grau de desvio à normalidade das observações fornecidas à ferramenta. Munidos deste ranking, os

utilizadores poderão direccionar os recursos disponíveis de forma mais racional, optimizando desta forma os resultados obtidos com os mesmos. É esta a assumpção base do trabalho aqui descrito.

2 Um Caso de Estudo

As companhias Portuguesas têm que declarar mensalmente as suas transacções comerciais com outros países da UE ao Instituto Nacional de Estatística (INE). Estes dados são fornecidos através de um formulário chamado INTRASTAT, cujo conteúdo fornece ao INE informações como o peso, o custo, a identificação dos itens transaccionados, etc.. Posteriormente, estes dados são inseridos numa base de dados para utilização na produção de estatísticas de comércio externo. Durante o preenchimento desses formulários podem ocorrer erros e a sua identificação é de extrema importância para que as estatísticas calculadas a partir deles sejam fidedignas. Os erros mais frequentes são por exemplo o da introdução de identificadores de artigos (IDs) incorrectos que vão associar uma transacção a atributos errados, ou o uso incorrecto de unidades de medida como declarar o peso em toneladas em vez de quilos. Estes erros podem ser vistos como anomalias relativamente ao que são as características típicas das transacções envolvendo o mesmo tipo de produtos. Neste contexto, e dada a sua raridade, eles têm todas as características de uma fraude no sentido descrito anteriormente[1].

Para saber se uma transacção contém um erro ou não, os especialistas do INE inspeccionam manualmente as informações de que dispõem. Dado que o número de transacções é da ordem dos milhares por mês, e que os recursos humanos disponíveis para esta tarefa são limitados, um sistema automático que dê apoio para esta tarefa de inspecção é altamente desejável. Este é o nosso objectivo operacional com este trabalho.

O INE forneceu-nos uma base de dados com informação sobre este tipo de transacções. Os dados dizem respeito a oito meses do ano de 1998. Para cada mês, cada artigo possui um determinado número de transacções. O número de transacções varia dependendo do artigo e do mês. Por exemplo, o artigo com ID "2013000" possui 70 transacções no mês de Setembro mas em Outubro não possui qualquer transacção.

Ao longo do tempo, os especialistas do INE adquiriram algumas estratégias para lidar com este problema. Em seguida apresentamos alguns dos conselhos que nos foram dados:

- Utilizar a variável custo por quilo para a identificação das transacções com erros - de acordo com os peritos do INE a variável custo por quilo, que é obtida dividindo-se os custos de cada transacção pelos seus respectivos pesos, é a variável mais eficiente na detecção dos erros.
- Inspeccionar os artigos separadamente - este conselho decorre da diversidade dos itens. Por exemplo, considerar o preço padrão do vinho só faz sentido quando estamos a tratar de transacções de vinho.
- Processar os dados mensalmente - esta é a forma como os dados eram tratados no INE na época em que obtivemos os dados.
- Enviar para inspecção todas as transacções dos artigos com poucas transacções.

No conjunto de dados fornecidos, existe uma coluna com o resultado da inspecção feita pelos peritos do INE. Esta coluna só indica quais as transacções que foram identificadas como erro. Nas transacções não sinalizadas pode dar-se uma de duas situações: ou foram inspeccionadas e não foram consideradas erro; ou não foram inspeccionadas. Infelizmente, na base de dados fornecida não havia qualquer distinção entre estas duas situações. Assim, possuímos informação somente sobre os erros detectados. Isto tem impacto nas medidas de avaliação que vamos usar para qualificar as metodologias que iremos tentar nesta tarefa de análise de dados.

[1]Na realidade algumas dessas situações podem ser mesmo tentativas de fraude, embora possam também ser simplesmente erros de introdução dos dados.

Qualquer metodologia que seja usada nestes dados vai pegar num conjunto de dados referentes às transacções de um determinado mês e vai sugerir um sub-conjunto destas como sendo as com maior potencial para serem erros. Seja N_i o conjunto de transacções no mês i, e n_i o conjunto de erros identificados por um qualquer modelo. Vamos definir a percentagem de transacções seleccionadas para inspecção como uma das medidas de avaliação, $\%S = n_i/N_i$. De acordo com o INE, para uma metodologia ser aceitável tem que ter $\%S \leq 50\%$. Seja E_i o conjunto de observações etiquetadas como erros pelos peritos do INE no mês i, e seja e_i o subconjunto de n_i que pertence a E_i. Neste contexto definimos uma outra estatística de performance como sendo o *Recall*, $\%R = e_i/E_i$. Mais uma vez o INE define como critério mínimo para uma metodologia ser aceitável atingir um valor de $\%R = 90\%$. Resumindo iremos qualificar a performance de qualquer modelo por um par de valores, $< \%S, \%R >$, sendo que o INE define como critérios mínimos que $\%S \leq 50\% \wedge \%R \geq 90\%$.

3 A Metodologia Proposta

A nossa metodologia para abordar este problema é baseada em métodos de agrupamento hierárquico. Os métodos de agrupamento hierárquico podem ser usados para identificar outliers como efeito lateral do processo de agrupamento (e.g. [10]). A maioria dos métodos de agrupamento baseiam-se na informação de uma matriz de distâncias e portanto podem ser classificados como métodos de detecção de outliers baseados em distâncias (e.g. [7]). Todavia métodos iterativos como os métodos hierárquicos de agrupamento (e.g. [6]) conseguem lidar com regiões de diferente densidade o que é um dos maiores problemas dos métodos de detecção de outliers baseados em distâncias. De facto, olhando por exemplo para os métodos hierárquicos aglomerativos, eles procedem de forma iterativa juntando dois dos grupos actuais baseados num critério relacionado com a sua proximidade, sendo esta decisão tomada localmente, i.e. levando em conta unicamente os dois grupos em causa.

A ideia geral do método que foi desenvolvido tem a ver com usar o percurso de cada observação pelo dendrograma como fonte de informação para obter um grau de outlier para a mesma. A motivação resulta na assumpção que, tendo em conta os critérios usados para o processo de aglomeração nos algoritmos de agrupamento hierárquico, os outliers deverão oferecer grande "resistência" a serem juntos num grupo com outras observações, nomeadamente se este grupo for formado por um grande número de observações "normais". Em resumo, a ideia chave da metodologia desenvolvida [12] é a de usar a altura em que uma observação é junta a outras durante o processo aglomerativo de agrupamento hierárquico, como indicador do grau de desvio da mesma, e desta forma obter um ranking de outliers. Cada observação pode participar em várias junções ao longo do processo iterativo. Levando isso em conta definimos o grau de outlier de uma observação como,

$$OF_H(x) = \max_i of_i(x) \tag{1}$$

em que i toma valores entre 1 e $N-1$ e representa o conjunto de passos de junção em que x participa, sendo N o número de observações do conjunto de dados.

A motivação para esta fórmula tem a ver com tentar capturar tanto outliers locais como outliers globais. Os outliers locais são observações que tipicamente ofereceram uma grande resistência a serem juntos com os seus vizinhos mais próximos, dado serem outliers neste contexto local. Todavia, assim que se juntam a estes, passando portanto a fazer parte de um conjunto maior de observações, deixam de ser vistos como outliers.

A função $of_i()$ determina o grau de outlier de uma observação que participa na junção que teve lugar no passo i do processo iterativo de agrupamento. Definimo-la como sendo função da diferença entre os tamanhos dos dois grupos envolvidos na junção no passo i. A ideia é a de que sempre que estamos em presença de uma junção entre dois grupos com tamanho muito diferente, então existem fortes suspeitas que os membros do grupo menor poderão ser outliers locais. Neste contexto, definimos a função como,

$$of_i(x) = \max\left(0, \frac{|g_{y,i}| - |g_{x,i}|}{|g_{y,i}| + |g_{x,i}|}\right) \quad (2)$$

em que $g_{x,i}$ e $g_{y,i}$ são os 2 grupos de observações envolvidos na junção no passo i, e $g_{x,i}$ é o grupo a que a observação x pertence.

De notar que nesta formulação se a observação x pertence ao grupo maior (i.e. $|g_{x,i}| > |g_{y,i}|$) o valor de $of_i(x)$ é 0 uma vez que a fracção terá um valor negativo.

A implementação destas ideias no R foi fácil pelo carácter de código aberto e programável deste ambiente. A nossa implementação actual usa o resultado da função `hclust()` que é baseada em código Fortran desenvolvido por F. Murtagh [9]. Esta função de agrupamento hierárquico produz como resultado uma estrutura de dados que fornece vária informação sobre o processo de agrupamento. A estrutura em causa é um objecto da classe `hclust` que tem, entre várias outras componentes, uma componente chamada `merge` que é uma matriz $(N-1) \times 2$. Cada linha i desta matriz descreve a junção que foi levada a cabo no passo i do processo de junção do algoritmo. Cada linha da matriz possui dois números representando os grupos que são juntos no passo em causa. Se algum desses números é negativo, e.g. $-k$, então isso significa que o respectivo grupo é formado unicamente pela observação k. Por sua vez, se algum dos 2 números é positivo, e.g. m, então significa que a junção é com o grupo que resultou da junção descrita na linha m da matriz `merge`. Usando a informação desta matriz `merge` é fácil implementar as ideias descritas pelas Equações 1 e 2, praticamente sem qualquer custo computacional adicional ao já incorrido no processo de agrupamento. Todavia, é altamente provável que este processo de agrupamento, levado a cabo pela função `hclust()` do R, contenha código que seja desnecessário para a obtenção dos graus de outlier que pretendemos. Assim, será provavelmente possível optimizar o processo computacional em causa. Isto pode ser conseguido uma vez que todo o código do R, e logo também da função `hclust()`, está disponível para todos os utilizadores. Este tipo de trabalho é difícil, senão impossível, em muitas ferramentas alternativas ao R, em que o utilizador está restringido a tentar ele próprio desenvolver o código dos outros, no caso de pretender realizar pequenas modificações ao mesmo. Isso não só é muito mais trabalhoso, como muitas vezes pode não ser fiável pois nem sempre os artigos que descrevem uma qualquer metodologia são (ou podem ser) exaustivos quanto a detalhes de implementação que muitas vezes se revelam cruciais em termos dos resultados obtidos. Neste contexto, o uso do R revelou-se crucial neste trabalho.

