Universidad de San Carlos de Guatemala
Centro de Telemática (CETE)
Facultad de Agronomía - USAC

MODELOS DE REGRESSÃO NÃO LINEARES APLICADOS A PESQUISA AGRONÔMICA

Prof. Dr. Tiago Almeida de Oliveira

Março de 2022

- MEDIDAS DE DIGNÓSTICOS
- REFERÊNCIAS BIBLIOGRÁFICAS

Seleção e qualidade do modelo

- A seleção de um modelo deve ser coerente com o evento biológico em estudo, além de basear-se em um adequado ajuste aos dados amostrais.
- A seleção e avaliação de modelo são de extrema importância, pois determinarão o modelo que melhor prediz a variável dependente com um menor número de parâmetros. Para isso, ferramentas analíticas, tais como testes de hipóteses e critérios de informação são utilizados.

- Em situações onde existem repetições dos níveis de x é possível testar a adequação do modelo.
- O modelo não linear será comparado com outro modelo que é o modelo de médias, pois assume que x é um fator qualitativo.
- Esse modelo é o maior modelo, considerando a especificação para o parâmetro de média, que pode ser proposto para esses dados.

O teste da falta de ajuste consiste em verificar se o modelo não linear é tão bom quanto o modelo de médias, que seria o modelo com melhor ajuste possível, porém sem interpretação em termos de função e sem permitir predição de valores.

H_0: falta de ajuste não significativa (modelo adequado)
H_1 : falta de ajuste significativa (modelo não adequado)

Fonte	GL	SQ	QM	F
Falta de Ajuste	k - 2	SQ_{FA}	$SQ_{FA}/(k-2)$	QM_{FA}/QM_{EP}
ErroPuro	n - k	SQ_{EP}	$SQ_{EP}/(n-k)$	
Resíduos	n - 2	SQRes		

Tabela: Tabela da ANOVA

O coeficiente determinação (R^2) não é facilmente definido nestes modelos, pois um dos problemas com a sua definição é que este requer a presença de intercepto no modelo, parâmetro que nem sempre compõe os modelos não lineares.

- $$R^2 = 1 - \frac{SSE\,(\beta)}{SSE\,(\beta_0)} = 1 - \frac{\|y - \hat{y}\|}{\|y - \bar{y}\|} \tag{1}$$

Uma medida relativamente próxima ao R^2, no caso dos modelos não lineares é o coeficiente de determinação ajustado

- $$R^2_{adj} = 1 - \frac{n-1}{n-p}\left(1 - R^2\right) \tag{2}$$

- log-verossimilhança (maior é melhor)

$$łł = \frac{n}{2} log\left(2\pi\right) - \frac{n}{2} log\left(\hat{\sigma}^2\right) - \frac{\left\|y - X\hat{\beta}\right\|}{2\hat{\sigma}^2} \tag{3}$$

$$łł = \frac{n}{2} log\left(2\pi\right) - \frac{n}{2} log\left(SSE/n\right) - \frac{n}{2}$$

$$\hat{\sigma}^2 = SSE/n = \left\|y - X\hat{\beta}\right\|/n$$

- Critérios de Informação de Akaike

$$AIC = 2\left(p + 1\right) - 2łł \tag{4}$$

- Critérios de Informação de Akaike

$$BIC = log\left(n\right)\left(p + 1\right) - 2łł \tag{5}$$

- Matriz de projeção

$$\hat{y} = Hu$$

$$H = X\left(X'X\right)^{-1}X' \tag{6}$$

H é simétrica e idempotente. O posto de H é $tr(H) = p$

- Alavancagem

$$h_i = H_{ii}$$

$$h = diag(H)$$

- Resíduos ordinários $V(\hat{e}) = \sigma^2(I - H)$

$$\hat{e}_i = y_i - \hat{y}_i$$

$$\hat{e} = y - \hat{y}$$

$$\hat{e} = y - X\hat{\beta}$$

- Resíduos padronizados (ou internamente studentizados),

$$r_i = \frac{\hat{e}_i}{s(\hat{e}_i)} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}$$

- Resíduos studentizados (ou externamente studentizados),

$$t_i = \frac{\hat{e}_i}{s(\hat{e}_{i(-i)})} = \frac{\hat{e}_i}{\hat{\sigma}_i\sqrt{1-h_i}}$$

$$\hat{\sigma}^2_{-i} = \frac{(n-p)\hat{\sigma}^2 - \frac{\hat{e}_i^2}{1-h_i}}{(n-1)-p}$$

Residuos Projetados

De acordo com Cook e Tsai (1985) pode-se empregar uma expansão em série de Taylor para aproximar e investigar o comportamento dos resíduos ordinários na regressão não linear.

