



LARISSA RIBEIRO DE ANDRADE

***BOOTSTRAP* COM MEDIDAS DE INFLUÊNCIA
NA ESTIMAÇÃO DOS PARÂMETROS DOS
MODELOS GOMPERTZ E LOGÍSTICO**

LAVRAS – MG

2012

LARISSA RIBEIRO DE ANDRADE

***BOOTSTRAP* COM MEDIDAS DE INFLUÊNCIA NA ESTIMAÇÃO DOS
PARÂMETROS DOS MODELOS GOMPERTZ E LOGÍSTICO**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador
Dr. Marcelo Angelo Cirillo

**LAVRAS – MG
2012**

**Ficha Catalográfica Elaborada pela Divisão de Processos Técnicos da
Biblioteca da UFLA**

Andrade, Larissa Ribeiro de.

Bootstrap com medidas de influência na estimação dos parâmetros dos modelos Gompertz e logístico / Larissa Ribeiro de Andrade. – Lavras : UFLA, 2012.

74p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2012.

Orientador: Marcelo Angelo Cirillo.

Bibliografia.

1. Monte Carlo. 2. CovRatio. 3. Viés. 4. Precisão. 5. Regressão não linear. I. Universidade Federal de Lavras. II. Título.

CDD – 519.54

LARISSA RIBEIRO DE ANDRADE

***BOOTSTRAP* COM MEDIDAS DE INFLUÊNCIA NA ESTIMAÇÃO DOS
PARÂMETROS DOS MODELOS GOMPERTZ E LOGÍSTICO**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 24 de fevereiro de 2012.

Dr. Joel Augusto Muniz	UFLA
Dr. Julio Sílvio de Sousa Bueno Filho	UFLA
Dr. Luiz Alberto Beijo	UNIFAL

Dr. Marcelo Angelo Cirillo
Orientador

**LAVRAS – MG
2012**

A meus pais, Roberto e Jussara.

DEDICO

AGRADECIMENTOS

À Universidade Federal de Lavras e ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes/REUNI), pela concessão de bolsa de estudos.

Ao professor Dr. Marcelo Angelo Cirillo, pela orientação, dedicação e paciência durante o desenvolvimento da dissertação.

Aos professores Dr. Joel Augusto Muniz, Dr. Julio Sílvio de Sousa Bueno Filho e Dr. Luiz Alberto Beijo pelas contribuições. Aos demais professores do Departamento de Ciências Exatas, bem como a todos os funcionários do DEX-UFLA.

A meus pais, Roberto e Jussara, grandes incentivadores de minha formação, e às minhas irmãs, Priscilla e Laís, pela preocupação e amizade. À minha querida vovó Arlete pelas orações, e ao meu cunhado Marcelo, por ser prestativo sempre que preciso.

Aos meus padrinhos, Vera e Maurício, e aos meus primos, em especial à Florence, pelo constante apoio à minha formação. Aos meus tios Andiara e Ademir pelo carinho.

Às minhas amigas Ana Célia, Cibelle, Kelly e Maria Rita, desde o início acompanhando minhas dificuldades e alegrias. À Danielle, Juscélia e Tábata, amigas e companheiras. Às minhas companheiras de república Danielle e Isabelle, obrigada pela confiança e amizade.

A EST 2010: Adriana, Adriele, André, Danielle, Elayne, Iábita, Juliano, Juracy, Juscélia, Leandro, Lourenço, Marcelo, Mariele e Tábata, o grupo de estudos mais divertido da UFLA.

LISTA DE TABELAS

Tabela 1	Resumo dos cenários avaliados na simulação Monte Carlo e <i>bootstrap</i>	43
Tabela 2	Percentual de amostras válidas que apresentaram convergência do Gauss-Newton para os modelos Logístico e Gompertz.	48
Tabela 3	Resultados referentes aos vieses relativos das estimativas do parâmetro α referente a assíntota superior dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.....	50
Tabela 4	Resultados referentes dos vieses relativos das estimativas do parâmetro β referente ao intercepto da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.	52
Tabela 5	Resultados referentes dos vieses relativos das estimativas do parâmetro γ referente a taxa média de crescimento dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.	54
Tabela 6	Resultados referentes ao desvio padrão das estimativas do parâmetro α referente a assíntota superior dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando	

	diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.....	55
Tabela 7	Resultados referentes ao desvio padrão das estimativas do parâmetro β referente ao intercepto da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.	57
Tabela 8	Resultados referentes ao desvio padrão das estimativas do parâmetro γ referente a taxa média de crescimento da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de <i>outliers</i> presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.	58

LISTA DE FIGURAS

Figura 1	Ilustração da ortogonalidade entre o vetor residual $y \sim - X\beta \sim$ e as colunas da matriz X	16
Figura 2	Representação da região de solução dos dados: (a) Modelo $EY = X\beta$; (b) Modelo $EY = X \log \phi$	18
Figura 3	Representação gráfica do modelo logístico e ponto de inflexão (linhas vermelhas)	23
Figura 4	(a) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro α ($\alpha = 1$ em preto, $\alpha = 2$ em vermelho e $\alpha = 4$ em azul, $\beta = 1$, $\gamma = 1$); (b) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro β ($\beta = 0,5$ em preto, $\beta = 1$ em vermelho e $\beta = 2$ em azul, $\alpha = 4$, $\gamma = 3$); (c) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro γ ($\gamma = 1$ em preto, $\gamma = 3$ em vermelho e $\gamma = 5$ em azul, $\alpha = 4$, $\beta = 2$).....	24
Figura 5	Representação gráfica das distribuições utilizadas na geração de <i>outliers</i> nos resíduos (a) Beta(6,2); (b) Beta(6,6) e (c) Beta(2,6).....	42

RESUMO

Em geral o método *bootstrap* é realizado supondo que cada unidade amostral apresente a mesma probabilidade de ser reamostrada. Contudo, ao considerar uma amostra que apresente *outliers*, a distribuição empírica gerada através da execução desse método pode ser influenciada, no sentido de não representar fielmente a amostra original. Tendo por base este problema, o objetivo desse trabalho consistiu em propor um método *bootstrap* que permita o pesquisador especificar uma medida de influência, a ser utilizada como critério para obter probabilidades de reamostragem distintas para cada observação. Tal método foi denominado por BMIC originado de “*Bootstrap* com medidas de influência corrigida pelo viés de Monte Carlo”. Com o propósito de validar esse método, estudos de simulação foram realizados considerando o modelo logístico e Gompertz, em alguns cenários caracterizados por diferentes tamanhos amostrais, distribuições simétricas e assimétricas nos resíduos e quantidade de *outliers* contida na amostra. Concluiu-se que a aplicação do método BMIC proporcionou uma melhoria na acurácia dos estimadores referente à assíntota superior (α) e ao intercepto (β) do modelo logístico. Em se tratando do modelo Gompertz, em geral apresentou os mesmos resultados, porém dada a baixa porcentagem de convergência observada, recomenda-se que novos estudos sejam realizados para obtenção de conclusões mais confirmatórias.

Palavras-chave: CovRatio. Viés. Precisão. Monte Carlo. Regressão não linear.

ABSTRACT

In general the *bootstrap* method is realized assuming that each sample shows the same probability of be resampled. However, considering a sample that contains *outliers*, the empirical distribution generated through the execution of this method can be influenced, in order not accurately represent the original sample. The purpose of this work consisted in propose a *bootstrap* method that allows the researcher to specify an influence measure, to be used as criterion to get different probabilities of resampling for each observation. This method was called BMIC, originated of “*Bootstrap* with influence measures corrected by Monte Carlo bias”. With the purpose of valuate this method, studies of simulation was realized considering the models Logistic and Gompertz, in some scenery characterized by different sample sizes, symmetric and asymmetric distributions in residues, and number of *outliers* in the sample. It was concluded that the application of BMIC method provided an improvement in accuracy of estimators relative to the superior asymptote (α) and to the intercept (β) of Logistic model. According to the Gompertz model, in general, shows the same results, however, given the low percentage of convergence observed, it is recommended that new studies are realized for attaining conclusions more confirmatory.

Keywords: CovRatio. Viés. Precision. Monte Carlo. Nonlinear regression.

SUMÁRIO

1	INTRODUÇÃO	12
2	REFERENCIAL TEÓRICO	14
2.1	Introdução aos modelos de regressão não lineares.....	14
2.1.1	Representação geométrica da não linearidade do modelo.....	15
2.1.2	Questões práticas a serem consideradas para a escolha de um modelo não linear	20
2.1.3	Modelos de regressão não linear Logístico e Gompertz.....	21
2.2	Método de estimação	24
2.2.1	Método dos mínimos quadrados	24
2.3	Identificação de <i>outliers</i> em modelos lineares e não lineares.....	28
2.3.1	Resíduo studentizado: modelo linear	29
2.3.2	Resíduo studentizado: modelo não linear.....	30
2.4	Análise de resíduos	31
2.5	Método <i>Bootstrap</i>	32
2.5.1	<i>Bootstrap</i> não paramétrico.....	32
2.5.2	<i>Bootstrap</i> Bayesiano.....	34
2.5.3	Reamostragem balanceada	36
2.5.4	Amostragem por importância	37
2.6	Medidas de Influência: <i>COVRATIO_i</i>	39
3	METODOLOGIA	41
3.1	Simulação Monte Carlo dos modelos Gompertz e Logístico	41
3.2	Incorporação da medida de influência <i>CovRatio</i> no método <i>bootstrap</i>	43
3.3	Avaliação da acurácia e precisão das estimativas dos parâmetros dos modelos Logístico e Gompertz.....	45
4	RESULTADOS E DISCUSSÃO	47
4.1	Estudo da acurácia das estimativas <i>bootstrap</i> com incorporação de medidas de influência	48
4.2	Estudo da precisão das estimativas <i>bootstrap</i> com incorporação de medidas de influência	54
5	CONCLUSÃO	59
	REFERÊNCIAS	60
	APÊNDICE	67

1 INTRODUÇÃO

Em se tratando de estudos de modelos não lineares, constantemente o pesquisador se depara com situações nas quais surge a impossibilidade de obter soluções analíticas ou fixar pressupostos que viabilizem o processo de derivação, mormente utilizado para construir a matriz hessiana.

Frente a esse problema, uma alternativa consiste na obtenção de resultados empíricos, que por sua vez, são obtidos utilizando técnicas de simulação Monte Carlo e métodos de reamostragem. Porém, salienta-se que, para a obtenção de estimativas precisas, o uso dessas técnicas requer um esforço computacional dispendioso. Desta forma, em muitas situações, a aplicação desses métodos torna-se inviável face a grandes bases de dados.

Decorrente a esta situação, a obtenção de estimativas precisas, torna-se mais agravante se considerarmos o fato de a amostra a ser submetida ao método *bootstrap* apresentar uma quantidade expressiva de *outliers*. Nesta situação, os métodos usuais, conhecidos na literatura por *bootstrap* paramétrico e não paramétrico (CHERNICK, 1999; DAVISON; HINKLEY, 1997), poderão proporcionar estimativas imprecisas, uma vez que, a probabilidade de uma unidade amostral ser sorteada para compor a amostra *bootstrap* segue uma distribuição uniforme discreta, em outras palavras, todas as unidades amostrais têm a mesma probabilidade de serem sorteadas. Desta forma, nota-se que a aplicação desses métodos, certamente irá gerar sub-amostras que poderão apresentar mais observações *outliers* do que a amostra original. Efron (1992) recomenda que em estudos de simulação deve-se observar com cautela os resultados obtidos, mantendo o foco na verificação de diferenças substanciais. Com este propósito o autor recomenda que seja realizado um procedimento

jackknife após o *bootstrap*. Maiores detalhes podem ser vistos em Efron (1982a), Hjorth (1994), Shao e Tu (1996).

Em termos práticos, surge a questão “A observação *outlier* deve ser retirada da amostra?”. A resposta correta a esta pergunta é dada em função do problema que está sendo estudado. Em problemas relacionados à genética, uma observação *outlier*, mediante ao conhecimento *a priori* do pesquisador, poderá revelar uma anomalia (contaminação na amostra ou algo do gênero). Nesta situação, geralmente é recomendável a retirada da observação. Entretanto, em problemas que envolvem custo elevado para obtenção da amostra, é viável que a observação classificada como *outlier* não seja descartada, assim sendo, neste contexto, uma alternativa consiste no uso de métodos de estimação robustos que proporcionam estimativas precisas mediante a uma pequena fração de observações *outliers* contidas na amostra.