4 Alguns Resultados

Apresentamos em seguida alguns dos resultados obtidos neste caso de estudo. Por limitação de espaço não poderemos abordar todas as questões que foram consideradas no nosso trabalho sobre este problema. O leitor mais interessado pode colher mais informação noutros artigos publicados [8, 12, 14, 13].

Foram realizadas várias experiências destinadas a aferir a eficácia do método proposto no caso de estudo que descrevemos. Procurou-se não só verificar se o método era capaz de satisfazer os critérios operacionais do INE, $\%S \leq 50\% \wedge \%R \geq 90\%$, mas também comparar este método com outras alternativas existentes para obter graus de outlier. Relativamente às alternativas consideradas foi usado o método LOF [3], considerado um estado da arte em termos de obtenção de rankings de outliers, e disponível na *package* **dprep** [1] do R. Tratando-se ambos de métodos de ranking de outliers, foi decido usar 5 níveis de esforço de inspecção ($\%S$) pré-determinados: 30%, 35%, 40%, 45% e 50%, todos dentro das restrições impostas pelo INE. Para cada um destes níveis de esforço de inspecção foram comparados os resultados obtidos pela nossa proposta e pelo método LOF. Tal comparação foi efectuada para cada um dos 8 meses, uma vez que o INE determina que a inspecção seja feita mensalmente. Os resultados desta comparação são apresentados na Figura 1, que foi obtida usando os gráficos disponíveis na *package* **lattice** do R.

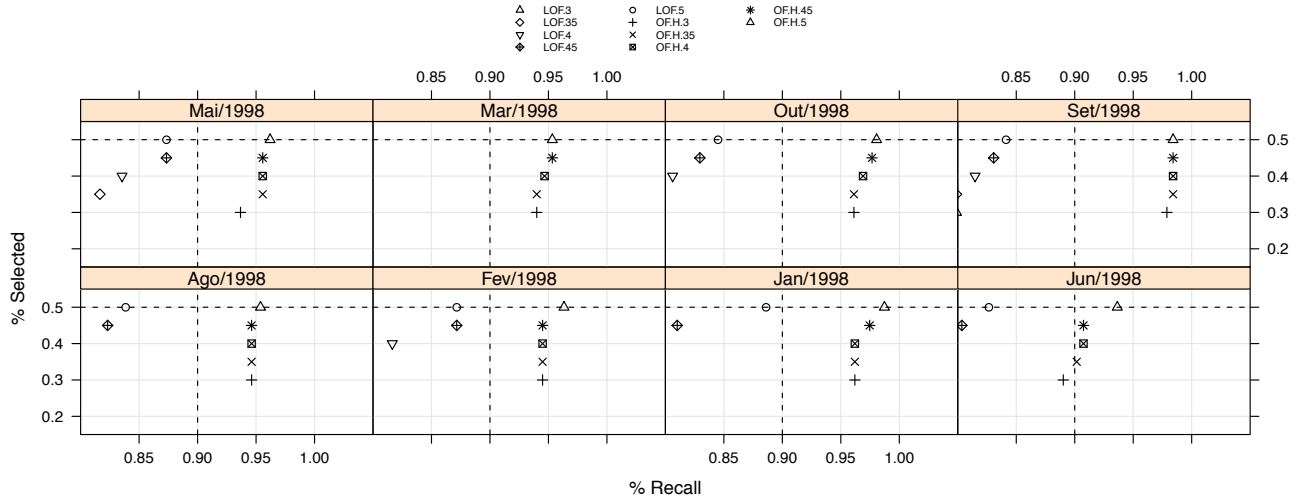


Figura 1: Método OF_H comparado com o método LOF.

Os gráficos mensais apresentam duas linhas a tracejado que delimitam a região onde deverão estar os resultados para satisfazerem os critérios do INE. Qualquer ponto fora do quadrante inferior direito está foram destes critérios. Conforme se poderá verificar, alguns pontos nem são visíveis nos gráficos por saírem fora da escala comum usada em todos eles, que permite uma mais fácil leitura comparativa dos resultados mensais. Isto só acontece para alguns *setups* do método LOF. A nossa proposta consegue "meter" quase todos os seus resultados nos critérios exigidos. Note-se também a superioridade clara do nosso método de ranking em relação ao LOF, neste problema em particular. Os resultados globais são notáveis uma vez que com um esforço de 30% é na maioria dos meses possível apanhar cerca de 95% dos erros detectados pelos peritos do INE, sendo a única excepção o mês de Junho.

5 Conclusões

Neste artigo descreveu-se uma abordagem de obtenção de rankings de outliers desenvolvida em R, tirando partido quer das potencialidades desta ferramenta, quer do facto de o seu código estar disponível gratuitamente, o que permite a criação de novos métodos trabalhando de forma incremental sobre ferramentas já existentes, como foi o caso.

O método desenvolvido foi aplicado a um caso de estudo concreto de detecção de erros/fraudes em registos de transacções de comércio externo das empresas Portuguesas. Esta é uma aplicação que requer grandes recursos de inspecção dado o volume de dados envolvido. Nestes contextos, e dada a usual limitação destes recursos, é de particular importância a utilização de métodos de ranking de outliers. O nosso método provou obter resultados bastante bons nesta aplicação particular, conseguindo ao mesmo tempo satisfazer os critérios operacionais do cliente e bater o estado da arte em ranking de outliers. O método desenvolvido é genérico e poderá ser aplicado a outros problemas de detecção de fraude em contextos de recursos limitados.

Relativamente a desenvolvimentos futuros, encontramo-nos neste momento a estudar e desenvolver formas alternativas de obter os rankings, nomeadamente usando critérios de ordenação mais flexíveis que possam ser ajustados mais facilmente aos objectivos operacionais dos utilizadores finais destas ferramentas.

Agradecimentos

Parte do trabalho descrito neste documento foi feito em colaboração com Carlos Soares (LIAAD) e Welma Pereira (LIAAD). Este trabalho enquadra-se no projecto oRANKI (PTDC/EIA/68322/2006), financiado pela FCT. Agradece-se ainda ao INE pelo fornecimento dos dados usados neste estudo.

Referências

[1] Edgar Acuna, , members of the CASTLE group at UPR-Mayaguez, and Puerto Rico. *dprep: Data preprocessing and visualization functions for classification*, 2008. R package version 2.0.

[2] V. Barnett and T. Lewis. *Outliers in statistical data, 3rd edition*. John Wiley, 1994.

[3] M. M. Breunig, H. P. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proceedings of ACM SIGMO 2000 International Conference on Management of Data*, 2000.

[4] D. M. Hawkins. *Identification of Outliers*. Chapman and Hall, 1980.

[5] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

[6] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[7] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of 24rd International Conference on Very Large Data Bases (VLDB 1998)*, pages 392–403. Morgan Kaufmann, San Francisco, CA, 1998.

[8] A. Loureiro, L. Torgo, and C. Soares. Outlier detection using clustering methods: a data cleaning application. In *Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector*, 2004.

[9] F. Murtagh. Multidimensional clustering algorithms. *COMPSTAT Lectures 4, Wuerzburg: Physica-Verlag*, 1985.

[10] R. Ng and J. Han. Efficient and efective clustering method for spatial data mining. In *Proc. of VLDB'94*, 1994.

[11] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[12] L. Torgo. Resource-bounded fraud detection. In Neves et. al, editor, *Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA'07)*, LNAI, pages 449–460. Springer, 2007.

[13] L. Torgo, W. Pereira, and C. Soares. Detecting errors in foreign trade transactions: dealing with insufficient data. In Lopes et. al, editor, *Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA'09)*, LNAI-5816. Springer, 2009.

[14] L. Torgo and C. Soares. *Data Mining for Business Applications*, chapter Resource-bounded outlier detection using clustering methods. IOS Press, (to appear in 2009).

Tutorial: Inferência bayesiana no R através do WinBUGS

Valeska Andreozzi, *valeska.andreozzi@fc.ul.pt*

Maria Antónia Amaral Turkman, *antonia.turkman@fc.ul.pt*

Faculdade de Ciências da Universidade de Lisboa
Centro de Estatística e Aplicações da Universidade de Lisboa

Introdução

Modelos estatísticos desenvolvidos para compreender os problemas do mundo real são cada vez mais complexos incluindo estruturas temporais, espaciais e hierárquicas, e muito frequentemente apresentam misturas de pelo menos duas dessas três estruturas. Por este motivo, a inferência bayesiana vem sendo cada vez mais utilizada, não só por estatísticos bayesianos, mas também por outros investigadores que trabalham em aplicações estatísticas. Este facto deve-se ao desenvolvimento tecnológico na década de 80 que permitiu que métodos de integração de Monte Carlo via cadeias de Markov (MCMC) pudessem ser utilizados para gerar amostras da distribuição a *posteriori* para os parâmetros do modelo, permitindo realizar inferências adequadas. BayesX, WinBUGS e R são alguns exemplos de softwares gratuitos que estimam modelos simples e complexos através de uma abordagem bayesiana. Todos os três softwares estão disponíveis para o ambiente Windows com uma documentação razoável e com exemplos incluídos.