Entretanto, em alguns casos, esses resíduos podem produzir resultados enganosos quando usados em métodos de diagnósticos análogos aos da regressão linear, não refletindo de forma correta a distribuição dos erros.

Sugere-se, então a utilização dos resíduos projetados, sendo ele superior em qualidade de diagnóstico.

Seja $\tilde{X} = \frac{\partial \mu(\beta)}{\partial \beta_r}$, i = 1, . . . , n e r = 1, . . . , p a matriz de derivadas de primeira ordem avaliada em θ. Em que, $H_1 = S(S^T S)^{-1} S$ é operador de projeção ortogonal do espaço gerado sobre a projeção $S = (I - H)T$, T é uma matriz $n \times q$ obtida pelos vetores não nulos, e derivadas de segunda ordem e $H = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$.

Sendo V uma matriz $n \times (p + q)$ definida como $V = (\tilde{X}, S)$ e $H_2 = V(V^T V)^{-1} V^T$ o operador de projeção ortogonal em C(V).

- Resíduos projetados $(I - H_2)_r = (I - H)_\epsilon (I - H_1)_\epsilon$

onde H, H_1 e H_2 são matrizes, r o resíduo estimado e ϵ é erro experimental

$$E = (I - H_2)_r = 0,$$

$$Var = \{(I - H_2)_r\} = \sigma^2 (I - P_2)$$

- Distância de Cook

$$D_i = \frac{(\hat{y} - \hat{y}_{i(-i)})^T}{p\hat{\sigma}^2} = \frac{1}{p} \cdot \frac{h_i}{(1-h_i)} \cdot \frac{\hat{e}_i^2}{\hat{\sigma}_i^2(1-h_i)} \tag{7}$$

- DFfits,

$$Dffits_i = \frac{(\hat{y} - \hat{y}_{i(-i)})}{\hat{\sigma}_{(-i)}\sqrt{h_i}} = t_i\left(\frac{h_i}{1-h_i}\right)^{1/2} \tag{8}$$

- DFbetas,

$$dbetas_i = \frac{\hat{\beta} - \hat{\beta}_{-i}}{\hat{\sigma}_{(-i)}\sqrt{diag((X^T X)^{-1})}} \tag{9}$$

$$\hat{\beta}_{-i} = \hat{\beta} - \frac{\hat{e}_i}{1-h_i} \cdot (X^T - X)^{-1} x_i$$

Medidas de não linearidade

Indicam se o grau de linearidade num problema de estimação não linear é pequeno o suficiente para justificar o uso da teoria dos modelos lineares como aproximação para os não lineares.

- Box (1971) desenvolveu uma fórmula para estimar o viés dos estimadores de máxima verossimilhança.
- Bates e Watts (1980) desenvolveram medidas de não linearidade baseadas no conceito geométrico de curvatura.

