Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Fernanda Matias de Araújo

Uso da técnica *bootstrap* em modelos de regressão não linear

Campina Grande
Dezembro de 2012

Fernanda Matias de Araújo

Uso da técnica *bootstrap* em modelos de regressão não linear

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:
Ana Patricia Bastos Peixoto

Campina Grande
Dezembro de 2012

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

A658u Araújo, Fernanda Matias de.
Uso da técnica bootstrap em modelos de regressão não linear [manuscrito] / Fernanda Matias de Araújo. – 2012.
45 f. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2012.
"Orientação: Profa. Ma. Ana Patrícia Bastos Peixoto, Departamento de Estatística".

1. Modelos de regressão. 2.Modelo de crescimento.3. Método bootstrap. I. Título.

21. ed. CDD 519.536

Fernanda Matias de Araújo

Uso da técnica *bootstrap* em modelos de regressão não linear

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.

Aprovado em: 07 / 12 / 12

Banca Examinadora:

Ana Patricia Bastos Peixoto
Profa. Msc. Ana Patricia Bastos Peixoto
Orientadora

Ricardo Alves de Olinda
Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Diago Almeida de Oliveira
Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Agradecimentos

Agradeço em primeiro lugar à Deus, força minha, por sempre me fortalecer e sempre estar à frente de minhas vitórias.

Agradeço aos meus pais, Ana Lúcia e Herivelto, e ao meu irmão Everton, pelo amor, apoio e compreensão oferecidos, no qual me ensinaram a ir em busca da realização dos meus sonhos com sabedoria, perseverança e humildade, e nunca desistir dos meus objetivos.

Agradeço a minha orientadora Ana Patricia Bastos Peixoto pela atenção, estímulo, confiança e apoio na realização deste trabalho.

Aos professores Ricardo, Tiago, Juarez, Edwirde, Kátia e Diana por seus ensinamentos, sendo exemplos de seriedade, compromisso e profissionalismo.

A todos os meus colegas de graduação, em especial a Edlaine, Erasnilson, Alessandra e Priscilla pelo companheirismo, amizade, dedicação e alegria que me proporcionaram.

Resumo

Modelos de regressão não linear são geralmente utilizados em dados com curvas de crescimento. Nesse contexto, o objetivo deste trabalho é ajustar curvas de crescimento em dados que estão associados a altura(cm) da cernelha de cavalos pantaneiros ao longo dos dias (idade). Nesse estudo, serão utilizados 15 indivíduos, aos quais, foram submetidos ao ajuste dos modelos: Weibull com quatro parâmetros, Von Bertalanffy e Logístico com três parâmetros, respectivamente. Os parâmetros estimados foram α (valor assintótico), β (maturidade do animal), τ (taxa de maturação) e γ(ponto de inflexão). A interpretação biológica dos parâmetros, o coeficiente de determinação e o AIC, foram utilizados como critérios para a seleção dos modelos ajustados, sendo o modelo Weibull o que apresentou melhor ajuste, a aplicação do método *bootstrap* foi bem sucedida validando o modelo de regressão. Contudo, o método de simulação *bootstrap* é utilizado em amostra consideravelmente pequena, dessa forma, procura-se um modelo que melhor represente os dados. Nesse sentido, o modelo Weibull apresentou o melhor desempenho, devido ao menor valor do crítério AIC e possuir parâmetros não viesados.

Palavras-chave: Modelos Não Lineares, Weibull, Von Bertalanffy, Logístico, *Bootstrap*

Abstract

Nonlinear regression models are generally used in data curves growth. In this context, the aim of this work and adjust growth curves data that are associated with height (cm) of Cernia horse pantaneiros over days (aged). In this study, 15 subjects are used, to which were subject to adjustment models: Weibull with four parameters, and Von Bertalan Logistic with three parameters, respectively. The parameters were estimated (value asymptotic) (maturity of the animal), k (maturation rate) in point (excision). The biological interpretation of the parameters, the coefficient of determination and AIC were used as criteria for the selection of the adjusted models, with the Weibull model showed that best fit the application of the bootstrap method was successful Validating the regression model. However, the simulation method used in sample bootstrape Pretty small, thus aiming for a model that best represents data. In this sense, the Weibull model showed the best performance due to lower value of AIC criteria and parameters have not versed.

Key-words: Nonlinear Models, Weibull, Von Bertalany, Logistic, *Bootstrap*.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A análise de regressão linear estuda o comportamento de variáveis numéricas e os efeitos lineares produzidos por elas. Este estudo é feito por meio de uma função matemática, onde se têm as variáveis explanatórias representadas pelo x_i, ou seja, variáveis que explicam o comportamento da variável resposta, representada por y (HOFFMANN, 2006). Essa relação pode ser analisada por meio dos modelos de regressão, os quais se dividem em duas classes distintas: os lineares e os não lineares (MAZUCHELI; ACHCAR, 2002).

Segundo Gujarati (2006), alguns modelos podem parecer não lineares nos parâmetros, mas são intrinsecamente lineares, ou seja, através das transformações, surgirá novos modelos com diversos parâmetros gerando-se modelos lineares. Uma vez feita a transformação de um modelo não linear em um linear, é estudada a adequação do modelo para verificar se as suposições da regressão linear não foram violadas. Na prática, um modelo não linear é linearizado para facilitar a obtenção das estimativas dos parâmetros, devido ao fato dos estimadores em regressão não linear não possuírem certas propriedades que os estimadores em regressão linear têm (BATES; WATTS, 1988). No entanto, o estudo de não linearidade tem como foco, verificar se o modelo não linear possui propriedades próximas dos modelos lineares, de modo que, se o estimador tem um viés pequeno, uma distribuição próxima da normal e uma variância próxima da variância mínima, parece razoável falar que o estimador tem um comportamento próximo do linear (SOUZA, 2008).

A técnica *bootstrap* foi introduzida por Efron e Tibshirani (1993), é baseado na construção de sub-amostras a partir de uma amostra inicial quando se deseja avaliar, para certo estimador, o seu erro padrão, o seu viés, ou ainda quando se quer estimar a distribuição de probabilidade do estimador. A técnica tenta realizar o que se gostaria de fazer na prática, se tal fosse possível: repetir a experiência.

Esta técnica é uma alternativa de estimação mais aceitável para se verificar a confiabilidade de resultados que provém de uma amostra consideravelmente pequena, por meio de várias reamostragens com reposição. Uma das aplicações do método *bootstrap* é obter

intervalos de confiança corretos, onde é possível também obter a distribuição amostral de um parâmetro a partir da amostra original, que é a primeira amostra selecionada. É necessário também estimar o vício das estimativas dos parâmetros e assim efetuar as correções necessárias.

Segundo Souza (2008), o *bootstrap* pode ser paramétrico e não paramétrico. O paramétrico empregado neste trabalho, é utilizado, fazendo-se suposição sobre a distribuição dos dados que gerou a amostra original, usando-se os valores das estimativas dos parâmetros no processo de geração de pseudo-amostras. Já o não paramétrico, o processo de reamostragem se dá a partir da função de distribuição empírica dos dados, quando se têm uma amostra de tamanho n, na qual, as variáveis aleatórias y_i são independentes e identicamente distribuídas.

O objetivo deste trabalho é ajustar os modelos não lineares Weibull, Von Bertalanffy e Logístico, obter as estimativas para os parâmetros aos dados de crescimento animal obtidos em Souza (1998) e, aplicar a técnica *bootstrap* aos modelos ajustados, para verificar nos estimadores, o seu erro padrão, o seu viés, com o propósito de validar os modelos com confiança nos resultados. Posteriormente compara-los entre si e identificar o modelo que melhor se ajustou aos dados.

2 Fundamentação Teórica

Nesta seção, serão apresentados os modelos de regressão linear e os modelos de regressão não linear. Serão discutidos os tipos de modelos não lineares e o método de estimação dos parâmetros. Será apresentado a medida de vício de Box, os modelos de crescimento utilizados neste trabalho juntamente com os critérios adotados para seleção do melhor modelo ajustado aos dados, e por fim, será abordado a técnica de reamostragem *bootstrap*.

2.1 Modelos de regressão linear

Ajustar modelos de regressão linear, é um dos métodos estatísticos utilizados quando se quer estudar o comportamento de variáveis numéricas e os efeitos lineares produzidos por elas. O modelo de regressão explica a existência da relação funcional de uma ou mais variáveis de interesse em função de outras variáveis explicativas, sendo a variável resposta a principal variável em estudo no modelo de regressão. Em geral, estima-se uma reta de regressão e, a partir desta reta, são feitas possíveis previsões sobre o comportamento da variável de interesse, para isto é necessário estimar os parâmetros do modelo de regressão. Este estudo é feito por meio de uma função matemática, onde se têm as variáveis explanatórias representadas pelos x_i, ou seja, variáveis que explicam o comportamento da variável resposta, representada por y (HOFFMANN, 2006).

As variáveis y e x podem estar relacionadas de forma linear, polinomial, exponencial, logarítmica, entre outras. Uma forma simples de se avaliar o tipo de relação entre as duas variáveis é utilizar o gráfico de dispersão bivariado entre y e x. Matematicamente, o modelo linear será apresentado da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, ..., n, \tag{2.1}$$

em que x_i representa cada observação da variável explicativa x; β_0 representa o coeficiente linear da reta, ou seja, representa o ponto inicial para a variável y; β_1 representa o

coeficiente angular da reta, ou seja, o grau que a reta faz com o eixo x, e define também o quanto aumenta, ou diminui, o valor de y em relação a x; e ε_i é o erro associado a cada observação em relação à reta de regressão linear. Para conhecer os valores de β_0 e β_1 deve-se estimar os parâmetros, mas para que esse modelo seja aceito, precisa-se fazer algumas suposições sobre o modelo:

i) Existe relação linear entre x e y;

ii) A média do erro é nula, ou seja, $E(\varepsilon_i) = 0$;

iii) A variância do erro, ou variância residual, é uma constante igual a σ^2, para todos os valores de x;

iv) Os erros não são correlacionados entre si;

v) Os erros tem distribuição normal.

