Regressão por *splines* aplicada a dados longitudinais

Breno Gabriel da Silva[1], Naiara Caroline Aparecido dos Santos[2], Yana Miranda Borges[3], Terezinha Aparecida Guedes[4]

1. Introdução

Quando o objetivo é investigar a relação entre uma ou mais variáveis dependentes e independentes, tem – se que a análise de regressão é uma alternativa para o estudo do conjunto de dados, em que esta permite modelar e identificar a possível existência da relação já citada. Os modelos de regressão podem ser classificados em linear e não-linear, é dito linear se for considerado que a relação da variável resposta (variável dependente) às variáveis regressoras (variáveis independentes) é uma função linear de alguns parâmetros, caso contrário é dito não-linear, ou seja, se pelo menos uma das derivadas parciais em relação ao parâmetro é função de parâmetros que não são conhecidos e mais, quando não podem ser linearizados por meio de transformações. Devido ao fato de não ser possível muitas vezes descrever um fenômeno através de um modelo de regressão linear, novas técnicas estatísticas foram surgindo, como por exemplo, uma nova classe de modelos, os modelos de regressão não-lineares.

Uma alternativa de solucionar problemas de não-linearidade é utilizar a categorização dos preditores quantitativos, técnica não recomendada por muitos autores, devido à perda de informação Turner et al. (2010); Bennett et al. (2012). No entanto, já existem metodologias alternativas disponíveis na literatura. De acordo com Paulson (2007); Gottschall (2010), modelos não-lineares têm sido utilizados para descrever curvas de crescimento, tais como: Gompertz, von Bertalanffy e Brody, pois estes modelos constituem-se de parâmetros de fácil interpretação biológica, mas as regressões segmentadas, denominadas funções *splines*, também podem ser utilizadas. Pode-se citar ainda que uma das vantagens para se utilizar modelos não-lineares, em relação aos lineares, é que fornecem um bom ajuste com um número menor de parâmetros, ou seja, ameniza a dificuldade de interpretação em diversas situações.

Advoga Keele (2008), que *splines* são funções polinomiais segmentadas, unidas por pontos de corte denominados "nós", utilizados para ajustar uma curva a um conjunto de dados. De acordo com Oliveira (2011), os nós da regressão segmentada podem ser determinados pelo diagrama de dispersão ou do conhecimento prévio pelo pesquisador da situação em que este está inserido. Ainda de acordo com Oliveira (2011), tem-se um modelo não-linear quando os nós são estimados e um modelo linear quando os nós são determinados pelo pesquisador.

A vantagem de se utilizar *splines* é a facilidade de convergência, manipulação e são úteis quando um polinômio de grau baixo não se ajusta aos dados e quando se trata de dados biológicos

[1] Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. email: *omatematico.breno@gmail.com*.

[2] Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. email: *naicaroline2@gmail.com*.

[3] Programa de Pós-graduação em Bioestatística, Universidade Estadual de Maringá - UEM. email: *borges.yana@gmail.com*.

[4] Professora do Departamento de Estatística, Universidade Estadual de Maringá - UEM e orientadora no Programa de Pós-graduação em Bioestatística. email: *taguedes@uem.br*.

cuja natureza é oscilatória. As funções *splines*, têm sido amplamente utilizadas na análise de dados longitudinais, como por exemplo, no ajuste de curvas de lactação e crescimento, devido à praticidade nas interpretações dos parâmetros. Assim, neste trabalho, objetiva-se utilizar a regressão segmentada com o intuito de identificar a forma funcional da relação da variável regressora com a resposta, sendo esta metodologia aplicada a um conjunto de dados referente a 57 camundongos suíços, infectados por *Trypanosoma cruzi*, machos com 56 dias de idade, provenientes do biotério central da Universidade Estadual de Maringá. Por fim, as análises estatísticas serão realizadas no software R versão 3.5.0.

2. Materiais e métodos

2.1. Materiais

Os dados utilizados para a aplicação da metodologia supracitada foram coletados por Ferreira (2018), em que de acordo com a autora citada anteriormente os experimentos foram realizados por ensaios cegos, controlados e aleatorizados, cujo objetivo geral do experimento foi avaliar o efeito de bioterápico de soro de galinha em roedores experimentalmente infectados por *Trypanosoma cruzi* (agente etiológico da doença de chagas). O bioterápico utilizado foi produzido a partir de soro de *Gallus gallus domesticus* (galinha) sob parâmetro parasitológicos, clínicos e imunológicos. O projeto para a execução dos experimentos foi aprovado pelo Comitê de Ética em Pesquisa com Animais da Universidade Estadual de Maringá, Paraná, parecer CEUA 2401220716/2016.