Considere o modelo

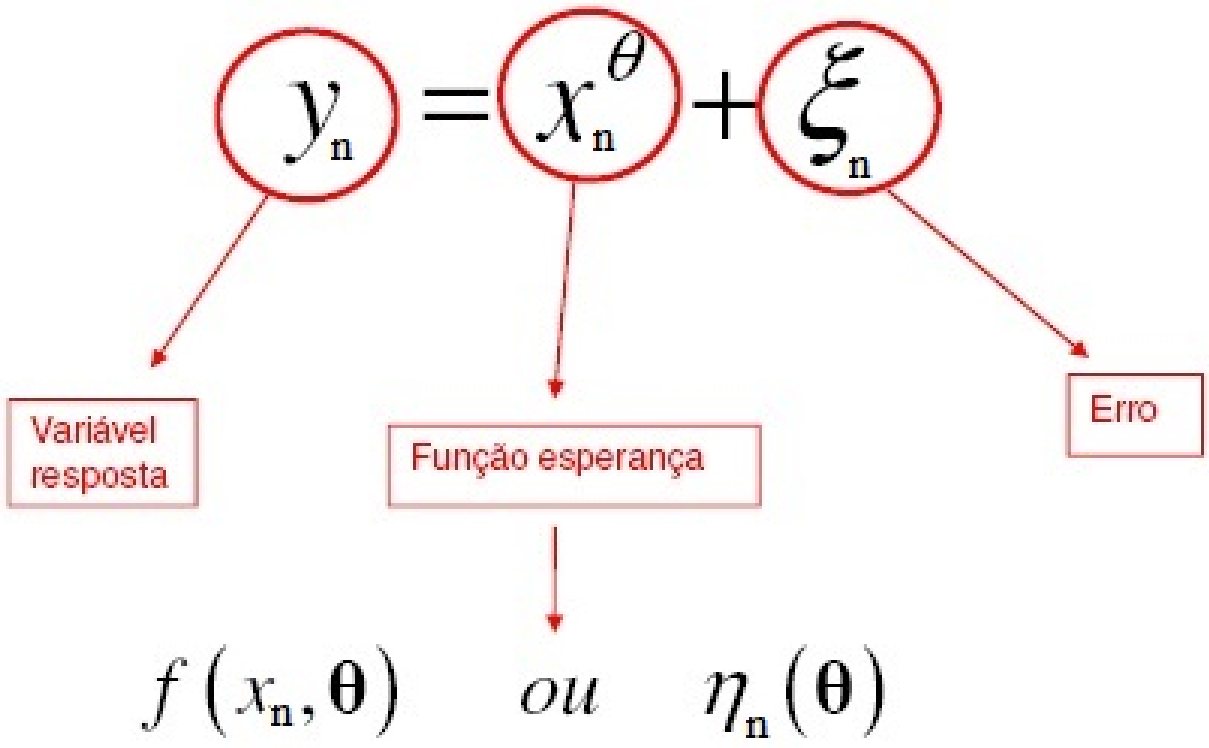

Figura: Representação dos dados no espaço amostral (RATKOWSKY, 1983)

Curvatura de Bates e Watts (1980)

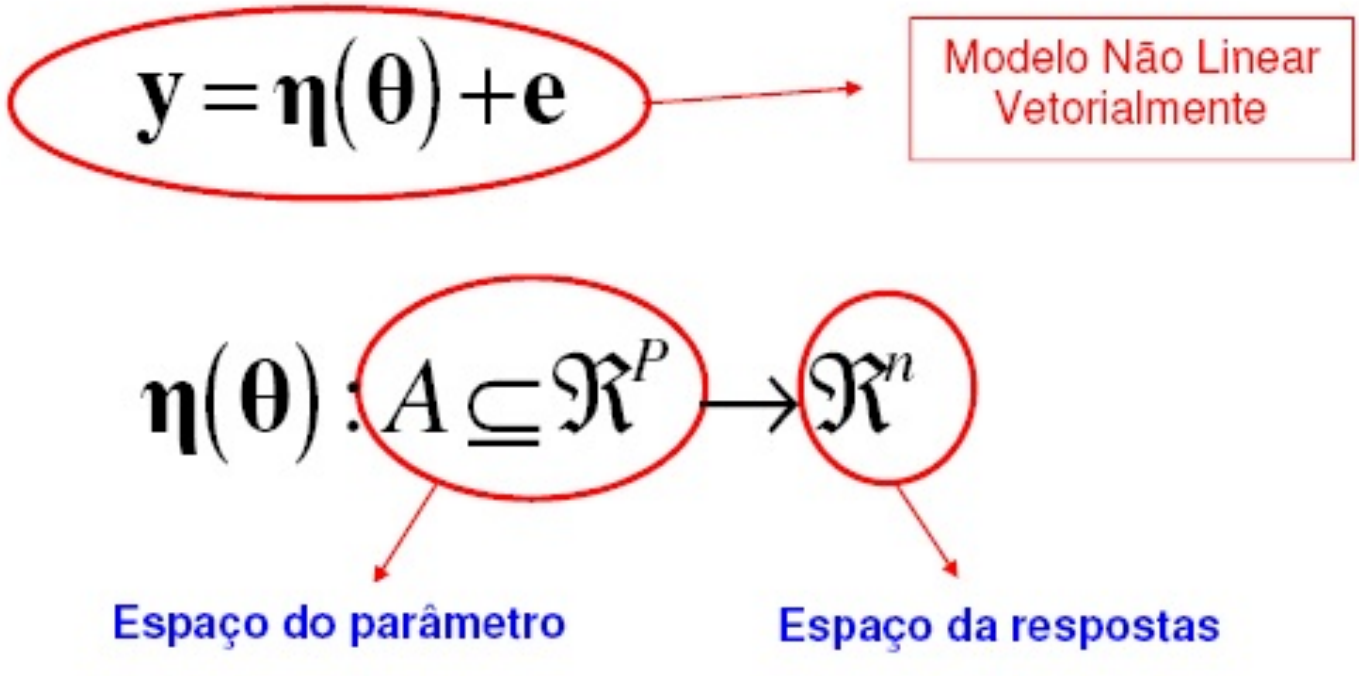

Figura: Representação dos dados no espaço amostral (RATKOWSKY, 1983)

Serão comparados, a seguir, um modelo linear e um modelo não linear para o caso de $n = 2$ e $p = 1$.

$$Y_i = \beta x_i, i = 1, 2$$

Formada pelos pontos $x\beta = (x_1, x_2)\beta, \beta \in R$ e as soluções possiveis para $x\beta$ são:

$$x\beta^{(i+1)} - x\beta^{(i)} = (x_1, x_2)\Delta, \ i = 1, 2, ...,$$

portanto, se as soluções para β
forem igualmente espaçadas, então os valores ajustados correspondentes serão, também, igualmente espaçados