O BayesX (Belitz *et al*. 2009) foi desenvolvido para estimar modelos aditivos generalizados mistos englobando diversos tipos de modelos complexos de regressão. Funciona através de funções pré-definidas que são executadas sob forma de "linha de comando".

O WinBUGS (Lunn *et al.* 2000), que pode ser considerado o software mais amplamente utilizado no meio científico, difere do primeiro, pois tem a vantagem de permitir ao utilizador a estimação de modelos por ele desenvolvido. Contudo, o utilizador precisa de escrever o respectivo código do programa definindo o modelo probabilístico para a função de verosimilhança e também para a distribuição a *priori* dos parâmetros, requerendo assim um conhecimento da linguagem que não é assim tão simples. Um ponto de partida para os iniciantes está nos códigos dos programas para os modelos bem conhecidos da literatura que podem ser encontrados nos exemplos do próprio WinBUGS, como também em diversos livros/artigos e na web.

O R (R Development Core Team 2009) não é um software específico para estimar modelos através de uma abordagem bayesiana. Não obstante, vêm sendo cada vez mais encontrados na sua página da internet uma grande quantidade de bibliotecas que permitem efectuar inferências bayesianas e que podem ser utilizadas nas mais diversas áreas e para os mais específicos modelos/métodos. Alguns exemplos são as bibliotecas para estimar modelos na área do Marketing e Micro-econometria (*bayesm*), análise de sobrevivência (*bayesSurv*) e Ciências Socias (*MCMCpack*). Outras bibliotecas mais específicas incluem métodos para modelos de valores extremos (*evdbayes*) e CGH micro-arranjos (*RJaCGH*). Uma lista completa das bibliotecas que utilizam inferência bayesiana pode ser encontrada em Park 2009. Contudo, no desenvolvimento pleno da modelação bayesiana, o R deve ser encarado

como uma potente ferramenta adicional na etapa de ligação com outros softwares específicos para inferência bayesiana e na etapa de diagnóstico das amostras simuladas das distribuições a *posteriori*. As bibliotecas *coda* (*Convergence Diagnosis and Output Analysis*) e *boa* (*Bayesian Output Analysis*) são as mais utilizadas na fase pós estimação dos modelos e a biblioteca *R2WinBUGS* para a ligação com o WinBUGS.

Este artigo propõe-se apresentar o potencial do R na estimação de modelos estatísticos com inferência bayesiana através do WinBUGS ilustrando a facilidade com que esta ligação oferece para a entrada de dados, que através do R se torna muito mais simples, e para a etapa do diagnóstico. Pressupõe-se que o leitor deste artigo tenha familiaridade tanto com o WinBUGS como com o R e o que vamos ensinar é como podemos tornar a vida mais simples através da ligação de ambos.

Preliminares

Para acompanhar este artigo como um tutorial, tenha em seu computador o R e o WinBUGS instalados. Será também necessário instalar as seguintes bibliotecas do R: *R2WinBUGS*, *boa* e *coda*. Os utilizadores do Windows Vista devem ter atenção às restrições de escrita impostas no directório *"C:\Program Files"*. Para facilitar a execução deste tutorial, sugere-se a instalação do WinBUGS em outro diretório que não seja o "*Program Files"*, por exemplo, diretamente no *"C:\"*, ou unidade equivalente.

Motivação

Foi seleccionado um exemplo bastante didáctico para que a utilização do R pudesse ser explorada ao máximo. Os dados que são apresentados são relativos à percentagem total de calorias resultantes de carbohidratos complexos, ingeridas por 20 indivíduos diabéticos do sexo masculino, nos quais se testou um novo regime alimentar. Como se pode verificar também se registou a idade (x_1) e o peso dos indivíduos (x_2), assim como a percentagem de calorias resultantes das proteínas presentes na dieta (x_3) .

Tabela 1. Primeiras 6 observações dos dados que se encontra no ficheiro *dadosexemplo.dat*

y	x1	x2	x3
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15

Pensa-se que a percentagem de carbohidratos (Y) está relacionada linearmente com as variáveis X_1, X_2 e X_3. De forma a verificar esta afirmação, foi elaborado um programa no WinBUGS, considerando que $Y_i \sim \text{Normal}(\mu,\sigma^2)$, $i = 1, 2, \ldots, 20$ e sem qualquer conhecimento a *priori* acerca dos parâmetros do modelo como mostra o código do programa no Quadro 1.

Tutorial

Os passos necessários para estimar o modelo bayesiano no R através do WinBUGS são:

a) Escrever o código do programa do modelo estatístico no WinBUGS
   Para escrever o código do programa utilize o próprio WinBUGS. Aproveite para, ainda no WinbBUGS, verificar a sintaxe do seu programa no menu *MODEL>Specification>Check model*. Depois salve o ficheiro com extensão *.bug*
   No nosso exemplo o modelo probabilístico está no ficheiro texto *modeloexemplo.bug*

b) Escrever o *script* do R para:
   i. ler os dados
   ii. chamar o Winbugs
   iii. guardar os resultados das simulações das distribuições a *posteriori*
   iv. carregar as bibliotecas *boa* ou *coda* para diagnóstico das cadeias

Quadro 1. Código do programa do modelo estatístico em WinBUGS (*modeloexemplo.bug)*.

```
model{
  for (i in 1:N){
    y[i] ~ dnorm (mu[i], tau)

    #Modelo
    #As covariáveis foram centralizadas
    mu[i] <- beta[1] + beta[2]*(x1[i]-mean(x1[])) +
    beta[3]*(x2[i]-mean(x2[])) + beta[4]*(x3[i]-mean(x3[]))

    #Resíduo padronizado
    r[i]<-(y[i]-mu[i])*sqrt(tau)

    #Predição
    z[i]~dnorm(mu[i],tau)

    #Resíduo absoluto para calcular p-valor
    #da medida de discrepância T
    absry[i]<-abs(y[i]-mu[i])
    absrz[i]<-abs(z[i]-mu[i])

    #O objeto inv será utilizado para o cálculo
    inv[i]<-1/(sqrt(tau)*exp(-tau/2*(y[i]-mu[i])*(y[i]-mu[i])))
  }

  #Prioris
  tau ~ dgamma(0.0001,0.0001)
  for (j in 1:4){
    beta[j]~ dnorm(0,0.0001)
  }

  #Cálculo do intercepto na escala original
  alpha<- beta[1] - beta[2]*(mean(x1[])) -
  beta[3]*(mean(x2[])) - beta[4]*(mean(x3[]))

  sigma2<-1/tau

  #Estatística T
  ty<-sum(absry[])
  tz<-sum(absrz[])
  pvalor<-step(tz-ty)
}
```

A seguir encontra-se o tutorial com os comandos do R.

1. Início

Antes de começar não se esqueça de alterar o diretório de trabalho, isto é, suponha que o ficheiro dos dados encontra-se no diretório *C:\exemplo*

```
> setwd("C:\\exemplo")
```

Carregando a biblioteca

```
> library(R2WinBUGS)
```

2. Leitura dos dados

```
> dados <- read.table("dadosexemplo.dat",header=T)
```

Criação de objectos separados para cada variável

```
> N <- nrow(dados) #número de indivíduos no banco de dados
> y <- dados$y     #variável resposta percentagem de carbohidratos
> x1 <- dados$x1   #covariável idade
> x2 <- dados$x2   #covariável peso em kg
> x3 <- dados$x3   #covariável percentagem de calorias das proteínas da dieta
```

Criação de uma lista com os dados que serão fornecidos ao programa do WinBUGS.

```
> data <- list("N", "y", "x1", "x2", "x3")
```

3. Declaração dos valores iniciais dos parâmetros

```
> inits <- list(list(tau = 1, beta = c(0,0,0,0),z=y))
```

uma alternativa é permitir que o R gere aleatoriamente os parâmetros da seguinte forma:

```
 #inits <- function()
  #   {list(tau = rgamma(1,shape=1000,scale=1/1000),
  #   beta = rnorm(4,0.001,sqrt(1/0.001)),z=y)}
```

Atenção à parametrização da distribuição Normal no R que difere da do Winbugs.
Temos para o R os parâmetros μ e σ e no WinBUGS μ e $\tau=1/\sigma^2$ (veja o help da função **`Normal`**)

4. Declaração dos parâmetros a serem monitorizados

```
> parameters <- c("tau","beta","alpha","mu","inv","pvalor")
```

5. Executando o WinBUGS

Listar o código do modelo dentro do R. O ficheiro com o código do modelo encontra-se no diretório de trabalho *C:/exemplo*

```
> file.show("modeloexemplo.bug")
```

Agora já temos todos os elementos para executar o WinBUGS através do R utilizando a função **`bugs()`**

```
> result <- bugs (data=data, inits=inits, parameters.to.save=parameters,
  + model.file="modeloexemplo.bug", n.chains=1, n.iter=10000,n.burnin=2000,
  + bugs.directory="C:\\WinBUGS14",
  + debug=FALSE,save.history=FALSE,DIC=TRUE)
```

Para maiores detalhes de todos os argumentos da função **`bugs()`** pesquise o *help file* (**`?bugs`**) ou o guia (*vignette*) da biblioteca através do comando **`vignette(R2WinBUGS)`**. Um rápido resumo dos principais argumentos encontra-se a seguir:
`data=` objecto que contém os dados
`inits=` objecto que contém os valores iniciais dos parâmetros
`parameters.to.save=` objeto que contém os nomes dos parâmetros a serem monitorizados
`model.file=` ficheiro que contém o modelo do WinBUGS
`n.chains=` número de cadeias
`n.iter=` número total de iterações por cadeia incluindo a fase de aquecimento
`n.burnin=` número de iterações que serão descartadas na fase de aquecimento
`bugs.directory=` directório onde se encontra instalado o WinBUGS

debug= o valor **FALSE** indica que o WinBUGS será fechado automaticamente após o término da estimação do modelo
save.history= se **TRUE** gráficos do traço de cada parâmetro será gerado no WinBUGS. É aconselhável não exigir que o WinBUGS faça esses gráficos, pois como em geral têm-se muito parâmetros, às vezes o WinBUGS pode deixar de responder. Além disso os traços poderão ser todos feitos posteriormente no R
DIC= se **TRUE** então deviance, pD e DIC são calculados.

