ISSN: 2317-0840

Modelos aditivos generalizados para locação, escala e forma na análise do número de lesões corticais em pacientes com esclerose múltipla

Ricardo Rasmussen Petterle[1†], Lucas Formighieri[1].

[1] *Universidade Federal do Paraná - UFPR*

Resumo: *A esclerose múltipla (EM) é uma doença inflamatória crônica autoimune que causa desmielinização e neurodegeneração no sistema nervoso central. Há crescente interesse em investigar variáveis associadas ao número de lesões corticais na EM. Nesse contexto, o objetivo deste artigo é analisar a relação entre o número de lesões corticais com a idade, graduação EDSS (expanded disability status scale) e tempo de doença em pacientes com este quadro. Para tanto, foram analisados 30 pacientes portadores de EM. A análise foi conduzida por meio dos modelos aditivos generalizados para locação, escala e forma (GAMLSS). Foram testadas quatro distribuições de probabilidade para a variável resposta. Os resultados mostraram que o modelo baseado na distribuição Poisson Gaussiana inversa foi o mais adequado para a análise dos dados, sendo possível modelar seus parâmetros de média e de dispersão em função das covariáveis em estudo. De maneira geral, os resultados indicam que os pacientes com maior quantidade de déficits neurológicos (representado por valor mais elevado na graduação EDSS), mais jovens e com maior tempo de doença são os que mais apresentaram lesões corticais.*

Palavras-chave: Dados de contagem; distribuição Poisson Gaussiana inversa; modelos aditivos generalizados; GAMLSS.

Abstract: *Multiple sclerosis (MS) is a chronic autoimmune inflammatory disease that causes demyelination and neurodegeneration in the central nervous system. There is a growing interest in investigating variables associated with the number of cortical lesions in MS. In this context, the aim of this paper is to analyze the relationship between the number of cortical lesions with age, graduation EDSS (expanded disability status scale) and disease duration in patients with MS. Therefore, we analyzed 30 patients with MS. The analysis was conducted using the generalized additive models for location, scale and shape (GAMLSS). Four probability distributions were tested for the response variable. The results showed that the Poisson inverse Gaussian was the most suitable distribution for the data analysis, where it was possible to model the mean and dispersion parameters as functions of some covariates. Overall, the results indicate that patients with higher neurological deficits (represented by higher value in EDSS graduation), younger and presenting longer disease duration are the ones that showed the most cortical lesions.*

Keywords: Count data; Poisson inverse Gaussian distribution; generalized additive models; GAMLSS.

[†]Autor correspondente: estatisticoufpr@gmail.com

Introdução

A esclerose múltipla (EM) é uma doença inflamatória crônica que causa desmielinização e neurodegeneração do sistema nervoso central (Compston e Coles, 2008). Suas causas são desconhecidas, mas sabe-se que é o resultado da combinação de fatores genéticos, infecciosos e ambientais (Ascherio; Munger; Lünemann, 2012). O achado histopatológico clássico desta doença são as lesões desmielinizantes focais da substância branca do cérebro e da medula espinhal, que se acumulam ao longo do tempo e são acompanhadas pela atrofia do cérebro e da medula espinhal (Lassmann, 2014).

Para investigar e analisar o relacionamento entre variáveis faz-se uso de modelos de regressão. Nesse contexto, o clássico modelo de regressão linear (gaussiano) é um dos mais adotados. Porém, para o ajuste desse modelo, é necessário supor que os erros sejam independentes e identicamente distribuídos segundo a distribuição Normal com média zero e variância constante. Na prática, isso nem sempre acontece. Segundo King (1989) a má especificação desse modelo pode resultar em erros padrões inconsistentes, além de outros problemas que invalidam todo o processo de inferência. Além disso, para respostas não contínuas, como dados de contagem, esse modelo pode não apresentar ajuste satisfatório, uma vez que ele não considera a natureza discreta da variável resposta e nem distribuições assimétricas.

Para contornar esses problemas e fornecer mais flexibilidade na modelagem por regressão, diversos modelos foram propostos na literatura. Por exemplo, os modelos lineares generalizados (*generalized linear models* - GLM) foram introduzidos por Nelder e Wedderburn (1972), sendo posteriormente desenvolvidos por McCullagh e Nelder (1989). Essa classe de modelos assume que a distribuição de probabilidade da variável resposta pertence à família exponencial e permite que apenas a média μ seja modelada em função de covariáveis. Além dessa classe de modelos, os modelos aditivos generalizados (*generalized additive models* - GAM) introduzidos por Hastie e Tibshirani (1990) são modelos mais flexíveis do que os GLM, pois permitem incorporar a soma de funções de suavização no preditor linear, mas também apresentam limitações como as citadas anteriormente.