Para Hoffmann (2006) têm-se uma regressão linear múltipla (2.2) quando existe uma relação de dependência da variável resposta (y) com mais de uma variável. O modelo estatístico de uma regressão linear múltipla com k variáveis explanatórias é:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, ..., n \tag{2.2}$$

Usando a notação matricial, tem-se o seguinte modelo:

$$Y_{n\times 1} = X\beta_{n\times h} + \varepsilon_{n\times 1}, \tag{2.3}$$

com

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{h1} \\ 1 & x_{12} & x_{22} & \cdots & x_{h2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{hn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_h \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

As suposições sobre o modelo de regressão são as mesmas descritas no modelos simples, apenas com algumas adaptações:

i) Existe relação linear entre y e $x_j, j = 1, 2, ..., k$;

ii) Os valores dos x_j são sempre fixos, ou seja, eles não são variáveis aleatórias;

iii) As variáveis aleatórias ε_i têm distribuição normal;

iv) $E(\varepsilon) = 0$, em que 0 representa o vetor nulo;

v) $Var\,(\varepsilon) = \sigma^2$, para todos os valores de x_j ;

vi) Os erros são não correlacionados dois a dois.

Para fazer as inferências sobre os parâmetros é necessário a utilização do método dos mínimos quadrados ou a máxima verossimilhança, que permitem encontrar uma reta que minimize a distância entre os pontos observados e a reta, fazendo-se, em média, a soma dos desvios quadráticos ser igual a zero.

2.2 Modelos de regressão não linear

Um modelo é não linear quando uma variável dependente y não pode ser escrita como funções lineares de seus parâmetros. O modelo não linear é definido e apresentado pela maioria dos autores, como Draper e Smith (1998) e Bates e Watts (1988), quando pelo menos uma derivada parcial da variável dependente, dependa de algum parâmetro. Por exemplo os modelos:

$$\begin{aligned} E\,(y) &= \exp\,(\theta_1 + \theta_2 x)\,, \\ E\,(y) &= \theta_1 + \theta_2 \exp\,(-\theta_3 x)\,, \\ E\,(y) &= (\theta_1 + \theta_2 x)^{-1}\,, \\ E\,(y) &= (\theta_1 - \theta_2)^{-1}\,[\exp\,(-\theta_1 x) + \exp\,(-\theta_2 x)]\,, \end{aligned}$$

são todos não lineares e o operado E$(\cdot)$ denota a função esperança ou função de regressão.

Considerando-se a situação em que os dados consistem de uma resposta y que depende de k variáveis independentes de x, as respostas y_t obedecem ao modelo de regressão não linear

$$y_t = f\,(x_t, \theta) + \varepsilon_t, \quad t = 1, ..., n \tag{2.4}$$

em que a função resposta $f\,(x, \theta)$ é conhecida e não necessariamente linear, y_t representa a observação da variável dependente, x_t representa um vetor de observações em k variáveis

explicadas e determinadas fora do modelo, θ é um parâmetro p dimensional e ε_t é um erro experimental.

2.2.1 Modelos de regressão intrinsecamente lineares

Segundo Bates e Watts (1988) e Draper e Smith (1998), os modelos de regressão são intrinsecamente lineares, quando à princípio parecem ser não lineares nos parâmetros, mas em consequência de transformações se tornam modelos de regressão lineares com novos parâmetros.

Considere o seguinte exemplo:

$$Y_i = \beta_1 X^{\beta_2}, \tag{2.5}$$

Empregando-se o logaritmo neperiano em ambos os membros da Equação (2.5), obtém-se o novo modelo:

$$\ln(Y) = \alpha + \beta_2 \left(\frac{1}{X}\right)$$

Com $\alpha = \ln \beta_1$ as derivadas parciais em relação aos novos parâmetros são:

$$Z_i = \ln(Y_i); \quad \frac{\partial Z_i}{\partial \alpha} = 1 \quad e \quad \frac{\partial Z_i}{\partial \beta_2} = \frac{1}{X}$$

não contém nenhum parâmetro, logo o modelo é linear segundo os parâmetros α e β_2, portanto, este é um modelo intrinsecamente linear.

Santos (2011) aborda em seu trabalho que, os erros do modelo original satisfazem ás suposições básicas da análise de variância, enquanto os erros do novo modelo, poderão não satisfazer tais suposições, sendo que as transformações farão com que os parâmetros percam sua explicação intrínseca, e além do mais, altera-se a estrutura e as distribuições dos erros.

2.2.2 Modelos de regressão intrinsecamente não lineares

Segundo Draper e Smith (1998), um modelo de regressão é intrinsecamente não linear nos parâmetros, quando ele não é linear e nem intrinsecamente linear. De acordo com Gallant (1987), quando dados observados estão associados a uma variável independente, pode ser representada pela seguinte expressão:

$$y_t = f(x_t, \theta) + \varepsilon_t, \qquad t = 1, 2, ..., n,$$

em que, $\theta = (\theta_1, \theta_2, ..., \theta_p)$ é um vetor P-dimensional de parâmetros desconhecidos, ε_t é o erro aleatório aditivo, cujos erros são independentes identicamente distribuídos com média zero e variância σ^2. A série de valores de x_t é o ajuste experimental, e junto com o vetor de parâmetros determinam o valor da função modelo $f(x_t, \theta)$, sobre as suposições de que $E[\varepsilon_t] = 0$, é o valor esperado de y_t condicional a x_t e θ (BATES; WATTS, 1988):

$$E[y_t|x_t, \theta] = f(x_t; \theta).$$

Têm-se o seguinte exemplo de modelo não linear:

$$y_i = \theta_1 \left(1 - e^{-\theta_2 x_i}\right) + \varepsilon_i,$$

o qual não pode ser linearizado por uma transformação conveniente. Assim, uma função é considerada não linear nos parâmetros se y for dependente apenas de uma variável x.

2.2.3 Método de estimação por mínimos quadrados

Segundo Campos (2011), as inferências de interesse para os modelos não lineares, é realizada por meio de processo iterativo, para encontrar os estimadores de mínimos quadrados e os intervalos de confiança, são baseados em resultados aproximados.

Segundo Mazucheli e Achcar (2002), o método de estimação por mínimos quadrados é usado na análise de dados em que as observações são constituídas por variáveis resposta y_i obtidas em diferentes níveis da variável independente x_i, (i = *1*,...,n). A relação variável resposta/variável independente pode ser adequadamente representada por uma equação da forma:

$$Y = f(X; \theta) + \varepsilon \tag{2.6}$$

em que $Y = (y_1, ..., y_n)^t$ e $X = (x_1, ..., x_n)^t$ são os vetores de variáveis resposta e variável explicativa, respectivamente, $\theta = (\theta_1, ..., \theta_p)^t$ é o vetor de parâmetros desconhecidos, $f(x; \theta) = [f(x_1; \theta), ..., f(x_n; \theta)]^t$ é uma função das variáveis regressoras e dos parâmetros chamada de função esperança ou função de regressão e $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^t$ é o vetor de erros aleatórios. Geralmente, assume-se que os erros são variáveis aleatórias independentes e identicamente distribuídas, normais com media 0 e variância constante σ^2.

Draper e Smith (1998), definem a soma de quadrados dos erros para o modelo não

linear provido pelos dados como:

$$S(\boldsymbol{\theta}) \quad = \quad \sum_{t=1}^{N} \{y_t - f(x_t, \boldsymbol{\theta})\}^2. \tag{2.7}$$

Sendo y_t e x_t observações fixas, a soma dos quadrados é uma função de $\boldsymbol{\theta}$. A estimativa de mínimos quadrados $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ é o valor que minimiza $S(\boldsymbol{\theta})$. Draper e Smith (1998) mostram que a estimativa de mínimos quadrados de $\boldsymbol{\theta}$ é a mesma estimativa obtida pela máxima verossimilhança. Para descobrir o estimador de mínimos quadrados $\hat{\boldsymbol{\theta}}$ faz-se necessário diferenciar a Equação (2.7) com relação a $\boldsymbol{\theta}$, gerando p equações normais que precisam ser resolvidas para $\hat{\boldsymbol{\theta}}$. As equações normais podem ter a seguinte expressão:

$$\sum_{t=1}^{N} \{y_t - f(x_t, \boldsymbol{\theta})\} \left[\frac{\partial f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \right]_{\theta=\hat{\theta}} = 0, \tag{2.8}$$

e quando $\frac{\partial f(x_i;\theta)}{\partial \theta_i}$ não depender de $\boldsymbol{\theta}$, tem-se as equações normais de um modelo de regressão linear.

Segundo Gallant (1987), o método dos mínimos quadrados pode ser utilizado na estimação dos parâmetros em modelos não lineares da mesma maneira que é utilizado em modelos lineares. No entanto, para o sistema de equação normais não linear, não existe uma solução explícita, mas uma série de soluções adequadas que serão adquiridas por meio de processo iterativo.

2.2.3.1 Aproximação linear

A aproximação linear está associado a não linearidade do modelo. Mazucheli e Achcar (2002) afirmam que os resultados da regressão não linear somente serão válidos assintoticamente, sendo esses baseados em uma aproximação linear de primeira ordem. De acordo com Zeviani (2009), uma das vantagens da boa aproximação linear está em poder ter estimadores não viesados, normalmente distribuídos, com variância mínima, mesmo em pequenas amostras. Pode-se perceber que todos os procedimentos inferencias para modelos não lineares admitem suposição de adequada aproximação linear e ainda fazem uso de propriedades assintóticas. A função esperança da série de Taylor em torno de uma vizinhança de θ^0 (considere θ^0 como sendo o verdadeiro valor do parâmetro θ), essa aproximação linear será dada por:

$$f(x_t, \theta^0) \cong f(x_t, \theta^0) + \sum_{i=1}^{P} (\theta_i - \theta_i^0) \left[\frac{\partial f(x_t, \theta)}{\partial \theta_i} \right]_{\theta=\theta^0} \tag{2.9}$$

ou matricialmente

$$f(\theta) \approx f\left(\theta^0\right) + F\left(\theta - \theta^0\right)$$

Por meio da equação (2.9) pode-se escrever a soma de quadrado dos erros na forma:

$$\begin{aligned} S(\theta) &= [y - f(\theta)]^t [y - f(\theta)] \\ &= \|y - f(\theta)\|^2 \\ &= \left\|y - f\left(\theta^0\right) - F\left(\theta - \theta^0\right)\right\|^2 \\ &= \|\theta - F(\theta)\|^2 . \end{aligned}$$

De acordo com Mazucheli e Achcar (2002), quando o tamanho da amostra for consideravelmente grande e sob certas condições de regularidade, $\hat{\theta}$ estará praticamente em uma vizinhança de θ^0. A afetividade do algoritimo de mínimos quadrados e a validade das inferências com respeito aos parâmetros serão afetadas pela aproximação linear.