Caso linear

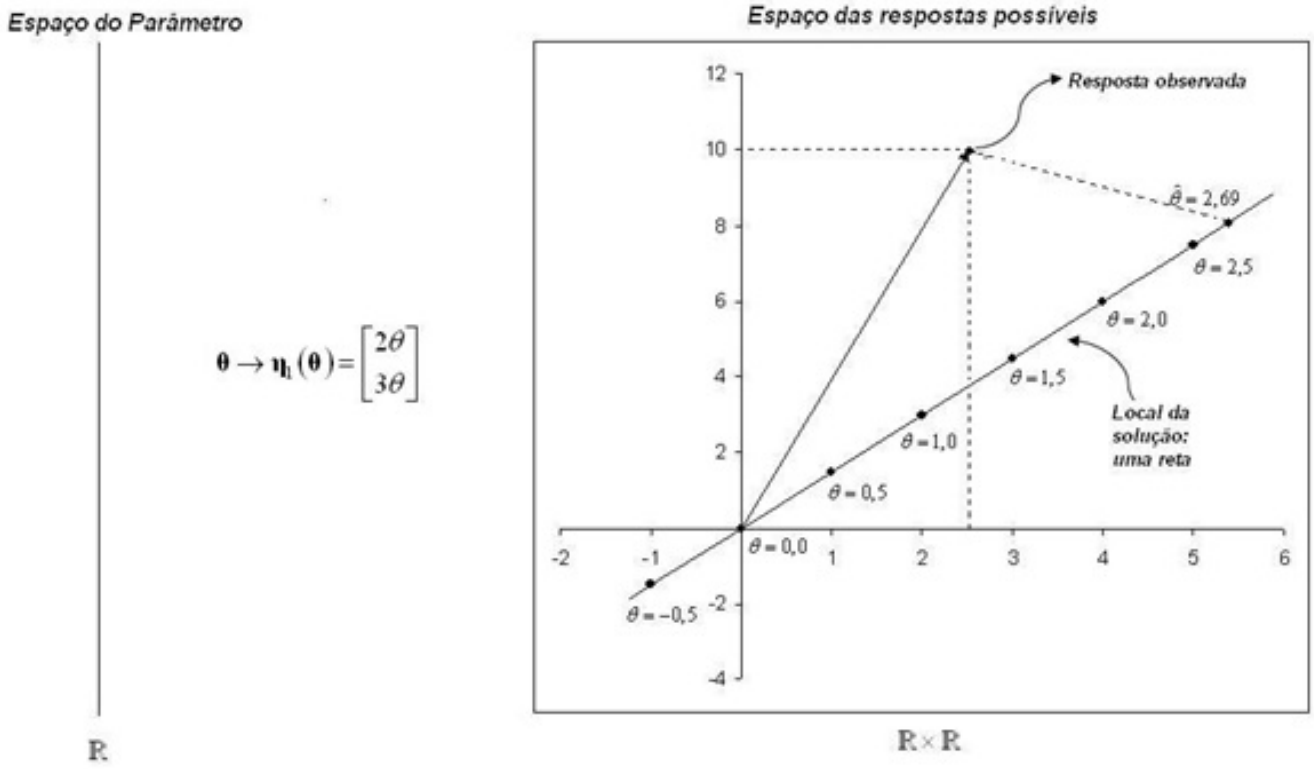

Figura: Espaço do parâmetro, espaço resposta e local da solução para um modelo linear com n=2 e p=1

Caso não linear

Considere o modelo $y_i = x_i^{\beta} + \epsilon_i, \quad i = 1, 2$

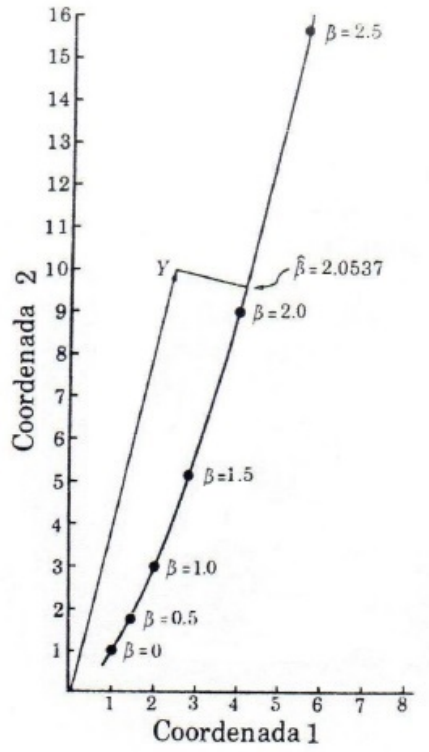

Figura: Espaço do parâmetro, espaço resposta e local da solução para modelos não lineares n=2, p=1

Nesse caso, o espaço de estimação não é mais uma reta, e sim uma curva ao redor da estimativa de máxima verossimilhança $\hat{\beta} = 2,05$.