Com o intuito de superar algumas limitações dos modelos supracitados (GLM e GAM) e apresentar uma classe de modelos mais flexíveis, Rigby e Stasinopoulos (2001, 2005) e Akantziliotou, Rigby e Stasinopoulos (2002) introduziram os modelos aditivos generalizados para locação, escala e forma (*generalized additive models for location scale and shape* - GAMLSS). Essa classe de modelos é considerada semi-paramétrica. São paramétricos porque precisam de uma distribuição de probabilidade para descrever o comportamento da variável resposta y em função de uma ou mais covariáveis e "semi" no sentido de ser possível modelar os parâmetros da variável resposta por meio de funções de suavização não-paramétricas como *splines* cúbicos, polinômios fracionários, funções de suavização dentre outras. Nos GAMLSS, a suposição de distribuição pertencente à família exponencial para a variável resposta é relaxada e substituída por uma família de distribuições mais geral, incluindo distribuições de probabilidades, contínuas e discretas, para modelagem de dados com diferentes graus de assimetria e curtose. A parte sistemática do modelo é expandida, permitindo modelar não somente a média (ou locação), mas todos os outros parâmetros da distribuição de y como funções lineares e/ou não-lineares, paramétricas e/ou não-paramétricas das covariáveis.

Neste contexto, o principal objetivo deste artigo é analisar, utilizando os GAMLSS, a relação entre o número de lesões corticais com a idade, graduação EDSS e tempo de doença de pacientes com esclerose múltipla (EM).

Material e métodos

Conjunto de dados

Foi realizado um estudo transversal prospectivo com 30 pacientes portadores de EM. Para ser incluído neste estudo, o paciente deveria ser adulto voluntário (maior de 18 anos) com a forma surto-remissão da EM, independente do sexo.

Cada paciente realizou avaliação clínica na qual foram coletadas as informações de idade, sexo, tempo de evolução da doença e graduação EDSS. Em seguida cada paciente realizou um exame de ressonância magnética em equipamento de 3 Tesla com sequências específicas para a investigação de lesões corticais. Estes exames foram avaliados por uma dupla de neuroradiologistas que registrou a quantidade de lesões corticais em cada paciente. O estudo foi aprovado pelo Comitê de Ética em Pesquisa do Hospital de Clínicas da Universidade Federal do Paraná (HC-UFPR).

A escala EDSS (expanded disability status scale) é uma gradução que mede o nível de déficits neurológicos e de limitações funcionais dos pacientes com EM. Ela varia de 0 (sem déficits) a 10 (morte por EM), com incrementos de 0,5.

Modelos aditivos generalizados para locação, escala e forma

Na especificação de um GAMLSS, as variáveis aleatórias $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ são observações independentes com função (densidade) de probabilidade $f(y_i|\boldsymbol{\theta}_i)$ condicional a $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)$ um vetor composto por quatro parâmetros. Esses parâmetros estão associados a uma distribuição de probabilidade D podendo ela ser discreta ou contínua. Geralmente, os dois primeiros parâmetros da distribuição (μ_i e σ_i) são caracterizados como parâmetros de locação e escala, respectivamente, enquanto que os outros dois parâmetros (ν_i e τ_i) são usualmente caracterizados como parâmetros de forma. Nos GAMLSS, é possível relacionar covariáveis a cada um dos parâmetros da distribuição. Isso é feito através de termos aditivos ligados por meio de uma função monótona e duplamente diferenciável chamada de função de ligação $g_k(\cdot)$ com $k = 1, 2, 3, 4$. Assim, tem-se a estrutura geral definida por,

