R é confiável para estatística computacional?

Marcelo G. Almiron
DCC – UFMG

Eliana S. Almeida
CPMAT & LCCV
IC – UFAL

Antonio C. Medeiros
LCCV – UFAL

Alejandro C. Frery
CPMAT & LCCV
IC – UFAL

Resumo

Este trabalho avalia a precisão numérica da plataforma R em duas arquiteturas de processador (i386 e amd386), rodando sistemas operacionais Microsoft Windows 7, GNU/Linux Ubuntu 9.10 e MAC OS X Leopard (este último apenas em i386). A avaliação consiste em calcular os valores da média, do desvio padrão, da correlação de primeira ordem e o estatístico F de ANOVA, empregando conjuntos de dados com comportamento conhecidamente problemático. Os valores reportados por R são contrastados com outros certificados, e o número de dígitos significativos corretos é informado para cada situação. Com excepção de uma situação onde R é incapaz de produzir resultados aceitáveis, esta plataforma se mostra precisa e portável, duas propriedades essenciais em estatística computacional.

1 Introdução

Este trabalho apresenta indícios de que a resposta à pergunta do título é **sim**.

R é uma linguagem de alto nível e ambiente interativo para análise de dados e geração de gráficos. Esta ferramenta, disponível gratuitamente em `http://www.r-project.org`, contem uma quantidade muito grande de pacotes associados para realizar as mais diversas tarefas em análise e visualização de dados.

Lamentavelmente, existe uma comunidade ainda maior da citada acima, que utiliza outro tipo de ferramentas que podem comprometer resultados em pesquisas e provocar grandes perdas de dinheiro, dentre outras desvantagens (ver Su, 2008; Croll, 2009; Powell et al., 2009). Este é o caso de Microsoft Excel (a ferramenta mais utilizada em estatística) e outras distribuições de planilhas eletrônicas (Almiron et al., 2010). Finalmente existe outra comunidade que está formada por usuários de outras ferramentas para análise de dados e computação numérica, e simplesmente não acha um motivo significativo para dedicar esforço em apreender a utilizar outras alternativas.

Neste trabalho apresentamos um motivo para aprender a utilizar a ferramenta R. Mas, por que R? Podemos responder esta pergunta comentando os resultados de Almiron et al. (2009). Esses autores apresentaram uma análise detalhada da precisão numérica de cinco ferramentas gratuitas para computação numérica. Esta análise compara a confiabilidade das implementações de funções estatísticas, obtendo resultados muito bons com R. Porém, o trabalho de Almiron et al. (2009) apresenta a avaliação de uma única dentre várias plataformas onde esta ferramenta está disponível; eles utilizaram um computador com arquitetura i386 e sistema operacional GNU/Linux Ubuntu 7.10. Mas existem fatores do ambiente computacional que podem influir na precisão numérica aferida, portanto avaliações mais completas, no sentido mencionado acima, são necessárias.

Uma das principais motivações para avaliar R em múltiples plataformas foi a análise feita por Almiron et al. (2010). Aqui os autores observam que, no caso de várias planilhas eletrônicas, as implementações em diferentes plataformas podem apresentar diferenças de precisão muito grandes. Este artigo visa avaliar a portabilidade da plataforma R, dado que esta característica é de fundamental importância para conformar uma alternativa confiável, onde os usuários possam empregá-la seja qual for a plataforma que estiverem utilizando.

A metodologia aplicada neste trabalho (descrita na seção 2) é um procedimento consolidado e amplamente aplicado na avaliação de precisão numérica de várias ferramentas (ver Keeling & Pavur, 2007; McCullough & Heiser, 2008; Almiron et al., 2010; Bustos & Frery, 2006).

Este artigo apresenta uma primeira avaliação multiplataforma que demonstra a robustez de R, como veremos na seção 3. A continuação, na seção 2, apresentamos uma descrição da metodologia de avaliação. Os resultados em forma de tabelas são apresentados na seção 3, junto com os comentários pertinentes. Finalmente, concluímos este artigo na seção 4 com as discussões dos resultados aqui obtidos, e comentamos os projetos futuros.