- A curva correspondente aos pontos $(2^{\beta}, 3^{(\beta)})^T$ com β variando em espaçamentos iguais a 0,5. Os pontos do espaço de estimação não são igualmente espaçados como ocorre no modelo linear.
- Quanto mais essa curva se afasta da reta tangente em $\hat{\beta}$ maior será a não linearidade intrínseca do modelo, e quanto mais desiguais forem os espaçamentos entre os pontos do espaço de estimação, maior será a não linearidade aparente causada pela parametrização do modelo.

A não linearidade de um modelo pode ser devida a duas causas.

- A primeira é a curvatura real do modelo ou intrínseca, que é invariante com qualquer tipo de reparametrização.
- A segunda é a curvatura devida à forma como os parâmetros aparecem no modelo. Essa última pode ser eliminada ou pelo menos reduzida por meio da reparametrização.

Seja o modelo normal não-linear descrito anteriormente com a seguinte reparametrização:

$$y_i = x_i^{\log \phi} + \epsilon_i, \; i - 1, 2$$

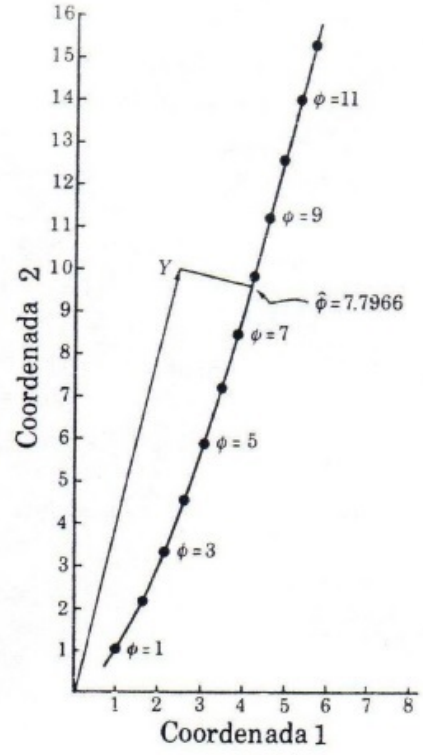

Figura: Espaço do parâmetro, espaço resposta e local da solução para modelos não lineares n=2, p=1

- Os pontos da curva $(2^{\log \phi}, 3^{\log \phi})^T$ com espaçamentos iguais a $1,0$ para ϕ.

- Os espaçamentos entre os pontos correspondentes são praticamente iguais, indicando que o grau de não linearidade aparente foi, substancialmente, reduzido com essa reparametrização.

- A curvatura do espaço de estimação continua com a mesma forma anterior, com era de se esperar.

O vício de Box (1971)

Box (1971) chegou à seguinte fórmula para o cálculo de viés em casos uni variados homocedásticos

$$B\left(\hat{\theta}\right) = -\frac{\sigma^2}{2}\left[\sum_{i=1}^{n} \mathrm{F}^j \mathrm{F}^{jT}\right]^{-1} \sum_{i=1}^{n} \mathrm{F}^j tr\left[\left(\sum_{i=1}^{n} \mathrm{F}^j \mathrm{F}^{jT}\right)^{-1} \mathrm{H}^j\right] \qquad (10)$$

em que, $F_i (= F_u)$ é o vetor $(p \times 1)$ da primeira derivada do modelo e H_u é a matriz $p \times p$ de segunda derivada, ambos com respeito aos elementos β, avaliados em $x_i, i = 1, \ldots, n$. Na prática, são usados $\hat{\theta}$ e $\hat{\sigma^2}$ no lugar das quantidades desconhecidas e *tr* indica a operação traço.

É comum expressar o valor da estimativa do vício em porcentagem, ou seja,

$$\%B\left(\hat{\theta}\right) = \frac{100 \times B\left(\hat{\theta}\right)}{\hat{\theta}} \tag{11}$$

valores acima de 1%, em valor absoluto, indicam comportamento não linear.

- A importância de se avaliar o vício reside em indicar quais parâmetros do modelos mais contribuem para o afastamento do comportamento linear.

Reamostragem bootstrap

Introduzido por Efron (1979), o método *bootstrap* é um procedimento muito geral de reamostragem para estimar distribuições de estatísticas com base em observações independentes.

