Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento Estatística

José Reinaldo de Lima Silva

Ajuste do modelo multiresposta na determinação espectroscópica do nitrogênio e fósforo na composição do solo

Campina Grande/PB
Agosto de 2014

José Reinaldo de Lima Silva

Ajuste do modelo multiresposta na determinação espectroscópica do nitrogênio e fósforo na composição do solo

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientadora:
Prof[a]. Dr[a]. Ana Patricia Bastos Peixoto

Campina Grande/PB
Agosto de 2014

S586a Silva, José Reinaldo de Lima.
Ajuste do modelo multiresposta na determinação espectroscópica do nitrogênio e fósforo na composição do solo [manuscrito] / Jose Reinaldo de Lima Silva. - 2014.
35 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Profa. Dra. Ana Patricia Bastos Peixoto, Departamento de Estatística".

1. Espectroscopia. 2. Infravermelho. 3. Modelo não linear. I. Título.

21. ed. CDD 547

José Reinaldo de Lima Silva

Ajuste do modelo multiresposta na determinação espectroscópica do nitrogênio e fósforo na composição do solo

Trabalho Acadêmico Orientado apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: 12/08/2014

Banca Examinadora:

Ana Patricia Bastos Peixoto

Profa. Dr^{a}. Ana Patricia Bastos Peixoto
Universidade Estadual da Paraíba-DE/CCT
Orientadora

Ricardo Alves de Olinda

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba-DE/CCT
Examinador

Tiago Almeida de Oliveira

Prof. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba-DE/CCT
Examinador

Dedicatória

A minha mãe, Maria Celia, minha inspiração maior em buscar de um futuro melhor
A minha filha, Maria Eduarda, razão da minha vida
Com muito amor dedico.

Agradecimentos

Agradeço sempre em primeiro lugar a Deus, pois nada é possível sem a força dele.

A minha mãe, Maria Celia Barboza de Lima, pois foi pensando em dar uma velhice com mais dignidade para ela que me inspirou a buscar, inicialmente, um curso de nível superior.

A minha filha, Maria Eduarda de Lima Silva, razão de viver, pois por você serei a cada dia mais forte.

A minha esposa, Rutinéia, porque ela sabe de toda luta, das noites sem dormir e da minha muita ausência no aconchego do nosso lar.

A minha filha Mayara, pois sei que sirvo de espelho e de inspiração para ela e isso reflete em mim como fonte ilimitada de força.

Ao meu pai, José Pereira, e aos meus irmãos que carinhosamente os chamo pelo apelido da nossa infância, Beto, Cado, Fia e Danda.

Aos meus grandes amigos, Jorge Oliveira, Severino pai, Severino filho, Bruno Carneiro, Josilene Cordeiro e outros que aqui cometo a injustiça de não citar.

A Ana Patricia Bastos Peixoto, mais que uma professora, mais que uma orientadora, uma amiga que por muitas vezes acreditou em mim quando nem eu mesmo acreditava.

Todos vocês estão em meu coração. À vocês, o meu amor dedico!

Resumo

A técnica de espectroscopia infravermelho é utilizada na análise orgânica qualitativa. De forma à contribuir com essa e outras técnicas, os modelos de regressão não linear são utilizados quando se deseja descrever a relação entre uma variável resposta Y e uma variável explicativa X. Para se encontrar as estimativas dos parâmetros desses modelos, seja por mínimos quadrados ou por máxima verossimilhança entre outros, é necessário o uso de métodos iterativos como os de Gauss-Newton, Quasi-Newton e Newton-Raphson, bem como verificar a qualidade do ajuste do modelo realizando-se uma análise dos resíduos. Os modelos multirespostas baseiam-se em uma relação para o qual os modelos são geralmente especificados como formas não lineares nos parâmetros. Nesse contexto, foi analisada sete amostras de (100g) de solo misturadas com fertilizantes nitrogênio, fósforo, na gama de concentrações de 1-15% para serem analisadas por meio de refletância de infravermelho com objetivo de ajustar dois modelos de regressão logístico e um modelo de regressão multiresposta. Verificou-se que os parâmetros do modelo logístico foram todos significativos ao nível de 5% de significância e que o modelo se ajustou bem aos dados. No modelo multiresposta, todos os parâmetros da função foram significativos a 95% de confiança, pois os intervalos assintóticos não incluem a valor zero. Tendo em vista os resultados encontrados para o modelo multiresposta, pode-se afirmar que os resultados inferencias obtidos foram satisfatórios.

Palavras-chave:Espectroscopia; Infravermelho; Modelo não linear; Multiresposta

Abstract

The technique is infrared spectroscopy used in qualitative organic analysis. In order to contribute to this and other techniques, models of nonlinear regression is used when you want to describe the relationship between a response variable Y and an explanatory variable X. To find the parameter estimates of these models, either by least squares or by maximum likelihood among others, it is necessary to use iterative methods such as Gauss-Newton, Quasi-Newton and Newton-Raphson and check the quality of fit of the model and analyze conduct a waste. The models multiresponse are based on a relationship to which the models are usually specified as non-linear forms in the parameters. In this context, we analyzed seven samples (100g) mixed with soil fertilizer nitrogen, phosphorus concentrations in the range of 1-15 % to be analyzed by means of infrared reflectance in order to fit two logistic regression models and multiresponse regression model. It was found that the parameters of the logistic model are all significant at the 5 % significance and that the model fits well to the data. In multiresponse model, all function parameters were significant at 95 % confidence intervals for the asymptotic not include the zero value. Considering the results for the multiresponse model, it can be stated that the results were satisfactory inferences.

Key-words: Spectroscopy; Infrared; Nonlinear model; Multiresponse

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

As técnicas espectroscópicas têm sido cada vez mais utilizadas na agricultura e na industria de alimentos nas últimas décadas. Os métodos de análise clássicos, ou seja, laboratórios de análises das amostras de alimentos, do solo ou da planta são caros, são demorados e exigem muito trabalho tanto na coleta da amostra, quanto pela avaliação em laboratório. Para suprir estes problemas, várias técnicas instrumentais, como reflectância, espectroscopia, espectroscopia de fluorescência, etc., tem sido utilizadas para a determinação da composição dos nutrientes, estas técnicas analíticas são relativamente vantajosas, uma vez que são rápidas e não destrutivas (SEN, 2003).

A utilização de modelos de regressão não lineares é bastante requerida quando se deseja descrever uma relação entre uma variável resposta Y e uma variável explicativa X de modo a se avaliar o comportamento de uma variável em função da outra. Dessa forma, modelos de regressão não lineares são bastante utilizados em processos como crescimento, decaimento, nascimento, mortalidade, etc. Para Ratkowsky (1983) um modelo de regressão é não linear nos parâmetros, se pelo menos uma derivada parcial da variável dependente, com relação a algum parâmetro existente no modelo, depende de algum parâmetro. Existe ainda os modelos que a princípio são ditos não lineares mas que ao ser realizado algum procedimento de transformação se tornam modelos de regressão lineares com novos parâmetros.

Sempre que o pesquisador se depara com a necessidade de compreender algum fenômeno o mesmo passa a obervar uma variável específica que traga informações a cerca do fenômeno que está sendo estudado, porém, nem sempre as indagações são respondidas analisando-se apenas uma variável mas, um conjunto de variáveis. Os modelos de regressão de resposta múltipla são aqueles em que as variáveis dependentes possuem relações funcionais diferentes com as variáveis independentes. Mesmo sendo difíceis de serem executados esses modelos são importantes, principalmente, porque o mesmo baseia-se em uma relação para o qual os modelos são geralmente especificados como formas não lineares nos parâmetros.

Sendo assim, esse trabalho tem por objetivo ajustar, inicialmente, um modelo de regressão logístico para as variáveis nitrogênio e fósforo, ambos separadamente, e em

seguida realizar uma análise de regressão não linear multiresposta com as duas variáveis de forma conjunta, com o objetivo de verificar se o comportamento da variável nitrogênio esta correlacionado ao comportamento da variável fósforo.

2 Fundamentação Teórica

Esta seção tem como objetivo retratar de forma clara e pontual os aspectos históricos referentes a modelos de regressão não linear, bem como modelos de regressão não linear multiresposta por meio de livros, artigos, teses e dissertações.

2.1 Marco Histórico

A expressão "regressão" foi sugerida por Sir Francis Galton (1885) em um estudo que demonstrou que a altura dos filhos não tende a ser igual a altura dos pais, mas que regride para a média da população. No modelo de regressão linear existe uma releção, ou não, entre uma variável dependente Y e uma variável independente X.