2 Metodologia

Existem três fontes de erros em funções numéricas: erro de truncamento, erro de cancelamento e erro acumulado. O erro de truncamento é ocasionado pelos dígitos decimais que não podem ser representados com a palavra binária em uso. Este erro pode se propagar, ou, no melhor caso, se tornar constante, mas não é possível eliminá-lo.

Ocorre um erro de cancelamento quando temos dados com variabilidade relativa pequena e com maior quantidade de dígitos significativos constantes; isto leva a complicações para o cálculo do desvio padrão dentre outros. Os erros acumulados são aqueles que se apresentam em proporção ao número total de operações realizadas McCullough (1998); National Institute of Standards and Technology (2010).

Uma maneira de medir estes erros é comparar o resultado obtido ao aplicar uma função de uma ferramenta, por exemplo a que calcula a média, e compará-lo com um valor certificado, isto é, um valor que seja conhecidamente correto. Para isto, o NIST (National Institute of Standards and Technology, 2010) provê conjuntos de dados (*datasets*), acessíveis pela Internet em `http://www.itl.nist.gov/div898/strd/general/dataarchive.html`. Esses dados contêm os argumentos da função a ser aplicada junto com o valor certificado que deveria ser obtido ao aplicar essa função. Os valores certificados, fornecidos pelo NIST com os *datasets*, têm quinze dígitos significativos.

Para se obter a quantidade de dígitos de coincidência entre o valor obtido e o certificado, McCullough McCullough (1998) introduz a função LRE (*Log-Relative Error*) definida por

$$LRE(x,c) = -\log_{10}\left(\frac{|x-c|}{|c|}\right) \tag{1}$$

onde $c \neq 0$ representa o valor certificado e x o valor calculado pela função da ferramenta. Para exemplificar, sejam $x_1 = 0.0001234$ e $c_1 = 0.0001254$, temos $LRE(x_1, c_1) = 1.797268$, cuja interpretação é que aproximadamente um dígito foi calculado corretamente pela função. Adicionalmente, podemos dizer que, aproximadamente, os primeiros $6 \approx 1.797268/\log_{10}(2)$ bits de x_1 coincidem com os bits de c_1.

Esses procedimentos foram aplicados para avaliar as funções que calculam a média, o desvio padrão, o coeficiente de autocorrelação de primeira ordem, a estatística F do procedimento ANOVA e regressões lineares em diversos conjuntos de dados. Os resultados da função LRE para cada plataforma nos diferentes *datasets* são apresentados a seguir.

3 Resultados

Avaliamos a versão 2.10.1 de R nas arquiteturas i386 e amd64 com sistemas operacionais Microsoft Windows 7 e GNU/Linux Ubuntu 9.10. Também consideramos o sistema operacional MAC OS X Leopard versão 10.5.6 na arquitetura i386, sendo que não avaliamos a arquitetura amd64 neste caso por falta de disponibilidade da versão 64 bits do MAC OS X Leopard.

Na avaliação da média amostral, utilizando a função `mean`, em todos os casos foram obtidos os maiores valores de precisão possíveis com a metodologia empregada, isto é, 15 dígitos significativos.

A tabela 1 apresenta os resultados de precisão numérica do desvio padrão amostral (utilizando a função `sd`) e o coeficiente de autocorrelação de primeira ordem (empregando a função `acf`). Para o caso do desvio padrão, todas as plataformas levaram à mesma precisão numérica. Observemos que para os *datasets* NumAcc3 e NumAcc4, a quantidade de dígitos significativos estimados corretamente são consideravelmente menores que no outros casos.