$$\begin{aligned}
\mathbf{y} &\overset{ind}{\sim} D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\
g_1(\boldsymbol{\mu}) &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{Z}_{11}\boldsymbol{\gamma}_{11} + \ldots + \mathbf{Z}_{1k_1}\boldsymbol{\gamma}_{1J_1} \\
g_2(\boldsymbol{\sigma}) &= \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{Z}_{21}\boldsymbol{\gamma}_{21} + \ldots + \mathbf{Z}_{2k_2}\boldsymbol{\gamma}_{2J_2} \\
g_3(\boldsymbol{\nu}) &= \mathbf{X}_3\boldsymbol{\beta}_3 + \mathbf{Z}_{31}\boldsymbol{\gamma}_{31} + \ldots + \mathbf{Z}_{3k_3}\boldsymbol{\gamma}_{3J_3} \\
g_4(\boldsymbol{\tau}) &= \mathbf{X}_4\boldsymbol{\beta}_4 + \mathbf{Z}_{41}\boldsymbol{\gamma}_{41} + \ldots + \mathbf{Z}_{4k_4}\boldsymbol{\gamma}_{4J_4}
\end{aligned}$$

em que:

- $D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ é a distribuição da variável resposta $\mathbf{y}$;
- $\mathbf{X}_k$ com $k = 1, 2, 3, 4$ é a matriz de delineamento que incorpora os termos lineares no modelo;
- $\boldsymbol{\beta}_k$ com $k = 1, 2, 3, 4$ o vetor de parâmetros associado à parte linear do modelo;
- $\mathbf{Z}_{jk}$ com $k = 1, 2, 3, 4$ também é uma matriz de delineamento usada para modelar termos de suavização, efeitos aleatórios, dentre outros;
- $\boldsymbol{\gamma}_{jk}$ é um vetor aleatório de dimensão q_{jk}, assumindo $\boldsymbol{\gamma}_{jk} \sim \mathcal{N}_{qjk}(\mathbf{0}, \boldsymbol{G}_{jk}^{-1})$, em que $\boldsymbol{G}_{jk}^{-1}$ é a matriz inversa de $\boldsymbol{G} = \boldsymbol{D}^\top\boldsymbol{D}$, em que $\boldsymbol{D}$ é a matriz de diferanças de ordem k.

A estimação dos parâmetros é feita via método de máxima verossimilhança penalizada. A partir da maximização 1, que representa a função de verossimilhança penalizada (l_p), tem-se as estimativas dos parâmetros. Assim, sua expressão é dada por:

$$l_p = l - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k}\lambda_{kj}\boldsymbol{\gamma}_{jk}^{\top}\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{1}$$

em que l é a função de log-verossimilhança expressa por $\sum_{i=1}^{n}\log f(y_i|\mu_i,\sigma_i,\nu_i,\tau_i)$ e λ_{kj} os hiperparâmetros de suavização.

Modelagem dos dados

Na especificação dos modelos, foram consideradas as distribuições Poisson (PO), Binomial negativa tipo I (NBI), Binomial negativa tipo II (NBII) e Poisson Gaussiana inversa (PIG) para a variável resposta. Para maiores detalhes, sobre esses modelos probabilísticos, ver Rigby e Stasinopoulos (2009). Nas três últimas distribuições, modelou-se os parâmetros de média (μ) e de dispersão (σ) em função das covariáveis, usando *splines* cúbicos. No ajuste dos modelos, considerou-se a função de ligação logarítmica tanto para μ como para σ, que é a ligação usual nos GAMLSS para dados de contagem. Para seleção do modelo final, foi usado o critério de informação Bayesiano (BIC; Schwarz, 1978) e o critério de informação de Akaike (AIC; Akaike, 1974), bem como a medida de deviance global. Adotou-se nível de significância $\alpha = 0,05$ sendo considerado significativo p-valor $< 0,05$. Por fim, a adequação do ajuste foi verificada por meio de uma análise visual dos resíduos quantílicos (aleatorizados) normalizados (Dunn e Smyth, 1996), por meio do gráfico *worm*-plot apresentado por Buuren e Fredriks (2001), comumente usado para diagnóstico nos GAMLSS.

Toda análise dos dados foi conduzida via *software* `R` (R Core Team, 2015) com auxílio do pacote `gamlss` (Stasinopoulos e Rigby, 2007).