É muito comum em diversos trabalhos científicos ocorrer uma relação funcional entre as variáveis estudadas, e como na grande maioria dos experimentos o pesquisador trabalha com dados amostrais, ocorre uma necessidade de verificação da significância dessa relação funcional entre as variáveis. Para a verificação estatística dessa relação, os pesquisadores usam com frequência os modelos de regressão, que consistem no estudo estatístico de uma equação que explica a variação da variável dependente pela variação dos níveis das variáveis independentes.

O estudo da relação entre duas variáveis X e Y, pode ser representado por uma função de X que explique Y

$$\mathrm{X, Y} \rightarrow \mathrm{Y} \simeq \mathrm{f(X)}.$$

Há muitas situações em que não é desejável, ou até mesmo possível, descrever um fenômeno por meio de um modelo de regressão linear. Nessas situações é indicado o uso de modelos de regressão não lineares, tendo em vista que o mesmo obtém uma relação teórica entre as variáveis de interesse. Modelos de regressão não lineares são conhecidos desde 1920 por meio de estudos de Fisher R. A. e Mackenzie. Entretanto, esse tipo de modelo foi melhor análisado e por consequência mais utilizado por volta de 1970 quando ocorreu um avanço dos cálculos computacionais (DODGE, 2008).

Durante as duas últimas décadas, duas abordagens para a análise estatística de mo-

delos de multiresposta foram usadas em paralelo. A primeira é o método dos mínimos quadrados generalizados proposto por Zellner (1962), que inicialmente trata os modelos de forma linear. Enquanto que, Gallant (1987), usou uma aproximação para modelos não lineares, cuja abordagem baseia-se na soma dos quadrados generalizado, que é dependente dos segundos momentos dos termos de erro. Também a abordagem, desenvolvida com um argumento bayesiano por Box e Draper (1965), utilizaram um critério determinante da matriz da soma dos quadrados e produtos cruzados dos resíduos. Esse critério dá estimativas dos parâmetros mais precisos do que o critério de mínimos quadrados generalizado no sentido de que, no contexto bayesiano, o método dos mínimos quadrados generalizados é derivado a partir da densidade marginal *posteriori* enquanto o bayseano é a partir da densidade condicional *posteriori* dos parâmetros do modelo.

2.2 Modelos de regressão não lineares

Modelos de regressão não lineares são muito utilizados quando se deseja descrever a relação entre uma variável resposta Y e uma variável explicativa X. De acordo com Gomes (2009), um estudo sobre adubação, a produção média em toneladas (ton) pode ser modelada como sendo uma função quadrática da dose de um nutriente em quilogramas (kg)

$$f(X) = a + bX + cX^2. \tag{2.1}$$

Em 2.1, o modelo é linear nos parâmetros a, b e c. Entretanto, por meio da lei de *Mitscherlich*, a produção média pode ser modelada da seguinte forma

$$f(X) = a\{1 - exp[-b(X - c)]\}, \tag{2.2}$$

em que, a representa a produção máxima teórica em (ton) possível quando se aumenta infinitamente a dose do nutriente, b em (kg^{-1}) deve ser interpretado como sendo o coeficiente de eficácia do nutriente e c em (kg) é o conteúdo do solo. Um grande número de pesquisadores acreditam que as relações entre variáveis biológicas são melhores interpretadas por modelos não lineares. Processos como crescimento, decaimento, nascimento, mortalidade, abundância, competição e produção raramente são relacionadas linearmente às variáveis explicativas (SCHABENBERGER; PIERCE, 2002).

Um modelo é considerado não linear nos parâmetros, quando uma variável dependente Y não pode ser escrita na forma de funções lineares de seus parâmetros. Muitos autores, por exemplo, Ratkowsky (1983) e Bates e Watts (1988) definem que um modelo é dito

ser não linear se pelo menos uma derivada parcial da variável dependente, com relação a algum parâmetro existente no modelo, depende de algum parâmetro. Na maioria das vezes, a quantidade de interesse é a média de Y. Desta forma, considere a situação em que

$$E(Y|x) = \eta(X, \boldsymbol{\theta}), \tag{2.3}$$

é um modelo cujo a quantidade de interesse é a média de uma variável explicativa X e do seu vetor p parâmetros $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_p)$ por meio de uma função não linear η. Para dar sequência, se faz necessário realizar algumas mudanças de notação em (2.1) e (2.2) de acordo com (2.3). Sendo assim, tem-se que

$$\begin{aligned} \eta(X, \theta_0, \theta_1, \theta_2) &= \theta_0 + \theta_1 X + \theta_2 X^2 & (2.4) \\ \eta(X, \theta_a, \theta_b, \theta_c) &= \theta_a(1 - exp\{-\theta_b(X - \theta_c)\}), & (2.5) \end{aligned}$$

são modelos não lineares nos parâmetros, onde os índices a,b e c na Equação (2.5) são interpretados respectivamente como à assíntota, eficiência e conteúdo. Como foi dito anteriormente, e agora utilizando a notação em (2.3), sabe-se que um modelo não linear é assim caracterizado quando pelo menos uma das derivadas de η com relção à um dos parâmetros em θ envolver parâmetros. Ao ser realizadas as derivadas parciais da Equação (2.4), é possível verifica que isso não ocorre, sendo assim, o modelo é linear nos parâmetros. O mesmo não é constatado, ao ser avaliada a Equação (2.5)

$$\begin{aligned} \frac{\partial \eta}{\partial \theta_a} &= 1 - exp\{-\theta_b(X - \theta_c)\} \\ \frac{\partial \eta}{\partial \theta_b} &= -\theta_a(\theta_c - X)exp\{-\theta_b(X - \theta_c)\} \\ \frac{\partial \eta}{\partial \theta_c} &= -\theta_a\theta_b exp\{-\theta_b(X - \theta_c)\}, & (2.6) \end{aligned}$$

haja visto que, os parâmetros θ_b e θ_c foram encontrados nas derivadas. Portanto, o modelo é não linear nos parâmetros. Uma observação importante sobre modelos não lineares está relacionado quanto a sua classificação. Modelos de regressão que a priori são considerados não linear nos parâmetros, mas que ao sofrerem algum tipo de transformação se tornam modelos de regressão lineares com novos parâmetros, são chamados de modelos de regressão intrinsecamente lineares segundo Bates e Watts (1988) e Draper e Smith (1998). É o que acontece em

$$\eta(X, \boldsymbol{\theta}) = exp\{\theta_0 + \theta_1 X\}, \tag{2.7}$$

ao ser aplicada a transformação por meio da função logaritmica (log) o modelo se tran-

forma em um modelo linear

$$log[\eta(X, \boldsymbol{\theta})] \quad = \quad \theta_0 + \theta_1 X. \tag{2.8}$$

Observe que isso não ocorre em (2.5). Esse tipo de modelos são chamados de modelos de regressão intrinsecamente não lineares nos parâmetros de acordo com Draper e Smith (1998), pois o mesmo é não linear e nem intrinsecamente linear nos parâmetros.

Até aqui, o que foi dito sobre modelos de regressão não lineares, não foi considerado um fator aditivo. Esse componente do modelo é o erro aleatório que será parte permanente do modelo. Sendo assim, faz-se necessário definir que, em dados cujas as observações são contituídas de respostas univariadas Y_i dependentes de uma contribuição X_i fato comum em análises estatísticas, Gallant (1987). Define,

$$Y_i \quad = \quad \eta(X_i, \boldsymbol{\theta}) + \varepsilon_i \qquad i = 1, ..., n, \tag{2.9}$$

em que $Y = (y_1, \ldots, y_n)^\top$ é o vetor de observações ou de variáveis independentes, $\eta(X_i, \boldsymbol{\theta}) = [\eta(x_1, \theta), \ldots, \eta(x_n, \theta)]$ é o vetor de funções de regressão ou função das variáveis regressoras e dos parâmetros chamada de função esperança, $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_p)^\top$ é o vetor p-dimensional de parâmetros desconhecidos e $\varepsilon_i = (\varepsilon_1, \ldots, \varepsilon_n)$ é o vetor de erros representando por fenômenos não observáveis ou ainda o erro experimental.