Os animais foram distribuídos em grupos de tratamentos de modo que a média inicial dos pesos dos animais em cada grupo fossem aproximadamente iguais, em que, os grupos experimentais foram os seguintes: G1) CNI - (Controle não infectado) - Animais não infectados e não tratados (n=5); G2) G13cH - Animais tratados com bioterápico de soro de galinha 13cH (n=13); G3) G6cH - Animais tratados com bioterápico de soro de galinha 6cH (n=13); G4) GCI - (Controle infecção). Os camundongos foram infectados e não receberam tratamento (n=13) e G5) G3cH - Animais tratados com bioterápico de soro de galinha 3cH (n=13). A evolução do peso corporal foi acompanhada durante 12 semanas, que foi mensurado em uma balança semi-analítica.

Foram consideradas como variáveis de estudo a variável peso e a variável tempo da evolução do peso corporal a partir da 2ª semana, uma vez que após este período verifica-se a natureza oscilatória dos dados.

2.2. Métodos

Conforme Keele (2008), seja *x* o eixo dividido em *n* parcelas, uma *spline* é definida como uma função de regressão segmentada em que é imposta uma restrição de que os pontos de cada segmento se unam nos pontos de corte (nós). Assim, o modelo *spline* é dado por:

$$y = \sum_{j=1}^{n} \beta_{0j}\ x^j + \sum_{i=1}^{c} \beta_i\ (x - x_0)^n$$

Onde:

1. n é o grau da função *spline*;
2. c é a quantidade de nós;
3. x_0 o nó;

4. $(x - x_0) = \begin{cases} x - x_0, & x > x_0 \\ 0, & se\ x \leq x_0 \end{cases}$

A função descrita anteriormente é composta por boas propriedades, em que uma delas é de ser contínua, já que é composta por um polinômio contínuo em todo o seu domínio e em cada nó. Assim, tem-se que o modelo pode ser reescrito de acordo com as condições estabelecidas anteriormente por:

$$f(x_0) = \sum_{j=0}^{n} \beta_{0j}\ {x_0}^j + \sum_{i=1}^{c} \beta_i\ (x - x_0) = \sum_{j=0}^{n} \beta_{0j}\ {x_0}^j, para\ x > x_0$$

$$f(x_0) = \sum_{j=0}^{n} \beta_{0j}\ {x_0}^j + \sum_{i=1}^{c} \beta_i\ (0) = \sum_{j=0}^{n} \beta_{0j}\ {x_0}^j, para\ x \leq x_0$$

Devido à presença de alterações no ganho de peso corporal após a 2º semana de acompanhamento, será utilizada a regressão por *splines*, pois com esta metodologia é possível modelar fenômenos em transição e mudança. Em relação à qualidade do modelo, será utilizado o coeficiente de correlação ao quadrado entre os valores observados e os preditos. Já em relação à igualdade dos parâmetros das funções *splines*, será utilizado o método da razão de verossimilhança com aproximação dada pela estatística qui-quadrado. O ajuste dos nós assim como a estimação dos parâmetros será realizada por meio do método dos mínimos quadrados, cujas soluções serão obtidas pelo processo iterativo de Gauss-Newton, em que, as soluções são obtidas com algoritmos computacionais que se baseiam em uma aproximação por série de Taylor de primeira ordem para produzir uma linearização da função não-linear.

3. Resultados esperados

Com a aplicação dessa técnica espera-se identificar a forma funcional da relação da variável regressora com a resposta, considerando os diferentes grupos de tratamentos e o tempo de tratamento. Espera-se ainda identificar o melhor tratamento, ou seja, o(s) tratamento(s) que fornecerá a melhor sobrevida para os camundongos em vista de seu peso observado durante o tempo de estudo.

4. Agradecimentos

Agradecemos a CAPES pelo suporte financeiro.

5. Referencias Bibliográficas

BENNETT, S., BISHOP, A., DALGAMO, B., WAYCOTT, J.. & KENNEDY,G. Implementing web 2.0 technologies in higher education: A collective case study. Computers & Education, 59(2), 524-534, 2012.

OLIVEIRA, D.C. Funções splines para estudo de curvas de crescimento em ovinos cruzados. 68 f. Dissertação (Mestrado) - Curso de Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Viçosa, 2011.

FERREIRA, E. C. et al. Phosphorus protects cardiac tissue by modifying the immune response in rats infected by Trypanosoma cruzi. Cytokine, v. 102, p. 102-106, 2018.

GOTTSCHALL, C.S. Impacto nutricional na produção de carne-curva de crescimento. In: LOBATO, J.F.P.; BARCELLOS, J.O.J.; KESSLER, A.M. Produção de bovinos de corte. Porto Alegre: EDIPUCRS, 1999. p. 169-192.

KEELE, L.J. Semiparametric regression statistic for the science. John Wiley and Sons, 1º edição, 2008.

PAULSON, D.S. Handbook of regression and modeling – Application for the clinical and pharmaceutical industries. Chapman & Hall / CRC Biostatistics Series, 2007.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

TURNER, J.R., HUEMANN, M., ANBARI, F.T. and BREDILLET, C.N.Perspectives on projects. London and New York: Routledge, 2010.