Reportando à aplicação do método *bootstrap*, e seus problemas mencionados (a presença de observações *outliers*), a obtenção das estimativas dos parâmetros de modelos não lineares, considerando amostras pequenas, proporciona um campo de pesquisa no que tange ao uso de recursos computacionais que tornam o esforço computacional reduzido.

Em virtude do que foi mencionado, este trabalho tem por objetivo propor um algoritmo *bootstrap* que permita utilizar medidas de influência no cálculo das probabilidades de reamostragem. Com este propósito, a ilustração desse método será dada nos modelos não lineares Gompertz e Logístico em alguns cenários de simulação.

2 REFERENCIAL TEÓRICO

2.1 Introdução aos modelos de regressão não lineares

Um modelo é não linear quando uma variável dependente Y não pode ser escrita como funções lineares de seus parâmetros. Alguns modelos não lineares são considerados “intrinsecamente lineares”. Isso ocorre quando o modelo puder ser linearizado via alguma transformação. O inconveniente dessa transformação é que, além do parâmetro perder sua interpretação intrínseca, pode-se alterar a estrutura e distribuição do erro. Ou seja, se os erros do modelo original satisfizerem as suposições usuais de normalidade, independência e homogeneidade da variância, os erros do novo modelo, em geral, não satisfarão tais suposições (KHURI; CORNELL, 1987). Caso o modelo não possa ser reduzido à forma linear, têm-se os chamados modelos “intrinsecamente não lineares”.

Um modelo de regressão não linear pode ser expresso por

$$\tilde{Y} = f(\tilde{\mathbf{X}}, \tilde{\beta}) + \tilde{\varepsilon} \quad (1)$$

em que \tilde{Y} é um vetor de n observações que representa a variável resposta, $\tilde{\mathbf{X}}$ é uma matriz $(n \times p)$ de variáveis de entrada, $\tilde{\beta}$ é um vetor $(k \times 1)$ de parâmetros desconhecidos, $\tilde{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^t$, $\tilde{\varepsilon}$ é um vetor $(n \times 1)$ de erros aleatórios e $f(\tilde{\mathbf{X}}, \tilde{\beta})$ é um vetor $(n \times 1)$ de funções não lineares em $\tilde{\beta}$, contínuas e diferenciáveis no espaço paramétrico Θ . Em geral, assume-se que os erros são variáveis aleatórias independentes e identicamente distribuídas, normais com média 0 e variância constante $\sigma^2 \mathbf{I}_n$ em que \mathbf{I}_n representa a matriz identidade de ordem n .

Em modelos lineares, assumindo como verdadeiras essas suposições sobre o erro, podemos afirmar que os estimadores de mínimos quadrados são não viciados, normalmente distribuídos e possuem variância mínima entre qualquer outra classe de estimadores. No caso não linear, essas propriedades são válidas somente quando o tamanho da amostra é suficientemente grande (MAZUCHELI; ACHCAR, 2002).

2.1.1 Representação geométrica da não linearidade do modelo

A não linearidade de um modelo pode ser decomposta em duas componentes: a não linearidade intrínseca e a não linearidade devida ao efeito dos parâmetros.

Para compreender a diferença entre os modelos lineares e os modelos não lineares, e o comportamento das componentes de não linearidade do modelo, Ratkowsky (1983) usou uma maneira alternativa de descrever os dados e desenhar a superfície esperada do parâmetro. Diferente da representação tradicional, em que cada eixo explica uma das variáveis (Y versus X), considerou-se um eixo coordenado para cada uma das n observações amostrais.

Nesse gráfico, à medida que o parâmetro β é modificado, determina-se uma superfície p -dimensional chamada de “região de solução” ou “superfície esperada” no espaço amostral n -dimensional (BOX; LUCAS, 1959), e a estimativa de mínimos quadrados $\hat{\beta}$ é um ponto dessa superfície. Em um modelo linear, essa região de solução é um hiperplano gerado pelas colunas de \mathbf{X} (em casos específicos pode ser uma reta ou um plano).

Observe ainda que a estimativa de mínimos quadrados é encontrada através da projeção ortogonal do vetor de observações \tilde{y} na superfície gerada pelas colunas de \mathbf{X} (Figura 1).

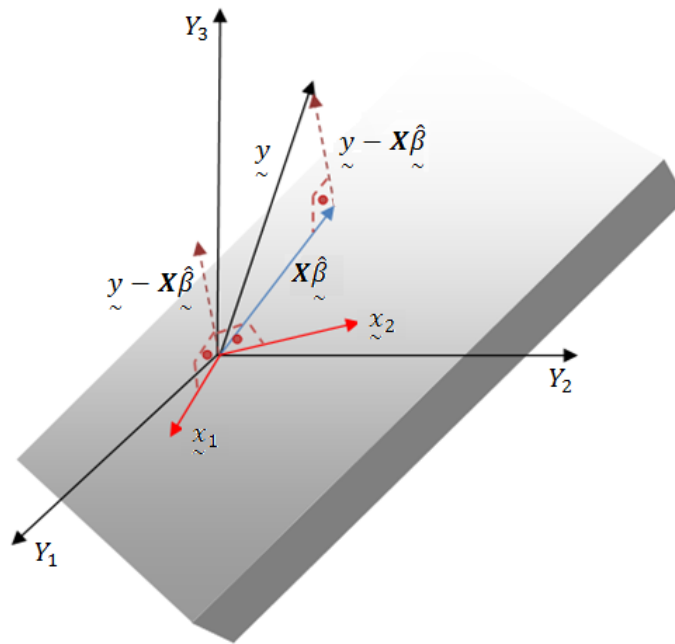


Figura 1 Ilustração da ortogonalidade entre o vetor residual $\tilde{y} - \mathbf{X}\hat{\beta}$ e as colunas da matriz \mathbf{X}

De fato, observe que dessa forma o vetor residual, $\tilde{y} - \mathbf{X}\hat{\beta}$, é ortogonal ao espaço gerado pelas colunas de \mathbf{X} , e, portanto, satisfaz a seguinte equação:

$$\mathbf{X}'(\tilde{y} - \mathbf{X}\hat{\beta}) = 0,$$

que é equivalente a

$$\mathbf{X}'\tilde{y} = \mathbf{X}'\mathbf{X}\hat{\beta},$$

ou seja, $\hat{\beta}$ satisfaz o sistemas de equações normais, e é portanto, a estimativa de mínimos quadrados.

As componentes de não linearidade de um modelo não linear podem ser identificadas geometricamente através do formato da região de solução. Considere, por exemplo, no caso bidimensional, uma amostra com $n = 2$ observações, (x_i, y_i) , dadas por: (2; 2,5) e (3; 10). Neste caso, são necessários dois eixos coordenados para representar os dados. Como $y_1 = 2,5$ e $y_2 = 10$, considere um Y -vetor representado da origem ao vértice (2,5; 10). Considere agora o modelo não linear

$$y_i = x_i^\beta + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (2)$$

Segue que:

$$E(Y) = X^\beta. \quad (3)$$

Podemos determinar a região de solução para este modelo, isto é, a região de pontos da forma (x_1^β, x_2^β) , neste caso $(2^\beta, 3^\beta)$, que representa todos os valores possíveis para β . Ratkowsky (1983) determina a estimativa de mínimos quadrados $\hat{\beta}$, como o ponto da região de solução com menor distância ao vetor y . Esta distância mínima é representada pela Figura 2(a) e corresponde a $\hat{\beta} = 2,0537$.

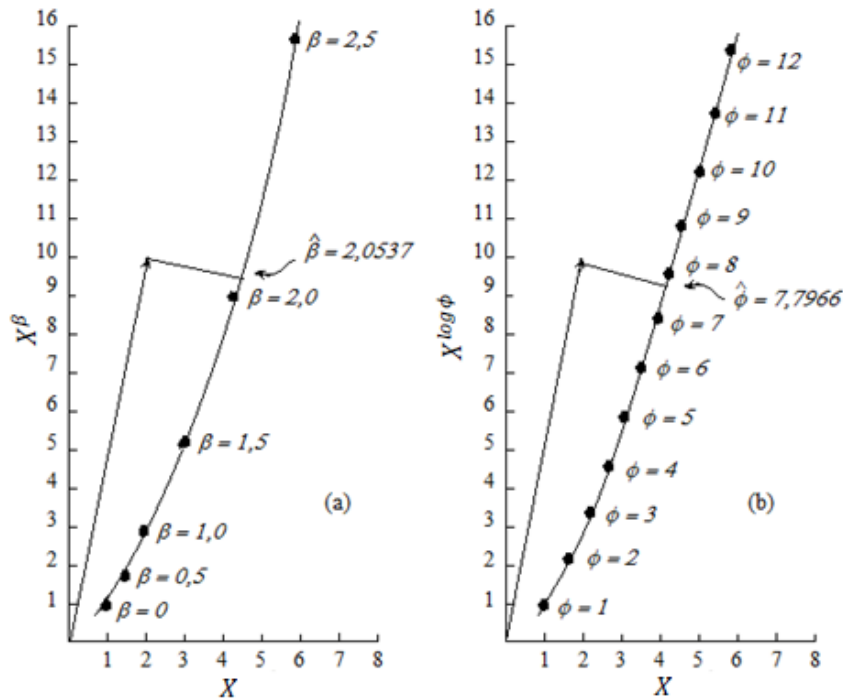


Figura 2 Representação da região de solução dos dados: (a) Modelo $E(Y) = X^\beta$; (b) Modelo $E(Y) = X^{\log \phi}$

Ratkowsky (1983) afirma que o formato da região de solução na vizinhança de $\hat{\beta}$ (não linearidade intrínseca) e o espaçamento entre os pontos da região de solução, obtidos através de um aumento constante do parâmetro β (efeito dos parâmetros), podem ser usados como medida do grau de não linearidade do modelo, ou seja, do quanto o modelo não linear difere do modelo linear. De fato, em modelos lineares de um parâmetro, a região de solução é representada por uma linha reta, um plano ou um hiperplano. Sendo assim, um modelo não linear poderia ser considerado como uma aproximação linear quando a região de solução esperada for plana (reta no caso bidimensional). Isto

é, quanto mais suave for a curvatura, na vizinhança de $\hat{\beta}$, menor a curvatura intrínseca.

Quanto ao espaçamento entre os pontos da região de solução, tomados a partir do mesmo incremento de β , observe na Figura 2(a) que os pontos em evidência (obtidos com $\Delta\beta = 0,5$), não estão igualmente espaçados. O que indica a não linearidade do modelo devida ao efeito dos parâmetros. Por outro lado, considere a seguinte reparametrização do modelo (3),

$$E(Y) = X^{\log\phi}, \quad (4)$$

em que $\phi = e^\beta$.

A Figura 2(b) mostra que os pontos da região de solução do modelo (4), tomados com mesmo incremento de ϕ ($\Delta\phi = 1$), possuem aproximadamente o mesmo espaçamento, além de apresentar uma curvatura bem suave. Portanto, podemos dizer que a reparametrização indicada que é uma aproximação linear para o modelo.

De acordo com o autor, apesar de a representação utilizada ser restrita a amostras com $n \leq 3$ e número de parâmetros $p \leq 2$, este conceito também é válido para maiores dimensões. Um modelo de regressão linear possui uma região de solução linear, que é representado por uma reta para $p = 1$, um plano para $p = 2$, e um hiperplano para $p = 3$. Neste caso, as linhas da região de solução, chamadas de linhas dos parâmetros, que representam valores constantes de β_i , são linhas retas, paralelas e igualmente espaçadas para a mesma variação de β .

Já em modelos com não linearidade intrínseca, ou seja, modelos que não possuem reparametrização que os reduza a forma linear, a região de solução é uma superfície curva e, em geral, as linhas dos parâmetros (ou projeções destas linhas no plano tangente à região de solução em $\hat{\beta}$, como usado em Bates e

Watts (1980)) não são retas, não são paralelas, nem são igualmente espaçadas. Mais detalhes sobre o tratamento geométrico para a não linearidade de um modelo pode ser visto em Zeviani (2009).