2.2.1 Método de estimação por mínimos quadrados

Sabe-se que os métodos mais frequentimentes utilizados para a estimação de parâmetros em modelos de regressão não linear são os métodos de mínimos quadrados e o da máxima verossimilhança. Para Gallant (1987), o método de estimação por mínimos quadrados utilizado para encontrar os parâmetros de um modelo de regressão não linear é o mesmo utilizado em modelos lineares. Draper e Smith (1998) afirmam que a estimativa de mínimos quadrados de $\boldsymbol{\theta}$ é a mesma estimativa encontrada por meio do método de máxima verossimilhança. Os mesmos definem ainda, que a soma de quadrados dos erros de $\boldsymbol{\theta}$ para o modelo não linear oriundo de dados cujo objetivo é o de encontrar a estimativa $\hat{\boldsymbol{\theta}}$, é dada minimizando-se a função quadrática

$$SQE(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{n} \{y_i - \eta(X_i, \boldsymbol{\theta})\}^2 \tag{2.10}$$

em seguida, deriva-se $SQE(\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$ para se obter

$$\frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\sum_{i=1}^{n}\{y_i - \eta_i(x_i, \boldsymbol{\theta})\}\frac{\partial \eta}{\partial \boldsymbol{\theta}}.$$

Com essa derivação foi gerado p equações normais que devem ser resolvidas para $\hat{\boldsymbol{\theta}}$. Essas equações normais podem ser expressas da seguinte forma

$$\sum_{i=1}^{n}\{y_i - \eta(x_i, \boldsymbol{\theta}_r)\}\left[\frac{\partial \eta_i(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r}\right]_{\theta=\hat{\theta_r}} = 0.$$

A estimativa $\hat{\boldsymbol{\theta}}$ para o vetor de parâmetros $\boldsymbol{\theta}_r$ é encontrado ao ser igualado $\frac{\partial SQE(\boldsymbol{\theta}_r)}{\partial \boldsymbol{\theta}_r}$ a zero para $r = 1, ..., p$. Comumente as equações $\frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} = 0$, não são lineares e portanto não tem uma solução explícita. Sendo assim, devem ser resolvidas numericamente por meio de processos iterativos.

A maioria dos algoritmos para computação das estimativas de mínimos quadrados $\hat{\boldsymbol{\theta}}$ e a maioria dos métodos inferências para modelos não lineares são baseados em métodos iterativos que consideram uma aproximação linear local para o modelo (BATES; WATTS, 1980).

2.2.2 Método de estimação por máxima verossimilhança

O método de estimação por máxima verossimilhança foi proposto por Ronald Aylmer Fisher em seu artigo "*On an absolute criterion for fitting frequency curves*", no qual propôs a máxima verossimilhança como um método de modelagem de curvas de frequência, Fisher (1912).

Sob a suposição de normalidade dos erros, tem-se que os estimadores de máxima verossimilhança são idênticos aos de mínimos quadrados, pois maximizar a função de verossimilhança corresponde a minimizar a soma de quadrados de resíduos (CHIACCHIO.E, 1993). Isso significa que, os erros aleatórios ε_i são

i) aditivos e de média zero;

ii) normais;

iii) de variância contante;

iv) independentes.

Seja a função de verossimilhança L dada por

$$\begin{aligned} L(\boldsymbol{\theta}, y_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{ -\frac{[y_i - \eta(X_i, \boldsymbol{\theta}_r)]^2}{2\sigma^2} \right\} \\ &= (2\pi)^{\frac{-n}{2}} (\sigma^2)^{\frac{-n}{2}} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - \eta(X_i, \boldsymbol{\theta}_r)]^2 \right\} \end{aligned}$$

em que, os ε_i sejam independentes e identicamente distribuídos (i.i.d.) e que $\eta(X_i, \boldsymbol{\theta})$ é a função de densidade, ambos seguem distribuição $N \sim (\boldsymbol{\eta}(X_i, \boldsymbol{\theta}), \boldsymbol{I}\sigma^2)$. Os estimadores de máxima verossimilhança dos parâmetros de $\boldsymbol{\theta}$ são obtidos ao maximizar a função de verossimilhança $\ell(\boldsymbol{\theta}, y_i) = log[L(\boldsymbol{\theta}, y_i)]$ assim, a função passa a ser

$$\ell(\boldsymbol{\theta}_r, y_i) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - \eta(x_i, \boldsymbol{\theta}_r)]^2. \tag{2.11}$$

Ao observar a expressão (2.11) é possível constatar que para maximizar uma função dessa natureza em modelos não lineares, se faz necessário a utilização de algum método iterativo. Alguns desses métodos iterativos serão discutidos a partir da próxima seção.

2.3 Métodos de otimização

Quando discutido sobre os métodos de mínimos quadrados e o método da máxima verossimilhança para a estimação de um parâmetro θ, foi destacado o fato de que estimar um parâmetro em modelos de regressão não linear não é uma tarefa simples, pois como em modelos multiparamétricos, as soluções das equações normais podem ser extremamente difíceis, não apresentando solução explícita, sendo necessário o uso de algum método iterativo de resolução para equações não lineares (RATKOWSKY, 1983; BATES; WATTS, 1988). Além disso, uma dificuldade presente para a análise estatística de modelos de regressão não linear, emerge do fato de que os algoritmos existentes podem não convergir e por consequência os resultados inferenciais que são baseados numa aproximação linear podem ser pouco confiáveis. Existem muitos métodos iterativos/de otimização para a obtenção de estimadores, a exemplo do método de Gauss-Newton ou método da linearização, Quasi-Newton, Newton-Raphson, Steepest Descent, método de Marquardt entre outros. Porém, neste trabalho, será realizada uma discussão sobre os três primeiros métodos que foram citados.

2.3.1 Método de Gauss-Newton

Foi visto anteriormente que as equações do tipo $\frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_r} = 0$ são não lineares e por isso devem ser resolvidas por meio de algum método de iteração. O algoritmo de Gauss-Newton é um dos métodos mais utilizados para obtenção de estimativas de mínimos quadrados não lineares, sendo que, esse método utiliza uma expansão em série de Taylor de primeira ordem com o objetivo de aproximar o modelo de regressão não linear em termos lineares para em seguida aplicar mínimos quadrados ordinários e assim, estimar os parâmetros. Na verdade o que se pretende, inicialmente, é encontrar um determinado $\boldsymbol{\theta}^{(0)}$ em que, $\boldsymbol{\theta}^{(0)} = [{\theta_1}^{(0)}, ..., {\theta_p}^{(0)}]$ é o vetor inicial de parâmetros, isto é,

$$\eta(\boldsymbol{X}, \boldsymbol{\theta}) \approx \eta(\boldsymbol{X}, \theta^{(0)}) + \boldsymbol{\Delta^{(0)}}(\theta - \theta^{(0)})$$

em que, $\boldsymbol{X}$ é denominada como a matriz de delineamento do experimento, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ é o vetor dos verdadeiros parâmetros de regressão e $\boldsymbol{\Delta}^{(0)}$ é o vetor de derivadas de primeira ordem da SQE($\boldsymbol{\theta}$) em relação aos parâmetros do modelo não linear, calculado em $\boldsymbol{\theta}^{(0)}$. Assim, definindo-se $\boldsymbol{g}$ como vetor de resíduos, pode-se escrever:

$$\begin{aligned} \boldsymbol{g} &= \eta(\boldsymbol{X}, \boldsymbol{\theta}) \\ \boldsymbol{g} &\approx \eta(\boldsymbol{X}, \boldsymbol{\theta}^{(0)}) + \boldsymbol{\Delta}^{(0)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) \\ \boldsymbol{g} - \eta(\boldsymbol{X}, \boldsymbol{\theta}^{(0)}) &\approx \boldsymbol{\Delta}^{(0)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}). \end{aligned}$$

Utilizando-se do mesmo algebrismo para fazer $\boldsymbol{g} - \eta(\boldsymbol{X}, \boldsymbol{\theta}^{(0)}) = \boldsymbol{g}^{(0)}$ e $\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)} = \boldsymbol{\varphi}^{(0)}$, tem-se uma aproximação para um modelo de regressão linear:

$$\boldsymbol{g}^{(0)} = \boldsymbol{\varphi}^{(0)}\boldsymbol{\Delta}^{(0)} + \boldsymbol{\varepsilon}, \tag{2.12}$$

em que, $\boldsymbol{\varphi}^{(0)}$ é o vetor de coeficientes de regressão obtido da diferença entre o vetor dos verdadeiros parâmetros de regressão e o vetor dos valores iniciais atribuídos a eles. Na verdade, o que se pretende obter com o modelo ajustado em (2.12) é estimar o vetor dos coeficientes de regressão $\boldsymbol{\varphi}^{(0)}$. Essa estimativa é obtida do método de mínimos quadrados em que $\boldsymbol{\varphi}^{(0)} = ({\boldsymbol{\Delta}'}^0\boldsymbol{\Delta}^0)^{-1}{\boldsymbol{\Delta}'}^0 g^{(0)}$ é utilizada para corrigir as estimativas iniciais dos parâmetros de correção para que o modelo de regressão não linear seja agora aproximado à termos lineares.

