Um estudo de regressão não linear

Aline Edlaine de Medeiros[1], Terezinha Aparecida Guedes[2] e Beatriz Regina Brum[3]

[1] Mestranda em Bioestatística PBE - UEM
[2] Programa de Pós-graduação em Bioestatística PBE - UEM
[3] Mestranda em Bioestatística PBE - UEM

RESUMO

Este estudo tem por objetivo revisar os conceitos dos modelos de regressão não linear, bem como, alguns métodos de estimação dos parâmetros e algumas características destes modelos. Será apresentado o ajuste de um modelo de regressão não linear, em que o objeto de estudo foi o envelhecimento precoce do milho híbrido OC705. Concluímos que em modelos não lineares existem certas peculiaridades na escolha do modelo e do chute inicial, que impactam diretamente na qualidade do ajuste.
Palavras chave: Regressão Não Linear; Milho híbrido OC705; Método do Mínimos Quadrados Ordinários, Método de Gauss-Newton.

1 INTRODUÇÃO

Uma das atividades mais importantes no estudo da Estatística é realizar a modelagem de dados associados a problemas reais. Existem distintas formas de realizar este estudo, através dos métodos de Regressão Linear Simples e Múltipla, Modelos Mistos e Regressão Não Linear, entre outros. Em geral, a forma de coleta dos dados e a forma da informações coletadas influenciam diretamente na escolha do modelo. Conjuntos de dados com características especificas, como crescimento e rendimento, cuja curva tenha formato parabólico ou senoidal, sugerem a utilização de modelos de Regressão Não Linear.

Existe uma teoria bem fundamentada sobre os modelos de Regressão Não Linear, sendo relativamente particular a metodologia utilizada para determinar os estimadores, a soma do quadrado dos resíduos, assim como, a própria obtenção dos gráficos associados a tal modelo. Neste sentido, este estudo apresentará algumas considerações sobre os aspectos de cunho teórico, com o objetivo de realizar uma breve revisão de literatura a respeito dos modelos de regressão não linear, bem como, realizar a devida aplicação desta modelagem em um conjunto de dados.

A aplicação realizada visa compreender o processo de envelhecimento precoce do milho híbrido OC705 em um experimento agronômico, bem como, a análise do ajuste e dos resíduos obtidos na modelagem, fornecendo ao leitor uma contextualização da relação entre a teoria e a modelagem em Regressão Não Linear.

2 METODOLOGIA

O modelo de regressão não linear (MRNL) visa descrever o comportamento de uma variável aleatória Y em função de X, em que X é denominada variável explicativa, variável independente, preditora ou covariável e Y é chamada de variável resposta ou variável dependente. Uma representação para este modelo é dada por:

$$y_i = f(x_i, \beta) + \epsilon_i, \tag{1}$$

em que y_i é uma observação da variável Y para o valor x_i da variável X, β é o parâmetro a ser estimado, ϵ_i são os erros aleatórios, que assumem-se ter média zero e variância desconhecida constante σ^2 seguindo uma distribuição Gaussiana, ou seja, $\epsilon_i \sim N(0, \sigma^2)$, e $f(.)$ é a função que liga a variável resposta com a variável preditora. Além disso, $f(.)$ deve ser não linear em relação aos parâmetros, isto é, a derivada da função f com relação aos parâmetros depende de pelo menos um dos parâmetros presentes no modelo (ZEVIANI, 2013) [4], e $E(y_i|x_i) = f(x_i, \beta)$.

Uma representação alternativa para o MRNL é a forma matricial, indicada abaixo:

$$y = f(x, \beta) + \epsilon \tag{2}$$

em que y é o vetor das observação da variável Y para a matriz de covariáveis X, β representa o vetor $[\beta_1, ..., \beta_k]'$ dos k parâmetros, e ϵ o vetor dos erros aleatórios.