2.1.2 Questões práticas a serem consideradas para a escolha de um modelo não linear

É muito comum em problemas de estatística o estudo da relação entre duas ou mais variáveis. Segundo Ross (1990), para escolher um modelo que explique de maneira satisfatória essa relação deve-se considerar alguns fatores como os citados a seguir:

- a. Definir o propósito do ajuste do modelo: O propósito do ajuste é previsão, estimativa de parâmetros, encontrar explicação satisfatória para os dados ou simplesmente mostrar que o modelo se ajusta aos dados?
- b. Qual o motivo da escolha do modelo? O modelo foi escolhido por causa de alguma pressuposição básica, porque os dados se assemelham ao modelo, porque é um modelo convencionalmente utilizado por outros autores, ou porque é um modelo de fácil interpretação e utilização?
- c. Modelos alternativos. Quais outros modelos poderiam ser escolhidos e quais diferenças existem entre os ajustes?
- d. Distribuição dos erros. As observações são independentes e normalmente distribuídas com mesma variância?

- e. Aplicabilidade dos resultados. Em quais outras circunstâncias o modelo pode ser aplicado? Quais parâmetros dependem dos dados e quais parâmetros poderiam ser considerados em outras aplicações?

Dadas essas considerações, a seguir os modelos Logístico e Gompertz encontram-se descritos na seção 2.1.3.

2.1.3 Modelos de regressão não linear Logístico e Gompertz

Processos que produzem curvas com aspecto sigmoidal são muito comuns nas áreas de agronomia, biologia, educação, economia e engenharia. Segundo Cordeiro e Demétrio (2009), uma curva sigmóide começa em algum ponto fixo, cresce monotonicamente até um ponto de inflexão e, dependendo da localização deste ponto, as curvas sigmoidais serão simétricas ou assimétricas.

Para o modelo logístico de crescimento que tem como propriedade a simetria em relação ao ponto de inflexão, admite-se que a taxa de crescimento dy/dx é proporcional a y (quantidade presente) e a $\alpha - y$ (complemento para alcançar a quantidade máxima teórica da variável resposta), isto é

$$\frac{dy}{dx} = \beta y(\alpha - y)$$

cuja solução é,

$$y = \frac{\alpha}{1 + \delta \exp(-\alpha \gamma x)}$$

Essa equação pode ser reparametrizada e conduz ao modelo logístico de crescimento na forma mais conhecida (CORDEIRO; DEMÉTRIO, 2009), apresentada por

$$y = \frac{\alpha}{1 + \exp(\beta - \gamma x)}$$

em que α é o valor máximo esperado para a resposta, ou seja, assíntota superior. O parâmetro γ está relacionado com a taxa média de crescimento. O parâmetro β está relacionado com o intercepto, isto é, o valor de $\mu = E(Y)$ correspondente a $x = 0$. O ponto de inflexão ocorre em $x = \beta/\gamma$ e $y = \alpha/2$. Com essas especificações Cordeiro e Demétrio (2009) cita as seguintes propriedades:

- a. $y = \alpha$ é uma assíntota horizontal à direita;
- b. $y = 0$ é uma assíntota horizontal à esquerda;
- c. É uma função estritamente crescente;
- d. Só assume valores no intervalo $(0, \alpha)$;
- e. y atinge metade da capacidade de sustentação em $x = \beta/\gamma$ e $y = \alpha/2$;
- f. Apresenta concavidade para cima se $x < \beta/\gamma$, para baixo se $x > \beta/\gamma$ e ponto de inflexão para $x = \beta/\gamma$ (ponto em que o crescimento atinge velocidade máxima);

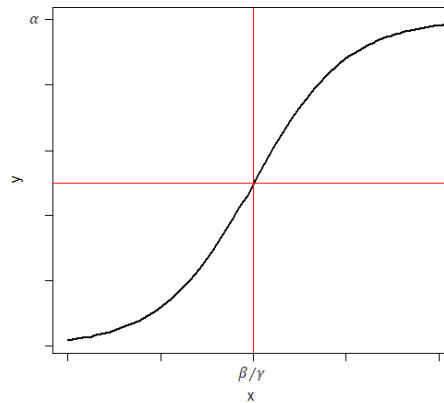


Figura 3 Representação gráfica do modelo logístico e ponto de inflexão (linhas vermelhas)

O modelo de Gompertz, diferente do logístico, é assimétrico em relação ao ponto de inflexão e admite-se que a taxa de crescimento dy/dx seja proporcional a y e a $\ln(\alpha) - \ln(y)$. Sua parametrização é representada por:

$$y = \alpha \exp\{-\exp[\beta(x - \gamma)]\}$$

em que α corresponde a assíntota superior; γ está relacionado à taxa média de crescimento e β está relacionado com o intercepto da curva. O ponto de inflexão ocorre em $x = \gamma$ e $y = \alpha/e$, em que e é definido como número de Euler ($e = 2,718$).

Algumas propriedades:

- a. $y = \alpha$ é uma assíntota horizontal à direita;
- b. $y = 0$ é uma assíntota horizontal à esquerda;
- c. É uma função estritamente crescente;
- d. Só assume valores no intervalo $(0, \alpha)$;

- e. y atinge menos da metade da capacidade de sustentação em $x = \gamma$, ou seja, $y(\gamma) = \alpha/e < \alpha/2$;
- f. Apresenta concavidade para cima se $x < \gamma$; para baixo se $x > \gamma$ e ponto de inflexão para $x = \gamma$;

Algumas representações gráficas do modelo de Gompertz, em função de vários cenários paramétricos, podem ser verificadas na Figura 4, de maneira que, em cada gráfico, dois dos parâmetros encontram-se fixos para visualizarmos o comportamento do terceiro parâmetro quando alterado.

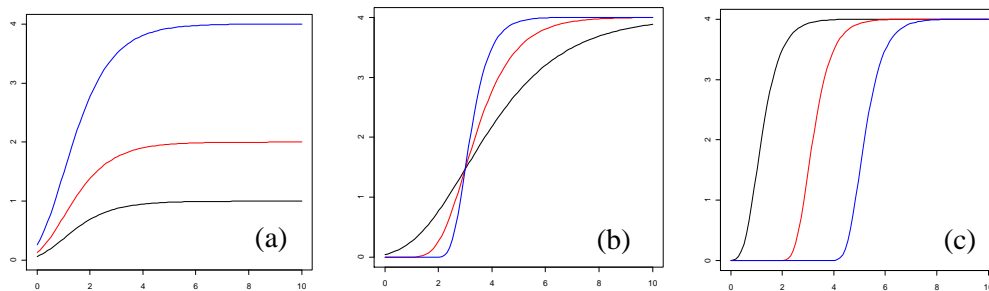


Figura 4 (a) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro α ($\alpha = 1$ em preto, $\alpha = 2$ em vermelho e $\alpha = 4$ em azul, $\beta = 1$, $\gamma = 1$); (b) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro β ($\beta = 0,5$ em preto, $\beta = 1$ em vermelho e $\beta = 2$ em azul, $\alpha = 4$, $\gamma = 3$); (c) Representação gráfica do modelo de Gompertz para diferentes valores do parâmetro γ ($\gamma = 1$ em preto, $\gamma = 3$ em vermelho e $\gamma = 5$ em azul, $\alpha = 4$, $\beta = 2$)

2.2 Método de estimação

2.2.1 Método dos mínimos quadrados

Considere o seguinte modelo não linear na forma matricial

$$\tilde{Y} = f(\mathbf{X}, \tilde{\beta}) + \tilde{\varepsilon}.$$

Segundo Gallant (1987), o estimador de mínimos quadrados do vetor de parâmetros $\tilde{\beta}$ deste modelo é o vetor $(k \times 1)$, $\tilde{\beta}$, que minimiza a soma de quadrados de resíduos

$$\begin{aligned} SSE(\tilde{\beta}) &= \tilde{\varepsilon}^t \tilde{\varepsilon} = [\tilde{Y} - f(\mathbf{X}, \tilde{\beta})]^t [\tilde{Y} - f(\mathbf{X}, \tilde{\beta})] \\ &= \left\{ \tilde{Y}^t - [f(\mathbf{X}, \tilde{\beta})]^t \right\} [\tilde{Y} - f(\mathbf{X}, \tilde{\beta})] \\ &= \tilde{Y}^t \tilde{Y} - \tilde{Y}^t f(\mathbf{X}, \tilde{\beta}) - [f(\mathbf{X}, \tilde{\beta})]^t \tilde{Y} + [f(\mathbf{X}, \tilde{\beta})]^t f(\mathbf{X}, \tilde{\beta}) \\ &= \tilde{Y}^t \tilde{Y} - [f(\mathbf{X}, \tilde{\beta})]^t \tilde{Y} - [f(\mathbf{X}, \tilde{\beta})]^t \tilde{Y} + [f(\mathbf{X}, \tilde{\beta})]^t f(\mathbf{X}, \tilde{\beta}) \\ &= \tilde{Y}^t \tilde{Y} - 2 [f(\mathbf{X}, \tilde{\beta})]^t \tilde{Y} + [f(\mathbf{X}, \tilde{\beta})]^t f(\mathbf{X}, \tilde{\beta}). \end{aligned}$$

Para minimizar a soma de quadrado de resíduos $SSE(\tilde{\beta})$, deve-se primeiro igualar a zero o vetor $(k \times 1)$ de derivadas parciais, ou seja,

$$\begin{aligned} \frac{\partial SSE(\tilde{\beta})}{\partial \tilde{\beta}} &= -2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t \tilde{Y} + \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t f(\mathbf{X}, \tilde{\beta}) + [f(\mathbf{X}, \tilde{\beta})]^t \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right] \\ &= -2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t \tilde{Y} + \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t f(\mathbf{X}, \tilde{\beta}) + \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t f(\mathbf{X}, \tilde{\beta}) \\ &= -2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t \tilde{Y} + 2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t f(\mathbf{X}, \tilde{\beta}). \end{aligned}$$

Sendo assim,

$$\left. \frac{\partial SSE(\tilde{\beta})}{\partial \tilde{\beta}} \right|_{\tilde{\beta}=\hat{\tilde{\beta}}} = \left\{ -2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t \tilde{Y} + 2 \left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t f(\mathbf{X}, \tilde{\beta}) \right\} \Bigg|_{\tilde{\beta}=\hat{\tilde{\beta}}} = 0,$$

em que $\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}}$ é uma matriz $(n \times k)$ de elementos do tipo $\frac{\partial f(\mathbf{X}_i, \tilde{\beta})}{\partial \beta_j}$; $f(\mathbf{X}_i, \tilde{\beta})$ é a i -ésima linha do vetor $f(\mathbf{X}, \tilde{\beta})$, com $i = 1, \dots, n$, e β_j é o j -ésimo elemento do vetor de parâmetros $\tilde{\beta}$, com $j = 1, \dots, k$. Essa matriz assume a seguinte forma:

$$\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} = \begin{bmatrix} \frac{\partial f(\mathbf{X}_1, \tilde{\beta})}{\partial \beta_1} & \dots & \frac{\partial f(\mathbf{X}_1, \tilde{\beta})}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X}_n, \tilde{\beta})}{\partial \beta_1} & \dots & \frac{\partial f(\mathbf{X}_n, \tilde{\beta})}{\partial \beta_k} \end{bmatrix}.$$

Por outro lado, para encontrarmos as soluções das equações do sistema de equações normais $\left[\frac{\partial f(\mathbf{X}, \tilde{\beta})}{\partial \tilde{\beta}} \right]^t \left[\tilde{Y} - f(\mathbf{X}, \tilde{\beta}) \right] \Bigg|_{\tilde{\beta}=\hat{\tilde{\beta}}} = 0$ de modelos não

lineares, é mais prático utilizar métodos numéricos de iteração.

O método de Gauss-Newton, ou método de linearização, usa uma expansão em séries de Taylor para fazer esta linearização e aplica o método de mínimos quadrados para estimar os parâmetros. Iterações sucessivas desses passos levam à solução aproximada do problema de regressão não linear.