Após a primeira iteração, partisse para atualizar o vetor de estimativas dos parâmetros dado por: $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}\boldsymbol{\varphi}^{(0)}$. Se a soma de quadrados dos desvios $SQE(\boldsymbol{\theta}^{(1)})$ for menor que $SQE(\boldsymbol{\theta}^{(0)})$ o processo deve ser repetido, agora utilizando $(\boldsymbol{\theta}^{(1)})$ no lugar de $(\boldsymbol{\theta}^{(0)})$, para

encontrar $\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)}\boldsymbol{\varphi}^{(1)}$ e assim sucessivamente até que se alcance uma convergência de acordo com algum critério pré-estabelecido. A esse processo, dar-se o nome de método de Gauss-Newton e um critério de parada é considerado a convergência na k-ésima iteração se,

$$|SQE({\theta_r}^{k-1}) - SQE({\theta_r}^{k})| < \delta$$

para $r = 1, ..., p$, em que p é o número de parâmetros e δ é o menor erro possível para o r-ésimo parâmetro.

2.3.2 Método de Quase-Newton

Os métodos Quase-Newton também conhecidos de classe de métodos de métricas variáveis, pertencem a uma classe de métodos de otimização inspiradas no método de Newton para minimização de uma função que são muito eficientes, pois elimina a necessidade do cálculo de segundas derivadas e tipicamente apresenta bom desempenho computacional. Alguns exemplos dessa classe são: o método BFGS (Broyden, Fletcher, Goldfarb e Shanno), método DFP (Davidon, Fletcher e Powell), Método de Shamanski, Método de Broyden e outros. Para a resolução de sistema de equações não lineares o método de Newton é bom, porém com alto custo computacional, portanto parece natural a introdução de métodos "quase bons", porém de custo relativamente baixo e os métodos Quase-Newton foram estabelecidos em sua maioria com este objetivo. Contudo, a redução de custos operacionais traz consigo a redução na velocidade de convergência. Isso acontece por que a solução do sistema é obtida com uma única avaliação da matriz Jacobiana, isto é, somente avalia-se esta matriz no ponto inicial, ao contrário do método de Newton, que avalia a Jacobiana em todas as iterações.

Sabe-se que o método de Newton, no qual a solução θ^* é aproximada por uma sequência de pontos $\{\theta^k\}$, é gerada por:

$$\boldsymbol{\Delta}(\theta^k)S^k = -\boldsymbol{\eta}(\theta^k), \qquad \theta^{k+1} = \theta^k + S^k$$

em que, $\boldsymbol{\Delta}(\theta)$ é a matriz Jacobiana da função $\boldsymbol{\eta}$ no ponto θ e S^k é a solução do sistema linear. Portanto, uma iteração de Newton exige basicamente que à avaliação da matriz Jacobiana em $\{\theta^k\}$ e a resolução do sistema linear

$$\boldsymbol{\Delta}(\theta^k)S^k = -\boldsymbol{\eta}(\theta^k).$$

Já para os métodos Quase-Newton, a sequência $\{\theta^k\}$, sendo que, o objetivo inicial é evitar

o cálculo de $\boldsymbol{\Delta}(\theta)$ em cada iteração aproximando-a de uma matriz B_k, é gerada através da fórmula:

$$B_k S^k = -\boldsymbol{\eta}(\theta^k), \qquad \theta^{k+1} = \theta^k + S^k.$$

A matriz B_{k+1} é obtida através de fórmulas de recorrência que envolvem θ^k, θ^{k+1}, $\boldsymbol{\eta}(\theta^k)$ e $\boldsymbol{\eta}(\theta^{k+1})$ como informações. Logo, a forma como B_{k+1} é obtida define um dos métodos Quase-Newton aos quais foram citados anteriormente.

2.3.3 Método de Newton-Raphson

O método de Newton-Raphson, que tem como base o método de Gauss-Newton, é um dos mais utilizados para obter solução de equações não lineares pois converge rapidamente. É importante ressaltar que esse método numérico é utilizado para encontrar as estimativas de mínimos quadrados e também para as estimativas de máxima verossimilhança. Resende (2007) trás em seu livro um estudo teórico de como proceder para encontrar os estimadores de máxima verossimilhança utilizando o método de Newton-Raphson. Aqui, será abordado a situação em que se deseja verificar como o método se desenvolve para se obter as estimativas de mínimos quadrados.

O método de Newton-Raphson vai procurar as raízes da derivada da função dada em (2.10) com relação a cada parâmetro, portanto, se $\boldsymbol{\theta}$ é um vetor de p parâmetros, a derivada de SQE($\boldsymbol{\theta}$) será uma função p-dimencional,

$$S_1(\boldsymbol{\theta}) = \frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \\ \vdots \\ \frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_p} \end{pmatrix}.$$

A ideia de DeVries (1994) é utilizada para descrever o método de Newton-Raphson, salvo as devidas mudanças de notação e observações de natureza estatísticas, que serão acrescentadas ao longo do texto. Sendo assim, pode-se afirmar que o método se baseia na aproximação de primeira ordem em torno da função $S_1(\boldsymbol{\theta})$, sendo que essa aproximação ocorre pela série de Taylor e está escrita como

$$S_1(\boldsymbol{\theta}) = S_1(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0) S_1{}'(\boldsymbol{\theta}_0) + \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2}{2!} S_1{}''(\boldsymbol{\theta}_0) + \cdots$$

conservando apenas os dois primeiros termos da série, tem-se

$$S_1(\boldsymbol{\theta}) \approx S_1(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0) S_1'(\boldsymbol{\theta}_0).$$

Essa equação forma uma reta que passa pelo ponto $S_1(\boldsymbol{\theta_0})$ com inclinação $S_1{}'(\boldsymbol{\theta}_0)$, ou seja, ela é tangente à curva no ponto $\boldsymbol{\theta_0}$. Por isso, o método é conhecido por alguns autores como método da tangente ou tangencial.

Sob a suposição de que $S_1(\boldsymbol{\theta}_0)$ é igual a zero, uma vez que $\boldsymbol{\theta}$ é a raiz da função ao ser considerada a expanção da série de Taylor de primeira ordem, tem-se:

$$\begin{aligned} 0 &= S_1(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)S_1'(\boldsymbol{\theta}_0) \\ -S_1(\boldsymbol{\theta}_0) &= (\boldsymbol{\theta} - \boldsymbol{\theta}_0)S_1'(\boldsymbol{\theta}_0) \\ \boldsymbol{\theta} &= \boldsymbol{\theta_0} - \frac{S_1(\boldsymbol{\theta}_0)}{S_1'(\boldsymbol{\theta}_0)} \end{aligned}$$

em que, $\boldsymbol{\theta}$ é o ponto que deve ser utilizado no lugar de $\boldsymbol{\theta}_0$ para se obter um novo valor inicial com uma melhor aproximação. Alterando rápidamente a notação para se calcular a $(i+1)$-ésimo valor, resultante das sucessivas iterações, ver-se que a atualização dos valores é dado por

$$\boldsymbol{\theta^{(i+1)}} = \boldsymbol{\theta_r} - \frac{S_1(\boldsymbol{\theta}_r)}{S_1'(\boldsymbol{\theta}_r)}.$$

Assim como no método de Gauss-Newton um critério de parada pode ser adotado para o processo iterativo afim de se obter bons valores para $\boldsymbol{\theta^{(i+1)}}$. Resende (2007) ao falar do método de Newton-Raphson para se obter valores de $\boldsymbol{\theta^{(i+1)}}$, quando o processo é extrutu-rado em torno da máxima verossimilhança, define que enquanto a diferença $\boldsymbol{\theta^{(i+1)}} - \boldsymbol{\theta^{(i)}}$, denominada de correção, não for nula, o processo de iteração deve continuar. A ideia pode ser extendida também para mínimos quadrados ou ainda, pode-se escolher um outro critério chamado de erro relativo cujo critério de parada é dada por $\frac{\boldsymbol{\theta^{(i+1)}} - \boldsymbol{\theta^{(i)}}}{\boldsymbol{\theta^{(i)}}}$.

O método de Newton-Raphson não é um método para máximizar/mínimizar funções com objetivo de estimar parâmetros mas, um processo que se baseia no fato equivalente de que a derivada no ponto crítico da função $SQE(\boldsymbol{\theta})$ dada em 2.10 é zero.