- *Bootstrap* não paramétrico - não são feitas pressuposições sobre a distribuição dos dados
- *Bootstrap* paramétrico - reamostragens são feitas a partir de uma função de distribuição conhecida

Quando as hipóteses de normalidade e da aproximação assintótica no modelo de regressão não linear se tornam questionáveis, recomenda-se o uso da técnica *bootstrap* de estimação.

- Ratkowsky (1983), Souza (1998) sugerem estudos de simulação como alternativa e ajuda em possíveis reparametrizações em modelos não lineares

- Ratkowsky (1983), Seber e Wild (1989) exploram o uso do *bootstrap* paramétrico no diagnóstico de regressão não linear

Modelos Multiresposta

Modelos de regressão com mais de uma variável resposta podem ser classificados em dois grupos.

- Os modelos de regressão linear multivariados, sendo aqueles em que cada variável dependente tem a mesma relação linear funcional com as variáveis independentes, mas com diferentes coeficientes.

- Os modelos de regressão multiresposta, em que as variáveis dependentes podem ter relações funcionais diferentes, lineares ou não lineares, com as variáveis independentes.

- As análises de regressão multiresposta são consideradas difíceis de serem executadas na prática, e em geral, os métodos de aproximação para obtenção das estimativas devem ser apropriados.

- Mesmo quando todas as funções do modelo são lineares, as estimativas dos parâmetros devem ser calculadas de forma iterativa e a distribuição exata das estimativas não é facilmente calculada.

Segundo Bates e Watts (1988), as análises de dados utilizando-se modelos multiresposta são caracterizadas por experimentos, com M respostas medidas em N observações e que os modelos das M respostas dependem de um total P de parâmetros θ.

$$y_{nm} = f_m(X_n; \theta) + z_{nm} \qquad \text{com} \;\; n = 1, \ldots, N \;\; , m = 1, \ldots, M, \quad (12)$$

em que y_{nm} são variáveis aleatórias associadas com as medidas dos valores da m-ésima resposta na n-ésima observação, f_m é a função do modelo para a m-ésima resposta dependendo de alguns ou todos os conjuntos experimentais x_n e alguns ou todos os parâmetros θ, e Z_{nm} são os termos dos erros.

O ajuste do modelo multiresposta se deu por meio da combinação de modelos logístico com três parâmetros representados da seguinte forma

$$f_1\left(X_n, \theta\right) = \frac{\theta_{11}}{1+\exp[(\theta_{21}-x_n)/\theta_{31}]}$$

$$f_2\left(X_n, \theta\right) = \frac{\theta_{12}}{1+\exp[(\theta_{22}-x_n)/\theta_{32}]}$$

- Para obtenção das estimativas dos parâmetros do modelo multiresposta é necessário o uso de valores iniciai.
- As estimativas dos parâmetros do modelo unirespostas foram combinadas, e por meio dessa combinação originou-se um vetor θ de parâmetros, com θ sendo um vetor que engloba todos os parâmetros.

Aplicação

- Os dados utilizados são referentes ao trabalho que foi desenvolvido no Laboratório de Física do Solo, do Departamento de Engenharia Rural, da Escola Superior de Agricultura ‘Luiz de Queiroz” - ESALQ/USP.

- O objetivo do trabalho foi montar um ensaio experimental em laboratório a fim de representar o comportamento do transporte dos teores de potássio, no solo Latossolo Vermelho Amarelo (LVA).

O ensaio experimental foi realizado utilizando-se uma coluna segmentada de acrílico (Figura 1).

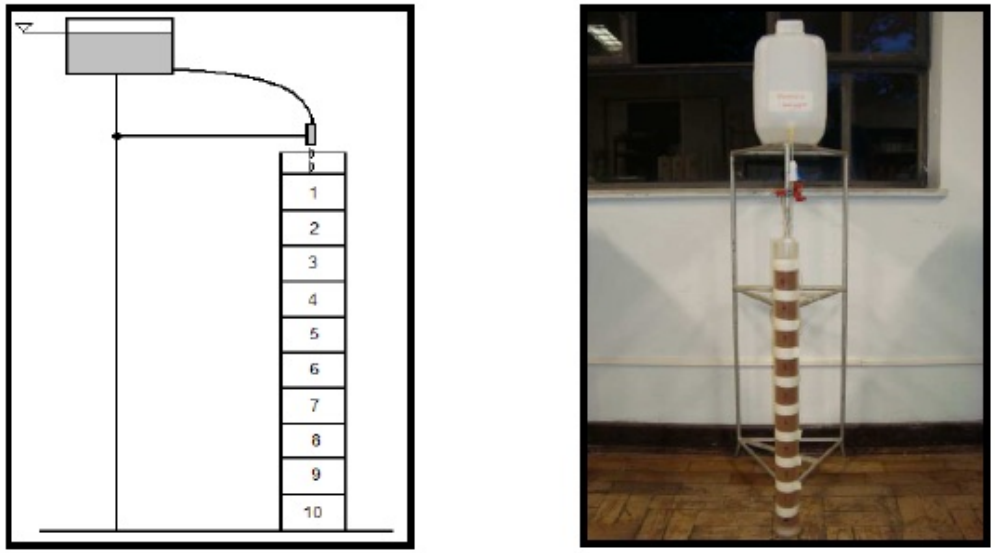

Figura: Esquema ilustrativo do ensaio experimental (a) montado em laboratório para elaboração dos perfis de água e solutos em coluna segmentada (b)