Resultados e discussão

Análise Exploratória

A partir da Figura 1 observa-se a variável resposta (número de lesões corticais) em função das possíveis covariáveis. O gráfico (A) mostra dois pontos discrepantes (*outliers*) referente aos pacientes com maior número de lesões corticais presentes na amostra (34 e 40 lesões corticais). O gráfico (B) se refere ao diagrama de dispersão do número de lesões corticais em função da idade. Esse gráfico sugere que ao avançar da idade as lesões diminuem lentamente. O gráfico (C) apresenta a variável resposta em função da graduação EDSS, sugerindo, portanto, que quanto maior o número de lesões corticais maior é o valor da graduação EDSS. Por fim, o gráfico (D) representa o número de lesões corticais *versus* o tempo de doença, apontando que, entre 10 e 15 anos de doença, o número de lesões é constante, diminuindo gradativamente até os 25 anos de doença. Ademais, os gráficos (B) e (D) têm formato de "cone" indicando a variabilidade presente nos dados, o que justifica a utilização dos GAMLSS.

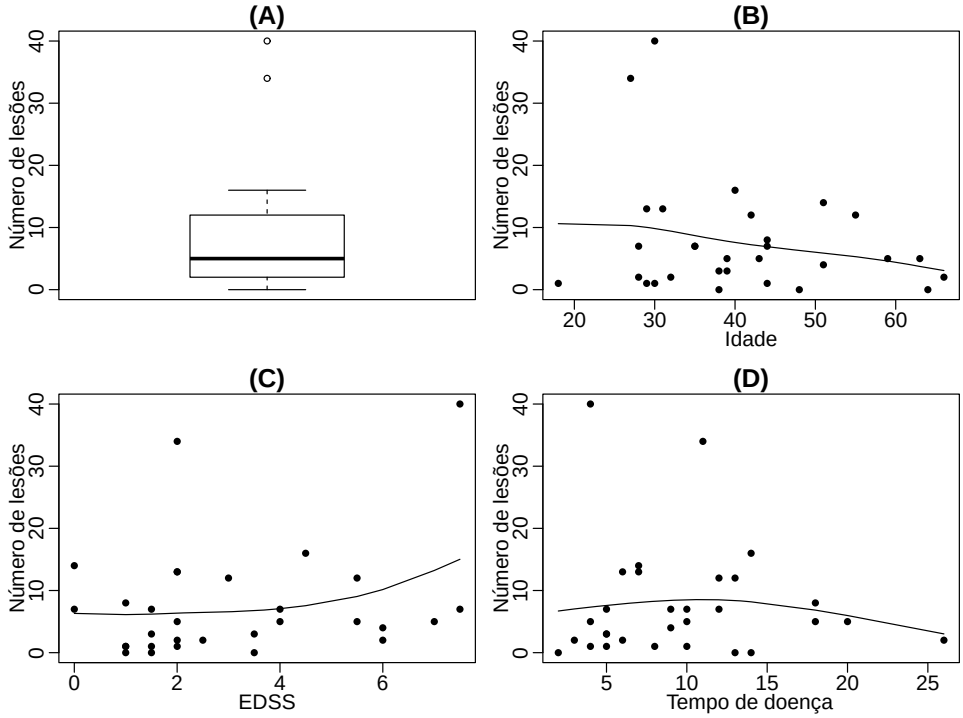

Figura 1: Gráfico box-plot (A) para o número de lesões corticais. Diagramas de dispersão (B, C e D) para o número de lesões corticais em função da idade, graduação EDSS e tempo de doença, respectivamente.

Na sequência, foram ajustados quatro modelos de regressão baseado em diferentes distribuições para a variável resposta. Ao considerar as distribuições discretas, para dados de contagem, disponíveis no pacote GAMLSS, verificou-se, com base nos critérios BIC e AIC, bem como pela deviance global que a distribuição Poisson Gaussiana inversa proporcionou melhor ajuste. Para efeito de comparação com os ajustes produzidos, a Tabela 1 apresenta a deviance global, o critério de informação Bayesiano (BIC), o critério de informação de Akaike (AIC) e graus de liberdade (G.L.) gerados pelos modelos correspondentes às quatro distribuições. Pode-se notar, com base nesses resultados, menores valores produzidos pelo modelo baseado na distribuição PIG em relação aos produzidos pelas demais distribuições (PO e NBI e NBII) para ambos os critérios.

Tabela 1: Medidas de qualidade de ajuste dos modelos de regressão.