Tabela 1: Valores de *LRE* para o desvio padrão amostral e o coeficiente de autocorrelação de primeira ordem

Função		Desvio padrão	Coeficiente de autocorrelação de primeira ordem				
		Todas as	MAC OS X	Ubuntu		Windows	
Dataset	Dificuldade	plataformas	32 *bits*	32 *bits*	64 *bits*	32 *bits*	64 *bits*
Lew	Baixa	15.0	15.0	15.0	15.0	14.8	15.0
Lottery	Baixa	15.0	15.0	15.0	15.0	15.0	15.0
Mavro	Baixa	13.1	13.8	13.8	13.8	13.8	13.8
Michelso	Baixa	13.8	13.4	13.4	13.4	13.4	13.4
NumAcc1	Baixa	15.0	15.0	15.0	15.0	15.0	15.0
PiDigits	Baixa	15.0	13.0	13.0	13.0	15.0	13.0
NumAcc2	Média	15.0	14.4	14.4	14.4	15.0	14.4
NumAcc3	Média	9.5	14.3	14.3	14.3	15.0	14.3
NumAcc4	Alta	8.3	14.5	14.5	14.5	15.0	14.5

A respeito do coeficiente de autocorrelação de primeira ordem, ainda na tabela 1, os resultados de precisão obtidos representam valores de *LRE* satisfatórios, em todos os casos. Observam-se pequenas variações de uma para outra versão. Em particular, a versão de 32 *bits* para Windows apresenta os melhores resultados de precisão numérica.

A seguir, a tabela 2 apresenta os valores de *LRE* correspondentes ao estatístico F do procedimento ANOVA. Aqui, como no caso do coeficiente de autocorrelação, os melhores resultados foram obtido utilizando a arquitetura i386 sobre Windows. Porém, nos resultados apresentados nesta tabela não observamos variações significativas que levem a concluir que essa versão é mais confiável do que as outras.

Para concluir com esta seção de resultados, apresentamos na tabela 3 os valores de precisão numérica correspondentes às regressões lineares. Nesta tabela podemos observar que existe um problema grave no *dataset* Filip. Este *dataset* é ajustado por meio do modelo polinomial $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_9 x^9 + \beta_{10} x^{10} + \epsilon$, onde os onze parâmetros são comparados com os valores certificados, e o valor da tabela corresponde ao menor dos valores de *LRE* dentre todos os valores correspondentes a cada parâmetro do modelo. Esse resultado é surpreendente, já que conforme Almiron et al. (2010) mostram, Excel nas versões 2007 e 2008 produz valores de *LRE* de 7,2 e 6,4, respectivamente, para esse conjunto de dados

Tabela 2: Valores de *LRE* para o estatístico *F* de ANOVA

		MAC OS X	Ubuntu		Windows	
Dataset	Dificuldade	32 *bits*	32 *bits*	64 *bits*	32 *bits*	64 *bits*
SiRstv	Baixa	13.3	13.3	13.3	13.3	13.3
SmLs01	Baixa	15.0	15.5	14.9	15.0	14.9
SmLs02	Baixa	13.7	14.2	14.0	15.0	14.1
SmLs03	Baixa	12.9	12.9	13.1	15.0	13.1
AtmWtAg	Média	9.7	8.8	9.2	9.2	9.7
SmLs04	Média	10.4	10.4	10.4	10.4	10.4
SmLs05	Média	10.2	10.2	10.2	10.2	10.2
SmLs06	Média	10.2	10.2	10.2	10.1	10.2
SmLs07	Alta	4.6	4.4	4.5	4.4	4.6

Tabela 3: Valores de *LRE* para o pior dos $\hat{\beta}$ em regressões lineares

		MAC OS X	Ubuntu		Windows	
Dataset	Dificuldade	32 *bits*	32 *bits*	64 *bits*	32 *bits*	64 *bits*
Norris	Baixa	12.5	12.6	12.0	12.3	12.5
Pontius	Baixa	12.1	11.9	12.7	12.3	12.7
NoInt1	Média	14.7	14.7	14.7	14.7	14.7
NoInt2	Média	15.0	15.0	15.0	15.0	15.0
Filip	Alta	0.0	0.0	0.0	0.0	0.0
Longley	Alta	13.0	12.9	12.8	13.3	13.0
Wampler1	Alta	9.8	10.1	9.0	9.3	9.8
Wampler2	Alta	13.6	13.4	12.6	12.9	13.5
Wampler3	Alta	9.3	9.8	9.0	9.6	9.2
Wampler4	Alta	7.5	8.6	7.5	8.3	7.5
Wampler5	Alta	5.5	5.9	5.5	6.3	5.5