2.2.3.2 O método de Gauss-Newton

O método de Gauss-Newton, também conhecido como método da linearização, usa as séries de Taylor sobre $\boldsymbol{\theta}_i$ (o subscrito i-ésima iteração) com o propósito de aproximar o modelo de regressão não linear com termos lineares e, a partir disso, aplica-se o método mínimos quadrados ordinários para estimar os parâmetros. Essas iterações levam a uma solução para o problema de regressão não linear. Uma característica interessante do método de Gauss-Newton ocorre quando a função esperança é linear.

O método de Gauss-Newton começa dando-se valores iniciais aos parâmetros, tais valores iniciais podem ser obtidos de estudos anteriores ou conhecimentos teóricos, a partir disso, aproximamos a função esperada $\mathrm{f}(\mathrm{x_i}; \theta)$ para os n casos por termos lineares da expansão em série de Taylor, assim:

$$f(x_i; \theta) \approx f\left(x_i; \theta^0\right) + \sum_{j=1}^{p} \left[\frac{\partial f(x_i; \theta)}{\partial \theta_j}\right]_{\theta - \theta^0} \left(\theta_j - \theta_j^0\right)$$

Cada coeficiente de regressão θ_j^0 compõem-se da diferença entre os verdadeiros parâmetros da regressão e as estimativas iniciais dos mesmos, representando uma correção que deve ser feita nos coeficientes de regressão iniciais.

O modelo matricialmente será denotado por:

$$f(\theta) \approx f\left(\theta^0\right) + F^0\left(\theta - \theta^0\right)$$

Definindo-se $r(\theta)$ como vetor de resíduos, pode-se escrever:

$$\begin{aligned} r(\theta) &= y - f(\theta) \\ &\approx y - f\left(\theta^0\right) - F^0\left(\theta - \theta^0\right) \\ &= r\left(\theta^0\right) - F^0\left(\theta - \theta^0\right). \end{aligned}$$

Portanto, $S(\boldsymbol{\theta})$, será minimizada quando:

$$\boldsymbol{\theta} - \boldsymbol{\theta^0} = \left[\boldsymbol{F^{t^0}}\boldsymbol{F^0}\right]^{-1}\boldsymbol{F^{t^0}} r\left(\boldsymbol{\theta^0}\right) \tag{2.10}$$

Então, em consequência da aproximação θ^0 , será dada por:

$$\boldsymbol{\theta^{(0+1)}} = \boldsymbol{\theta} + \left[\boldsymbol{F^{t^0}}\boldsymbol{F^0}\right]^{-1}\boldsymbol{F^{t^0}} r\left(\boldsymbol{\theta^0}\right), \tag{2.11}$$

em que, $\boldsymbol{F^0}$ é a matriz de derivadas avaliada em θ^0, tendo como colunas os vetores $\boldsymbol{F^{t^0}}$. Com os valores iniciais para θ^0 com $i = 1$, o processo continua até a convergência que ocorre quando $|\theta^{0+1} - \theta^0|$ é menor do que alguma quantidade pré-fixada. Sendo este, o efeito no processo iterativo conhecido do Método de Gauss-Newton ou método da linearização.

Segundo Mazucheli e Achcar (2002), apesar do método de Gauss-Newton ser numericamente estável, a convergência pode ser lenta se uma grande precisão for exigida; a matriz F^0 pode ser singular ou tornar-se singular durante o processo iterativo; a convergência pode ser para um mínimo local e não para o mínimo global.

De acordo com Draper e Smith (1998), pelo método de linearização, a estimativa fica imutável no processo interativo, logo o modelo de regressão linear para o estimador de mínimos quadrados ficará do lado direito da igualdade. Sendo assim, um modelo linear converge para o estimador de mínimos quadrados em uma única iteração para qualquer valor inicial.

2.2.4 Medida de vício de Box

Para Zeviani (2009), o vício de Box indica quais parâmetros mais contribuem para o desvio da lineridade.

Segundo Mazucheli e Achcar (2002), a estatística para analisar o vício dos estimadores de mínimos quadrados dos parâmetros, será definida pela expressão:

$$vicio\left(\hat{\boldsymbol{\theta}}\right) = -\frac{\sigma^2}{2}\left[\sum_{i=1}^{n}\boldsymbol{F}\left(\boldsymbol{\theta}\right)\boldsymbol{F^t}\left(\boldsymbol{\theta}\right)\right]^{-1}\sum_{i=1}^{n}\boldsymbol{F}\left(\boldsymbol{\theta}\right) \tag{2.12}$$

$$\text{traço} = \left[\left(\sum\nolimits_{i=1}^{n}\boldsymbol{F}\left(\boldsymbol{\theta}\right)\boldsymbol{F^t}\left(\boldsymbol{\theta}\right)\right)^{-1}H\left(\boldsymbol{\theta}\right)\right],$$

em que $\boldsymbol{F}\left(\boldsymbol{\theta}\right)$ representa o vetor $(p \times 1)$ de primeiras derivadas da $f\left(x_i;\theta\right)$, denominado de vetor velocidade e $\boldsymbol{H}\left(\boldsymbol{\theta}\right)$ é uma matriz $(p \times p)$ das derivadas segunda em relação a cada elemento de θ.

Para resolver a Equação (2.12), a princípio considera $\hat{\theta}$ e $\hat{\sigma}^2$, como sendo os verdadeiros valores de θ e σ^2. Assim, o vetor $\boldsymbol{(p \times 1)}$, representa a discrepância entre as estimativas e os verdadeiros valores dos parâmetros.

2.3 Modelos de crescimento

Conforme abordado em Santos (2011), existem vários modelos não lineares e dentre eles, existem os modelos de curvas de crescimento. Os modelos Weibull, Von Bertalanffy e Logístico são os mais conhecidos, no qual apresentam diversos parâmetros em comum, em que é possível agregar significado biológico a cada um deles. No entanto, o ajuste a essas curvas *altura* × *idade*, devem ser coerentes com as interpretações biológicas do crescimento do animal.

Segundo Campos (2011), as curvas de crescimento chamam a imagens de curvas sigmoidais, em que representam o tempo de vida de medidas de dimensão como altura e peso. Para modelar os dados de crescimento, deseja-se obter as informações físicas dos parâmetros, com a finalidade de construir um modelo padrão para as observações em estudo.

Os parâmetros dos modelos de curva de crescimento predizem as taxas de crescimento, podendo ser utilizados como critérios de seleção para programas de melhoramento animal. De acordo com o relato de Santos (2011), as curvas de crescimento refletem a relação entre

a idade do animal e o seu impulso de crescimento e maturidade, sendo importantes na produção em programas de melhoramento, assim podendo aumentar o lucro do produtor. Os modelos citados anteriormente são derivados da curva de Richards, definido por:

$$y_t = \alpha \left(1 - \left(\beta e^{-\tau t}\right)^{\gamma}\right) + \varepsilon, \tag{2.13}$$

sendo que a diferença entre eles consiste na variação do parâmetro de inflexão. O parâmetro α, definido como a altura em cm da cernelha dos cavalos pantaneiros, representa a estimativas da altura a maturidade. O parâmetro τ determina a eficiência do crescimento animal, observando que quanto maior for o valor desse parâmetro, mais precoce é o animal e vice-versa. O parâmetro γ é denominado parâmetro de inflexão, referindo-se ao ponto em que o animal passa de uma fase de crescimento acelerado para uma fase de crescimento inibitório e aponta o ponto que o animal passa a crescer com menor eficiência. O β é o parâmetro de interceptação com o eixo-y, é utilizado apenas para adequar o valor inicial da altura fazendo-se com que a curva passe pela origem quando $y \neq 0$ e/ou $t \neq 0$, sendo t a expressão da idade.

2.3.1 Modelo de regressão Weibull

Proposto por Ernest Hjalmar Wallodi Weibull (1887- 1979) que popularizou o seu uso para análise de confiabilidade, especialmente para os modos de falha metalúrgicos, este modelo representa uma generalização da distribuição exponencial e, de acordo com Lawless (1982), é bastante utilizada no ajuste de dados de confiabilidade nas diversas áreas do conhecimento.

A distribuição Weibull é correspondente à distribuição exponencial quando $y = 1$. Sendo assim, o modelo Weibull é composto por 4 parâmetros na forma:

$$y_t = \alpha - \beta e^{-\tau t^{\gamma}} + \varepsilon, \tag{2.14}$$

em que, y representa a altura no tempo, α representa a estimativa da altura assintótica à maturidade, β é o parâmetro de interceptação com o eixo-y, τ é a taxa de maturação e o γ é o parâmetro de inflexão. É fundamental acentuar que o modelo Weibull é bastante empregado em nossa realidade, por assumir variação na forma da função de risco ou taxa de falha, que refere-se a quantidade de risco relacionada a uma determinada unidade no tempo.

Para Bolfarine (2005), a estimativa dos parâmetros é obtida a partir do método da

máxima verossimilhança, que consiste em resolver o sistema de equações gerado a partir da diferenciação do logaritmo da função de verossimilhança. As equações resultantes do processo de derivação geralmente não apresentam uma solução analítica, portanto, devem ser resolvidas a partir de um método numérico.