Método Gauss-Newton

A seguir é apresentado o algoritmo de Gauss Newton, o qual é um dos métodos mais utilizados para obtenção de estimativas de mínimos quadrados não lineares (MAZUCHELI; ACHCAR, 2002).

Passo 1) Atribuir valores iniciais para o vetor de parâmetros:

$$\underset{\sim}{\beta}^0 = (\beta_1^0, \beta_2^0, \dots, \beta_k^0)$$

Passo 2) Aproximar a função esperada $f(\mathbf{X}, \underset{\sim}{\beta})$, para os n casos, por termos lineares de expansão em série de Taylor em uma vizinhança de $\underset{\sim}{\beta}^0 \in \Theta$, da seguinte forma: $f(\mathbf{X}, \underset{\sim}{\beta}) \approx f(\mathbf{X}, \underset{\sim}{\beta}^0) + \mathbf{Z}^0 (\underset{\sim}{\beta} - \underset{\sim}{\beta}^0)$, em que \mathbf{Z}^0 é a matriz de derivadas parciais $\frac{\partial f(\mathbf{x}, \underset{\sim}{\beta})}{\partial \underset{\sim}{\beta}}$, do modelo, avaliada que $\underset{\sim}{\beta} = \underset{\sim}{\beta}^0$ e Θ é o espaço paramétrico.

Assim, o modelo não linear $\underset{\sim}{Y} = f(\mathbf{X}, \underset{\sim}{\beta}) + \underset{\sim}{\varepsilon}$ pode ser escrito como

$$\underset{\sim}{Y} \approx f(\mathbf{X}, \underset{\sim}{\beta}^0) + \mathbf{Z}^0 (\underset{\sim}{\beta} - \underset{\sim}{\beta}^0) + \underset{\sim}{\varepsilon}.$$

Chamando $\underset{\sim}{Y} - f(\mathbf{X}, \underset{\sim}{\beta}^0) = \underset{\sim}{Y}^0$ e $\underset{\sim}{\beta} - \underset{\sim}{\beta}^0 = \underset{\sim}{\theta}^0$, temos:

$$\underset{\sim}{Y}^0 = \mathbf{Z}^0 \underset{\sim}{\theta}^0 + \underset{\sim}{\varepsilon}$$

que é uma aproximação linear para um modelo de regressão linear. Sendo assim, a estimativa de mínimos quadrados para $\underset{\sim}{\theta}^0$ é:

$$\underset{\sim}{\hat{\theta}}^0 = (\mathbf{Z}^{0'} \mathbf{Z}^0)^{-1} \mathbf{Z}^0 \underset{\sim}{Y}^0.$$

Passo 3) Atualizar o vetor de estimativas do parâmetro: $\underset{\sim}{\beta}^1 = \underset{\sim}{\beta}^0 + \underset{\sim}{\hat{\theta}}^0$. Note que se o algoritmo estiver na direção correta, a soma de quadrados dos resíduos considerando $\underset{\sim}{\beta}^1$, $SSE\left(\underset{\sim}{\beta}^1\right)$, será menor que a soma de quadrados dos resíduos considerando $\underset{\sim}{\beta}^0$, $SSE\left(\underset{\sim}{\beta}^0\right)$.

Passo 4) Repetir o passo 3 até atingir um critério de parada estabelecido pelo método de Gauss-Newton, especificado em uma das seguintes situações:

- a. Efetuar um número pré-determinado de iterações no processo;
- b. Quando a modificação $\underset{\sim}{\beta}^j \rightarrow \underset{\sim}{\beta}^{j+1}$ (onde j se refere à j -ésima iteração) for abaixo de algum limite pré-estabelecido;
- c. Quando a modificação na soma de quadrados dos resíduos atingir um valor inferior a um critério de tolerância especificado;
- d. Quando se verificar uma “quase ortogonalidade” entre o vetor de resíduos e as colunas da matriz de derivadas de \mathbf{Z}^j .

2.3 Identificação de *outliers* em modelos lineares e não lineares

A detecção de *outliers* é importante em análise de regressão. Draper e Smith (1998) sugerem o uso do máximo absoluto do resíduo studentizado para testes de *outlier* em modelos lineares. Vários autores têm discutido e fornecido valores críticos destes testes. As tabelas de valores críticos apresentados por Lund (1975) são os mais utilizados. Estes valores também foram reproduzidos em Barnett e Lewis (1993) e a precisão desses resultados foram discutidos por Doornbos (1981) e Tse (1988). Outros trabalhos nessa área incluem Ellenberg (1976), Prescott (1975), e Tietjen, Moore e Beckman (1973).

Cook e Weisberg (1982) e Atkinson (1985) pouco discutiram sobre o papel dos resíduos studentizados para a detecção de *outliers* em regressão não linear. Barnett e Lewis (1993) examinaram *outliers* em modelos não lineares específicos como o modelo misto, regressão logística e modelos lineares generalizados. Testes para *outliers* associados a valores críticos dificilmente são mencionados em livros de regressão não linear como Batters e Watts (1988), Ross (1990) e Seber e Wild (1989).

2.3.1 Resíduo studentizado: modelo linear

Consideremos o modelo de regressão linear na forma matricial

$$\tilde{Y} = \tilde{X}\tilde{\beta} + \tilde{\varepsilon}$$

em que \tilde{Y} é a variável resposta, \tilde{X} é uma matriz $(n \times p)$ de p variáveis de entrada, $\tilde{\beta}$ é um vetor $(p \times 1)$ de parâmetros desconhecidos $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_p)^t$, $\tilde{\varepsilon}$ é um vetor $(n \times 1)$ de erros distribuídos normalmente com média 0 e variância σ^2 . O resíduo studentizado é definido como (COOK; WEISBERG, 1982, p. 18)

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n,$$

em que $\hat{\varepsilon}_i = y_i - \tilde{x}_i^t \hat{\tilde{\beta}}$ é a estimativa do resíduo, \tilde{x}_i é um vetor $(p \times 1)$ definido pela i -ésima linha de \tilde{X} , $\hat{\tilde{\beta}}$ é a estimativa de mínimos quadrados de $\tilde{\beta}$, $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 / (n - p)$ e h_{ii} o i -ésimo elemento da diagonal da matriz “hat” $H = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$.

O valor $\max|r_i|$ é comumente usado para testar *outliers* em modelos de regressão. Pode ser mostrado que $r_i^2/(n-p)$ tem distribuição Beta. Utilizando a relação entre as distribuições F e Beta, Lund (1975) tabulou os limites superiores para os valores críticos de $\max|r_i|$ em modelos de regressão múltipla.

2.3.2 Resíduo studentizado: modelo não linear

Consideremos o modelo de regressão não linear na forma matricial

$$\tilde{Y} = f(\tilde{X}, \tilde{\beta}) + \varepsilon.$$

Seja $\hat{\tilde{\beta}}$ a estimativa de mínimos quadrados de $\tilde{\beta}$, $\hat{\tilde{Z}}$ a matriz de derivadas parciais

sob $\hat{\tilde{\beta}}$, $\hat{\tilde{Z}} = \left. \frac{\partial f(\tilde{x}, \tilde{\beta})}{\partial \tilde{\beta}} \right|_{\tilde{\beta} = \hat{\tilde{\beta}}}$, e h_{ii} o i -ésimo elemento da diagonal da matriz

$$H = \hat{\tilde{Z}}(\hat{\tilde{Z}}'\hat{\tilde{Z}})^{-1}\hat{\tilde{Z}}'.$$

O resíduo studentizado do modelo de regressão não linear (COOK; WEISBERG, 1982) é dado por

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}; \quad i = 1, \dots, n,$$

em que $\hat{\varepsilon}_i = y_i - f(\tilde{X}_i, \hat{\tilde{\beta}})$ é a estimativa do resíduo e $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 / (n-p)$.

O valor de $\max|t_i|$ pode ser usado para testar *outlier* em modelos de regressão não linear. Maiores detalhes poderão ser vistos em Fung e Ngai (1998).

2.4 Análise de resíduos

Dado um modelo ajustado, é importante investigar a veracidade da modelagem. A existência de algumas suposições permite evitar o emprego de modelos que acarretam pouca confiabilidade nos resultados.

Para detectar o problema podemos observar as discrepâncias entre os valores observados e os valores ajustados, ou seja, fazer a análise de resíduos.

Para cada observação i , temos um erro associado \hat{e}_i que é a diferença entre o valor observado e o ajustado, isto é

$$\hat{e}_i = y_i - \hat{y}_i, (i = 1, 2, \dots, n).$$

O vetor de erros $\hat{\tilde{e}} = \tilde{y} - \tilde{\hat{y}} = \tilde{y} - \tilde{\mathbf{X}}\tilde{\hat{\beta}}$, em que $\tilde{\hat{\beta}}$ é um estimador não viesado de $\tilde{\beta}$, tem esperança nula, ou seja, $E(\hat{\tilde{e}}) = 0$, porém, os resíduos estimados podem ser correlacionados e heterocedásticos.

Pode-se definir os resíduos através da matriz \mathbf{H} na expressão $\hat{\tilde{e}} = \tilde{y} - \tilde{\hat{y}} = \tilde{y} - \mathbf{H}\tilde{y}$, uma vez que $\tilde{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{y}$ e $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

A matriz “hat” possui propriedades importantes como ser simétrica ($\mathbf{H} = \mathbf{H}'$) e idempotente ($\mathbf{H}\mathbf{H} = \mathbf{H}$). O que implica em $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{H}$ (SCHABENBERGER; PIERCE, 2002).

A variância do vetor de erros estimados é dada por:

$$\begin{aligned} Var(\hat{\tilde{e}}) &= Var(\tilde{y} - \mathbf{H}\tilde{y}) = Var[\tilde{y}(\mathbf{I} - \mathbf{H})] \\ &= \sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H}). \end{aligned}$$

Para evitar que se trabalhe com erros heterocedásticos, recomenda-se usar os erros apropriadamente transformados, como os resíduos studentizados, ou padronizados, que diminuem a heterogeneidade dos erros.

Seja h_{ii} o elemento da diagonal da matriz \mathbf{H} . A variância do erro é:

$$Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$

e os resíduos padronizados, como média zero e variância um, são obtidos por:

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\sigma\sqrt{1 - h_{ii}}}; \quad (i = 1, \dots, n).$$

Como σ^2 é desconhecida, substituímos σ por $\hat{\sigma} = \sqrt{QM_{Erro}}$. Assim, o resíduo:

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

é chamado de resíduo padronizado e é mais apropriado para medidas de diagnóstico.

Métodos gráficos são úteis para verificar os pressupostos de normalidade. A análise da relação entre os quantis teóricos com os quantis esperados ou testes como o de Shapiro Wilk cuja aplicação com mais detalhes pode ser vistas em Shapiro e Wilk (1965).

2.5 Método *Bootstrap*

2.5.1 *Bootstrap* não paramétrico

Amostras *bootstrap* não paramétricas são obtidas reamostrando, com reposição, diretamente dos dados disponíveis $(y_1, x_{11}, x_{12}, \dots, x_{1k}), \dots, (y_n, x_{n1}, x_{n2}, \dots, x_{nk})$ (EFRON; TIBSHIRANI, 1993; DICICCIO; EFRON, 1996).

Tendo por base essas amostras, a obtenção de intervalo de confiança *bootstrap* para os coeficientes de regressão β é dada nos percentis da

distribuição empírica de $\hat{\beta}$. Uma recente discussão a respeito dos intervalos de confiança, obtidos via simulação *bootstrap*, é apresentada em Carpenter e Bithell (2000).

Para ilustrar o método *bootstrap*, Barros et al. (2008) apresentam o seguinte algoritmo considerando o modelo de regressão múltiplo especificando $U = (y, x_1, x_2, \dots, x_k)$ em que $y = (y_1, y_2, \dots, y_n)$ é o vetor de observações da variável dependente e $x_j = (x_{1j}, \dots, x_{nj}), j = 1, \dots, k$, são os vetores de observações das variáveis independentes, x_j é uma variável contínua para todo j .