Ao executar a função **bugs()** no R o WinBUGS será automaticamente aberto para simular amostras das distribuições a *posteriori* dos parâmetros. Se o argumento **debug=FALSE** o WinBUGS será fechado também automaticamente, retornando ao R e guardando os resultados no objecto que recebeu a função **bugs()**, que neste exemplo se chama **result.**

6. Resultado

Sumário dos parâmetros que foram declarados no vector **parameters** no passo 4. Este sumário é muito semelhante ao que se obtém no WinBUGS

```
> result$summary
```

Diversos outros elementos estão presentes no objeto **result.** Por exemplo, as simulações podem ser encontradas em diversos formatos (matriz (**result$sims.matrix**) , lista (**result$sims.list**), *array* (**result$sims.array**))

```
> names(result)
```

Lista dos nomes de todos os parâmetros que foram monitorizados

```
> dimnames(result$sims.matrix)
```

Para futuras comparações com outros modelos, pode obter-se o valor de DIC, BIC e do CPO (*condicional predictive ordinate*). O valor de DIC encontra-se no objeto **result**. BIC e CPO têm que ser calculados.

```
> #DIC
> dic <- result$DIC
> dic
> #cálculo do BIC
> bic <- unlist(result$mean["deviance"]) + 4*log(N)
> #4 é o número de parâmetros do modelo
> bic
> #cálculo do CPO
> #vector com os nomes dos parâmetros
> nomesinv <- paste("inv[",1:N,"]",sep="")
> #amostra dos parâmetros seleccionados
> simulainv <- result$sims.matrix[,nomesinv]
> cpo <- 1/apply(simulainv,2,mean)
```

Um gráfico do CPO versus o índice da observação (dados) também pode ser útil quando for necessário comparar modelos

```
> plot(1:N,cpo,type="h",xlab="observation index",ylab="CPO")
```

E uma medida global do CPO pode ser calculada através de:

```
> fitmed<-sum(log(cpo))
> fitmed
```

A partir daqui inicia-se a outra etapa da estimação do modelo, onde o R possui um papel essencial, pois as simulações de cada parâmetro estão automaticamente disponíveis para serem diagnosticadas.

7. Diagnóstico

7.1 Utilizando a biblioteca BOA

Em geral o número de parâmetros é sempre muito grande. Podemos guardar um objecto que inclui somente alguns parâmetros de interesse para fazer o diagnóstico

```
> dimnames(result$sims.matrix)
> #gerando um vector com os nomes dos parâmetros a serem diagnosticados
> nomespar <- paste("beta[",1:4,"]",sep="")
> nomespar
> #criando um objecto do tipo matriz com as simulações dos parâmetros escolhidos
> simulashort <- result$sims.matrix[,nomespar]
> #carregando a biblioteca
> library(boa)
> #carregando o menu
> boa.menu()
```

A biblioteca *boa* funciona com um menu incorporado. Para carregar o objecto **simulashort** que contém as simulações siga os seguintes passos: a) No *BOA MAIN MENU*, escolha opção *1:File;* b) No *FILE MENU*, escolha opção *3:Import Data;* c) No *IMPORT DATA MENU,* escolha opção *5:Data Matrix Object;* d) E entre com o nome do objecto que contém as simulações (**simulashort**).
Agora as amostras das distribuições a *posteriori* dos parâmetros estimados no WinBUGS estão disponíveis para a biblioteca *boa* e basta percorrer os menus para fazer o diagnóstico.

7.2 Utilizando a biblioteca CODA

```
> #carregando a biblioteca
> library(coda)
```

Caso tenha optado por não guardar os resultados das simulações no formato do WinBUGS, deixando o argumento **codaPkg=FALSE**, temos antes que transformar o objeto que recebeu a função **bugs()** da seguinte forma:

```
> resultcoda <- as.mcmc.list(result)
```

Também a biblioteca *coda* possui um menu

```
> codamenu()
```

Para ler os dados com as simulações basta, no menu principal, selecionar opção *2: Use a mcmc object*. Logo a seguir escreve-se o nome do objecto, que no nosso caso é o ***resultcoda***, e pronto, as amostras estão disponíveis para o diagnóstico.

Ao final deste tutorial espera-se que o leitor que o tenha experimentado fique com a impressão de que vale a pena executar o WinBUGS através do R. Desta forma acaba-se com aquela preocupação de ter que formatar os dados para o WinBUGS, guardar o ficheiro com as simulações das cadeias e depois ter que ler cada ficheiro no R para fazer o diagnóstico e outras tarefas para apresentação de resultados como gráficos e mapas. Outras situações em que essa ligação é muito útil surgem em estudos de simulações e no estudo de adequabilidade dos modelos usando jacknife. É claro que nem tudo são flores. Uma desvantagem dessa ligação deve-se ao facto da função **bugs()** guardar as simulações em diversos formatos (matrix, lista, array) e isso pode gerar um problema de memória para o R quando o número de parâmetros é excessivamente grande. Mas quando esse não é o problema, podemos tirar muito proveito desta ligação e se juntarmos as características do *Sweave* (Leisch 2002), que gera um relatório em Latex com as saídas do R automaticamente, temos o pacote completo para facilitar nosso trabalho e a apresentação dos resultados.

Todo o material deste tutorial, incluindo os ficheiros com os dados e o modelo e um pdf com os outputs do R, está disponível em *www.curso-r.wikidot.com/r2winbugs*.

Outros exemplos também podem ser encontrados na *vignette* da biblioteca *R2WinBUGS*, que pode ser visualizada em formato pdf através do seguinte comando no R:

```
> vignette("R2WinBUGS")
```

Referências

Belitz C, Brezger A, Kneib T, Lang S. (2009): BayesX - Software for Bayesian inference in structured additive regression models. Version 2.00 (6.5.2009) . Available from *http://www.stat.uni-muenchen.de/~bayesx*.

Leisch F, 2002. Sweave user manual. *http://www.ci.tuwien.ac.at/~leisch/Sweave/.*

Lunn DJ, Thomas A, Best N, and Spiegelhalter D. (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325-337. Available from *http://www.mrc-bsu.cam.ac.uk/bugs/*.

Park JH, 2009. CRAN Task View: Bayesian Inference.
Available from *http://cran.r-project.org/web/views/Bayesian.html*

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL *http://www.R-project.org*.

Bibliografia

Albert J, 2007. Bayesian Computational with R. Capítulo 11. Springer

Plummer M, Best N, Cowles K , Vines K, 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. Rnews, 6(1):7-11. *http://CRAN.R-project.org/doc/Rnews/*

Smith B, BOA (Bayesian Output Analysis) *http://www.public-health.uiowa.edu/boa/*

Sturtz S, Ligges U, Gelman A, 2005. R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software, 12(3), 1-16.

Ciência Estatística

• Artigos Científicos Publicados

Caeiro, F., Gomes, M.I. and Henriques Rodrigues, L. (2009). Reduced-bias tail index estimators under a third order framework. *Communications in Statistics - Theory and Methods*, 38(7), 1019-1040.

Caeiro, F. and Gomes, M.I. (2009). Semi-parametric second-order reduced-bias high quantile estimation. *Test*, 18(2), 392-413.

Caiado, J., N. Crato and D. Peña (2009) - Comparison of time series with unequal length in the frequency domain**,** *Communications in Statistics: Simulation and Computation*, 38, 527-40.

Gomes, M.I., Pestana, D. and Caeiro F. (2009). A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator. *Statistics and Probability Letters*, 79(3), 295-303.

Menezes, R. and J.Tawn (2009) - Assessing the effect of biased and clustered sampling on variogram estimation. *Environmetrics,* vol 20, issue 4, 445-459.

Valente, V. and T. Oliveira (2009) - Hierarchical Linear Models in Education Sciences: An Application. *Biometrical Letters* vol. 46(1), 71-86.