Modelo	Deviance Global	BIC	AIC	G.L.
Poisson	202,43	243,24	226,42	11,99
Binomial Negativa I	157,22	204,84	185,22	13,99
Binomial Negativa II	149,66	197,27	177,66	13,99
Poisson Gaussiana Inversa	144,79	192,40	172,78	13,99

Resultados do ajuste com distribuição Poisson Gaussiana inversa

A Tabela 2 apresenta o resumo do modelo de regressão baseado na distribuição Poisson Gaussiana inversa. Observa-se que todos os parâmetros do modelo são significativos, ao nível de 5% de significância. Desse modo, foram modelados os parâmetros de média e de dispersão em função de covariáveis. Assim, para a média, tem-se que a graduação EDSS, a idade e o tempo de doença foram importantes na explicação do número de lesões corticais, enquanto que para o

parâmetro de dispersão apenas a idade foi considerada.

Tabela 2: Estimativas dos parâmetros, erros-padrão e p-valores para $\log \mu$ e $\log \sigma$ do modelo de regressão baseado na distribuição Poisson Gaussiana inversa.

Parâmetro	Covariável	Coeficiente	Erro-Padrão	p-valor
μ	Intercepto	1,479	0,607	0,0268
	EDSS	0,157	0,047	0,0044
	Idade	-0,039	0,014	0,0151
	Tempo de doença	0,135	0,026	0,0001
σ	Intercepto	15,202	5,805	0,0186
	Idade	-0,486	0,199	0,0270

Com base nos coeficientes estimados (Tabela 2), pode-se escrever o modelo de regressão baseado na distribuição Poisson Gaussiana inversa em termos da média $\mu(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, conforme apresentado na Eq. 2,

$$\mu(\mathbf{x}) = \exp(1,479 + 0,157 \text{ EDSS} - 0,039 \text{ Idade} + 0,135 \text{ Tempo de doença}). \quad (2)$$

Na sequência, analisou-se os resíduos do modelo. A Figura 2 mostra o gráfico *worm*-plot, no qual é possível observar que os resíduos se encontram dentro das bandas de confiança, evidenciando que o modelo apresenta um bom ajuste.

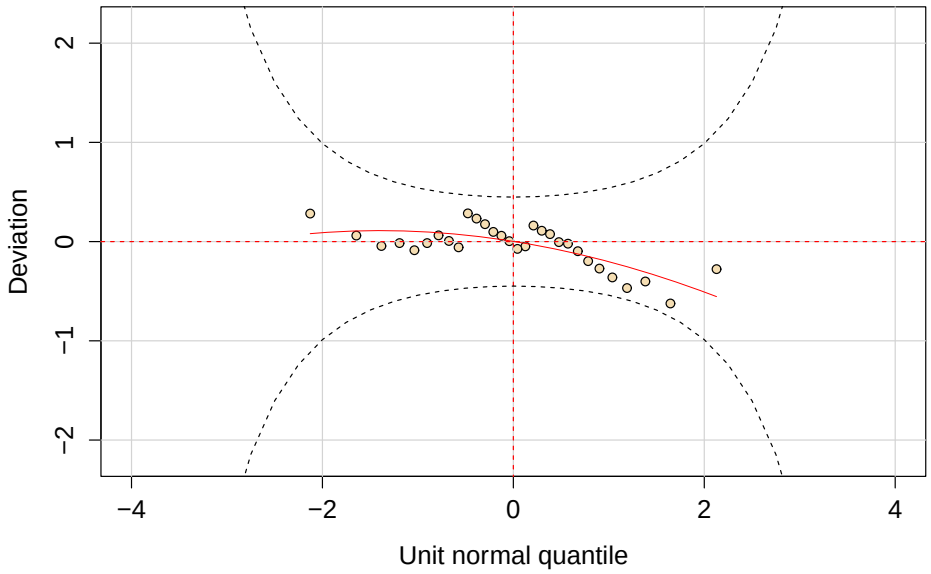

Figura 2: Gráfico *worm*-plot para o ajuste final com distribuição Poisson Gaussiana inversa.

A partir da Tabela 2, foi possível verificar que a graduação EDSS e o tempo de doença dos pacientes, influenciaram positivamente no número de lesões corticais. Desse modo, o sinal positivo referente ao coeficiente da graduação EDSS indica que um acréscimo nesta graduação corresponde a um aumento no número de lesões corticais, quando as demais covariáveis são mantidas constantes no modelo. Da mesma forma, o sinal positivo do coeficiente referente ao tempo de doença aponta que os pacientes com maior tempo de doença são os que mais apresentaram lesões corticais. Ressalta-se que essa interpretação, fornecida pelo modelo de regressão, foi

diferente da observada na análise exploratória. Possivelmente, isso ocorreu devido ao pequeno tamanho da amostra não tendo, portanto, informação suficiente para observar exatamente o que ocorre no decorrer do tempo de doença. Por outro lado, a covariável idade exerceu efeito negativo no número de lesões corticais. Assim, quanto maior for a idade do paciente espera-se que ele apresente um número menor de lesões corticais. Estima-se que a cada ano de idade do paciente, a média do número de lesões diminui em aproximadamente 4% $\left(e^{-0,039}\right)$, mantidas fixas as demais covariáveis no modelo.