Intervalos *bootstrap*

- I. Amostragem, com reposição de U , uma amostra *bootstrap* $(y_1^*, x_{11}^*, \dots, x_{1k}^*), \dots, (y_n^*, x_{n1}^*, \dots, x_{nk}^*)$.
- II. Da amostra *bootstrap* $(y_1^*, x_{11}^*, \dots, x_{1k}^*), \dots, (y_n^*, x_{n1}^*, \dots, x_{nk}^*)$, obter o estimador de mínimos quadrados de $\hat{\beta}$, representado por $\hat{\beta}^*$.
- III. Repetir os passos I e II um número B grande de vezes.
- IV. De $\hat{\beta}^* = (\hat{\beta}_{(1)}^* \leq \hat{\beta}_{(2)}^* \leq \dots \leq \hat{\beta}_{(B)}^*)$, para algum valor de α , ($0 < \alpha < 1$), o intervalo de confiança *bootstrap* $100 \times (1 - \alpha)\%$ é dado por

$$\left(\hat{\beta}_{(q_1)}^*; \hat{\beta}_{(q_2)}^* \right)$$

em que, $q_1 = [B \times (\alpha/2)]$ e $q_2 = B - q_1$.

Como exemplo, em IV, para ($\alpha = 0,05$ e $B = 1000$) temos ($q_1 = 25$ e $q_2 = 975$). Logo, o intervalo de confiança *bootstrap* de 95% é dado

por $(\hat{\beta}_{(25)}^*; \hat{\beta}_{(975)}^*)$. Os intervalos de confiança para quaisquer outros parâmetros de interesse são obtidos de maneira similar.

Quando uma variável independente x_j não é contínua, deve-se fazer o processo de reamostragem (passos I, II e III) dentro de cada nível da variável (WU, 1986; TIBSHIRANI, 1988). Outras alternativas ao intervalo de confiança *bootstrap* são discutidas, por exemplo, em Efron e Tibshirani (1993), Davison e Hinkley (1997) e Carpenter e Bithell (2000).

2.5.2 *Bootstrap* Bayesiano

Considere x_1, x_2, \dots, x_n uma amostra independente e identicamente distribuída de tamanho n de uma variável aleatória X com distribuição F e distribuição empírica denotada por \hat{F} . Lembre que a reamostragem *bootstrap* não paramétrico é feita com reposição de \hat{F} . Seja ϕ um parâmetro da distribuição F . Para simplificar a notação, considere x_i unidimensional e ϕ o único parâmetro, mas ambos poderiam ser multidimensionais. Seja $\hat{\phi}$ a estimativa amostral de ϕ baseada em x_1, x_2, \dots, x_n . O *bootstrap* não paramétrico pode ser usado para aproximar a distribuição de $\hat{\phi}$.

Ao invés de amostrar cada x_i com reposição e probabilidade $1/n$ o *bootstrap* Bayesiano usa uma distribuição de probabilidade posterior para cada x_i . Esta distribuição de probabilidade posterior é centrada em $1/n$ para cada x_i , mas varia de uma reamostragem para outra.

Especificamente, as reamostras *bootstrap* Bayesiano são definidas de acordo com Chernick (1999) da seguinte maneira: Defina $n - 1$ variáveis uniformes do intervalo $[0,1]$. Sejam $u_{(1)}, u_{(2)}, \dots, u_{(n)}$ seus valores em ordem crescente. Considere $u_{(0)} = 0$ e $u_{(n)} = 1$, e defina $g_i = u_{(i)} - u_{(i-1)}$ para

$i = 1, \dots, n$. Os g_{iS} são chamados de intervalos de estatísticas de ordem uniforme. O vetor

$$\underset{\sim}{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$$

é usado para definir as probabilidades de reamostragem do *bootstrap* Bayesiano. Ou seja, as n observações são reamostradas com reposição de x_1, x_2, \dots, x_n mas, agora, cada x_i tem probabilidade g_i de ser selecionada. Uma segunda reamostra *bootstrap* Bayesiana é gerada com o mesmo procedimento, mas com um novo conjunto de $n - 1$ números uniformes aleatórios e portanto com um novo vetor $\underset{\sim}{g}$.

Rubin (1981) ressalta que o *bootstrap* não paramétrico e o *bootstrap* Bayesiano são muito similares e, portanto, possuem propriedades em comum. Ele sugere que qualquer limitação atribuída ao *bootstrap* Bayesiano pode ser atribuída ao não paramétrico.

Uma vantagem do *bootstrap* Bayesiano é que pode ser usado para realizar inferências sobre o parâmetro ϕ baseadas na estimativa da distribuição posteriori de ϕ , enquanto que, o *bootstrap* não paramétrico tem somente uma interpretação frequentista sobre a distribuição da estatística $\hat{\phi}$ (CHERNICK, 1999).

Se denotarmos por $g_i^{(1)}$ o valor de g_i na primeira reamostra *bootstrap* Bayesiana e por $g_i^{(2)}$ o valor de g_i na segunda reamostra, encontramos, baseado em resultados elementares de estatísticas de ordem uniforme (DAVID, 1981), que

$$E(g_i^{(1)}) = E(g_i^{(2)}) = \frac{1}{n},$$

$$Var(g_i^{(1)}) = Var(g_i^{(2)}) = \frac{(n-1)}{n^3},$$

$$C(g_i^{(1)}, g_i^{(2)}) = -\frac{1}{n-1},$$

em que "E", "Var" e "C" denotam esperança, variância e correlação respectivamente. Pelas propriedades acima, a distribuição *bootstrap* para $\hat{\phi}$ e a distribuição *bootstrap* Bayesiana para ϕ serão aproximadamente iguais em muitas aplicações.

Rubin (1981) forneceu alguns exemplos e mostrou que o *bootstrap* Bayesiano leva a uma distribuição posterior para ϕ que é Dirichlet e é baseada na distribuição conjugada Dirichlet a priori. Ele passou então a criticar o *bootstrap* Bayesiano por causa da estranha distribuição a priori implícita. Com essa restrição sobre a distribuição a priori no *bootstrap* Bayesiano deixou de recomendá-lo como uma ferramenta geral de inferência, assim como o *bootstrap* não paramétrico, por ser tão similar ao *bootstrap* Bayesiano. Versões como o *bootstrap* paramétrico e o *bootstrap* suavizado superaram algumas dessas objeções.

2.5.3 Reamostragem balanceada

A ideia da reamostragem balanceada é controlar o número de vezes que a observação ocorre na amostra *bootstrap* de modo que em B amostras *bootstrap*, cada observação ocorre o mesmo número de vezes.

Reamostragem balanceada não força cada observação a aparecer uma vez em cada amostra, mas equilibra o número de ocorrências de cada observação

sobre o conjunto de amostras *bootstrap*. Ou seja, se uma observação ocorre duas vezes em uma amostra *bootstrap*, haverá outra amostra *bootstrap* em que esta observação não aparece.

Uma maneira simples de realizar reamostragem balanceada é criar uma sequência das observações X_1, X_2, \dots, X_n repetidas B vezes (ou seja, temos uma sequência Y_1, Y_2, \dots, Y_{nB} , em que $Y_i = X_j$ com j sendo o resto da divisão de i por n). Depois tome uma permutação aleatória π dos inteiros de 1 a nB . Tome $Y_{\pi(1)}, Y_{\pi(2)}, \dots, Y_{\pi(n)}$ como a primeira amostra *bootstrap*, $Y_{\pi(n+1)}, Y_{\pi(n+2)}, \dots, Y_{\pi(2n)}$ como a segunda amostra *bootstrap*, e assim por diante.

Hall (1992) mostrou que reamostragem balanceada produz uma estimativa com variância condicional na ordem de $B^{-1}n^{-2}$.

2.5.4 Amostragem por importância

Amostragem por importância é uma técnica de redução de variância. Um dos primeiros a sugerir seu uso no método *bootstrap* foi Johns (1988). Segundo a ideia de Johns (1988), a amostragem (ou reamostragem) por importância é uma ferramenta útil ao estimar as caudas de uma função distribuição ou para estimar os quantis de uma distribuição e tem valores limitados ao estimar viés e variância. Portanto, é especialmente aplicável para problemas de testes de hipóteses em que a estimação do valor- p é importante.

A ideia da amostragem por importância é controlar a amostragem de modo que tomamos um maior número de amostras da parte da distribuição que é importante para o problema de estimação em estudo. Por exemplo, quando estimamos as caudas de uma distribuição, ou seja, $1 - F(x)$, para valores

grandes de x , ou $F(x)$, para pequenos valores de x , precisamos observar valores maiores que x . No entanto, se a probabilidade $1 - F(x)$ for muito pequena, o tamanho da amostra n deve ser suficientemente grande para observarmos, também, os valores maiores que x . Nesse caso, mesmo para valores extremamente elevados de n e $1 - F(x) > 0$, em que F_n é a distribuição empírica, o número de observações maiores que x na amostra será pequeno.

Reamostragem por importância é uma ideia usada para melhorar as estimativas de um modelo incluindo determinadas observações com maior frequência nas amostras *bootstrap*. Como a distribuição de amostragem é distorcida por esse procedimento, é necessário um esquema de ponderação apropriado para assegurar que a estimativa irá convergir para a estimativa *bootstrap* à medida que o número B de reamostras *bootstrap* aumenta.

Podemos ver a reamostragem por importância como uma generalização da reamostragem uniforme. Em reamostragem uniforme cada x_i tem probabilidade de reamostragem $1/n$. Em geral, podemos definir uma reamostragem por importância por atribuir probabilidade p_i para X_i , em que a única restrição para p_i é que $p_i \geq 0$ para todo i e $\sum_{i=1}^n p_i = 1$.

Quando $p_i = 1/n$, a média da j -ésima amostra *bootstrap*

$$\bar{X}_j^* = \frac{1}{n} \sum_{i=1}^n \bar{X}_i^*$$

é uma estimativa não viesada de \bar{X} e a aproximação de Monte Carlo

$$\bar{X}_B^* = \frac{1}{B} \sum_{j=1}^B \bar{X}_j^*$$

tende a \bar{X} quando $B \rightarrow \infty$. Esta é uma propriedade desejável que é perdida se, para algum valor de i , $p_i \neq 1/n$. No entanto, desde que

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Define $\bar{X}_j^* = \sum_{i=1}^n (\omega_i \bar{X}_i^*)$ em que ω_i é escolhido de modo que se $\bar{X}_i^* = X_i$ então $\omega_i = 1/(np_i)$ para $i = 1, 2, \dots, n$. Esta ponderação garante que condicionada a X_1, X_2, \dots, X_n temos $E(\bar{X}_j^*) = \bar{X}$. Podemos então procurar por valores para p_i de modo que a variância do estimador (neste caso \bar{X}_j^*) seja minimizada.

2.6 Medidas de Influência: *COVRATIO_i*

As medidas de influência DFBETAS e DFFITS fornecem informação sobre o efeito das observações sobre a estimativa dos parâmetros $\hat{\beta}$ e dos valores ajustados \hat{y}_i . Tais medidas não fornecem informação sobre a precisão geral da estimação. Tendo por base essa questão, sabendo que o determinante da matriz de covariância é comumente utilizado como uma medida escalar de precisão, define-se a variância generalizada de $\hat{\beta}$ como

$$VG(\hat{\beta}) = \det[Var(\hat{\beta})] = \det[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}];$$

em que “det” significa determinante. Assim, utilizando-se desta medida, para expressar o efeito da i -ésima observação retirada da amostra sobre o determinante da matriz de covariância das estimativas dos parâmetros, seguindo esta interpretação, utiliza-se a medida de influência CovRatio definida por

$$COVRATIO_i = \frac{\det[QMRes_{(-i)}(\mathbf{X}'_{(-i)}\mathbf{X}_{(-i)})^{-1}]}{\det[QMRes(\mathbf{X}'\mathbf{X})^{-1}]}, \quad i = 1, 2, \dots, n \quad (5)$$

em que o índice $(-i)$ refere-se ao ajuste desconsiderando a i -ésima observação (MONTGOMERY; PECK; VINING, 2006).

Percebe-se que se $COVRATIO_i > 1$, a i -ésima observação aumenta a precisão de estimação, enquanto que se $COVRATIO_i < 1$, a inclusão da i -ésima observação diminui a precisão.