2.4 Qualidade do ajuste

Nem sempre as estimativas dos parâmetros de modelos não lineares são de interesse práticos. Isso ocorre porque as estimativas são obtidas de maneira assintótica por meio de algum preocesso iterativo, podendo levar o processo a convergir para um mínimo local e não para um mínimo global que é o desejado. Por esse motivo, se faz necessário observar alguns critérios quanto a seleção de modelos para o qual se pretende descrever um determinado fenômeno. Draper e Smith (1998) afirma que em qualquer análise, se

faz presente a necessidade de ser avaliado o ajuste do modelo para os dados. Sartorio (2013) em seu estudo, comenta sobre a necessidade de se observar esses aspectos quanto a qualidade do ajuste de um modelo não linear, quais sejam: o número de iterações para que haja a convergência de um processo de otimização, o coeficiente de determinação, a estimativa da variância, o erro de predição média e ainda, o erro padrão das estimativas.

2.4.1 Coeficiente de determinação

A definição do coeficiente de determinação R^2 em modelos de regressão não linear não é obtida de forma simples tendo em vista que um dos critérios de sua definição requer a presença de intercepto no modelo, parâmetro que nem sempre é parte integrante nesses modelos. Sendo assim, considere-se o R_j^2 definido mais adiante, como sendo uma medida próxima a R^2,

$$R_j^2 = 1 - \frac{SQE(\hat{\boldsymbol{\theta}})}{\|y - \hat{y}\|^2}$$

em que, $SQE(\hat{\boldsymbol{\theta}})$ é a soma dos quadrados dos resíduos avaliados em $\hat{\boldsymbol{\theta}}$, y é o valor observado e $\hat{y}$ indica o valor predito. Porém, ressalta-se que o R_j^2 não é, isoladamente, um critério adequado para discussão de ajuste de modelos, pois, geralmente, em ajuste de modelos não lineares, é comum a obtenção de R_j^2 assintóticos altos e similares (REZENDE *et al.*, 2007).

2.4.2 Análise de resíduos

A análise de dignóstico, também chamada de análise de resíduo, é uma etapa de grande importância no que se refere a analisar os pressupostos do modelo em estudo e para identificar alguma caracteristica inesperada nos dados. Existe a disposição uma grande variedade de métodos de diagnóstico para auxiliar na análise do modelo e, em geral, esses métodos são utilizados em modelos de regressão linear. Sendo assim, para muitos autores, com excessão dos resíduos, os métodos de diagnóstico utilizados para modelos de regressão linear podem ser facilmente estendidos para modelos não lineares. Todavia, para o desenvolvimento desses métodos é necessáio um estudo aprofundado do comportamento dos resíduos, principalmente quando se esta trabalhando com pequenas amostras.

A análise de resíduos num modelo estatístico pode ser baseada nos resíduos ordinários, ou em versões padronizadas, ou em resíduos construídos apartir dos componentes da função desvio (MCCULLAGH; NELDER, 1989), ou em resíduos generalizados (COX; SNELL,

1968), além dos resíduos projetados explorados por (CORDEIRO; PAULA, 1989). Usualmente os resíduos ordinário, normalizados, estudentizados e outros são os mais utilizados na elaboração de métodos de diagnóstico em modelos de regressão não linear sendo que sua definição pode ser observada em Cox e Snell (1968).

Outra maneira de se análisar os resíduos se dar por meio de gráficos informais que exibam características dos resíduos bem como testes formais baseados em hipóteses. Para Cook e Weisberg (1982) tanto as análises informais quanto formais se complementam e têm seu destaque na análise residual. Além disso, é através dos resíduos que pode-se encontrar pontos influentes e/ou até mesmo *outliers* nos dados quando se está análisando os resíduos.

Segundo Cordeiro e Paula (1989), embora as técnicas de diagnóstico da regressão não linear sejam simples extensões das técnicas da regressão linear, as interpretações não são diretamente aplicáveis, particularmente, em virtude de os resíduos ordinários não terem mais uma distribuição, aproximadamente, normal.

Por meio dos resíduos, também pode-se encontrar possíveis pontos influentes e/ou *outliers*. Dessa forma, uma ou mais observações são ditas discrepantes (*outliers*) se seus resíduos são muito grandes em relação aos demais (DRAPER; SMITH, 1998). Já os pontos influentes são observações que, embora não apresentem resíduos grandes, podem alterar significativamente as estimativas dos parâmetros do modelo escolhido. Para detectar a presença de pontos influentes ou de *outliers*, técnicas gráficas, como diagramas de dispersão, e gráficos da distância de Cook podem ser usadas.

2.5 Modelos de regressão multiresposta

Segundo Bates e Watts (1988), as análises de dados utilizando-se modelos multiresposta são caracterizadas por experimentos, nos quais há M respostas medidas em N observações experimentais e que os modelos das M respostas dependem de um total P de parâmetros $\boldsymbol{\theta}$, e pode ser escrito como

$$y_{nm} = f_m(X_n; \boldsymbol{\theta}) + z_{nm} \qquad \text{com } \; n = 1, \ldots, N \;\; , m = 1, \ldots, M, \tag{2.13}$$

em que y_{nm} são variáveis aleatórias associadas com as medidas dos valores da m-ésima resposta nas n-ésimas observações, 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 $\boldsymbol{\theta}$, e Z_{nm} são os termos dos erros.

Os modelos de regressão com mais de uma variável resposta, podem ser classificá-los em dois grupos, sendo que, no primeiro grupo estão os clássicos, ou seja, os modelos de regressão linear multivariados nos quais cada variável dependente tem a mesma relação linear funcional com as variáveis independentes, mas com diferentes coeficientes. No segundo grupo, os modelos de regressão de resposta múltipla (modelos multiresposta), em que as variáveis dependentes podem ter relações funcionais diferentes, lineares ou não lineares, com as variáveis independentes (PEIXOTO, 2013).

O modelo clássico de regressão linear múltipla fornece uma generalização multivariada da regressão linear univariada e análise de modelos de variância. Assim, a maioria das inferências estatísticas, incluindo estimação, testes de hipóteses e distribuição teórica, para estes modelos lineares são exatas e são extensões dos modelos de regressão linear univariada, como discutido, por Anderson (1984).

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. Além disso, 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.

No entanto, a utilização dos modelos de regressão de multiresposta é importante quando comparada com o modelo de regressão linear multivariada clássica, especialmente quando o modelo baseia-se em uma relação para o qual os modelos são geralmente especificados como formas não lineares nos parâmetros. As vantagens de combinar informações de várias respostas em comparação com o uso uma resposta de cada vez são tais que, o primeiro permite-nos obter uma melhor compreensão do estudo em questão, bem como estimativas mais explicativas dos parâmetros (BOX; DRAPER, 1965).

Durante as últimas décadas, duas abordagens para a análise estatística de modelos de multiresposta foram usadas em paralelo. A primeira é o método dos mínimos quadrados generalizados proposto por Zellner (1962), que inicialmente trata os modelos de forma linear. Enquanto que, Gallant (1987), usou uma aproximação para modelos não lineares, cuja abordagem baseia-se na soma dos quadrados generalizado, que é dependente dos segundos momentos dos termos de erro. Também a abordagem, desenvolvida com um argumento bayesiano por Box e Draper (1965), utilizaram um critério determinante da matriz da soma dos quadrados e produtos cruzados dos resíduos.

3 Aplicação

Encontram-se nesta seção as principais metodologias que serviram de base para este trabalho, utilizando-se de modelos não lineares e modelos multiresposta e nas inferências realizadas por meio de aproximação.

3.1 Material e métodos

Os dados para realização deste trabalho foram obtidos do trabalho de SEN (2003), no qual, as amostras de solo foram coletadas em Fazendas de pesquisa. As amostras foram preparadas para a análise, com o uso da técnica de espectroscopia infravermelho. Uma técnica de inestimável importância na análise orgânica qualitativa, sendo amplamente utilizada nas áreas de química de produtos naturais, síntese e transformações orgânicas. As amostras de solo foram secas ao forno e peneiradas por meio de uma peneira de 2 mm e misturadas com fertilizantes NPK na gama de concentrações de 1-15%, totalizando sete concentações (Tabela 1). Para este estudo utilizou-se somente os compostos nitrogênio e fósforo.