Os modelos de Regressão não linear se dividem em 3 classes: crescimento, rendimento e compartimentados conforme foi descrito por Mattos (2013, p.17-18) [2]. A tabela 1 apresenta algumas funções costumeiramente utilizadas nos modelos de crescimento e rendimento.

Tabela 1: Modelos de crescimento e rendimento

Classe	Nome da Função	Função
Crescimento	Exponencial	$f(x,\beta) = \beta_1(1 - \exp(-\beta_2 x))$
Crescimento	Miscellaneous	$f(x,\beta) = \beta_1(\beta_2 + x)^{-1/\beta_3}$
Rendimento	Hiperbólico	$f(x,\beta) = \frac{x}{\beta_1+\beta_2 x+\beta_3 x^2}$
Rendimento	Holliday	$f(x,\beta) = \frac{1}{\beta_1+\beta_2 x}$

Diversos são os métodos para estimação dos parâmetros do MRNL, os principais métodos são: Mínimos Quadrados Ordinários, Mínimos Quadrados Ponderados, Mínimos Quadrados Generalizados e Verossilhança, aliados a algum método iterativo, tais como, Gauss-Newton, Newton-Raphson e Marquardt.

Na análise dos dados deste estudo, utilizou-se o software R, versão 3.3.1, por meio da função *nls*, que realiza o ajuste do MRNL através do método dos mínimos quadrados ordinários e cujo procedimento iterativo *default* para a estimação dos parâmetros é o método Gauss-Newton (REIS,2012)[3].

O método dos Mínimos Quadrados Ordinários (MMQO), (CUNHA, 2014)[1], é amplamente utilizado em problemas que envolvem a estimação de curvas, em áreas como a Geografia, Física e Biologia. No estudo de MRNL, o MMQO é empregado para obter uma estimativa do vetor β, com este intuito, o procedimento inicial é minimizar a soma de quadrados dos resíduos, isto é, a expressão 3 deve ser minimizada com relação aos parâmetros.

$$\begin{aligned} SQR(\beta) &= \parallel y - f(x,\beta) \parallel^2 \\ &= [y - f(x,\beta)]'[y - f(x,\beta)] \end{aligned} \tag{3}$$

Seja, $F(\beta)$ a matriz Jacobiana de $f(x,\beta)$, realizando algumas manipulações algébricas em 3, tais como, derivação e a expansão em Série de Taylor, e supondo ser conhecida uma aproximação do vetor $\hat{\beta}_0$ para o vetor β (REIS,2012)[3], pode-se mostrar que:

$$\beta - \hat{\beta}_0 = [F'(\hat{\beta}_0)F(\hat{\beta}_0)]^{-1}F'(\hat{\beta}_0)(y - f(x,\hat{\beta}_0)) \tag{4}$$

Considere $\Delta\hat{\beta}_0$ a variação entre os vetores β e β_0 estimado, isto é, $\Delta\hat{\beta}_0 = \beta - \hat{\beta}_0$. Pode-se definir $\hat{\beta}_1$, tal que, $\hat{\beta}_1 = \Delta\hat{\beta}_0 + \hat{\beta}_0$, em que $\Delta\hat{\beta}_0$ é o valor que minimiza a SQR e $\hat{\beta}_1$ é uma estimativa para o vetor β. Repetindo este procedimento p vezes tem-se o método iterativo de Gauss-Newton, em que p é o número de iterações (repetições) necessárias para a convergência, sob um certo valor de erro σ estabelecido pelo pesquisador, como segue:

$$\hat{\beta}_{p+1} = \hat{\beta}_p + [F'(\hat{\beta}_p)F(\hat{\beta}_p)]^{-1}F'(\hat{\beta}_p)(y - f(x,\hat{\beta}_p)) \tag{5}$$