• Teses de Mestrado

Título: *A Satisfação no trabalho em Portugal: Uma análise longitudinal com recurso a Latent Growth Curve Models*
Autora: Ana Lúcia Teixeira Dias, *analuciadias@fcsh.unl.pt*
Orientadora: Maria de Fátima Salgueiro

Título: *Costumer Lifetime Value na Banca*
Autora: Ana Margarida Gomes Alexandre, *ana.margarida.alexandre@gmail.com*
Orientadores: Anabela Costa e Elson Filho

Título: *A situação económica e social na União Europeia. Análise de alguns indicadores*
Autora: Patrícia Pereira, *patricia_arp02@hotmail.com*
Orientadora: Manuela Magalhães Hill

Título: *Modelação longitudinal do bem-estar psicológico com modelos de trajectória latente*
Autora: Joana Malta Correia Guedes, *joanavmalta@gmail.com*
Orientadora: Maria de Fátima Salgueiro

Título: *Análise de Dados Longitudinais Discretos: uma Aplicação ao Estudo da Influência de Lípidos no Adenocarcinoma Mamário*
Autora: Eunice Isabel Ganhão Carrasquinha Trigueirão, *nicecarrasquinha@hotmail.com*
Orientadora: Salomé Cabral

Título: *Risco de Exposição Humana aos Contaminantes na Alimentação: o cádmio e o chumbo no peixe-espada preto*
Autora: Inês Alves Farias, *farias.ines@gmail.com*
Orientadora: Maria Isabel Fraga Alves

• Livros

Título: A Linguagem R, programação para a análise de dados
Autor: Luís Torgo
Ano: 2009. Editora: Escolar Editora. ISBN: 978-972-592-246-0

Título: COMPSTAT 2008 – Proceedings in Computational Statistics (com CD-ROM)
Edição: Paula Brito
Ano: 2008. Editora: Physica-Verlag. ISBN: 978-3-7908-2083-6

Título: Abordagem Estatística de Conjuntos Difusos
Autor: Abdul Suleman
Ano: 2009. Editora: Sílabo. ISBN: 978-972-618-544-4

Título: Estatística Descritiva e Probabilidades. Problemas resolvidos e propostos com aplicações em R
Autores: Fernanda Figueiredo, Adelaide Figueiredo, Alexandra Ramos, Paulo Teles
Ano: 2009 (2ª Edição). Editora: Escolar Editora. ISBN: 978-972-592-249-1

Título: Um mundo para conhecer os números
Autores: M. J. Ferreira, I. Tavares, P. Campos, L. Loura, M. E. G. Martins, A. A. da Silva, R. Sousa
Ano: 2009. Editora: INE. ISBN: 978-98925-0043-0

Título: Análise de Sobrevivência
Autores: Cristina Rocha e Ana Luísa Papoila
Ano: 2009. Edições SPE. ISBN: 978-972-8890-22-3

Título: Estatística. Arte de Explicar o Acaso
Editores: Irene Oliveira, Elisete Correia, Fátima Ferreira, Sandra Dias e Carlos Braumann
Ano: 2009. Edições SPE. ISBN: 978-972-8890-20-9

• Teses de Doutoramento

Título: *META-ANÁLISE – Harmonização de testes usando os valores de prova*
Autor: Fernando José Araújo Correia da Ponte Sequeira, *fjsequeira@fc.ul.pt*
Orientador: Dinis Duarte Ferreira Pestana

Na minha tese, apresento uma breve resenha de tópicos de Meta-Análise que mostram a importância dessa subdisciplina da Estatística na construcção do conhecimento científico, viabilizando sínteses dos factos e conclusões conhecidas, e explora-se com algum detalhe o problema das sínteses usando níveis de significância descritivos.

Apresentamos uma técnica artificiosa de calcular pseudo p-values, ampliando computacionalmente a amostra, e estudamos as implicações desses procedimentos na potência dos testes meta-analíticos usuais, de Fisher e de Tippett; usando os métodos de Stouffer, averiguamos o impacto do uso da amostra ampliada no número de estudos não significativos que seriam necessários para virar a decisão de rejeição global, uma das formas mais usadas de avaliar o efeito do enviesamento na publicação nas revisões sistemáticas e sínteses meta-analíticas.

Estabelecem-se ainda resultados sobre caracterização de uniformes, e estuda-se a distribuição exacta de funções de betas potencialmente úteis em eventuais extensões usando espaçamentos.

O objectivo fundamental é a ampliação do vector de valores de prova (em certas experiênçias de número muito reduzido) de forma a que se reforçe a probabilidade de aceitação de H_0 no caso desta ser verdadeira e o mesmo suceda na rejeição no caso de ser falsa. Os p-values "artificiais" são gerados a partir dos originais mas são independentes dos mesmos.

Fernando Sequeira

Título: *SETAR Nonlinearity, Nonstationarity and Forecasting*
Autor: Pedro Goubeia, *pgouveia@ualg.pt*
Orientador: Paulo M.M. Rodrigues

Na minha tese são realizados diversos desenvolvimentos no âmbito dos modelos SETAR (Self-Exciting Threshold Autorregressive).
Numa primeira fase, em contexto não sazonal, são derivados testes de raízes unitárias que têm por base o princípio do Multiplicador de Lagrange e obtidas as suas distribuições. Estes testes, de acordo com o estudo de Monte Carlo realizado, têm bom desempenho e não apresentam distorção do nível face a quebras na média, contrariamente ao que ocorre com a versão não-linear do teste DF proposta na literatura.

Também ao nível dos testes de raízes unitárias, são realizados alguns desenvolvimentos que têm por base a aplicação do método dos mínimos quadrados generalizados e a estimação recursiva das componentes determinísticas. Ainda em contexto não sazonal, são derivados os limites assimptóticos dos testes propostos na região de quase não estacionaridade.

Por outro lado, boa parte do interesse do estudo de séries económicas com sazonalidade está associado à presença de raízes unitárias nas frequências zero e sazonais. Nesta tese procede-se à derivação de testes de raízes unitárias em modelos SETAR sazonais. Estes testes apresentam a vantagem de permitir testar a eventual interferência do ciclo económico nos padrões de sazonalidade.

Finalmente, na componente empírica da tese é desenvolvido um estudo de previsão que tem por base a combinação de diversos modelos lineares e não-lineares e a aplicação de diferentes filtros à variável dependente de forma a ter ou não em conta os efeitos da sazonalidade e não-estacionaridade. Este estudo de previsão procura ainda ser pioneiro na utilização de modelos sazonais em metodologias de combinação de previsões. Esta componente aplicada tem por base o princípio segundo o qual diferentes modelos apresentam complementaridades na aproximação ao Processo Gerador de Dados.

Pedro Gouveia

Título: *Modelação Estatística com Misturas e Pseudo-Misturas*
Autor: Miguel Martins Felgueiras, *mfelg@estg.ipleiria.pt*
Orientador: Dinis Duarte Pestana

Na minha tese procurei estudar diferentes tipos de misturas de distribuições, que por permitirem uma miríade de combinações de achatamento, assimetria e multimodalidade, são extremamente eficazes na análise de dados.

Comecei por trabalhar com as misturas finitas e convexas mais habituais (com aplicações nas mais diversas áreas do conhecimento), que surgem quando um determinado atributo é observado numa população com várias subpopulações, não sendo possível classificar os elementos da amostra nestas. Para misturas unimodais, apresentei alguns resultados assintóticos, que poderão ser úteis em várias situações práticas. Questões de parcimónia foram igualmente analisadas neste contexto. Em misturas de gaussianas, as aproximações obtidas permitem testar a igualdade das médias e a igualdade das variâncias.

Através de uma generalização da teoria clássica de extremos (permitindo estabilidade para transformações de forma) explorei ainda um novo tipo de misturas, finitas mas não convexas, que permitem que as suas componentes tenham pesos negativos e pesos superiores a 1. Estas misturas são extremamente flexíveis, podendo ser uma séria alternativa na modelação, por exemplo, de tráfego na *internet*.

Finalmente, analisei misturas infinitas com parâmetro de escala Pareto. Ao aleatorizar o parâmetro de escala, consegui modelos baseados no original mas de caudas mais pesadas, e que generalizam as distribuições divididas usuais, com diversas aplicações em estudos de robustez. Devido à densidade polinomial da distribuição Pareto, várias densidades explícitas destas misturas foram obtidas.

Miguel Felgueiras

Título: *Métodos Robustos em Geoestatística*
Autor: Hilário Amílcar dos Santos Ribeiro Miranda, *hmiranda@ua.pt*
Orientadora: Maria Manuela Souto de Miranda

Na minha tese faz-se uma revisão dos métodos de estimação usuais em Geoestatística e propõe-se um estimador robusto do variograma, com boas propriedades de eficiência sob modelos Normais.

Como os métodos de estimação do variograma existentes ou não são robustos, ou têm pouca eficiência em modelos Normais, no trabalho apresenta-se um novo estimador do variograma, que se designou por estimador de múltiplos variogramas. Resumidamente, o novo método consiste em quatro etapas, nas quais prevalecem, alternadamente, critérios de robustez ou de eficiência. Com a amostra inicial e por questões de eficiência, são calculadas, de forma robusta, tantas estimativas pontuais do variograma quantos os parâmetros do modelo; com base nessas estimativas, os parâmetros são estimados pelo método dos mínimos quadrados; as duas fases anteriores são repetidas um elevado número de vezes, criando um conjunto de múltiplas estimativas da função variograma; por fim, a estimativa final do variograma é definida pelas medianas das estimativas dos parâmetros obtidas anteriormente. Assim se obtém um estimador robusto e com boa eficiência em processos Gaussianos.

A investigação desenvolvida revelou que, ao usar estimativas discretas na primeira fase da estimação do variograma, existem situações onde a identificabilidade dos parâmetros não está assegurada. Para os modelos mais comuns, foi possível estabelecer condições que garantem a unicidade de solução na estimação do variograma.

A estimação do variograma supõe sempre a estacionaridade da média do processo. Como não são conhecidos procedimentos objectivos para avaliar tal condição, no trabalho sugere-se um teste para validar essa hipótese. A estatística do teste é um estimador-MM, cuja distribuição é desconhecida nas condições assumidas. Tendo em vista a sua aproximação, propõe-se uma versão do método *bootstrap* adequada ao estudo de observações de processos espaciais, a qual preserva a estrutura de dependência do processo.

Finalmente, o estimador de múltiplos variogramas é avaliado em termos da sua aplicação prática. O trabalho contém um estudo com dados reais e outro de simulação, os quais confirmam as propriedades estabelecidas. Em todos os casos analisados, o estimador proposto produziu melhores resultados do que as alternativas usuais, tanto para a distribuição assumida, como para distribuições contaminadas.