2.3.2 Modelo de regressão Von Bertalanffy

Segundo Oliveira *et al.* (2007), o modelo não linear Von Bertalanffy, é bastante utilizado para o estudo de crescimento em peso de peixes, para explicar a variação de comprimento de uma determinada variável ao longo do tempo. No modelo de Von Bertalanffy, y cresce para uma assíntota horizontal superior, sendo a distância que falta percorrer dada por uma exponencial decrescente, logo a especificação do modelo na família Von Bertalanffy composta por 3 parâmetros será definida por:

$$y_t = \alpha \left(1 - \beta e^{-\tau t}\right)^3 + \varepsilon, \tag{2.15}$$

De acordo com Campos (2011), o modelo acima ajusta processos de crescimento sigmoidais, no qual o ponto de inflexão está localizado aproximadamente em 30% do ultimo valor medido. Os parâmetros α representam o valor assintótico da variável resposta, β refere-se a uma constante que esta relacionada ao valor observado inicial e τ é a taxa de crescimento da variável resposta, indica a velocidade com que o valor observado se aproxima do valor máximo, determina a eficiência do crescimento.

2.3.3 Modelo de regressão Logístico

Conforme em Campos (2011), o modelo Logístico é utilizado, a partir de uma função, em que, é possível modelar o crescimento populacional. Geralmente é utilizado para crescimentos sigmoidais em que o ponto de inflexão está localizado aproximadamente na metade do último valor medido.

De acordo com Hosmer e Lemeshow (2000), por meio de uma função de ligação é possível realizar um ajuste para a resposta média ao modelo linear, assim o modelo Logístico será definido por:

$$y_t = \alpha \left(1 + \beta e^{(-\tau t)}\right)^{-\gamma} + \varepsilon, \tag{2.16}$$

Os pontos de inflexão são $y = \ln \alpha k$ e $E(y) = 2\alpha$ e independem dos valores de y.

Os parâmetros podem ser estimados pelo método de estimação por máxima verossimilhança (MV), no qual, busca fornecer valores para os parâmetros, em que α representa o valor assintótico da variável resposta, β refere-se a uma constante que esta relacionada ao valor observado inicial e τ é a taxa de crescimento da variável resposta, indica a velocidade com que o valor observado se aproxima do valor máximo, determina a eficiência do crescimento.

2.3.4 Critério para seleção de modelos

Para comparar os modelos quando ao seu ajustamento aos dados, admite-se o teste da hipótese nula de que todos os modelos são igualmente bons para este fim, contra a hipótese alternativa de que um ou mais deles são melhores que os outros. Desta forma, foram estipulados alguns critérios para comparação dos parâmetros, que podem ser mais importantes que outros.

Segundo Santos (2011), o estatístico George E. P. Box afirma que todos os modelos são errados, mas alguns modelos são úteis, ele refere-se á importância da escolha de um modelo para se modelar um evento. Ao ajustar os modelos, a comparação entre eles por meio computacional será complexa, devido a quantidade de parâmetros no modelos. Nesse contexto, avaliar a qualidade do ajuste dos modelos não lineares, se faz, com o uso de diferentes critérios.

Em modelos não lineares, o coeficiente de determinação R^2 não é uma medida de critério de seleção confiável, devido a soma dos resíduos não necessariamente serem iguais a zero e a soma dos quadrados dos resíduos mais a soma dos quadrados da regressão não necessariamente serem iguais a soma dos quadrados total. Nesse contexto, o desempenho dos modelos é selecionado a partir de um método, que é útil para comparar modelos com diferentes números de parâmetros, este método conhecido como critério de informação, como o Akaike's Information Criterion (AIC). Segundo Campos (2011), o critério de informação Akaike foi desenvolvido por Horotsugu Akaike, onde oferece uma medida relativa da informação perdida quando um modelo é usado para descrever a realidade, descreve troca de vício e a variância na construção do modelo.

O AIC (critério de informação de Akaike), é definido por:

$$AIC = -2\log L + 2(p+1), \tag{2.17}$$

em que, L é o log de verossimilhança maximizado e p é o número de parâmetros. Segundo

este critério, o melhor modelo é o que possui menor valor de AIC.

2.4 O método *bootstrap*

Segundo Souza (2008), quando no modelo de regressão não linear as hipóteses de normalidade e da aproximação linear assintótico são controverso, devido, possivelmente, ao tamanho reduzido da amostra ou à curvatura excessiva da superfície resposta, a técnica *bootstrap* de estimação se torna uma alternativa para o processo inferencial, como também como ferramenta de diagnóstico (SOUZA, 1998).

Proposto por Efron (1979) o método *bootstrap* consiste na construção de distribuições amostrais com reposição por reamostragem, a partir de um conjunto de dados, diretamente ou via um modelo ajustado, a fim de criar repetições dos dados dos quais podemos avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos. O método de *bootstrap* também pode ser utilizado, para estimar o viés e a variância de estimadores ou de testes de hipóteses calibrados. Assim, a base da idéia do pesquisador consiste em poder tratar sua amostra como se ela fosse a população que deu origem aos dados e usar amostragem com reposição da amostra original para gerar pseudo-amostras, onde por meio destas mesmas é possível estimar características da população, tais como média, variância, percentis, entre outras.

A variabilidade presente no *bootstrap* é dada pela escolha da amostra mestre e pelas reamostras, sendo a variabilidade devido à escolha da amostra mestre, a mais significativa. A distribuição *bootstrap* usualmente tem aproximadamente a mesma forma e amplitude que a distribuição amostral, porém está centrada na estatística dos dados originais (amostra mestre), enquanto a distribuição amostral está centrada no parâmetro da população.

Quando as hipóteses de normalidade e da aproximação linear assintótica no modelo de regressão não linear se tornam questionáveis pela razão, possivelmente, do tamanho reduzido da amostra, a técnica *bootstrap* se torna uma alternativa para o processo inferencial, como também como ferramenta de diagnóstico da confiabilidade dos resultados (SOUZA, 1998).

O método de *bootstrap* foi preparado para fazer simulações a respeito de fazer inferências para um tal parâmetro. O processo acontece na repetição de inúmeras vezes até a obtenção de B valores. Calcula-se o erro padrão da média destes valores, cujo erro

padrão pode ser obtido por:

$$EP = \left[\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} . \tag{2.18}$$

Para realizar o teste utilizando a técnica *bootstrap* é preciso colher uma amostra de tamanho n, que será denominada amostra mestre. Essa amostra deve ser coletada de maneira planejada, uma vez que se esta amostra for mal tirada e não representar bem a população, a técnica de *bootstrap* não levará a resultados confiáveis.

Segundo Montgomery e Runger (2003), uma estatística utilizada para estimar um parâmetro é viciada quando a distribuição amostral não estiver centrada no verdadeiro valor do parâmetro. A técnica *bootstrap* nos permite verificar que a correção de vício em modelos autorregressivos causa uma melhora substancial nos intervalos de previsão.

Seguindo a idéia de Lana (2012), o vício nos estimadores dos parâmetros pode afetar a cobertura real de intervalos de previsão. Efron e Tibshirani (1993) propõem o uso do *bootstrap* para a correção de vício. Considere o seguinte vetor de parâmetros $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$ e seja, $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, ..., \hat{\beta}_k\right)$ o estimador de β. O vicio de $\hat{\beta}$ é definido como $\Psi = E\left(\hat{\boldsymbol{\beta}}\right) - \boldsymbol{\beta}$.

A correção *bootstrap* de vício consiste em estimar β^* de cada B repetições usando-se o mesmo estimador $\hat{\boldsymbol{\beta}}$. A partir disso, calcula-se a média aritmética dos β^* , $\bar{\beta}^*$. O vício de $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Psi}$, é então estimado como sendo $\hat{\boldsymbol{\Psi}} = \bar{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}$, uma aproximação para o vício real de $\hat{\boldsymbol{\beta}}$. O estimador corrigido de vício, $\tilde{\boldsymbol{\beta}}$, é então calculado como $\tilde{\beta} = \hat{\beta} - \hat{\Psi}$ (LANA, 2012).

Segundo Souza (2008), a partir de simulações do modelo de regressão não linear, é possível estudar as propriedades dos estimadores de mínimos quadrados. Esse método é conhecido como *bootstrap* paramétrico, sendo sugerido para se estudar o comportamento amostral em modelos de regressão não linear, usar um mínimo de 1.000 amostras de dados simulados pseudo-aleatoriamente, especialmente para estatísticas baseadas em momentos amostrais grandes, tais como coeficientes de assimetria e curtose.

2.4.1 Método *bootstrap* paramétrico

De acordo com Souza (2008), o método "bootstrap"paramétrico estuda as propriedades distribucionais dos estimadores de mínimos quadrados a partir de simulações do modelo de regressão não linear em estudo. O processo desse tipo de reamostragem baseia-se em modelos com parâmetros estimados via amostra original, no qual utiliza esses

parâmetros e os erros reamostrados para ajustar os dados ao modelo específico.

O *bootstrap* paramétrico, é utilizado, fazendo-se suposição sobre a distribuição dos dados que gerou a amostra original, usando os valores das estimativas dos parâmetros no processo de geração de pseudo-amostras (BERNARDINO, 2012). A partir de uma distribuição $F\left(t\left|\hat{\theta}\right.\right)$ conhecida, este método simula amostras de tamanho n.

Efron e Tibshirani (1993) apresentou um método *bootstrap* em que é feito suposição sobre a distribuição dos dados que gerou a amostra original. Uma distribuição $F\left(t\left|\hat{\theta}\right.\right)$ pode ser a função densidade da distribuição Normal com $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma}^2)$, em que $\hat{\mu}$ e $\hat{\sigma}^2$ são os estimadores de máxima verossimilhança de μ e σ^2 respectivamente.