Valores de corte para $COVRATIO$ não são simples de se obter. Belsley, Kuh e Welsch (1980) sugerem que se $COVRATIO_i > 1 + 3p/n$ ou se $COVRATIO_i < 1 - 3p/n$, em que p é o número de parâmetros e n o número de observações, então a i -ésima observação pode ser considerada uma observação influente. Isso permite que a medida Covratio atue como um indicador de observações que tem influência tanto positiva quanto negativa sobre o conjunto de coeficientes (HAIR et al., 2007).

3 METODOLOGIA

Em consonância com os objetivos propostos, a obtenção dos resultados foi feita seguindo a descrição metodológica organizada nas seguintes etapas:

- 3.1- Simulação Monte Carlo dos modelos Gompertz e Logístico;
- 3.2- Incorporação da medida de influência CovRatio no método *bootstrap* e
- 3.3- Avaliação da acurácia e precisão das estimativas dos parâmetros dos modelos Logístico e Gompertz.

3.1 Simulação Monte Carlo dos modelos Gompertz e Logístico

Os modelos não lineares Gompertz e Logístico, estudados neste trabalho, foram definidos respectivamente nas seguintes equações:

$$y_i = \alpha \exp\{-\exp[-\beta(x_i - \gamma)]\} + \varepsilon_i \quad (i = 1, \dots, n) \quad (6)$$

$$y_i = \frac{\alpha}{1 + \exp(\beta - \gamma x_i)} + \varepsilon_i \quad (i = 1, \dots, n) \quad (7)$$

considerando, os tamanhos amostrais especificados em $n=50$ e 150 , x_i , $i = 1, \dots, n$, como efeito fixo. O parâmetro α referiu-se a assíntota superior; β correspondeu ao intercepto da curva e γ indicou a taxa média de crescimento. Por fim ε_i referiu-se ao i -ésimo resíduo gerado por uma distribuição normal contaminada, assumindo a seguintes distribuições:

$$\varepsilon_i \sim (1 - \delta) \times N(0,1) + (\delta) \times \text{Beta}(6,2); \quad (8)$$

$$\varepsilon_i \sim (1 - \delta) \times N(0,1) + (\delta) \times \text{Beta}(6,6); \quad (9)$$

$$\varepsilon_i \sim (1 - \delta) \times N(0,1) + (\delta) \times \text{Beta}(2,6); \text{ sendo} \quad (10)$$

δ o percentual de *outliers* contido na amostra, previamente especificado nos valores de 5 e 10%. Os parâmetros definidos na distribuição Beta, descritos nas expressões (8-10), foram arbitrariamente fixados, de tal forma que a distribuição dos resíduos a qual os *outliers* foram gerados, apresentasse diferentes assimetrias, conforme as situações ilustradas na Figura 5.

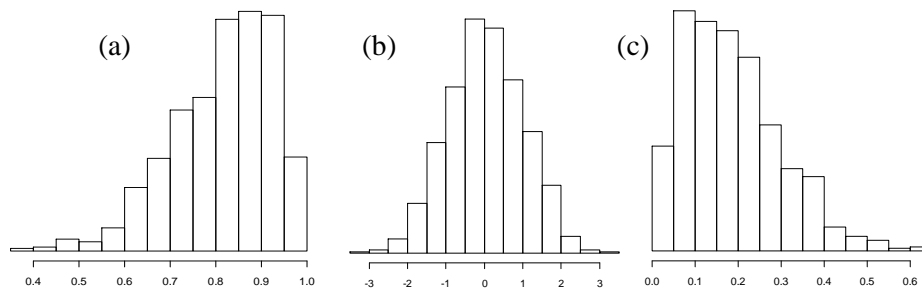


Figura 5 Representação gráfica das distribuições utilizadas na geração de *outliers* nos resíduos (a) Beta(6,2); (b) Beta(6,6) e (c) Beta(2,6)

Mantendo essas especificações, os cenários simulados foram dados pela combinação dos fatores, resumidos na Tabela 1.

Para cada situação descrita na Tabela 1, em ambos os modelos, os valores paramétricos foram fixados em $\alpha = 1$, $\beta = 1$ e $\gamma = 1,5$. Com a imposição desses valores paramétricos, para cada amostra gerada as estimativas de mínimos quadrados foram obtidas utilizando o algoritmo iterativo Gauss - Newton assumindo as seguintes configurações: número máximo de iterações definido $r = 1000$ e critério de convergência fixado em $\xi = 1e^{-100}$. Mantendo esses critérios, das 500 realizações Monte Carlo considerou apenas as

K realizações que apresentaram a convergência alcançada no processo Monte Carlo e no procedimento *bootstrap* descrito na Seção 3.2.

Tabela 1 Resumo dos cenários avaliados na simulação Monte Carlo e *bootstrap*.

	n	$\delta(\%)$	Distribuição dos resíduos utilizados para geração de outliers.		
Modelo	50	5	Beta (6,2)	Beta (6,6)	Beta (2,6)
		10	Beta (6,2)	Beta (6,6)	Beta (2,6)
Logístico	150	5	Beta (6,2)	Beta (6,6)	Beta (2,6)
		10	Beta (6,2)	Beta (6,6)	Beta (2,6)
	n	$\delta(\%)$	Distribuição dos resíduos utilizados para geração de outliers.		
Modelo	50	5	Beta (6,2)	Beta (6,6)	Beta (2,6)
		10	Beta (6,2)	Beta (6,6)	Beta (2,6)
Gompertz	150	5	Beta (6,2)	Beta (6,6)	Beta (2,6)
		10	Beta (6,2)	Beta (6,6)	Beta (2,6)

3.2 Incorporação da medida de influência CovRatio no método *bootstrap*

A incorporação da medida de influência CovRatio no procedimento *bootstrap*, em ambos os modelos não lineares (5) e (6) foi realizada inicialmente em duas etapas conforme o seguinte procedimento:

- (a) Obtenção das matrizes de derivadas parciais avaliadas na solução obtida através do algoritmo Gauss – Newton, representada por \mathbf{X} ;
- (b) Obtenção das estimativas de mínimos quadrados sem o intercepto considerando o modelo na forma $\tilde{Y} = \mathbf{X} + \tilde{\varepsilon}$.

Desta forma, assumindo os modelos linearizados, definiu-se a medida de influência F_i , para $i = 1, 2, \dots, n$ arbitrariamente definida por meio da medida Covratio. Assim sendo, procedeu-se com a execução do algoritmo descrito no Quadro 1, com o propósito de calcular as probabilidades para cada observação amostral.

Quadro 1 Algoritmo utilizado para cálculo das probabilidades de reamostragem incorporando as medidas de influência

<p>(1°) Considere uma amostra definida pelo conjunto $D = \{(y_1, x_1, \varepsilon_1), \dots, (y_n, x_n, \varepsilon_n)\}$ em que cada elemento foi gerado seguindo um modelo linearizado.</p> <p>(2°) Calcular a medida de influência F_i, para $i = 1, 2, \dots, n$ referente à medida CovRatio.</p> <p>(3°) Atribuir a cada observação o peso w_i obedecendo à regra:</p> <ul style="list-style-type: none"> • Se $F_i \geq r_i$ então $w_i = F_i$ • Se $F_i \leq r_i$ então $w_i = r_i$ <p>em que r_i indica o i-ésimo resíduo studentizado do modelo não linear linearizado.</p> <p>(4°) Calcular as probabilidades de reamostragem por $p_i = 1 - \frac{w_i}{\sum_{i=1}^n w_i}$, $i = 1, \dots, n$.</p> <p>(5°) Atribuir a cada elemento do conjunto D os pesos p_i, conforme segue a ilustração:</p> $D = \{(y_1, x_1, \varepsilon_1), (y_2, x_2, \varepsilon_2), \dots, (y_n, x_n, \varepsilon_n)\}$ <div style="text-align: center;"> $\downarrow \quad \quad \downarrow \quad \quad \downarrow$ $p_1 \quad \quad p_2 \quad \quad p_n$ </div>

Com a adaptação do algoritmo (Quadro 1) no método *bootstrap*, os subconjuntos formados a partir de $B = 500$ reamostragens *bootstrap* do conjunto D realizadas com reposição e representados por $D_b = \{(y_1^*, x_1^*, \varepsilon_1^*), (y_2^*, x_2^*, \varepsilon_2^*), \dots, (y_n^*, x_n^*, \varepsilon_n^*)\}$ com $b = 1, \dots, B$. Seguindo esta metodologia, tornou-se possível obter as probabilidades de reamostragem, ponderadas pela CovRatio. Tendo por base esse critério e a comparação do valor de F_i com o resíduo studentizado, uma determinada observação identificada como *outlier* e influente apresentou uma baixa ocorrência em cada subconjunto gerado. Frente ao exposto, o procedimento *bootstrap* mantendo esta modificação nas probabilidades de reamostragens foi executado, e para cada subconjunto D_b as estimativas de mínimos quadrados ordinário foram obtidas. Ao final da execução desse procedimento, os B subconjuntos foram considerados na composição da distribuição empírica de cada parâmetro.

3.3 Avaliação da acurácia e precisão das estimativas dos parâmetros dos modelos Logístico e Gompertz

A acurácia das estimativas dos modelos Logístico e Gompertz foi feita considerando duas situações: estimativas Monte Carlo obtidas e estimativas obtidas via *bootstrap* com medida de influência corrigido pelo viés Monte Carlo. Para tais situações, denominou-se viés (MC) e viés (BMIC), cujos estimadores foram definidos em (11) e (12).

Convém ressaltar que para a obtenção das estimativas dos vieses, considerou-se K amostras, interpretado como o número de amostras “válidas”, no sentido de que a amostra simulada apresentou convergência do método iterativo Gauss-Newton na simulação Monte Carlo e no procedimento *bootstrap*.

Frente ao exposto, para cada amostra “válida” as estimativas dos modelos mencionados na Seção 3.1 foram computadas, gerando a distribuição empírica, na qual, permitiu-se o cálculo dos vieses (11) e (12).

$$Viés(\widehat{MC})_j = \frac{1}{K} \left| \sum_{k=1}^K (\hat{\theta}_{mjk} - \theta_j) \right|, \quad (11)$$

$$Viés(\widehat{BMIC})_j = \frac{1}{K} \left| \left| \sum_{k=1}^K (\hat{\theta}_{bjk} - \theta_j) \right| - Viés(\widehat{MC})_j \right|, \quad (12)$$

em que θ_j indicou o j -ésimo ($j = 1, 2, 3$) parâmetro do vetor paramétrico $\Theta' = [\alpha \quad \beta \quad \gamma]$; $\hat{\theta}_{mjk}$ e $\hat{\theta}_{bjk}$ respectivamente referiu-se a estimativa Monte Carlo e *bootstrap* do j -ésimo parâmetro especificado em Θ' .

Em se tratando da precisão das estimativas dos modelos, computou-se o desvio padrão das estimativas obtidas via o método Monte Carlo e *bootstrap* com medidas de influência com correção de viés, seguindo as expressões (13) e (14).

$$D_{MC} = \sqrt{\frac{\sum_{k=1}^K (\hat{\theta}_{mjk} - \theta_j)^2}{K - 1}} \quad (13)$$

$$D_{BMIC} = \sqrt{\frac{\sum_{k=1}^K (\hat{\theta}_{bjk} - \theta_j)^2}{K - 1}} \quad (14)$$

4 RESULTADOS E DISCUSSÃO

Os resultados descritos na Tabela 2 indicam o percentual de convergência do método de Gauss-Newton para os modelos Logístico e Gompertz, nas situações em que as amostras foram classificadas como “válidas” conforme mencionado na Seção 3.3. Ressalta-se que o percentual de convergência é relativamente baixo, no entanto, se considerarmos as definições revistas por Monico et al. (2009) referentes a acurácia e precisão, sendo que o principal diferencial entre esses conceitos é o fato de que a precisão contempla apenas efeitos aleatórios e a acurácia contempla efeitos aleatórios e sistemáticos. Portanto, tendo em vista que em todo o processo para validar a metodologia proposta utilizou-se de um mecanismo gerador de dados caracterizando o efeito aleatório e que a componente sistemática, em ambos os modelos, proporciona a mesma interpretação aos parâmetros, o baixo percentual de amostras envolvidas nesse estudo não permite obter conclusões mais amplas. Porém, não invalida os resultados obtidos empiricamente, uma vez que, na abordagem *bootstrap*, os pesos das observações são diferentes.