O mesmo procedimento, descrito acima, foi aplicado em todos os *datasets* do conjunto para avaliar regressões lineares, mas os modelos variam de um *dataset* para outro. Para maior detalhe ver a página *Web* do StRD (*Statistical Reference Datasets*) no National Institute of Standards and Technology (2010).

De maneira geral, os resultados apresentam uma variabilidade pequena quanto as diferentes plataformas. A seguir na seção 4 fechamos este resumo com as discussões e trabalhos futuros.

4 Discusão e trabalhos futuros

Nos resultados apresentados na seção prévia, observamos que de maneira geral R mostra ser uma ferramenta robusta. Por outro lado, análises prévias (Almiron et al., 2009) mostram que R é uma ferramenta confiável quando comparada com outras ferramentas disponíveis na comunidade científica.

A partir destes resultados, pretendemos como trabalho futuro avaliar outros tipos de funções nas

diversas plataformas: (i) regressões não lineares, (ii) funções de distribuição e (ii) geração de números pseudo-aleatórios.

Pretendemos, também como trabalho futuro, concluir com a implementação de um pacote em R para geração automática de relatórios de precisão numérica. Este pacote pretende oferecer suporte ao usuário na escolha de uma plataforma para trabalhar com R sem depender de publicações em veículos científicos. A idéia é oferecer também ao usuário a possibilidade de avaliar várias funções que realizam a mesma tarefa sobre estruturas de dados diferentes, como é o caso de várias estruturas utilizadas em pacotes de uso específico.

Referências

Almiron, M. G., Almeida, E. S. & Miranda, M. N. (2009), 'The reliability of statistical functions in four software packages freely used in numerical computation', *Brazilian Journal of Probability and Statistics* **23**(2), 107–119.

Almiron, M. G., Lopes, B., Oliveira, A. L. C., Medeiros, A. C. & Frery, A. C. (2010), 'On the numerical accuracy of spreadsheets', *Journal of Statistical Software* **34**(4), 1–29. URL `http://www.jstatsoft.org/v34/i04`.

Bustos, O. H. & Frery, A. C. (2006), 'Statistical functions and procedures in IDL 5.6 and 6.0', *Computational Statistics & Data Analysis* **50**(2), 301–310.

Croll, G. J. (2009), Spreadsheets and the financial collapse, *in* 'Proceedings of the European Spreadsheet Risks Interest Group', pp. 145–161. URL `http://arxiv.org/pdf/0908.4420`.

Keeling, K. B. & Pavur, R. J. (2007), 'A comparative study of the reliability of nine statistical software packages', *Computational Statistics & Data Analysis* **51**(8), 3811–3831.

McCullough, B. D. (1998), 'Assessing the reliability of statistical software: Part I', *American Statistician* **52**(4), 358–366.

McCullough, B. D. & Heiser, D. A. (2008), 'On the accuracy of statistical procedures in Microsoft Excel 2007', *Computational Statistics & Data Analysis* **52**(10), 4570–4578.

National Institute of Standards and Technology (2010), 'Statistical reference datasets'. URL `http://www.itl.nist.gov/div898/strd/general/dataarchive.html`, última consulta em maio de 2010.

Powell, S. G., Baker, K. R. & Lawson, B. (2009), 'Impact of errors in operational spreadsheets', *Decision Support Systems* **47**, 126–132.

Su, Y. (2008), 'It's easy to produce chartjunk using Microsoft Excel® 2007 but hard to make good graphs', *Computational Statistics & Data Analysis* **52**(10), 4594–4601.