2.4.2 Método *bootstrap* não paramétrico

De acordo com Bernardino (2012), quando nenhum modelo matemático do tipo paramétrico é usado, a analise estatística é não paramétrica, e usa apenas o fato de que as variáveis aleatórias y_i são independentes e identicamente distribuídas. Mesmo se houver um modelo paramétrico plausível, uma analise não-paramétrica pode ainda ser útil para avaliar a robustez das conclusões de uma analise paramétrica. No *bootstrap* não paramétrico, o processo de reamostragem se dá a partir da função de distribuição empírica dos dados.

Por exemplo, seja $\boldsymbol{t} = (t_1, ..., t_n)$ uma amostra contendo n observações. Constrói-se então, B amostras $T^{*(1)}, ..., T^{*(B)}$ independentes, onde cada amostra é obtida a partir da reamostragem da amostra finita inicial $\boldsymbol{t} = (t_1, ..., t_n)$. Para cada uma das $T^{*(1)}, ..., T^{*(B)}$ amostras, estima-se os parâmetros de interesse.

2.4.3 Intervalos de confiança *bootstrap*

Segundo Manteiga (1994), uma das aplicações da metodologia *bootstrap* é obter intervalos de confiança confiáveis, sendo assim é essencial em regressão não linear a realização de intervalos de confiança para os parâmetros do modelo. Souza (2008) e Hall (1988) descrevem dois métodos para a obtenção de intervalos de confiança *bootstrap* - método percentil e método percentil t.

Segundo Souza (1998), os intervalos obtidos via método percentil tem por base unicamente os quantis e outras medidas de distribuição *bootstrap* do estimador de interesse

$\hat{\gamma}$. Os intervalos gerados via o método perecntil t tem a forma:

$$\hat{\gamma} - t_1 s(\hat{\gamma}) \leq \gamma \leq \hat{\gamma} - t_2 s(\hat{\gamma}), \tag{2.19}$$

em que $s(\hat{\gamma})$ é o desvio padrão (estimado) de $\hat{\gamma}$. Os t_i são determinados com base na distribuição *bootstrap* de $\hat{\gamma}$.

Considerando $A = (x_1, x_2, ..., x_n)$ uma amostra aleatória de tamanho n de uma população com função de distribuição $F(x)$, média μ e variância finita σ^2, obtêm-se o intervalo de confiança percentis t para μ, a $100(1-\alpha)\%$ tem a forma (SOUZA, 1998):

$$IC = \left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right] \tag{2.20}$$

em que $t_\beta^{(1)}$ satisfaz

$$P_{\hat{F}}\left\{\sqrt{n}\left(\frac{\bar{x}^* - \bar{x}}{s^*}\right) \leq t\right\} = \beta, \tag{2.21}$$

sendo $t_\beta^{(1)}$ os valores de cada parâmetro.

Rizzo e Cymrot (2006) propõem que um intervalo de confiança realizado por qualquer método *bootstrap* com $100(1-\alpha)\%$, devem-se rejeitar com $100\alpha\%$ as hipóteses nulas de que o parâmetro estimado seja igual a qualquer valor fora desse intervalo e deve se aceitar com $100\alpha\%$ as hipóteses nulas de que este mesmo parâmetro seja igual a qualquer valor dentro do intervalo de confiança.

2.4.4 Testes de normalidade *bootstrap*

De acordo com Samohyl (2009), o teste Shapiro-Wilk foi proposto em 1965, é baseado nos p-valores do teste, quando estes p-valores for menor que o nível de significância fixado, indica que há evidências de que os dados são normais. A hipótese do teste, é definida como:

$$\begin{cases} H_0 : \text{Os parâmetros provém de uma distribuição Normal} \\ H_1 : \text{Os parâmetros não provém de uma distribuição Normal.} \end{cases} \tag{2.22}$$

Segundo Samohyl (2009), assimetria é uma medida da simetria da distribuição em torno da mediana e da média. No caso da distribuição normal, exatamente a metade das observações fica de um lado da média, que coincide com a mediana. Quando as duas metades da distribuição são idênticas, uma espelhando a outra, então a distribuição é simétrica, e já poderia ser da classe de distribuições normais. O teste de assimetria é baseado nos p-valores do teste, aceitando a hipótese nula quando o p-valor for menor que

o nível de significância fixado. A hipótese do teste, é definida como:

$$\begin{cases} H_0 : \text{Coeficientes da assimetria } = 0 \\ H_1 : \text{Coeficientes da assimetria } \neq 0 \end{cases} \tag{2.23}$$

Seguindo a idéia de Samohyl (2009), o coeficiente da curtose deve ser igual a três para que a distribuição dos dados seja normal. A hipótese do teste, é definida como:

$$\begin{cases} H_0 : \text{Coeficientes da curtose } = 3 \\ H_1 : \text{Coeficientes da curtose } \neq 3 \end{cases} \tag{2.24}$$

3 Aplicação

Encontram-se nesta seção as principais metodologias que serviram de base para este trabalho, tanto na parte da descrição dos dados utilizando-se os modelos não lineares, quanto nas inferências realizadas via aproximação e simulação *bootstrap*.

3.1 Material e métodos

Com a intenção de aplicar a teoria abordada, foi utilizado uma amostra de 15 observações de crescimento animal obtidos em (SOUZA, 1998). Considerando-se que, a variável resposta y refere-se a altura na cernelha em cm de um cavalo pantaneiro em relação a idade em dias que esta dividida por 1800 com o propósito de limitar a medida de tempo ao intervalo (0,1). A Tabela 1 apresenta a relação dos 15 animais com seus respectivos valores, altura (cm) e idade (dias).

Tabela 1: Relação da altura(cm) da cernelha e da idade(dias) de cavalos pantaneiros

Animal	Altura	Idade
1	86	5
2	96	54
3	107	131
4	114	195
5	121	286
6	124	412
7	124	476
8	129	586
9	131	682
10	135	785
11	135	817
12	135	874
13	138	1494
14	138	1530
15	138	1711

Fonte: Oliveira e Souza (1996)

A Figura 1, apresenta a curva observada da altura em função do tempo para cada animal. Por meio desta figura nota-se que o crescimento é não linear, recomendando-se a necessidade de aplicação de um modelo não linear, para descrição dos dados.

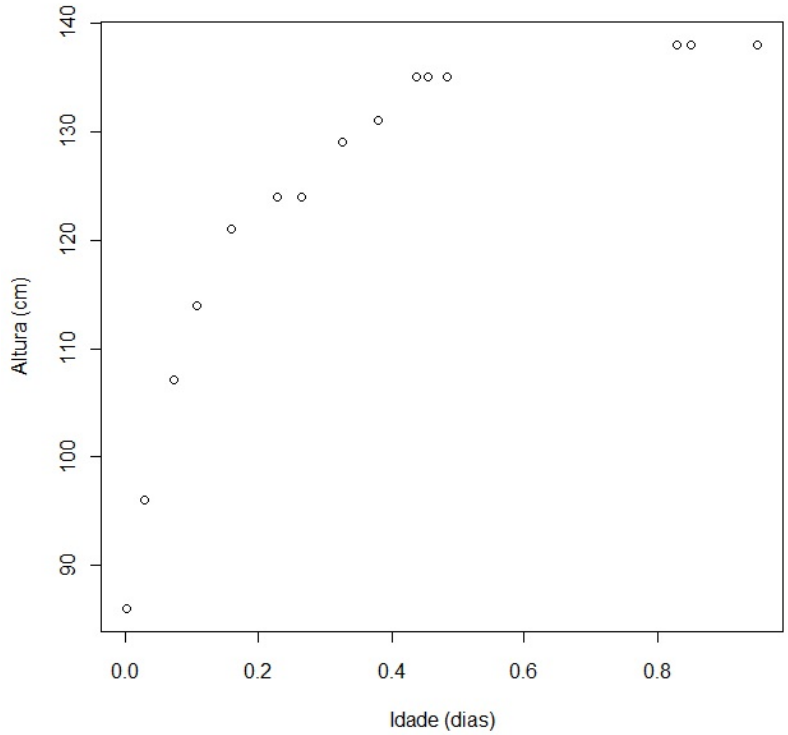

Figura 1: Relação entre a altura (cm) dos animais ao longo dos dias.

Foram ajustados os três modelos não lineares descritos anteriormente, o Weibull (2.14), o Von Bertalanffy (2.15) e o Logístico (2.16), com o propósito de estimar o crescimento dos animais em função da idade. Para os três modelos calculou-se as estimativas dos parâmetros, o erro padrão, a estatística do test t para verificar se os parâmetros são significativos e os intervalos de confiança dos respectivos modelos, nos quais foi aplicado a técnica *bootstrap* com a finalidade de verificar nos estimadores , o seu erro padrão, o seu viés, e verificar a qualidade do ajuste entre os modelos, validando-os com confiança nos resultados. Por meio da função "nls" do *Software* livre R versão 2.14.2, foi possível obter as estimativas dos parâmetros.

Para o ajuste do modelo foi utilizado o método iterativo de Gauss Newton. Este processo começa dando-se valores iniciais aos parâmetros, buscando-se aproximar o modelo não linear à termos lineares. Após supor os valores iniciais, utilizando a série de Taylor, ocorre várias iterações até convergir, em seguida é aplicado o método de mínimos quadrados para encontrar as estimativas aproximadas dos estimadores dos parâmetros, obtendo assim, o problema de regressão não linear. Por meio do vício de Box, pode-se verificar se

existe parâmetros que contribuam para o desvio da linearidade.

Como critério de seleção foi considerado as estatísticas do coeficiente de determinação R^2 e o critério de informação de Akaike (AIC), no qual relaciona a discrepância, medida que existe entre o modelo verdadeiro e o modelo aproximado por meio da máxima verossimilhança. Os modelos foram comparados por meio dos critérios de seleção, no qual, foi possível verificar qual modelo melhor se ajustou aos dados.

O método *bootstrap* paramétrico, iniciou-se a partir dos valores assintóticos estimados dos parâmetros da amostra original. Para as 1000 estimativas *bootstrap* obtidas na simulação, foi calculado o vício relativo *bootstrap*, os intervalos de confiança e, realizado os testes de normalidade dos parâmetros, os quais foram: Shapiro-Wilk, assimetria e curtose, com o objetivo de medir a dispersão dos dados e verificar a normalidade na distribuição para cada modelo utilizado. As medidas destes testes, tiveram como fonte os resíduos das distribuições, pois se assumir que os modelos a serem testados seriam os mais adequados, os resíduos refletiriam as propriedades assumidas pela variável de erro.