3 APLICAÇÃO

Com o intuito de realizar uma aplicação do MRNL, realizou-se a modelagem dos dados relativos ao envelhecimento precoce das sementes de milho híbrido OC 705. O experimento para análise da germinação foi desenvolvido a uma temperatura de 41°C, e as sementes foram analisadas ao longo do tempo. Após análise primária dos dados, observou-se que o comportamento das sementes ao longo do tempo pode ser descrito pela função:

$$f(x \mid \beta_1, \beta_2) = -\beta_1 \exp(\beta_1 x) + \beta_2 \tag{6}$$

A etapa seguinte foi determinar os chutes iniciais para os estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$, tais chutes surgem principalmente do conhecimento prévio do pesquisador acerca dos parâmetros envolvidos no problema, bem como de algumas implementações disponíveis no R. Neste sentido, adotou-se os valores $\hat{\beta}_1 = 0.05$ e $\hat{\beta}_2 = 96$, e utilizou-se do pacote *nls* do software R, versão 3.3.1, para ajustar o modelo, através da função *nls()*, que forneceu a seguinte modelo:

$$\hat{y} = -0.0397 \exp(0.0397x) + 87.3813 \tag{7}$$

Outros valores iniciais foram analisados, contudo, a medida em que os valores iniciais se distanciavam dos valores estimados, era necessário uma maior quantidade de iterações para obter a convergência pelo método de Gauss-Newton, e em alguns casos (por exemplo, $\beta_1 = 50$, $\beta_2 = 6$, ou $\beta_1 = 30$, $\beta_2 = 10$) o método não convergiu. Outros modelos também foram avaliados, por exemplo, a Weibull dois parâmetros, mas, não houve convergência a partir destes modelos.

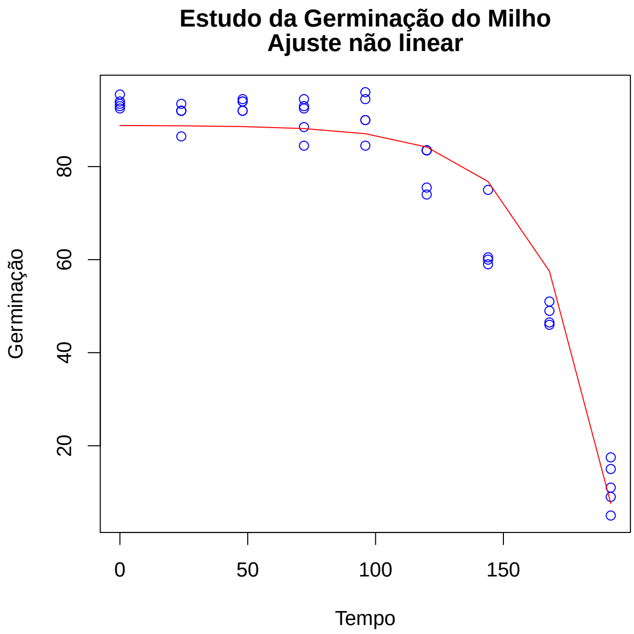

Figura 1: Dados vs Ajuste

O passo seguinte foi analisar a qualidade do ajuste, para isso, foram analisados informações como a normalidade e a independência dos resíduos, e análise gráfica conforme a figura 1. Nota-se que o modelo embora, graficamente pareça estar adequado, aspectos como a normalidade dos resíduos não se verificam, neste sentido, os outros aspectos inferênciais não foram analisados. A falta de normalidade dos resíduos torna os valores ajustados pelo modelo não confiáveis, sendo que uma alternativa seria buscar um novo modelo.

Referências

[1] da Cunha, J. C. V. O método dos mínimos quadrados: uma proposta ao ensino médio para o ajuste por retas.

[2] Mattos, T. d. B. Modelos não lineares e suas aplicações.

[3] Reis, R. M. d., et al. Modelos de regressão não linear para descrição do crescimento de plantas de alho.

[4] Zeviani, W. M., Ribeiro Jr, P. J., and Bonat, W. H. *Curso em modelos de regressão não linear*, 1 ed. ?, Campina Grande, 2013.