O modelo de logístico ajustado foi

$$f(x_n; \theta) = \frac{\theta_1}{1 + \exp\left[(\theta_2 - x_n)/\theta_3\right]}, \tag{13}$$

- $f(x_n; \theta)$ representa o teor do potássio ao longo da profundidade (m);
- x_n é a profundidade em m que define o perfil do solo;
- Se $\theta_3 > 0$, então θ_1 é a assíntota horizontal quando $x \to \infty$ e 0 é o assíntota horizontal quando $x \to -\infty$. Se $\theta_3 < 0$, esses papéis são invertidos.
- O parâmetro θ_2 é o valor de x para o qual a resposta é $\theta_1/2$. Este é o ponto de inflexão da curva.
- O parâmetro de escala θ_3 representa a distância no eixo x entre o ponto de inflexão e o ponto em que a resposta é $\theta_1/\left(1 + e^{-1}\right) \approx 0,73\theta_1$.

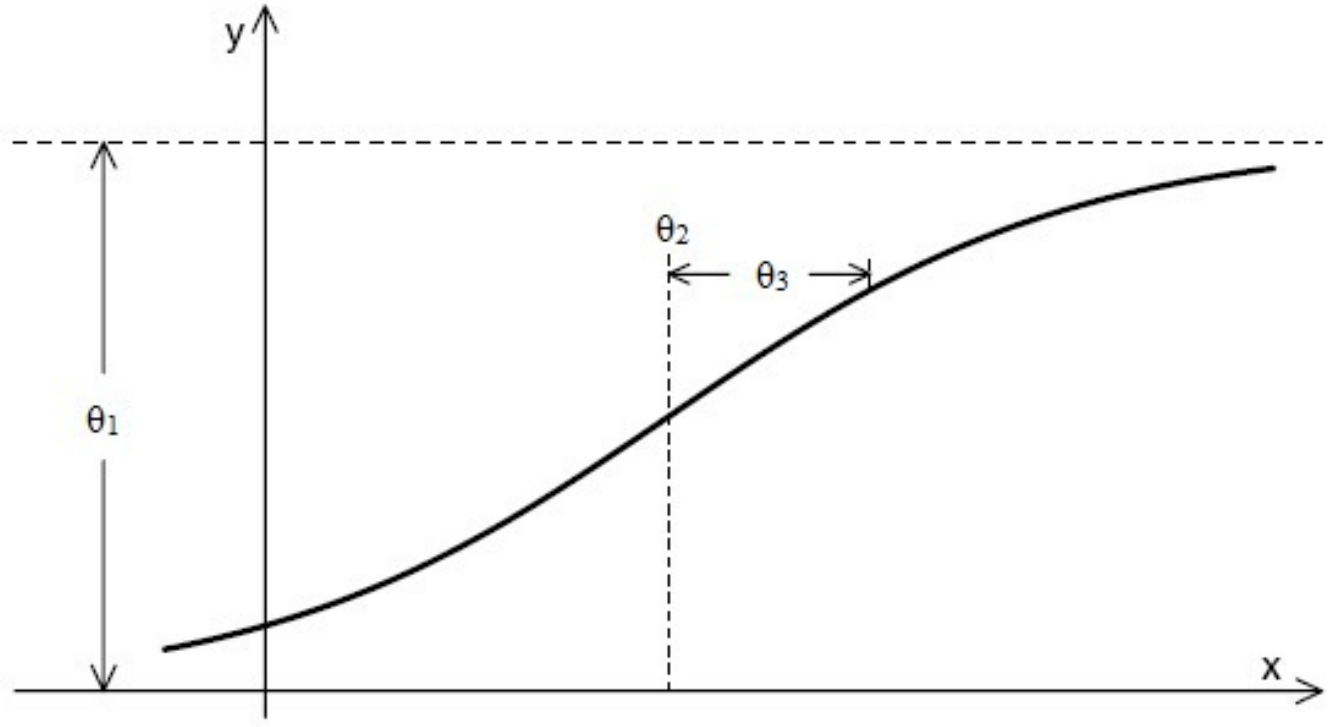

Figura: Representação gráfica do modelo logístico com três parâmetros.