Neste contexto, seguindo recomendações de Johns (1988), a aplicação do procedimento *bootstrap*, com probabilidade de reamostragem não uniforme, implica em uma redução do número de simulações para obter estimativas precisas. Desta forma, reduz o esforço computacional exigido ao aplicar o método *bootstrap* convencional.

Tabela 2 Percentual de amostras válidas que apresentaram convergência do Gauss-Newton para os modelos Logístico e Gompertz.

δ	Distribuição dos resíduos	n	Logístico	Gompertz
5%	Beta(6,6)	50	21%	4%
		150	28%	2%
	Beta(6,2)	50	19%	2%
		150	27%	2%
	Beta(2,6)	50	18%	3%
		150	27%	4%
10%	Beta(6,6)	50	18%	3%
		150	27%	2%
	Beta(6,2)	50	19%	3%
		150	23%	2%
	Beta(2,6)	50	20%	3%
		150	28%	3%

Dado que a principal motivação desse trabalho consiste em propor um método que altere as probabilidades de reamostragem, mantendo em foco a finalidade de reamostrar com menor frequência as observações mais influentes, prossegue-se com a discussão e interpretação dos resultados, organizados nas seguintes seções: 4.1 – Estudo da acurácia das estimativas *bootstrap* com incorporação de medidas de influência e 4.2 – Estudo da precisão das estimativas *bootstrap* com incorporação de medidas de influência.

4.1 Estudo da acurácia das estimativas *bootstrap* com incorporação de medidas de influência

Mantendo os cenários descritos na metodologia (Seção 3.1), os resultados descritos na Tabela 3 referem-se à acurácia das estimativas dos

parâmetros dos modelos Logístico e Gompertz obtidos via o método Monte Carlo e o método *bootstrap* proposto neste trabalho com a incorporação da medida de influência CovRatio, utilizada como critério para definir os pesos para cada observação amostral. Neste contexto, a acurácia foi avaliada utilizando o viés relativo em duas abordagens definidas por MC e BMIC referindo-se respectivamente a Monte Carlo e *bootstrap* com medida de Influência corrigida pelo viés de Monte Carlo.

Segundo Pereira (2004), esse viés permite inferir a confiabilidade das medidas resultante de variações controláveis, portanto, coerentemente, mantiveram-se fixos os valores das covariáveis com distribuição dos resíduos caracterizada por diferentes graus de simetria. Além do mais, é razoável supor que a obtenção de resultados provenientes de diferentes métodos de estimação, contextualizados, neste trabalho, pelos métodos Monte Carlo e *bootstrap* supostamente poderão ocasionar dúvidas em relação à acurácia dos estimadores de mínimos quadrados. Tendo por base esta afirmativa, os resultados são apresentados na Tabela 3.

Em função dos resultados descritos na Tabela 3, considerando diferentes graus de simetria para o resíduo, observou-se que o método BMIC apresentou uma melhoria na acurácia das estimativas referente ao parâmetro α relacionado à assíntota superior.

Especificamente em relação à distribuição dos resíduos simétrica (Beta (6,6)) em concordância com os resultados obtidos por Cook, Tsai e Wei (1986) referente ao estudo dos vieses das estimativas de máxima verossimilhança em modelos de regressão não linear, considerando os erros com distribuição normal, os autores concluíram que o baixo valor do viés, meramente poderá ser uma consequência da posição da covariável no espaço amostral,

portanto, nesse sentido, tendo por base os resultados obtidos em poucos cenários de simulação, não há fortes evidências estatísticas que permitam conclusões mais amplas sobre o desempenho do método BMIC, porém, ressalta-se a existência de uma coerência dos resultados obtidos empiricamente neste trabalho com os resultados obtidos pelos referidos autores.

Tabela 3 Resultados referentes aos vieses relativos das estimativas do parâmetro α referente a assíntota superior dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,2506	0,1823	0,2007	0,1131	0,1539	0,1588
	150	0,1090	0,0997	0,0624	0,0671	0,0715	0,0710
BMIC	50	0,0285	0,0167	0,0208	0,0327	0,0878	0,0539
	150	0,0065	0,0376	0,0507	0,0440	0,0618	0,0208
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,5424	0,0869	0,0869	0,0047	0,0930	0,0469
	150	0,0737	0,0722	0,0113	0,0386	0,0498	0,0329
BMIC	50	0,0616	0,5245	0,0829	0,0709	0,3176	0,0960
	150	0,0343	0,0020	0,0122	0,0441	0,0767	0,0038

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

Frente a essas considerações que corroborem com os resultados obtidos por Cook, Tsai e Wei (1986), dadas as perturbações provocadas nos resíduos, por diferentes graus de assimetria por meio das distribuições Beta (2,6) e Beta (6,2), há evidências para afirmar que a violação nas condições de regularidade

dos modelos não lineares contribuiu para elevados resultados dos vieses, observado principalmente em pequenas amostras ($n=50$). Mediante a este fato, observou-se que o método BMIC proporcionou uma redução na estimativa dos vieses. Tal fato não foi perceptível no modelo Gompertz, porém dado o baixo número de amostras validadas na simulação, este resultado poderá ser questionado.

Convém ressaltar que mesmo na limitação, os resultados obtidos para grandes amostras ($n=150$) foram adequados em algumas situações bem específicas como, por exemplo, os resultados referentes à distribuição Beta (2,6) em diferentes quantidades de *outliers* ($\delta=5\%$) e ($\delta=10\%$) em que os valores obtidos foram inferiores a 0,01 e concordantes com a recomendação de Box (1971) que considera um valor aceitável para o viés relativo igual ou inferior a 0,01.

Considerando o estimador de mínimos quadrados relacionado ao intercepto da curva, representado pelo parâmetro β , por meio dos resultados descritos na Tabela 4, observou-se que o método BMIC novamente proporcionou uma redução de viés para ambos tamanhos amostrais avaliados, nas situações em que a distribuição dos resíduos foi assimétrica. Contudo, mesmo ocorrendo essa redução, os resultados obtidos são empíricos e carecem de resultados inferenciais que confirmem tais evidências. Nesse sentido, uma alternativa a ser empregada em estudos futuros, consiste na aplicação de técnicas de correção de vieses, sugeridos por Cox e Snell (1968) e Mackinnon e Smith (1998).

Tabela 4 Resultados referentes dos vieses relativos das estimativas do parâmetro β referente ao intercepto da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,1854	0,5729	0,4607	0,2838	0,3124	0,2129
	150	0,6067	1,0903	0,6270	0,7730	0,8470	1,1090
BMIC	50	0,2908	0,0913	0,1778	0,3765	0,1782	0,1594
	150	0,0137	0,5854	0,1527	0,3514	0,2736	0,6737
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,3939	0,7611	0,4209	0,1423	0,3988	0,2394
	150	0,3861	0,1383	0,0578	0,4218	0,5821	0,0434
BMIC	50	0,8767	0,2896	0,0891	0,2428	0,0646	0,0915
	150	0,2080	0,0131	0,0063	0,5628	0,5288	0,3725

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

Outra opção seria o desenvolvimento de fórmulas que permitam o cálculo dos vieses de segunda ordem, sendo úteis nas situações de tamanhos amostrais menores. Nesse aspecto, justificam-se novos estudos relacionados aos métodos de estimação, que propiciem aproximação da série de Taylor até o termo de segunda ordem, ou métodos quasi-newton com o propósito de avaliar a aproximação da matriz hessiana (EVERITT, 1987). Neste caso, a taxa de convergência em ambos os modelos seria supostamente melhorada, principalmente no que tange ao modelo de Gompertz. Tal afirmativa é corroborada com os estudos de Mazzini et al. (2005) referente aos modelos

Logístico e Gompertz aplicados a curvas de crescimento de novilhos hereford com resíduos heterocedásticos e autoregressivos. Nesta aplicação, os autores concluíram que para todos os métodos de estimação: Quadrados Mínimos Ponderados, Quadrados Mínimos Ponderados Generalizados com erros autoregressivos de primeira e segunda ordem, os modelos Logístico e Gompertz, apresentaram maiores percentuais de convergência.

Em se tratando da contaminação dos resíduos dado por uma distribuição simétrica (Beta (6,6)), há evidências para afirmar que o uso do método BMIC não proporcionou resultados condizentes, no sentido de alcançar a redução do viés. Mediante a este fato, recomenda-se que o uso de alguma transformação na variável resposta, no sentido minimizar o efeito de *outliers*. Rodrigues et al. (2010) sugerem o uso de regressão isotônica com pesos diferentes. Na linha de modelos robustos, alternativas à suposição de erros normais têm sido propostas na literatura.

Uma dessas alternativas consiste em assumir para os erros distribuições com caudas mais pesadas do que a normal, a fim de reduzir a influência de pontos aberrantes. Maiores detalhes poderão ser vistos em Cordeiro (2004), referente a correção de viés dos estimadores de máxima verossimilhança na classe dos modelos de regressão não lineares simétricos homocedásticos.

A fim de corroborar com o uso dessas metodologias, os resultados ilustrados na Tabela 5, correspondendo aos valores de viés, evidenciaram que a violação da suposição de normalidade por meio da contaminação de *outliers* em diferentes quantidades, afetou seriamente as estimativas dos vieses relacionadas ao parâmetro γ , interpretado como a taxa média de crescimento. Tal fato é notório principalmente para pequenas amostras, uma vez que, ao considerar

tamanhos amostrais pequenos ou até mesmo moderados, esse viés poderá tornar-se grande quando comparado ao erro padrão do estimador.

Tabela 5 Resultados referentes dos vieses relativos das estimativas do parâmetro γ referente a taxa média de crescimento dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,7949	0,7836	0,8138	0,8020	0,8019	0,8028
	150	0,8122	0,8548	0,8419	0,8449	0,8684	0,8308
BMIC	50	1,8578	0,6463	1,8239	0,7527	1,3738	2,5804
	150	0,3760	2,0029	0,4386	1,9275	1,6052	0,1141
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,0390	0,7216	0,1213	0,0188	0,5878	0,7899
	150	0,1554	0,2033	0,3648	0,2074	0,8509	0,8669
BMIC	50	0,8497	5,7129	0,0401	0,1200	3,7395	1,4689
	150	0,4658	1,4162	2,0225	4,5126	7,0415	1,0995

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

4.2 Estudo da precisão das estimativas *bootstrap* com incorporação de medidas de influência

Seguindo as especificações descritas na metodologia e avaliadas em relação a acurácia na Seção 1.1, os resultados encontrados nas Tabela 6

corresponderam à precisão das estimativas relacionada ao parâmetro α , interpretado como à assíntota superior dos modelo Logístico e Gompertz.

Tabela 6 Resultados referentes ao desvio padrão das estimativas do parâmetro α referente a assíntota superior dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,7130	0,2954	0,5303	0,2383	0,2718	0,3293
	150	0,2256	0,2427	0,1513	0,2172	0,1797	0,2270
BMIC	50	0,4138	0,4435	0,3293	0,4794	0,6221	0,5011
	150	0,2498	0,4691	0,3647	0,3977	0,6015	0,5075
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,3603	0,3925	0,3134	0,3255	0,3877	0,3405
	150	0,3132	0,3043	0,2488	0,3643	0,3412	0,2753
BMIC	50	0,3822	1,6112	0,3988	0,3575	1,1256	0,4723
	150	0,3476	0,3522	0,2896	0,4493	0,4738	0,2802

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

Por meio dos resultados descritos na Tabela 6, considerando todos os cenários avaliados, observou-se que o uso do método BMIC, de um modo geral, resultou em estimativas imprecisas. Naturalmente, tal resultado é influenciado pelo baixo número de amostras consideradas válidas no processo de simulação Monte Carlo, porém há referências que sugerem que a violação das condições de regularidade imposta no ajuste dos modelos não lineares com erro normais

implica que a precisão das estimativas seja afetada. Maiores detalhes podem ser encontrados em Fang, Kotz e Ng (1990) e Fang e Anderson (1990). Meyer et al. (2006) apontam que a superestimativa da falta de precisão está amplamente relacionada à falta de acurácia (viés relativo alto). Neste sentido, o autor recomenda que outras medidas complementares, como por exemplo, a raiz quadrada do erro médio de predição possa ser empregada, uma vez que esta medida apresenta uma correção para a falta de acurácia.