Hilário Miranda

Título: *Extremos em séries temporais max-autorregressivas*
Autora: Marta Ferreira, *msferreira@math.uminho.pt*
Orientadora: Luísa Canto e Castro

A minha tese tem como motivação inicial a continuação do estudo do comportamento extremal de níveis que persistem por um período de tempo fixo, introduzido em Draisma (2001). De uma análise sob o pressuposto de que esses níveis constituem uma série de observações i.i.d., rapidamente se passa à hipótese mais realista de considerar dependência entre as observações, embora mantendo a estacionaridade. Uma vez que é sempre uma modelação dos valores extremos que está em mente, muito naturalmente se pensa nos modelos autorregressivos de máximos, como os MARMA (Davis e Resnick, 1989), em particular, os MARMA(1,0), também designados ARMAX (Alpuim 1989a, 1989b e Canto e Castro 1992). Na sequência do interesse em contemplar situações de dependência, surge a questão de avaliar se existe uma dependência ou independência exactas entre observações consecutivas consideradas nas caudas, ou se é uma dependência que vai desaparecendo gradualmente. Ledford e Tawn (1996) introduzem um modelo, no qual surge um novo parâmetro que permite ``medir o grau'' de dependência na cauda, designado *coeficiente de dependência assintótica na cauda* ou *coeficiente de Ledford e Tawn*. É no decurso do cálculo do valor deste coeficiente para os usuais max-autorregressivos, que surge a construção do processo pARMAX, o qual inclui um parâmetro potência, que faz com que o coeficiente de Ledford e Tawn se relacione directamente com esse mesmo parâmetro.

De modo a atenuar o carácter um tanto determinístico do processo pARMAX e, assim, torná-lo mais aplicável na modelação de dados reais, considera-se uma generalização do mesmo, com a introdução de um factor aleatório.

Surge assim um novo processo max-autorregressivo potência, que designamos pRARMAX, o qual mantém a particularidade do seu parâmetro potência se relacionar com o coeficiente de Ledford e Tawn, de modo análogo ao do processo pARMAX. Aproveitando a maleabilidade permitida num processo pRARMAX, desenvolve-se uma metodologia de análise do seu ajustamento a uma série de dados.

Marta Ferreira

Título: *Feira dos Momentos – Planeamento Experimental e Investigação de Localização e Escala em Populações não Gaussianas*
Autor: João Paulo Oliveira Martins, *jpmartins@estg.ipleiria.pt*
Orientadores: Dinis Duarte Ferreira Pestana e Sandra Maria Freitas Mendonça

Na minha tese os desenvolvimentos clássicos relativos a planeamentos discriminantes óptimos e planeamentos robustos óptimos serviram de inspiração para a definição de planeamentos mistos óptimos, que têm em conta quer a estimação do grau da regressão bem como dos seus coeficientes, sendo quase óptimos no que se refere aos critérios discriminante e robusto. O caso dos planeamentos mistos óptimos até grau 4 é caracterizado detalhadamente, e a investigação computacional mostra que a perda de eficiência comparativamente com os planeamentos discriminantes óptimos e robustos óptimos é inferior a 2%, enquanto a perda de eficiência dos planeamentos robustos óptimos comparada com os planeamentos discriminantes óptimos, ou vice-versa, pode atingir os 15%. A teoria dos momentos canónicos serve de suporte à apresentação dos resultados relativos a planeamentos óptimos. Discutem-se as truncaturas da série de Taylor sugeridas pela aplicação do método delta. São apresentadas algumas extensões e aplica-se o método à avaliação da variância da soma de n variáveis aleatórias eventualmente correlacionadas, um problema com aplicações ao nível da Química. Ainda neste contexto, são consideradas as transformações estabilizadoras da variância apresentando-se uma extensão da definição para variáveis aleatórias univariadas com n parâmetros desconhecidos. O caso multivariado é também abordado e são dadas algumas pistas para compreender as condições para a existência ou não dessas transformações. O método delta é também usado para estimar os primeiros quatro momentos da estatística de Student T_n. Mostra-se que o comportamento da distribuição de T_n é próximo do comportamento de uma distribuição do tipo IV do sistema de Pearson. Discute-se o papel da assimetria na atracção e repulsão da média amostral e variância amostral. A escolha do título pretende transmitir desde logo que, ainda que haja uma linha condutora, os resultados assimptóticos e os métodos utilizados para os deduzir recorrem a um variado espectro de valores esperados e funções de valores esperados, que com alguma latitude de linguagem podemos apodar de momentos e de transformações integrais.

João Paulo Martins

Título: *Concepção de um Modelo Multicritério de Suporte à Avaliação de Agências Bancárias*
Autor: Fernando Alberto Freitas Ferreira, *fernando.ferreira@esg.ipsantarem.pt*
Orientadores: Sérgio Pereira dos Santos e Paulo Manuel Marques Rodrigues

A minha tese insere-se no domínio MCDA – *Multiple Criteria Decision Analysis* – e propõe a concepção de um modelo multicritério de suporte à avaliação de agências bancárias com base num processo que conjuga mapas cognitivos com a técnica MACBETH – *Measuring Attractiveness by a Categorical Based Evaluation Technique*. Após analisar as tendências evolutivas do sector bancário em Portugal (e das respectivas unidades tradicionais de retalho), bem como os principais métodos de avaliação utilizados, foi possível conceber um modelo de avaliação de agências que, assente numa análise integrada das vertentes *potencial* e *efectiva*, permite distinguir (global e localmente) diferentes

agências segundo múltiplos critérios. Este modelo, ao qual se atribuiu a designação *M-M_4BE – Multicriteria Model for Branch Evaluation*, visa mensurar a performance de agências bancárias através da conjugação de variáveis quantificáveis (*i.e.* aspectos objectivos) com variáveis oriundas da esfera intangível das agências (*i.e.* aspectos subjectivos). Além disso, assume uma base complementar, por nutrir-se das mais-valias dos métodos genéricos analisados. Naturalmente, para que esta concepção fosse possível, outros objectivos intermédios tiveram de ser atingidos, como por exemplo: (a) identificar e validar critérios de avaliação relevantes, segundo os juízos dos decisores; (b) estruturar hierarquicamente esses critérios com base em processos metodológicos adequados; (c) construir escalas numéricas necessárias à quantificação dos critérios e ponderá-los segundo as apreciações semânticas dos decisores; (d) aplicar o modelo junto de uma amostra de agências, revelando os seus perfis de desempenho e (e) realizar análises de sensibilidade e robustez. Como resultado, o ensaio desenvolvido revelou ser útil como base de reflexão para a definição e implementação de políticas de desenvolvimento que, uma vez amadurecidas pela aquisição de novos conhecimentos, proporcionem melhorias na performance das agências avaliadas. Na prática, o estudo corrobora, e como tal fortalece, os contributos de outros investigadores relativamente às vantagens de utilizar técnicas de mapeamento cognitivo e metodologias multicritério de apoio à decisão, quer individualmente quer de forma integrada, para apoiar a concepção e implementação de sistemas de avaliação do desempenho. De resto, para além do ensaio experimental do qual resultou o M-M_4BE, deve testemunhar-se a mais-valia obtida com o trabalho efectuado junto de profissionais da Banca em Portugal.

Fernando Ferreira

Título: *Extremum Estimators and Stochastic Optimization Methods*
Autor: Miguel de Carvalho, *mb.carvalho@fct.unl.pt*
Orientadores: João Tiago Mexia e Manuel L. Esquível

A minha tese incide sobre estimadores extremais (*extremum estimators*). Estes métodos unificam uma ampla classe de estimadores, que podem ser formulados através da solução de um problema de optimização. O método dos mínimos quadrados, o método generalizado dos momentos, bem como os métodos de máxima verosimilhança resultam da solução de um problema de optimização, sendo consequentemente especificações particulares de estimatores extremais. Um problema relevante no cálculo de estimativas deste tipo, está relacionado com as propriedades de convergência do método utilizado para obter a solução óptima. Com efeito, se o método utilizado convergir, eventualmente, para uma solução local, deixam de ser garantidas a consistência e a normalidade assimptótica do estimador extremal.

Esta tese contribui para o estado da arte através da introdução de um método de pesquisa estocástica, com vista à obtenção de estimativas extremais. O método proposto – doravante designado por *método mestre* – é extremamente geral, incluindo como caso particular o algoritmo conceptual de pesquisa aleatória simples, bem como a variante estocástica do algoritmo zigzag de Mexia *et al.* (1999). São apresentadas duas variantes do método mestre: uma relativa a uma formulação algorítmica; outra com uma estrutura matricial inerente. A formulação matricial permite uma melhor compreensão do ponto de vista conceptual do método introduzido. Além disso, esta formulação pode ainda possibilitar uma implementação mais simples, conforme evidenciamos através da decomposição Kronecker-zigzag. A formulação matricial torna também claro como se pode tirar partido da teoria dos valores extremos. Com efeito, através da aplicação de resultados assimptóticos da teoria dos valores extremos à primeira coluna da matriz das iteradas, é possível construir intervalos de confiança para o máximo da função dos parâmetros. Um dos grandes triunfos desta tese reside na prova de convergência estocástica do método mestre. Efectivamente, da demonstração de convergência deste método, sai como corolário a convergência do método estocástico zigzag, bem como a convergência de todos os remanescentes casos particulares deste método. Com efeito, a designação do método é devida ao facto de este funcionar como uma "chave mestra" no que concerne ao estabelecimento da convergência de uma vasta classe de métodos de optimização.