Tabela 1: Concentração de fertilizantes (%) e nutrientes nas amostras contendo os fertilizantes Nitrogênio, Fósforo e Potássio (NPK) em (g)

Concentração(%)	Nitrogênio (N)	Fósforo (P)	Potássio (K)
1,0	0,15	0,07	0,13
2,5	0,38	0,16	0,31
5,0	0,75	0,33	0,62
7,5	1,13	0,49	0,93
10,0	1,50	0,66	1,25
12,5	1,88	0,82	1,56
15,0	2,25	0,98	1,87

A aplicação de fertilizantes é normalmente realizada em conjunto com a adição de água para dissolvê-los na solução de solo. Deste modo, após a adição de fertilizantes foi adicionado a quantidade de água perdida durante a secagem. Realizou-se, também a mistura dos componentes para se obter a dispersão homogênea, antes de realizar as análises de infravermelhos. A mistura foi armazenada, durante três dias e lavadas com água para

equilibrar o solo. As concentrações dos nutrientes nas amostras foram calculadas a partir da porcentagem dos fertilizantes e usando fatores de conversão. Neste experimento, sete amostras de (100g) foram preparadas para serem analisadas por meio de refletância de infravermelho.

De posse dessas informações, utilizou-se um modelo de regressão não linear, afim de inferir sobre a relação existente entre as concentrações dos fertilizantes e os nutrientes presentes no solo obtidas por refletância de infravermelho. Pinheiro e Bates (2000), assumiram a seguinte estrutura não linear

$$f(x_n;\boldsymbol{\theta}_r) = \frac{\theta_1}{1+\exp\left[(\theta_2 - x_n)/\theta_3\right]}, \quad \boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3). \tag{3.1}$$

Se $\theta_3 > 0$, então θ_1 é a assíntota horizontal quando $x \to \infty$ e 0 é a 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 no eixo x entre o ponto de inflexão e o ponto em que a resposta é $\theta_1/(1+e^{-1}) \approx 0,73\theta_1$.

Para a estimação dos parâmetros do modelo não linear foi utilizada a técnica dos mínimos quadrados ordinários, com o uso do método iterativo de Gauss-Newton (BATES; WATTS, 1988). Os valores iniciais utilizados para obtenção das estimativas dos parâmetros, em cada um dos conjuntos de dados, foram obtidas por uma função no *software* R versão 3.1.1 (R DEVELOPMENT CORE TEAM, 2012) que gera os valores iniciais, encontrada na *self-starting nonlinear models* desenvolvido por Pinheiro e Bates (2000).

Dentre as estatísticas fornecidas pelo procedimento de estimação, foram obtidos intervalos de confiança para os parâmetros e a correlação entre eles, bem como o coeficiente de determinação assintótico (R_j^2). A análise de resíduos foi feita para validar o modelo em estudo.

Pelo que foi visto em Bates e Watts (1988), as análises de dados utilizando modelos multiresposta são caracterizadas por experimentos, nos quais há M respostas medidas em N observações experimentais e que os modelos das M respostas dependem de um total p de parâmetros $\boldsymbol{\theta}$, e para o caso experimental em questão o modelo pode ser escrito como

$$y_{nm} = f_m(X_n;\boldsymbol{\theta}) + z_{nm} \quad \text{com} \quad n = 1,\ldots,7 \ \ , m = 1,2 \tag{3.2}$$

em que $n = 1,\ldots,7$ são sete concentrações (%) aplicadas ao solo, e $m = 1,2$ duas respostas, a concentração de nitrogênio (g) e a concentração de fósforo (g).

Para obtenção das estimativas dos parâmetros do modelo multiresposta é necessário o uso de valores iniciais, sendo assim, as estimativas dos parâmetros do modelo unirespostas foram combinadas, e por meio dessa combinação originou-se um vetor $\boldsymbol{\theta}$ de parâmetros, com $\boldsymbol{\theta}$ sendo um vetor que engloba todos os parâmetros. O ajuste do modelo multiresposta se deu por meio da combinação de modelos logístico com três parâmetros representados da seguinte forma

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

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

Obtidos os ajustes procedeu-se a análise dos resíduos do modelo para para verificar a sua adequação.

3.2 Resultados e discussão

Ao analizar-se a Tabela 2 é possível verificar que os parâmetros dos modelos do nitrogênio e do fósforo são significativos ao nível de 5% de significância pelo teste t tendo em vista que todos os valores-p são menores que $0,05$. Isso é um indicativo de que o modelo de regressão logístico com três parâmetros se ajustou bem aos dados. Os intervalos de confiança são bem parecidos para os mesmos parâmetros em modelos diferentes e a amplitude do intervalo do parâmetro $\boldsymbol{\theta}_2$ é maior que os demais. Isso ocorre porque o erro padrão da estimativa para esses parâmetro é maior que as demais. Nota-se ainda que os intervalos de confiança são significativos uma vez que não contém o zero.

Tabela 2: Estimativas dos parâmetros para o modelo logístico com três parâmetros, erro padrão da estimativa (E.P.E.), valor-p para o teste t e intervalos de confiança (IC) de 95% para os nutrientes nitrogênio e fósforo

Nutrientes	Parâmetros	Estimativas	E.P.E.	Valor-p	IC (95%)
Nitrogênio	θ_1	2,5502	0,2160	<0,0003	[1,9503; 3,1500]
	θ_2	8,5513	0,7981	<0,0004	[6,3353; 10,7674]
	θ_3	3,4771	0,4497	<0,0015	[2,2284; 4,7257]
Fósforo	θ_1	1,1146	0,0948	<0,0003	[0,8515; 1,3778]
	θ_2	8,5537	0,8019	<0,0004	[6,3274; 10,7801]
	θ_3	3,4827	0,4514	<0,0015	[2,2294; 4,7360]

Na Figura 1 é possível comprovar que o modelo logístico se ajustou bem aos dados uma vez que os valores ajustados encontram-se proxímos aos valores observados.

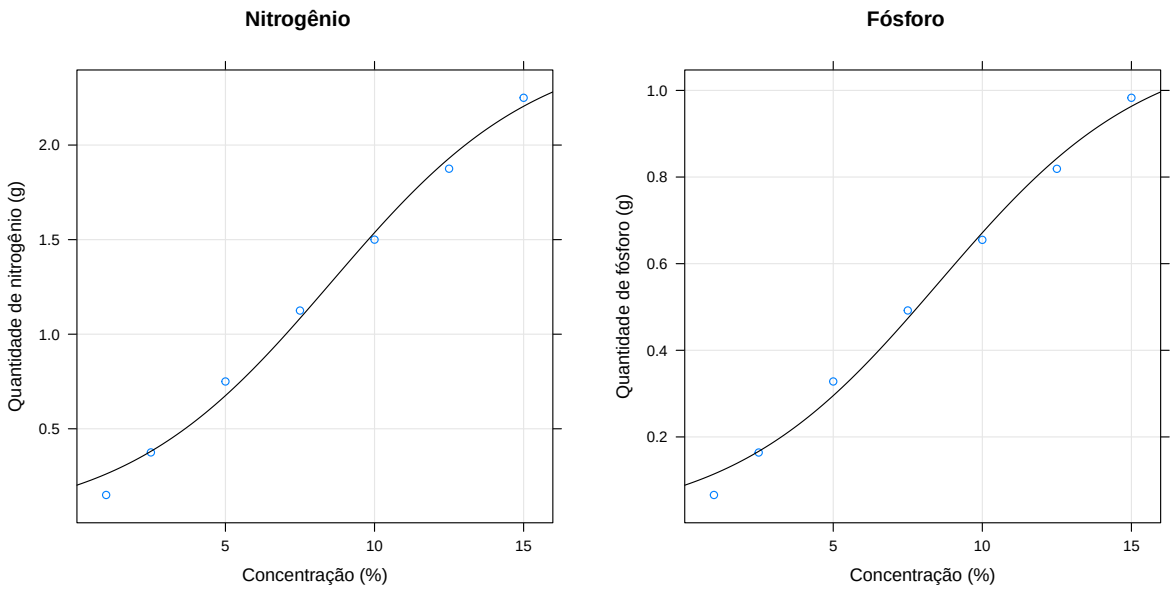

Figura 1: Gráfico de ajuste do modelo logístico para os nutrientes nitrogênio e fósforo

O estudo da normalidade dos dados é importante para verificar se a amostra que foi coletada realmente segue uma distribuição normal. Isso porque todo estudo da análise de resíduos se baseia no fato de que a amostra provém de uma população normalmente distribuída. Para o diagnóstico de heteroscedasticidade, tentou-se encontrar alguma tendência nos gráficos. Os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento ou tendência, tem-se indícios de que a variância dos resíduos é homoscedástica (Figuras 2 e 3).