3.1.1 Resultados e Discussão

Considerando-se todos os dados, foi possível ajustar os modelos não lineares em função do comportamento dos dados e obter as estimativas dos parâmetros para cada um dos modelos (2.14), (2.15) e (2.16), que relaciona altura com a idade do animal. A Tabela 2 apresenta os resultados das estimativas dos estimadores dos parâmetros, para os três modelos não lineares ajustados, com seus respectivos erro padrão, teste t com p-valor e o intervalo de confiança.

Tabela 2: Estimativas (E) dos parâmetros (P), o erro padrão (E.P.), teste t e seus respectivos intervalos de confiança IC a 95%

Modelo	P.	E.	E.P.	valor t	p-valor	IC
Von Bertalanffy	α	137,4631	0,9580	$143,4800^*$	$< 0,001$	[135,3757 ; 139,5505]
	β	6,1852	0,4810	$12,8600^*$	$< 0,001$	[5,1370 ; 7,2333]
	τ	-0,1626	0,0147	$-11,0300^*$	$< 0,001$	[-0,1947 ; -0,1304]
Logístico	α	136,7737	1,1161	$122,5300^*$	$< 0,001$	[134,3417 ; 139,2056]
	β	-0,0782	0,0119	$-6,5300^*$	$< 0,001$	[-0,1044 ; -0,0521]
	τ	0,1277	0,0119	$10,7200^*$	$< 0,001$	[0,1017 ; 0,1537]
Weibull	α	139,5219	1,3313	$104,8000^*$	$< 0,001$	[136,5915 ; 142,4522]
	β	56,1989	2,7365	$20,5400^*$	$< 0,001$	[50,1759 ; 62,2219]
	τ	4,1889	0,5908	$7,0900^*$	$< 0,001$	[2,8884 ; 5,4894]
	γ	0,7747	0,0699	$11,0800^*$	$< 0,001$	[0,6208 ; 0,9286]

Por meio da Tabela 2, é possível perceber que todos os parâmetros dos três modelos são significativos ao nível de 0,05 de significância pelo teste t, devido aos p-valores serem menor que 0,05. As estimativas dos parâmetros foram distintas, portanto variam de modelo para modelo, mas, é possível notar que o modelo Weibull teve uma melhor aproximação do valor inicial suposto. Os limites inferiores e superiores dos intervalos de confiança a 95% de cada estimativa possuem estreita amplitude de estimação. O modelo Weibull foi o que apresentou a amplitude menos estreita, devido o tamanho do erro padrão, dentre os modelos ajustados.

A Figura 2 apresenta as curva do crescimento animal do ajuste dos modelos Weibull, Von Bertalanffy e Logístico, respectivamente, em que pode-se observar que o ajuste foi semelhante entre os 3 modelos.

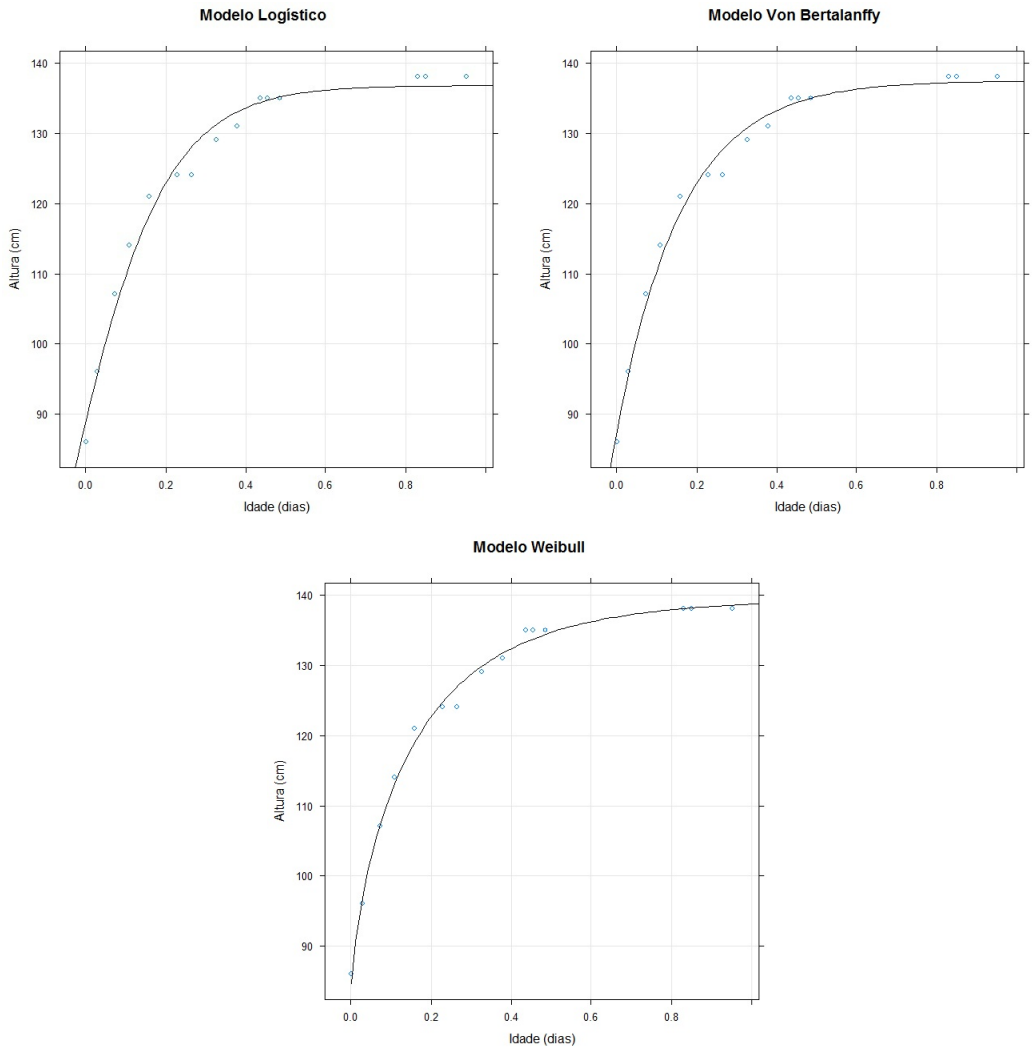

Figura 2: Ajuste da altura do animal em relação a altura para os modelos Weibull, Von Bertalanffy e o Logístico

A qualidade de ajuste dos modelos aos dados, foi analisada por dois critérios: Informação de Akaike (AIC) e coeficiente de determinação R^2. Na Tabela 3 encontram-se os valores das estatística R^2 e AIC. Assim o melhor modelo ajustado é aquele que apresentar menor valor para AIC e maior valor para R^2.

Tabela 3: Estatísticas da qualidade de informação de ajuste, coeficiente de determinação (R^2) e AIC (Informação de Akaike).

	Von Bertalanffy	Logístico	Weibull
R^2	0,9886	0,9818	0,9818
AIC	65,7267	72,8286	57,9341

Pode-se observar na Tabela 3 que os coeficientes de determinação obtidos foram altos, em torno de 98%, contudo, embora os coeficientes de determinação tenham sido praticamente iguais para o três modelos, o modelo que apresentou o menor valor para AIC, foi o modelo Weibull. Nesse contexto, pode-se notar que o Weibull apresenta mais parâmetros no modelo em relação ao Von Bertalanffy e Logístico, em que, a presença de mais parâmetros pode ser um indicativo de melhor interpretação dos dados.

Após a escolha do melhor modelo ajustado aos dados, deu-se a continuidade com a análise gráfica referentes a homocedasticidade e normalidade dos resíduos, apresentados na Figura 3, dos modelos Weibull, Von Bertalanffy e Logístico, respectivamente. Nestas Figuras apresenta-se os resíduos *versus* idade (dias) e valores observados *versus* valores preditos.

Na Figura 3, resíduos *versus* valores ajustados, estão referentes a homocedasticidade dos resíduos, no qual pode-se observar que a variância é constante, os resíduos se distribuem aleatoriamente em torno da média zero, apresentando-se uma homogeneidade de variâncias ao longo do tempo, no qual, mostra que o modelo de regressão não linear está adequado. Pelo comportamento gráfico da Figura 3, valores preditos *versus* valores observados, referentes a normalidade, observa-se que os pontos caem próximos da reta, no qual existe evidências que os resíduos segue uma distribuição normal. Pode-se notar que o ajuste e a análise de resíduos dos três modelos são bastante semelhantes.