Miguel de Carvalho

Título: *Métodos Analíticos em Probabilidade e Métodos Probabilísticos em Análise Fractalidade Associada aos Modelos Beta(p,q), Evolução de Populações e Dimensões de Hausdorff*

Autora: Sandra Maria da Silva Figueiredo Aleixo, *sandra.aleixo@dec.isel.ipl.pt*
Orientadores: José Leonel Linhares da Rocha e Dinis Duarte Ferreira Pestana

Na minha tese, deduzi modelos de crescimento populacional proporcionais a densidades beta com parâmetros de forma p e 2, onde $p \geq 1$, cuja complexidade dinâmica está relacionada com o parâmetro malthusiano r. Usando técnicas de dinâmica simbólica, investiguei o comportamento caótico destes modelos, em termos de entropia topológica, no espaço de parâmetros (r,p), identificando diferentes comportamentos dinâmicos.

Verifiquei a universalidade da constante de Feigenbaum nos modelos apresentados, usando uma fórmula diferente daquela que é usualmente apresentada na literatura.

O efeito de Allee foi analisado nestes modelos. Para $p>2$, eles exibem uma dinâmica populacional onde o efeito de Allee surge naturalmente. No entanto, no caso onde $1<p\leq 2$, os modelos propostos não incluem este efeito. Para invocá-lo, apresentei alguns modelos alternativos e investiguei as suas dinâmicas.

Analisei também a negatividade da derivada de Schwarz em todos os modelos propostos.

Defini poeira de Cantor aleatória, um fractal obtido por eliminação recursiva do espaçamento central que é definido entre o mínimo e máximo de duas observações aleatórias uniformemente distribuídas, de cada intervalo da iteração anterior. A designação atribuída ao fractal é justificável, uma vez que os valores esperados dos extremos dos intervalos de cada iteração, coincidem com os extremos dos intervalos da correspondente iteração na construção da poeira de Cantor determinista.

Calculei a dimensão de Hausdorff (que intuitivamente avalia a que ponto um conjunto é denso) da poeira de Cantor aleatória, e verifiquei que apesar de a poeira de Cantor ser o "fractal médio", da poeira de Cantor aleatória, é mais denso (a dimensão de Hausdorff da poeira de Cantor C é superior dimensão de Hausdorff da poeira de Cantor aleatória F_U).

Este resultado levou-me a uma definição mais geral de conjuntos de Cantor aleatórios F_X, onde X é uma variável aleatória com distribuição Beta(p,q), ao cálculo das suas dimensões de Hausdorff, e das dimensões de Hausdorff dos fractais deterministas que são a esperança daqueles fractais aleatórios, num sentido similar ao de a poeira de Cantor determinista ser a esperança da poeira de Cantor aleatória.

O fenómeno é geral, e para essa diferença entre dimensões de Hausdorff encontrei uma explicação probabilista que reforça a interpretação de dimensão de Hausdorff como reveladora da abundância de pontos do fractal.

Sandra Aleixo

PRÉMIOS "ESTATÍSTICO JÚNIOR 2009"

Trabalho classificado em 1º lugar (Ensino Básico)

Título: *"Como se ocupam os nossos avós"* Autoria: Mariana Branco Farinha, Henrique Manuel T. Manso Vinhas Nunes, Mariana Sofia das Neves Cruz. Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves, Torres Novas. Professor orientador: Maria Alice da Silva Martins.

Trabalhos classificados em 2º lugar (exæquo) (Ensino Básico)

Título: *"Um olhar sobre a Estatística nos 2º e 3º ciclos"* Autoria: Abel Filipe Santiago Nicolau, António Manuel de Jesus Ferreira, Mário António Ferreira Esteves da Silva Leal. Estabelecimento de Ensino: Colégio Internato dos Carvalhos, Vila Nova de Gaia. Professor orientador: Sandra Maria de Sousa Campelos.

Título: *"A Escola e a Família"* Autoria: Leonor Oliveira Pedro, Inês Oliveira Pedro dos Santos. Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves, Torres Novas. Professor orientador: Teresa de Jesus Poço Isabel.

Trabalho classificado em 3º lugar (Ensino Básico)

Título: *"Futuros Eleitores da Grão Vasco"* Autoria: Rafael José Gonçalves de Melo, Carlos Miguel Cardoso Garrido, Henrique Miguel Afonso Domingos. Estabelecimento de Ensino: Escola E. B. 2,3 Grão Vasco – Viseu. Professor orientador: Cheila Isabel Ferreira Nunes e Sá Pereira.

Trabalho classificado em 1º lugar (Ensino Secundário)

Título: *"A educação para os Barcelenses - "Barcelos call: sondagem sobre a educação"* Autoria: Andreia Sofia Ferraz Araújo, José Emanuel da Silva Boavista, Pedro Manuel Costa Magalhães. Estabelecimento de Ensino: Escola Secundária de Barcelos, Barcelos, Professor orientador: José Eduardo Fernandes da Cunha

Trabalhos classificados em 2º lugar (exæquo) (Ensino Secundário)

Título: *"A Cultura Geral no Ensino Secundário"* Autoria: José Pedro Gomes Marques da Silva, António Gil Cabral Azevedo. Estabelecimento de Ensino: Externato Ribadouro, Porto. Professor orientador: Susana Luzia Machado Gonçalves Moreira Gomes Antunes da Silva.

Título: *"Pesos e alturas das crianças do J.I. de Santa Maria"* Autoria: Kayla Pires Pereira, Maria Inês da Luz Ferreira, Sara Filipa Alves Pina dos Santos. Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves, Torres Novas. Professor orientador: Maria Alice da Silva Martins.

Trabalho classificado em 3º lugar (Ensino Secundário)

Título: *"Pokémon - Estudo Estatístico para Matemática"* Autoria: Rita Pereira Casmarrinha, Diogo Chotas Arsénio Dias. Estabelecimento de Ensino: Escola Secundária de Cacilhas – Tejo, Cacilhas. Professor orientador: Luís Miguel Fonseca Nunes.

PRÉMIOS "ESTATÍSTICO JÚNIOR 2009"

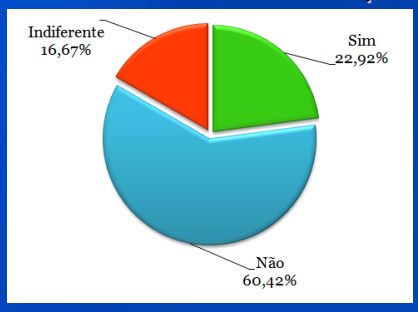
Trabalho classificado em 1º lugar (Ensino Secundário)

SOCIEDADE PORTUGUESA DE ESTATÍSTICA
desde 1980

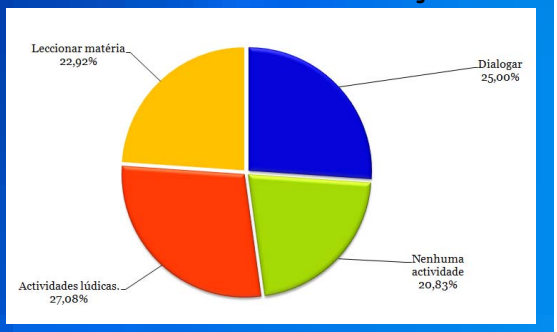
ESCOLA SECUNDÁRIA DE BARCELOS

A educação para os barcelenses
" Barcelos call: sondagem sobe a educação"

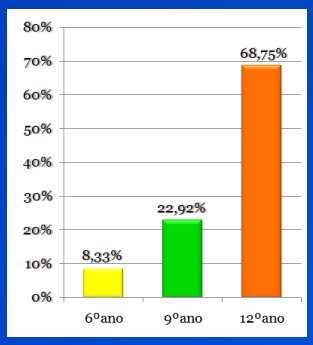
Professores das escolas privadas, profissionalmente melhores do que os das escolas públicas?



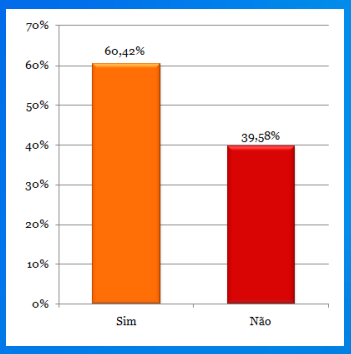
O que é feito nas aulas de substituição?



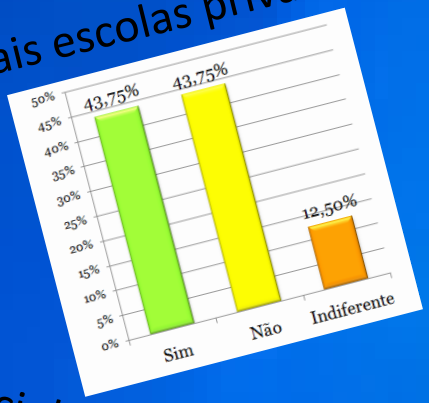
Escolaridade obrigatória, até quando?



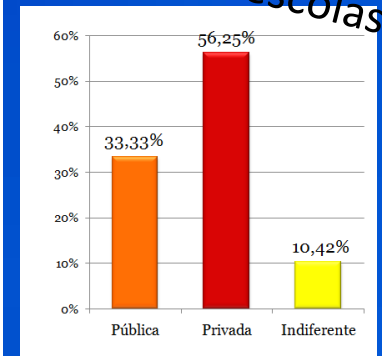
Os CURSOS PROFISSIONAIS preparam adequadamente os jovens?



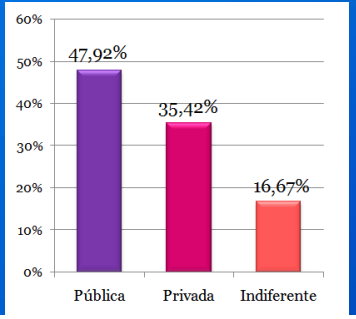
Mais escolas privadas?



Maior rigor disciplinar, onde? Nas escolas públicas? Nas privadas?