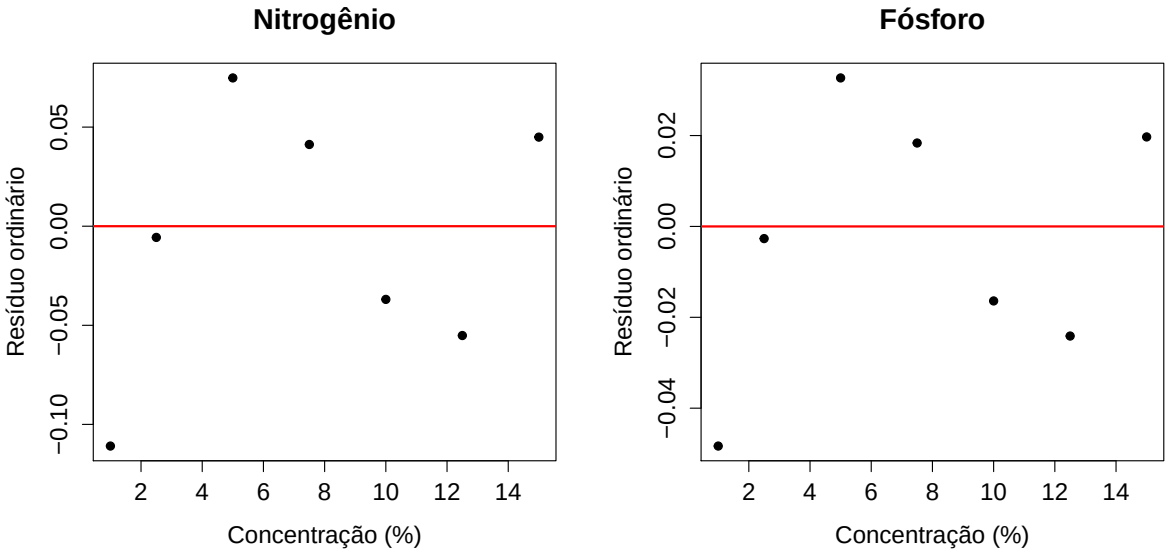

Figura 2: Gráfico dos resíduos ordinário *versus* valores preditos para os nutrientes nitrogênio e fósforo

De acordo com o que foi visto em Martin e Storck (2008) a falta de normalidade nos resíduos do modelo de regressão compromete a confiança nas estimativas dos parâmetros, podendo substimar ou sobrestimar as estimativas dos parâmetros, porêm a falta de normalidade não foi detectada nos ajustes dos modelos, como pode ser observado na Figura 3.

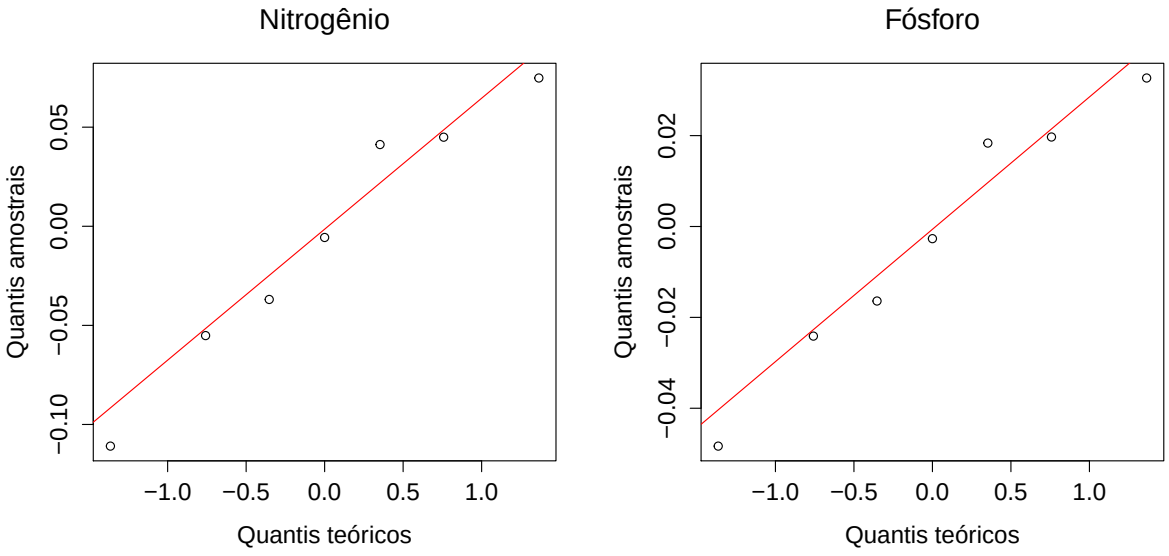

Figura 3: Gráfico dos quantis amostrais *versus* quantis teóricos para os nutrientes nitrogênio e fósforo

A técnica abordada para o ajuste multiresposta foi baseada em aproximações sugeridas por Kang e Bates (1990), segundo os autores análises estatísticas exatas são difíceis na prática e tem-se que confiar em métodos de aproximação apropriadas.

O vetor de resposta esperada para o ajuste multiresposta, foi obtido por meio do modelo logístico considerando-se duas respostas simultâneas. Para o caso analisado foram consideradas uma aproximação linear, utilizando como valores iniciais, as estimativas dos parâmetros dos modelos univariados ajustados. Os valores dos ajustes estão apresentados na Tabela 3. Todos os parâmetros da função foram significativos a 5%, pois os intervalos assintóticos não incluem a constante zero.

Na Figura 4 é possível visualizar, no ajuste do modelo, a relação entre o nitrogênio e fósforo. O ajuste simultâneo das variáveis exibe, o comportamento dos nutrientes no solo. A medida que a concentração de nitrogênio aumenta com o perfil, a concentração de fósforo também aumenta.

A matriz de correlação dos parâmetros do modelo multiresposta é observada por meio

Tabela 3: Estimativas dos parâmetros (P) do modelo multiresposta, erro padrão (E.P.), valor-p, intervalo de confiança IC (95%) para os nutrientes nitrogênio e fósforo

Parâmetro	Estimativa	E.P.	Valor-p	IC (95%)
θ_{11}	2,5500	0,1310	<0,0001	[2,5418; 2,5582]
θ_{12}	1,1146	0,1315	<0,0001	[1,1063; 1,1228]
θ_{21}	8,5508	0,4887	<0,0001	[8,5202; 8,5815]
θ_{22}	8,5537	1,1230	<0,0001	[8,4832; 8,6241]
θ_{31}	3,4768	0,2629	<0,0001	[3,4603; 3,4933]
θ_{32}	3,4827	0,6036	<0,0001	[3,4449; 3,5206]

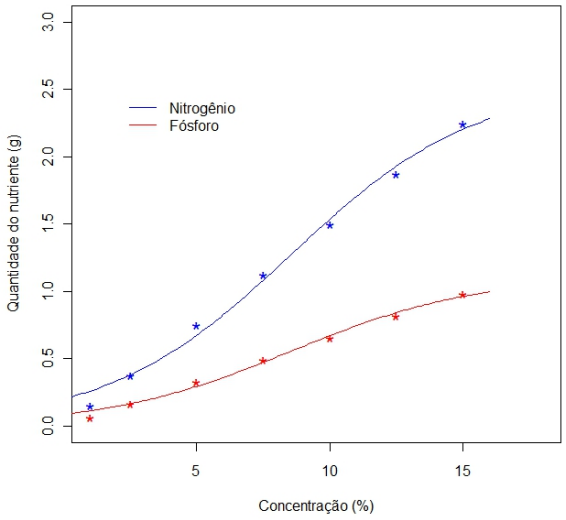

Figura 4: Ajuste do modelo multiresposta para os nutrientes nitrogênio e fósforo

Tabela 4: Matriz de correlação dos parâmetros do modelo para os nutrientes nitrogênio e fósforo

Parâmetro	θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}
θ_{11}	1,0000					
θ_{12}	-0,0899	1,0000				
θ_{21}	-0,9891	0,0948	1,0000			
θ_{22}	0,0919	-0,9977	-0,0969	1,0000		
θ_{31}	-0,8535	0,0402	0,7682	-0,0411	1,0000	
θ_{32}	0,0714	-0,9622	-0,0753	0,9417	-0,0319	1,0000

da Tabela 4. É possível verificar uma correlação positiva e alta entre os parâmetros (θ_{21} e θ_{31}), e entre (θ_{22} e θ_{32}). Ao se comparar a correlação entre os parâmetros (θ_{11} com θ_{21} e θ_{31}, a correlação foi alta e negativa. Já os demais parâmetros apresentaram uma correlação baixa entre si. Nas situações em que os valores das correlações entre os parâmetros foram negativos, indica-se que à medida que um determinado parâmetro cresce o outro tende

a diminuir, o que se explica devido a natureza das variáveis, indicando a necessidade da análise multiresposta.