Considerando o parâmetro de dispersão, temos que à medida que a covariável idade aumenta, a dispersão diminui, ou seja, os pacientes mais velhos tendem a apresentar respostas menos dispersas. Note que o parâmetro σ está na escala logarítmica. Dessa forma, a estimativa pontual para a idade é dada por $\hat{\sigma} = \exp(\hat{\beta}_{\sigma_{\text{idade}}}) = \exp(-0,486) = 0,615$.

Conclusões

Este artigo analisou a relação entre o número de lesões corticias com a idade, a graduação EDSS e o tempo de doença de pacientes com EM. A análise foi conduzida por meio dos GAMLSS. Foram consideradas as distribuições de probabilidade Poisson, Binomial negativa tipo I, Binomial negativa tipo II e Poisson Gaussiana inversa para a variável resposta. Os resultados mostraram que a distribuição Poisson Gaussiana inversa foi mais adequada para análise dos dados, sendo possível modelar seus parâmetros de média e de dispersão em função das covariáveis em estudo.

De maneira geral, nossos resultados apontam que os pacientes com maior quantidade de déficits neurológicos (representado por valor mais elevado na graduação EDSS), menor idade e maior tempo de doença são os que mais apresentaram lesões corticais.

A principal limitação do trabalho é o tamanho da amostra. Dessa forma, sugere-se que futuros trabalhos confirmem nossos resultados, sobretudo, com um dimensionamento amostral superior ao nosso.

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. Notre Dame, v.19, n.6, p.716-723, 1974.

AKANTZILIOTOU, C.; RIGBY, R. A.; STASINOPOULOS, D. M. *"The R Implementation of Generalized Additive Models for Location, Scale and Shape." In M Stasinopoulos, G Touloumi (eds.), "Statistical Modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling,"* p.75-83. Chania, Greece, 2002.

ASCHERIO, A.; MUNGER, K. L.; LÜNEMANN, J. D. The initiation and prevention of multiple sclerosis. *Nature Reviews Neurology*, v.8, n.11, p.602-612, 2012. Nature Publishing Group.

BUUREN, S. v.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, v.20, n.8, p.1259-1277, 2001.

COMPSTON, A.; COLES, A. Multiple sclerosis. The Lancet, 2008. Elsevier Ltd.
CONFAVREUX, C.; VUKUSIC, S. *The clinical epidemiology of multiple sclerosis.* Neuroimaging clinics of North America, v.18, n.4, p.589-622, 2008.

DUNN, P.K.; SMYTH, G.K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, v.5, n.3, p.236-244, 1996.

HASTIE, T.J.; TIBSHIRANI, R.J. *Generalized Additive Models*. Chapman and Hall, London, 1990.

KING, G. Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator. *American Journal of Political Science*, v.33, n.3, p.762-784, 1989.

LASSMANN, H. Mechanisms of white matter damage in multiple sclerosis. *Glia*, v.62, n.11, p.1816-1830, 2014.

McCULLAGH, P.; NELDER, J.A. *Generalized linear models*. 2.ed. London: Chapman & Hall, 1989. 511p.

NELDER, J. A.; WEDDERBURN, W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A*, v.135, n.3, p.370-384, 1972.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RIGBY, R. A.; STASINOPOULOS, D. M. *"The GAMLSS project: a Flexible Approach to Statistical Modelling". In B Klein, L Korsholm (eds.), "New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling,"* p.249-256. Odense, Denmark, 2001.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society. Series C*, v.54, n.3, p.507-554, 2005.

RIGBY, R. A.; STASINOPOULOS, D. M. *A flexible regression approach using GAMLSS in R*. London Metropolitan University, London, 2009. 282p.

SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, v.6, n.2, p.461-464, 1978.

STASINOPOULOS, D.M.; RIGBY, R.A. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, Los Angeles, v.23, p.1-10, 2007.