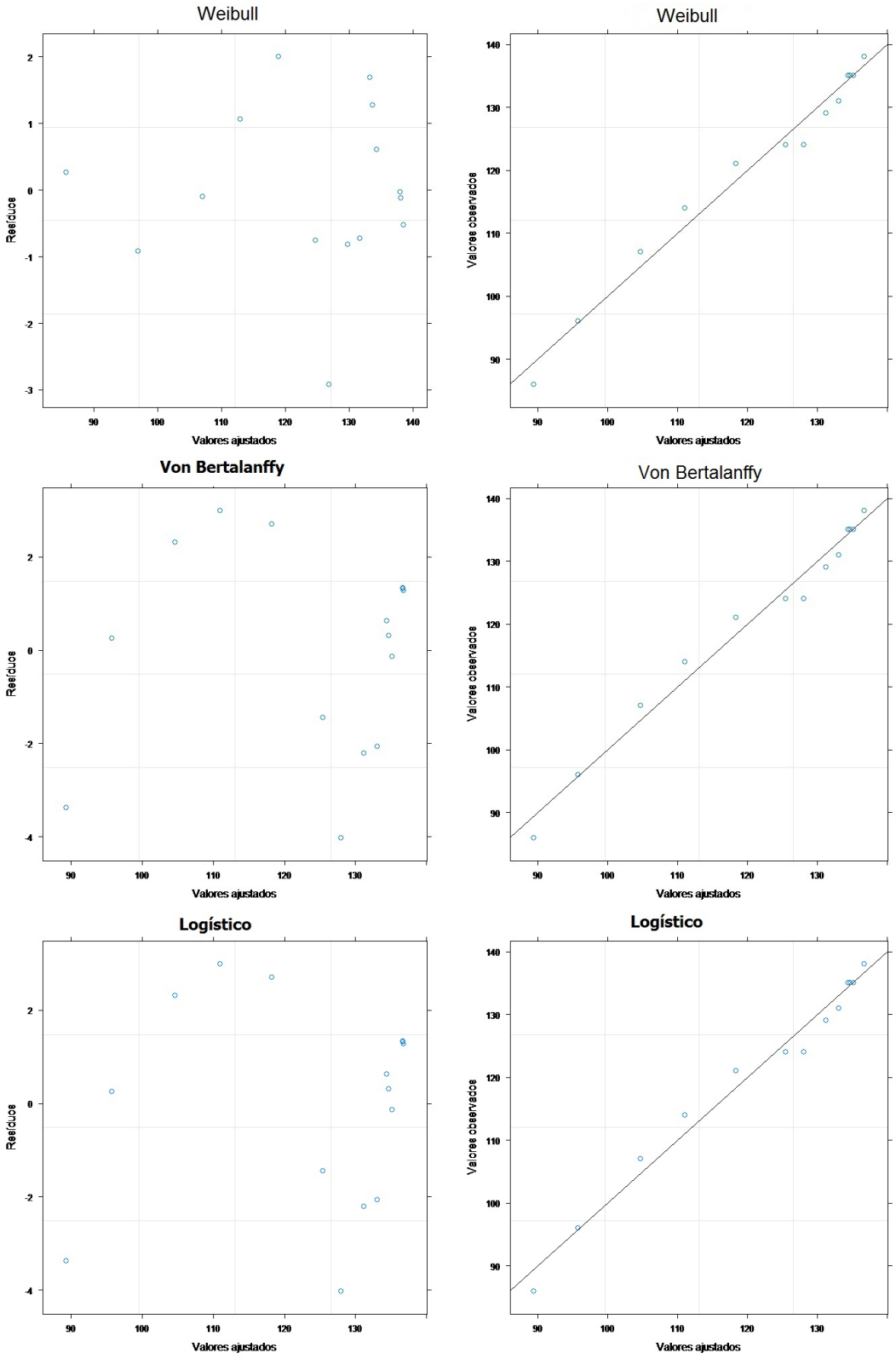

Figura 3: Gráfico da Análise dos Resíduos

3.1.2 Simulação *bootstrap*

A Tabela 4 apresenta as estimativas dos parâmetros pelo *bootstrap* nos três modelos, seus respectivos valores do vício relativo a estimativa *bootstrap* (V.R.B.), vício de Box e o intervalo de confiança.

Tabela 4: Estimativas *bootstrap* (E.B.), vício relativo a estimativa *bootstrap*(V.R.B.), vício de Box, intervalo de confiança *bootstrap* a 95% para os parâmetros (p)

Modelo	p	E.B.	V.R.B	Vício de Box	Intervalo de Confiança
Von Bert.	α	137,4669	3,4144	0,0252	[135,7577; 139,0044]
	β	6,1956	2,7324	0,2748	[5,4096; 7,1150]
	τ	-0,1630	3,4013	0,3465	[-0,1895; -0,1373]
Logístico	α	136,7679	-0,0062	0,0342	[134,8209; 138,6026]
	β	-0,0788	0,8561	0,4501	[-0,1010; -0,0588]
	τ	0,1276	0,3161	0,4030	[0,1081; 0,1497]
Weibull	α	139,5836	0,0908	0,1041	[137,6181; 142,1690]
	β	56,2649	0,3616	0,4606	[52,0811; 62,0539]
	τ	4,1856	0,8134	1,0908	[3,2880; 5,3225]
	γ	0,7750	0,2016	0,2515	[0,6501; 0,8973]

Pode-se observar por meio da Tabela 4, que os valores das estimativas dos parâmetros e seus respectivos intervalos de confiança via *bootstrap*, são muito semelhantes com os valores assintóticos. No qual, a amplitude dos intervalos de confiança assintóticos e *bootstrap* aponta não haver indicativo de excesso de variância para os estimadores de mínimos quadrados. O vício *bootstrap* apresentou valores próximos ao vício de Box no modelo Weibull, porém, os três modelos apresentam vício pequeno inferior a 1%, com excessão do parâmetro τ do modelo Weibull, com vício de Box maior que 1%. Portanto este parâmetro pode contribuir à não linearidade do modelo. Tal resultado indica que o vício de Box estima bem o vício nas estimativas dos parâmetros.

A Tabela 5, logo abaixo, apresenta os resultados dos testes de normalidade para os parâmetros via método *bootstrap*, no qual, encontra-se os valores do teste de Shapiro-Wilk, do teste de assimetria e do teste de curtose, com seus respectivos p-valores. Nota-se para o teste Shapiro-Wilk, que os parâmetros não significativos aceitam a hipótese nula do teste, já os parâmetros significativos rejeitam a hipótese nula, no qual, esta hipótese nula evidencia que os parâmetros seguem uma distribuição normal. Neste caso, o modelo Weibull apresenta mais parâmetros com indícios de normalidade. Para o teste de

Tabela 5: Testes de normalidade *bootstrap*, Shapiro-Wilk, assimetria e curtose e seus respectivos p-valores para os parâmetros (p)

Modelo	p	Shapiro-Wilk	p-valor	Assimetria	p-valor	Curtose	p-valor
Von Bert.	α	$0,9971^{*}$	0,0697	$-0,1095_{ns}$	0,3490	$2,6835^{*}$	0,0193
	β	$0,9967_{ns}$	0,0276	$0,1518_{ns}$	0,1954	$2,8632_{ns}$	0,3997
	τ	$0,9976^{*}$	0,1759	$-0,1198_{ns}$	0,3060	$2,8502_{ns}$	0,3468
Logístico	α	$0,9973^{*}$	0,1066	$-0,0646_{ns}$	0,5796	$2,6489^{*}$	0,0076
	β	$0,9961_{ns}$	0,0129	$-0,1356_{ns}$	0,2472	$2,7786_{ns}$	0,1316
	τ	$0,9961_{ns}$	0,0001	$0,2989^{*}$	0,0121	$3,0053_{ns}$	0,8818
Weibull	α	$0,9927_{ns}$	$< 0,0001$	$0,4896^{*}$	$< 0,0001$	$3,8559^{*}$	$< 0,0001$
	β	$0,9930_{ns}$	$< 0,0001$	$0,5716^{*}$	$< 0,0001$	$3,9778^{*}$	$< 0,0001$
	τ	$0,9835_{ns}$	0,0008	$0,2968^{*}$	0,0126	$3,1056_{ns}$	0,4417
	γ	$0,9957^{*}$	0,4070	$-0,0196_{ns}$	0,8664	$3,0466_{ns}$	0,6806

* siginificativo ao nível de 5% de probabilidade e ns não significativo

assimetria, os parâmetros significativos ao teste, rejeitam a hipótese nula, já os parâmetros não significativos ao teste, aceitam a hipótese nula que é referente a normalidade, sendo os coeficientes nulos assim representando uma simetria, e portanto seguem uma distribuição normal. Para o teste de curtose, aceitamos a hipótese de igualdade, quando os parâmetros são não significativos, em que a curtose é muito próxima ou igual a 3, ou seja, essa curtose tem o mesmo achatamento que a distribuição normal. Portanto, para estes dois testes, os três modelos apresentam pelo menos dois parâmetros com indícios de normalidade.

Logo abaixo a Figura 4 apresenta o gráfico de ditribuição de frequências para cada parâmetro de cada modelo não linear ajustado, em que, pode-se fazer jus aos testes de normalidade que encontram-se na Tabela 5.

Observando-se os gráficos da Figura 4, pode-se verificar que, analisar o teste de assimetria e curtose pelos seus respectivos coeficientes e p-valores, é de fácil diagnóstico em comparação a visualização gráfica, devido, os histogramas aparentarem uma leve assimetria e, em alguns parâmetros aparentam uma simetria. Os parâmetros com calda alongada à direita, apresentam assimetria positiva e os parâmetros com calda alongada à esquerda, apresentam assimetria negativa. Contudo, os desvios de assimetria não são muito grandes.