BATES, D. M.; WATTS, D. G. Relative curvature measures of nonlinearity (With discussion). **Journal of the Royal Statistical Society**, Ser. B, v.42, n. 1, p. 1-25, 1980.

BATES, D.M.; WATTS, D.G. **Nonlinear Regression Analysis and its Applications**. New York: Wiley series in probability e mathematical statistics, 1988. 365p.

BOX, M. J. Bias in nonlinear estimation. **Journal of royal statistical society**. Serie B. Methodological, London, v. 33, n. 2, p. 171-201, Apr. 1971.

COOK, R.D. Detection of influential observations in linear regression. **Technometrics**, Alexandria, v.19, p.15-18, 1977.

COOK, R.D.; WEISBERG, S. **Residuals and influence in regression**. New York: Chapman & Hall, 1982. 280p.

EFRON, B. Bootstrap methods: another look at the jackknife. **Ann. Stat**., Beachood, v.7, p. 1-26, 1979.

MARCHI, G.; GUILHERME, L.R.G.; LIMA, J.M.; CHANG, A.C.; FONTES, R.L. Adsorption/desorption of organic anions in Brazilian Oxisols. **Communications in Soil Science and Plant Analysis**, v.37, p.1367-1379, 2006b.

MARTINEZ, M.A. **Modeling subsurface drainage in clermont silt loam using finite element technique.** West Lafayette, 1989. 173 p. Thesis (Ph.D) - Purdue University.

MARTINEZ, E. Z.; NETO, F. L. Estimação intervalar via bootstrap. **Rev. Mat. Estat**., São Paulo, v. 19, p. 217-251, 2001.

MAZUCHELI, J.; ACHCAR, J. A. Algumas considerações em regressão não linear. **Acta Scientiarum**, Maringá, v. 24, n. 6, p. 1761-1770, 2002.

MIRANDA, J.H. **Modelo para simulação da dinâmica de nitrato em colunas verticais de solo não saturado**. 2001. 79 f. Tese (Doutorado em Irrigação e Drenagem) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2001.

MIRANDA, J.H.; DUARTE, S.N.; LIBARDI, P.L.; FOLEGATTI, M.V. Simulação do deslocamento de potássio em colunas verticais de solo não-saturado. **Engenharia Agrícola**, Jaboticabal, v. 25, n. 3, p. 677-685, set./dez. 2005.

HUTSON, J.L.; WAGENET, R.J. LEACHM: Leaching estimation and chemistry model: a process-based model of water and solute movement, transformations, plant uptake and chemical reactions in the unsaturated zone; version 3.0. New York: **Cornell University**. 1992. 131p.

PEREIRA, J. M.; MUNIZ, J. A.; SÁFADI, T. SILVA, C. A. comparação entre modelos para predição do nitrogênio mineralizado: uma abordagem bayesiana. **Ciênc. agrotec**., Lavras, v. 33, Edição Especial, p. 1792-1797, 2009.

PINHERO, J.C.; BATES, D.M. **Mixed-Effects Models in S and S-PLUS**, Springer-Verlag, New York. 528 p. 2002.

R Development Core Team (2012). R: A language and enviroment for statistical computing. Disponível em: <http://www.R-project.org>. Acesso em: 10 mar. 2013.

RATKOWSKY, D.A. **Nonlinear Regression Modeling**: a Unified Practical Approach. New York: Marcel Dekker, 1983.

SEBER, G. A. F.; WILD, C. J. **Nonlinear regression**. New York: J. Wiley, 1989.

SIMUNEK, J.; SUAREZ, D.L.; SEJNA, M. The UNSATCHEM software package for simulating the one-dimensional variably saturated water flow, heat transport, carbon dioxide production and transport, and multicomponent solute transport with major ion equilibrium and kinetic chemistry, Version 2.0. **Research Report**, Riverside, n. 141, U.S. Salinity Laboratory, USDA, ARS, 186 p. 1996.

SIMUNEK, J.; VAN GENUCHTEN, M.Th.; SEJNA, M. **The HYDRUS-1D Software Package for Simulating the Movement of Water, Heat, and Multiple Solutes in Variably Saturated Media, Version 3.0, HYDRUS Software Series 1**. Riverside: Department of Environmental Sciences, University of California Riverside, 2005.

SOUZA, G. S. **Introdução aos modelos de regressão linear e não linear**. Brasília: EMBRAPA-SPI/Embrapa-SEA, 1998. 489p.

VAN GENUCHTEN, M.T. van.; WIERENGA, P.J. Solute dispersion coefficients and retardation factors. In: BLACK, C.A. (Ed.) Methods of soil analysis. Madison: **Soil Science Society of America**, pt. 1: Physical and mineralogical methods: p.1025-1054. (American Society of Agronomy, 9). 1986.