Na Figura 5 é possível verificar que o comportamento dos resíduos foram aparentemente satisfatório, devido a a aleatoridade dos resíduos e normalidade dos resíduos ordinários *versus* os quantis teóricos.

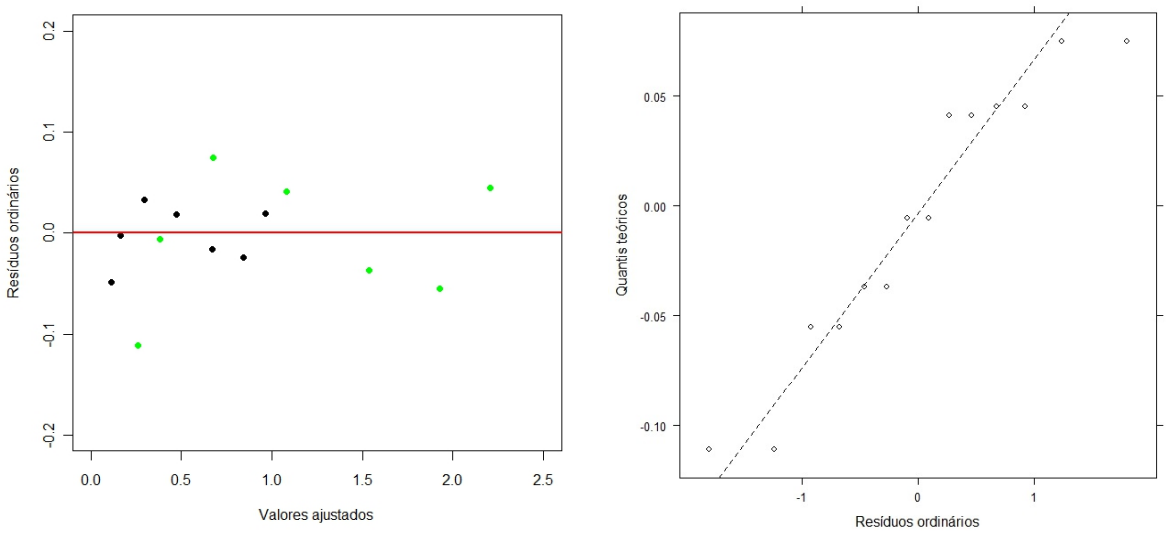

Figura 5: Gráfico dos resíduos ordinários *versus* valores ajustados e resíduos ordinários *versus* os quantis teóricos para os nutrientes nitrogênio e fósforo

BATES e WATTS (1987) ajustaram um modelo multiresposta com dois parâmetros para três conjuntos de valores, tendo como partida para o processo de estimação, o uso do método dos mínimos quadrados ordinários em cada resposta individual. Tendo em vista os resultados encontrados para o modelo multiresposta, pode-se afirmar que os resultados inferencias obtidos foram satisfatórios.

Peixoto (2013) utilizou um modelo multiresposta para analisar o comportamento das variáveis nitrato e potássio ao longo do perfil dos solos. Para esse modelo foi utilizado o método da máxima verossimilhança para encontrar as estimativas dos parâmetros e observou a adequação do modelo para descrever o comportamento dos solutos nos solos, sendo uma alternativa para os pesquisadores que trabalham com estudo de solos.

4 Conclusão

A escolha adequada do modelo de regressão não linear unidimencional no início do processo, foi de grande importância no decorrer do ajuste do modelo multiresposta uma vez que as estimativas dos parâmetros do modelo unidimencional foram todas estatíticamente significativas.

Ao ser verificado a qualidade do ajuste e a analisado os resíduos do modelo de regressão não linear logístico é possível constatar que o modelo se ajustou bem aos dados, tendo em vista que as análises realizadas para verificar as pressuposições dos modelos de regressão não linear foram atendidas de forma eficaz . Portanto, o modelo logístico unidimencional para as variáveis nitrogênio e fósforo se mostram apropriados para descrever a curva de crescimento para os dados apresentados.

O modelo de regressão não linear multiresposta, função dos modelos logísticos unidimencionais, foi capaz de exibir, simultâneamente, o comportamento dos nutrientes nitrogênio e fósforo nas amostras de solo e que, a medida que a concentração de nitrogênio aumenta com o perfil, a concentração de fósforo também aumenta.

Referências

ANDERSON, T. *An introduction to multivariate statistical analysis*. 2ed. ed. [S.l.]: New York, 1984.

BATES, D.; WATTS, D. Relative curvature measures of nonlinearity (with discussion). *Journal of the Royal Statistical Society*, Serie B, Methodological, London, v.42, n. 1, p. 1–25, 1980.

BATES, D.; WATTS, D. A generalized gauss-newton procedure for multiresponse parameter estimation. *SIAM Journal on Scientific and Statistical Computing*, Lysaker, v.8,, p. p.49–55,, 1987.

BATES, D. M.; WATTS, D. G. *Nonlinear regression analysis and its applications*. New York: New York: John Wiley & Sons., 1988. 365 p.

BOX, G.; DRAPER, N. R. The bayesian estimation of common parameters from several responses. *Biometrika*, v. 52, p. 355–365, 1965.

CHIACCHIO.E. *Regressão não linear desenvolvimento de um sistema computacional e aplicações*. Dissertação (Mestrado em Agronomia) — Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 1993.

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

CORDEIRO, G.; PAULA, G. Modelos de regressão para análise de dados univariados. In: *COLÓQUIO BRASILEIRO DE MATEMÁTICA*. [S.l.: s.n.], 1989.

COX, D. R.; SNELL, E. J. A general deninition of residuals. *Journal of the Royal Statistical Society*, v. 20, p. 248–275, 1968.

DEVRIES, P. L. *A first course in computational physics*. 1. ed. New York: Jhon Wiley & Sons, 1994. 424 p.

DODGE, Y. *The Concise Encyclopedia of Statistics*. Springer, 2008. (Springer reference). ISBN 9780387317427. Disponível em: <http://books.google.com.br/books?id=k2zklGOBRDwC>.

DRAPER, N. R.; SMITH, H. *Apllied regression analysis*. 3. ed. New York: J. Wiley, 1998. 706 p.

FISHER, R. On an absolute criterion for fitting frequency curves. *Messeng. Math*, v. 41, p. 155–160, 1912.

GALLANT, A. R. *Nonlinear Statistical Models*. New York: J. Wiley & Sons, 1987. 610 p.

GOMES, F. P. *Curso de estatística ´experimental*. 15. ed. Piracicaba: FEALQ, 2009.

KANG, G.; BATES, D. G. Approximate inferences in multiresponse regression analysis. *Biometrika*, v. 77, n. 2, p. 321–322, 1990.

MARTIN, T.; STORCK, L. Análise das pressuposições do modelo matemático em experimentos agrícolas no delineamento blocos ao acaso. *Sistemas de Produção Agropecuária. Curitiba: UTFPR.*, 2008.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. 2nd. ed. [S.l.]: Londo: Chapman and Hall, 1989. 511 p.

PEIXOTO, A. P. B. *Análise da dinâmica do potássio e nitrato em colunas de solo não saturado por meio de modelos não lineares e multiresposta*. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2013.

PINHEIRO, J.; BATES, D. *Mixed-Effects Models in S and S-PLUS*. [S.l.]: New York:, 2000.

RATKOWSKY, D. A. *Nonlinear regression modelling: a unified practical approach*. New York: Marcel Dekker, 1983. 276 p.

RESENDE, M. D. V. de. *Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético*. [S.l.: s.n.], 2007. 561 p.

REZENDE, D. M. L. C.; MUNIZ, J. A.; FERREIRA, D. F.; SILVA, F. F. e; AQUINO, L. H. de. Ajuste de modelos de platô de resposta para a exigência de zinco em frangos de corte. *Ciência e Agrotecnologia*, v. 31, p. 468–478, 2007.

SARTORIO, S. D. *Modelos não lineares mistos em estudos de degradabilidade ruminal in situ*. Tese (Doutorado) — Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2013.

SCHABENBERGER, O.; PIERCE, F. *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press, 2002. ISBN 9781584881117. Disponível em: <http://books.google.com.br/books?id=c2Rr_J7geGQC>.

SEN, I. *Spectroscopic Determination of Major Nutrients (N, P, K) of Soil*. Dissertação (Mestrado) — Partial Fulfillment Program: Food Engineering Department, 2003.

ZELLNER, A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, v. 58, 977-992 1962.