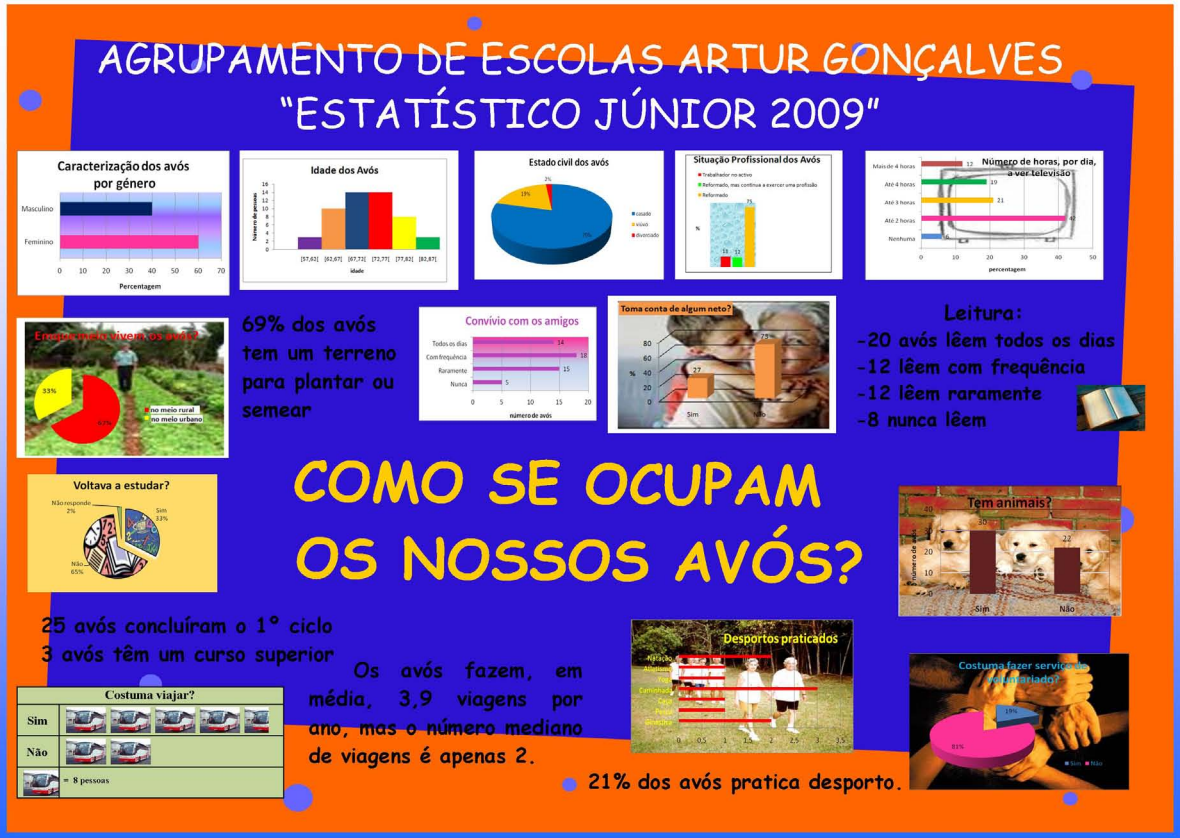
Onde é que os alunos são melhores preparados?



Autores: Emanuel, Andreia e Manuel

PRÉMIOS "ESTATÍSTICO JÚNIOR 2009"

Trabalho classificado em 1.º lugar (Ensino Básico)



"Estatísticos Júnior 2009"

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

PRÉMIOS "ESTATÍSTICO JÚNIOR 2010"

Está aberto, até 28 de Maio de 2010, o concurso para atribuição de prémios **"Estatístico Júnior 2010"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2010"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística.

2. Os candidatos a prémios **"Estatístico Júnior 2010"** devem ser alunos do 3.º Ciclo do Ensino Básico, do Ensino Secundário, ou dos Cursos de Educação e Formação de Adultos (EFA) no ano lectivo 2009/2010.

3. As candidaturas podem ser individuais ou em **grupo com um máximo de 3 alunos**. Do grupo pode ainda fazer parte um professor do ensino básico ou secundário ao qual caberá o papel de orientador.

4. Os candidatos devem apresentar um trabalho cuja temática deve estar relacionada com a teoria da Probabilidade e/ou Estatística.

5. O trabalho deverá ser constituído por um texto escrito em Português com um máximo de 10 páginas A4 dactilografadas e um poster formato A2 que resuma os principais aspectos do trabalho. O trabalho deverá ser enviado impresso em papel para efeitos da avaliação.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2010"** a 7 trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.º Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário, e um primeiro classificado de entre os trabalhos candidatos dos Cursos EFA. Os prémios são constituídos por produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 600 euros, 300 euros e 200 euros, a atribuir, respectivamente, aos grupos cujos trabalhos sejam classificados em 1.º, 2.º e 3.º lugar para as categorias Ensino Básico e Secundário e 600 euros para a categoria dos Cursos EFA.

7. Ao professor orientador do trabalho classificado em 1º lugar, em cada categoria, é ainda atribuída uma anuidade grátis como sócio da SPE, ajudas de custo para participação no XVII Congresso Anual da SPE e produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 500 Euros.

8. Aos grupos proponentes dos trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente poster que será colocado na Sessão de Posters do XVIII Congresso Anual da SPE.

9. O boletim de candidatura, acompanhado do trabalho concorrente, deverá ser dirigido ao Presidente da SPE para a morada abaixo indicada. O carimbo do correio validará a data de entrega.

Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa

O boletim de candidatura e este regulamento podem ser obtidos em
http://www.spestatistica.pt/static/docs/BoletimCandidaturaPEJ10.pdf
http://www.spestatistica.pt/static/docs/RegulamentoPEJ10.pdf

10. A admissibilidade e apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição e nomeação será da responsabilidade da Direcção da SPE.

11. O júri é soberano nas decisões, não havendo lugar a impugnação ou recurso.

12. A atribuição dos prémios **"Estatístico Júnior 2010"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada no XVII Congresso Anual da SPE.

13. Os prémios **"Estatístico Júnior 2010"** poderão não ser atribuídos.

Apoio da Porto Editora

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO SPE 2009

Método de Imputação Recorrente: Análise Espectral Singular com Valores Omissos

Miguel de Carvalho, *mb.carvalho@fct.unl.pt*
Paulo C. Rodrigues, *paulocanas@fct.unl.pt*
Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia e CMA

A Análise de Componentes Principais (ACP) é uma das ferramentas mais populares no domínio da análise multivariada. No entanto, o contexto original sob o qual a técnica foi desenvolvida torna a ACP inapropriada para o estudo de séries temporais. A Análise Espectral Singular (AES) surge precisamente como uma extensão da ACP para séries temporais univariadas (Golyandina *et al.*, 2001). A ideia basilar da AES consiste na decomposição da série temporal em diversos blocos distintos que possam ser identificados como componentes referentes a tendência, movimentos sazonais, ruído, etc. São também conhecidas na literatura técnicas para articular com a AES, por forma a permitir a condução de experiências de previsão. Essencialmente, a AES encontra a sua motivação originária na decomposição clássica de Karhunen-Loève, e outros resultados célebres sobre a representação ortogonal de processos estocásticos. As raízes deste procedimento são geralmente atribuídas ao trabalhos de Broomhead e King (1986). Algumas aplicações deste procedimento podem ser encontradas em Golyandina *et al*. (2001), e referências aí incluídas. Uma panorâmica consubstanciada da AES pode também ser encontrada na mesma referência.

Neste trabalho é proposto um Método de Imputação Recorrente (MIR) para séries temporais com valores omissos, baseado na AES. O MIR recorre a uma combinação ponderada de valores de previsão directa (*forecast*) e previsão inversa (*backcast*) por forma a imputar de modo recorrente os valores omissos. Com o intuito de ilustrar a mecânica do método referido foi usada a base de dados clássica na qual são consideradas observações mensais do número total de passageiros em diversas companhias aéreas internacionais. A utilização deste conjunto de dados permitiu-nos estabelecer comparações imediatas com um método alternativo proposto recentemente por Golyandina e Osipov (2007). Os resultados obtidos são extremamente apelativos superando mesmo o método de Golyandina e Osipov em algumas medidas de qualidade de previsão.

Referências:

[1] Broomhead, D.S. e King, G.P., 1986. Extracting qualitative dynamics from experimental data. *Physica D*, 20, 217–236.
[2] Golyandina, N. e Osipov E., 2007. The “Catterpillar”-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*, 137, 2642–2653.
[3] Golyandina, N., Nekrutkin, V. e Zhigljavsky, A., 2001. *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC, London.

Miguel de Carvalho, **galardoado com o Prémio SPE 2009**, licenciou-se em Matemática pela Universidade Nova de Lisboa e Mestre em Economia pela mesma Universidade. Concluiu o seu Doutoramento em Estatística Matemática sob a orientação de João Tiago Mexia e Manuel L. Esquível. Durante o presente ano lectivo realiza estudos de pós-doutoramento na Faculdade de Ciências da Universidade de Lisboa, sob a supervisão de Feridun Turkman e Antónia Turkman.

Paulo Canas Rodrigues, **galardoado com o Prémio SPE 2009**, licenciou-se em Matemática pela Universidade Nova de Lisboa. É Mestre em Estatística pelo Instituto Superior Técnico. Actualmente é aluno de doutoramento em Matemática (especialização em Estatística) na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, sob a orientação dos Professores Stanislaw Mejza e João Tiago Mexia. Neste momento é Investigador e Assistente convidado na Universidade de Wageningen, na Holanda.