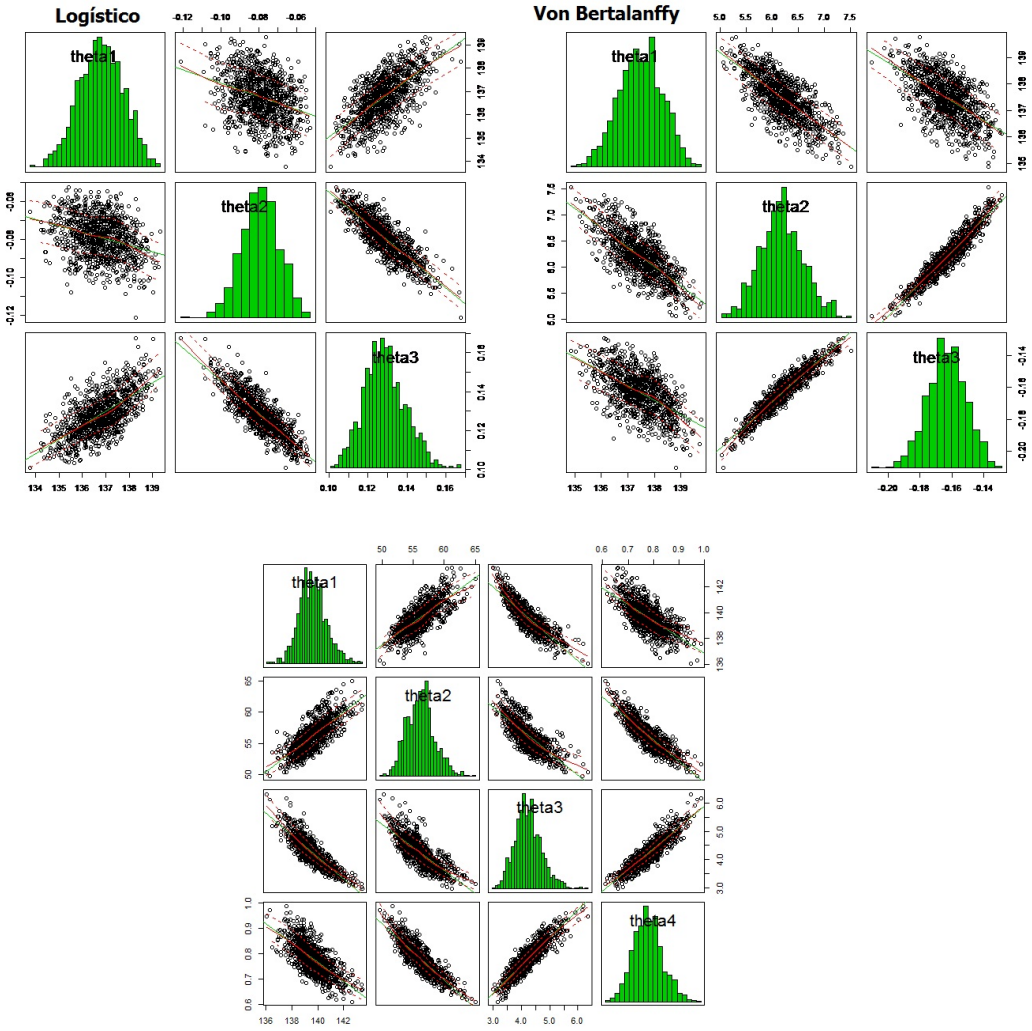

Figura 4: Matrizes de dispersão conjunta de estimativas *bootstrap* dos parâmetros e os histogramas de frequência para os modelos analisados

4 Conclusão

Com os estudos desenvolvidos ao longo deste trabalho, de maneira geral, nota-se que os modelos não linear Weibull, Von Bertalanffy e Logístico, apresentaram uma considerável qualidade de ajuste para os dados de curvas de crescimento. De acordo com o critério de seleção AIC, o modelo Weibull foi o mais adequado para interpretar a curva de crescimento dos dados. Do ponto de vista inferencial e pelo estudo de simulação *bootstrap*, os três modelos apresentaram valores assintóticos e *bootstrap* muito semelhantes, além de variância pequena na amplitude dos intervalos de confiança assintóticos.

Para os modelos Weibull, Von Bertalanffy e Logístico, os desvios de normalidade estiveram associados tanto à assimetria quanto à curtose, porém o modelo Weibull apresentou maiores desvios de assimetria, mesmo sendo considerado o melhor modelo pelo critério de seleção. Contudo, por meio do método *bootstrap*, os resultados assintóticos revelam-se confiáveis.

Referências

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

BERNARDINO, R. *Método Bootstrap*. Universidade Federal do Amazonas: [s.n.], 2012. Ebah. Acesso em: 25/11/2012. Disponível em: <http://www.ebah.com.br/content/ABAAAAr8cAA/metodo-bootstrap>.

BOLFARINE, H. *Testar a homogeneidade no Weibull - Regressão modelos*. São Paulo: Jornal Biométrica, 2005. 720 p.

CAMPOS, A. M. *Uma abordagem bayesiana para alguns modelos de crescimento na presença de assimetria e heteroscedasticidade*. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2011.

DRAPER, N. R.; SMITH, H. *Applied Regression Analyses*. 3. ed. New York: Wiley, 1998. 736 p.

EFRON, B.; TIBSHIRANI, R. J. *And Introduction to the Bootstrap*. New York: Chapman and Hall, 1993. 436 p.

GALLANT, R. A. *Nonlinear Statistical Models*. New York: Wiley series in probability e mathematical statistics, 1987. 624 p.

GUJARATI, D. N. *Econometria Básica*. 4. ed. São Paulo: Elsevier, 2006. 860 p.

HALL, P. *Theoretical Comparison of Bootstrap Confidence Intervals*. The Annals of Statistics, 1988. 927 p. Disponível em: <http://www.jstor.org/stable/2241604>.

HOFFMANN, R. *Análise de Regressão - Uma Introdução à Econometria*. 4. ed. São Paulo: Hucitec, 2006.

HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. *Jhon Wiley e Sons*, 2000.

CARVALHO LANA, G. de. *Intervalos de previsão em modelos ARFIMA utilizando a metodologia bootstrap*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: Wiley, 1982.

GONZáLEZ MANTEIGA, P. S. e. R. U. *The Bootstrap - a review*. [S.l.]: Computational statistics, 1994. 165-205 p.

MAZUCHELI, J.; ACHCAR, J. A. Algumas considerações em regressão não linear. *Acta Scientiarum*, 2002.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística Aplicada e Probabilidade para Engenheiros*. 2. ed. Rio de Janeiro: LTC, 2003.

OLIVEIRA, L. de; BRANDãO, A. J. V.; BASSANEZI, R. C. O modelo de von bertalanffy generalizado aplicado ao crescimento de suínos de corte. *Biomatemática*, 2007.

RIZZO, A. L. T.; CYMROT, R. Utilização da técnica de reamostragem bootstrap em aplicação na engenharia de produção. *Universidade Presbiteriana Mackenzie*, 2006.

SAMOHYL, R. W. *Controle Estatístico de Qualidade*. 1. ed. [S.l.]: Campus, 2009.

SANTOS, A. L. P. dos. *Estudo de modelo de crescimento via dados simulados*. Monografia (Graduação) — Universidade Estadual da Paraíba, Campina Grande, 2011.

SOUZA, E. M. de. *Modelagem não-linear da extração de zinco em um solo tratado com lodo de esgoto*. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras - Minas Gerais, 2008.

SOUZA, G. da Silva e. *Introdução aos modelos de regressão linear e não-linear*. [S.l.]: EMBRAPA-SEA, 1998.

ZEVIANI, W. M. *Avaliação de modelos de regressão não linear na cinética de liberação de potássio de resíduos orgânicos*. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras - Minas Gerais, 2009.

5 Anexo I

Rotina R para ajuste do modelo não linear Weibll

Pacotes utilizados:

```
library(lattice)

library(nlstools)

library(MASS)

library(moments)

library(car)

library(NRAIA)

library(nlme)

# Ajuste do modelo Weibull

t=c(5,54,131,195,286,412,476,586,682,785,817,874,1494,1530,1711)

y=c(y=c(86,96,107,114,121,124,124,129,131,135,135,135,138,138,138)

x=t/1800

#calcular as derivadas

mmcurve <- quoc.der <- deriv3(~(theta1-theta2*exp(-theta3*x^theta4)),"

function(theta1, theta2, theta3, theta4, x)

NULL) modelo.weibul <- nls(y~quoc.der(theta1, theta2, theta3, theta4, x),

 data=dados, start=c(theta1=140,theta2=56,theta3=5,theta4=1)),
```

```
summary(modelo.weibul)

#Gráfico do modelo ajustado

plotfit(modelo.weibul, xlab = Idade (dias)",",

ylab = Altura (cm)",main = "Modelo Weibull")"

#Critério de Seleção do modelo

AIC(modelo.weibul)

#coeficiente de determinação

R2 <- 1-deviance(m0qm)/deviance(lm(Amedio~1, data=lmedio));R2

teste das pressuposições da análise

shapiro.test(residuals(modelo.weibul))

#Gráfico dos resíduos x valores preditos

plot(modelo.weibul,resid(.)~fitted(.),

ylab=Resíduos", xlab="Valores ajustados")"

#Valores dos resíduos

resid(modelo.weibul)

#Valores ajustados

fitted(modelo.weibul)

plot(modelo.weibul,y~fitted(.),abline=c(0,1),

ylab=Valores observados", xlab="Valores ajustados")"

#intervalo de confiança assintótico

sm <- summary(modelo.weibul)$coef

cbind(sm[,1]-sm[,2]*qt(0.975, df=df.residual(modelo.weibul)),

sm[,1]+sm[,2]*qt(0.975, df=df.residual(modelo.weibul)))

#Função para cálculo do vício de Box (1971)
```

```
biasbox <- function(nls.obj){

theta <- summary(nls.obj)$coef[,1]

sd.theta <- summary(nls.obj)$coef[,2]

F <- attr(nls.obj$m$fitted(), gradient")"

H <- attr(nls.obj$m$fitted(), hessian")"

sig <- summary(nls.obj)$sigma

n <- dim(F)[1]

 FlFi <- t(F)%*%F

d <- -(sig^2/2)*sapply(1:n, function(x){

sum(diag(solve(FlFi)%*%H[x, , ]))})

bias <- as.vector(solve(FlFi)%*%t(F)%*%d)

names(bias) <- names(coef(nls.obj))

bias.sd <- 100*bias/sd.theta

bias.th <- 100*bias/theta

return(list(viés bruto"=bias,"

#vício de Box

biasbox(modelo.weibul)

#simulação bootstrap

m0q.sim <- nlsBoot(modelo.weibul, niter=1000)

#vício bootstrap relativo a estimativa

100*(apply(m0q.sim$coefboot, 2, mean)-c(sm[,1]))/c(sm[,1])

#teste de normalidade

apply(m0q.sim$coefboot, 2, function(x){data.frame

(W=shapiro.test(x)$statistic,pval=shapiro.test(x)$p.value)})

#teste de assimetria
```

```
apply(m0q.sim$coefboot, 2, function(x){data.frame
(A=agostino.test(x)$statistic[1],pval=agostino.test(x)$p.value)})

#teste de curtose

apply(m0q.sim$coefboot, 2, function(x)
{data.frame(C=anscombe.test(x)$statistic[1],pval=anscombe.test(x)$p.value)}

#gráfico de frequências

scatterplot.matrix(m0q.sim$coefboot, diagonal=histogram")"
```