Em se tratando do estudo da precisão referente ao parâmetro β relacionado com o intercepto da curva, os resultados encontrados na Tabela 7, corresponderam à precisão das estimativas obtidas. Para tal situação, é possível observar que o método BMIC apresentou uma redução no desvio padrão médio, para os dois tamanhos amostrais avaliados em ambas as quantidades de *outliers* avaliadas em $\delta=5\%$ e 10% .

Quanto ao modelo Gompertz, embora o mesmo apresente um número baixo de amostras simuladas, é conveniente mencionar que Mikhail e Ackermann (1976) mencionam que a diferença da acurácia e precisão, está associada à presença ou ausência de erros sistemáticos. Neste contexto, os autores ratificam que a precisão considera-se apenas efeitos aleatórios, enquanto que a acurácia inclui não apenas os efeitos aleatórios, mas também sistemáticos. Desta forma, tendo por base esta afirmativa, mesmo com baixo número de simulações para as situações avaliadas neste trabalho, o modelo Gompertz apresentou estimativas mais precisas em relação ao modelo Logístico.

Em se tratando do desvio padrão médio, relativo à taxa média de crescimento representada pelo parâmetro γ , os resultados encontrados na Tabela 8 indicaram que nestes cenários de avaliação, a aplicação do método BMIC

resultou em baixa precisão. Tal fato foi detectado em todas as situações avaliadas, incluído amostras maiores ($n=150$).

Tabela 7 Resultados referentes ao desvio padrão das estimativas do parâmetro β referente ao intercepto da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	1,9535	2,4357	1,7542	1,6458	1,5706	1,1813
	150	1,7344	2,9820	1,6555	2,6015	1,1162	2,8225
BMIC	50	0,7019	0,7519	1,5189	1,6382	0,8350	0,7937
	150	0,5931	0,9043	1,0720	0,8608	0,6108	0,8357
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	1,6254	0,2560	0,4970	0,5902	0,6249	0,6275
	150	0,4900	0,4887	0,5628	0,9458	0,9990	0,4057
BMIC	50	0,5698	1,0390	0,9891	1,3368	0,7312	0,9889
	150	0,3680	0,6650	1,0259	0,1252	0,3890	1,3224

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

Sendo concordantes com os estudos realizados por Evans (1996) referente ao efeito do grau de assimetria de duas curvas de crescimento sobre a precisão das estimativas de mínimos quadrados de seus parâmetros, através da estatística t de Student. Para isso, o autor considerou o modelo Logístico de 4 parâmetros desenvolvido por Stone (1980). Tal modelo foi gerado adicionando um parâmetro θ à equação Logística. A função desse parâmetro no modelo é dada por relacionar a altura que o ponto de inflexão se encontra na curva, em

função do parâmetro α , sendo que este representa a assíntota superior do modelo. Com esta especificação, θ é interpretado como o grau de assimetria do modelo. O autor concluiu que a presença da assimetria possui um efeito mais prejudicial nas variâncias dos parâmetros do modelo Logístico modificado, sendo que o parâmetro mais afetado é a taxa média de crescimento.

Tabela 8 Resultados referentes ao desvio padrão das estimativas do parâmetro γ referente a taxa média de crescimento da curva dos modelos Logístico e Gompertz obtidas pelo método MC e BMIC considerando diferentes tamanhos amostrais (n), percentual de *outliers* presente nos resíduos (δ) provenientes de distribuições com diferentes graus de simetria.

		Modelo Logístico					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	0,3057	0,6745	0,3107	0,3064	0,3147	0,3092
	150	0,5041	0,3429	0,2546	0,2955	0,2303	0,4360
BMIC	50	13,7013	10,6441	8,2448	10,0699	11,7297	8,3438
	150	34,3643	38,2027	41,7933	23,5487	46,1629	35,0767
		Modelo Gompertz					
Método	n	$\delta = 5\%$			$\delta = 10\%$		
		Beta(6,6)	Beta(6,2)	Beta(2,6)	Beta(6,6)	Beta(6,2)	Beta(2,6)
MC	50	2,2107	2,6511	1,9985	1,4973	2,7456	3,0269
	150	2,9576	2,2813	2,8765	3,4600	1,2611	4,2824
BMIC	50	5,3719	17,2933	4,3105	2,6300	23,2762	7,5451
	150	7,6566	3,9115	15,2174	12,8765	28,4606	9,6148

MC refere-se ao método Monte Carlo.

BMIC refere-se ao método *bootstrap* com medida de influência corrigida pelo viés de Monte Carlo.

5 CONCLUSÃO

Em função dos cenários avaliados, concluiu-se que a metodologia proposta é factível de ser utilizada permitindo ao pesquisador definir outras medidas de influência a serem consideradas como critério para obtenção das probabilidades de reamostragem.

Em relação aos resultados empíricos, a aplicação do método BMIC proporcionou uma melhoria na acurácia dos estimadores referente à assíntota superior (α) e ao intercepto (β) do modelo logístico. Em se tratando do modelo Gompertz, em geral apresentou os mesmos resultados, porém dada a baixa porcentagem de convergência observada, recomenda-se que novos estudos sejam realizados para obtenção de conclusões mais confirmatórias.

REFERÊNCIAS

ATKINSON, A. C. **Plots, Transformations and Regression**. Oxford: Oxford University Press. 1985.

BARNETT, V.; LEWIS, T. **Outlier in Statistical Data**. 3. ed. New York: John Wiley. 1993.

BARROS, E. A. C.; SIMÕES, P. A.; ACHCAR, J. A.; MARTINEZ, E. Z.; SHIMANO, A. C. Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. **Revista Colombiana de Estatística**. v. 31, n. 1, p. 111 – 129, Jun. 2008.

BATES, D.M.; WATTS, D.G. **Nonlinear Regression Analysis and its Applications**. New York: J. Willey, 1980. 365p.

BATES, D. M.; WATTS, D. G. **Nonlinear Regression Analysis and its applications**. New York: J. Willey, 1988. 384p.

BELSLEY, D. A.; KUH, E.; WELSCH, R. E. **Regression Diagnostics: Identifying Influential Data and Sources of Collinearity**. New York: J. Willey, 1980.

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

BOX, G. E. P.; LUCAS, H. L. Design of experiments in nonlinear situations. **Biometrika**, Oxford, n. 46, p. 77-90, 1959.

CARPENTER, J.; BITHELL, J. *Bootstrap* Confidence Intervals: When, which, what? A Practical Guide for Medical Statistician. **Statistics in Medicine**. v. 19, p. 1141 – 1164. 2000.

CHERNICK, M. R. ***Bootstrap Methods***: A Practitioner's Guide. New York: John Wiley & Sons. 1999. 264p.

COOK, R. D.; WEISBERG, S. **Residual and Influence in Regression**. London. Chapman and Hall, 1982. 229p.

COOK, R.; TSAI, C.L; WEI, B. **Bias in nonlinear regression**. *Biometrika*, 73, p. 615-623, 1986.

CORDEIRO, G.M. Corrected LR tests in symmetric nonlinear regression models. **Journal of Statistical Computation and Simulation**. v. 74, n. 8, p. 609 - 620, 2004.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados e Extensões**. Piracicaba. 2009. 413p.

COX, D. R.; SNELL, E.A general definition of residuals. **Journal of the Royal Statistical Society, B**, v. 30, p. 248-275,1968.

DAVID, H. A. **Order Statistics**. New York: Wiley, 1981. 157 p.

DAVISON, A.C.; HINKLEY, D.V. ***Bootstrap Methods and their Application***. New York: Cambridge University Press, 1997. 592p.

DICICCIO, T. J.; EFRON, B. *Bootstrap* Confidence Intervals. **Statistical Science**. v. 11, n. 3, p. 189 – 228. 1996.

DOORNBOS, R. **Testing for a single outlier in linear model.** *Biometrics*. 1981. p. 705-712.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis.** 3. ed. New York: John Wiley & Sons. 1998. 706 p.

ELLENBERG, J. H. Testing for a single *outlier* from a general linear regression. **Biometrics**. p. 637-645. 1976.

EFRON, B. **The Jackknife, the Bootstrap, and Other Resampling Plans.** Philadelphia: SIAM. 1982a. 199p.

EFRON, B. Jackknife-after-*bootstrap* standard errors and influence functions. **Journal of the Royal Statistical Society**. v. 54, p. 83-127. 1992.

EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Bootstrap.** New York: Chapman & Hall, 1993.

EVANS, M. Aspects of maximum likelihood estimation of asymmetric growth curves. **Journal of Applied Statistics**. 23:5, p. 467-492. 1996.

EVERITT, E.S. **Introduction to Optimization Methods and their Application in Statistics.** Chapman and Hall, 1987.

FANG, K. T.; KOTZ, S.; NG, K. W. **Symmetric multivariate and related distributions.** London: Chapman and Hall, 1990.

FANG, K. T.; ANDERSON, T. W. **Statistical inference in elliptically contoured and related distributions.** New York: Allerton Press, 1990.

FUNG, W. K.; NGAI, M. C. Critical values for testing for a single *outlier* in a nonlinear regression model. **Communications in Statistics – Simulation and Computation**. 27:1,95-105. 1998.

GALLANT, A.R. **Nonlinear Statistical Models**. New York: J. Willey, 1987. 624p.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise Multivariada de Dados**. 5. Ed. Porto Alegre: Bookman, 2007.

HALL, P. **The *Bootstrap* and Edgeworth Expansion**. New York: Springer-Verlag, 1992. 294p.

HJORTH, J. S. U. **Computer Intensive Statistical Methods. Validation, Model Selection and *Bootstrap***. Chapman & Hall. 1994.

JOHNS, M. V. Jr. **Importance sampling for *bootstrap* confidence intervals**. J. Am. Statistic. Associ. 1988. 354p.

KHURI, A. I.; CORNELL, J. A. **Response Surface: Designs and Analyses**. New York: Marcel Dekker. 1987.

LUND, R. E. Tables for an approximate test for *outliers* in linear models. **Technometrics**. 1975. p. 473-476.

MACKINNON, J.; SMITH, S. A. A. Approximation bias correction in econometrics. **Journal of Econometrics**, 85, p. 205-230, 1998.

MAZZINI, A.R.A.; MUNIZ, J.A.; SILVA, F.F.; AQUINO, L.H. Curva de Crescimento de novilhos Hereford: heterocedasticidade e resíduos Autoregressivos. **Ciência Rural**, Santa Maria, v.35, n.2, p.442-427, 2005.

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

MEYER, P. M.; MACHADO, P. F; COLDEBELLA, A.; CASSOLI, L. D.; COELHO, K. O.; RODRIGUES, P. H. M. Validação de modelos de predição de nitrogênio uréico no leite, estimando-se o consumo individual pelo consumo no rebanho. **Acta Scientiarum**, v. 28, n.1, p.73-79, 2006.

MIKHAIL, E.; ACKERMAN, F. **Observations and Least Squares**. University Press of America, 1976.497 p.

MONICO, J.F.G.; PÓZ, A.P.D.; GALO, M.; SANTOS, M.C.; OLIVEIRA, L.C. **Acurácia e precisão: Revendo os conceitos de forma acurada**. Bol. Ciênc. Geod., sec. Comunicações, Curitiba, v. 15, no 3, p.469-483, jul-set, 2009.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. John Willey. 2006.

PEREIRA, J. C. R. **Análise de dados qualitativos: Estratégias metodológicas para as ciências da saúde, humanas e sociais**. São Paulo: Edusp. 2004.

PRESCOTT, P. An approximate test for *outliers* in linear models. **Technometrics**. 1975. p. 129-132.a

R DEVELOPMENT CORE TEAM R: A language and environment for statistical computing: versão 2.7.0.v. Vienna: R Foundation for Statistical Computing. Disponível em: <<http://www.r-project.org/>>. Acesso em 10 fev. 2011.

RATKOWSKY, D. A. **Nonlinear Regression Modeling**. New York: M. Dekker. 1983. 276p.

RODRIGUES, A.; CHAVES, L.M.; SILVA, F.F.; ZEVIANI, W.M. Utilização da Regressão Isotônica em Estudos de Curvas de Crescimento, **Revista Brasileira de Biometria**, v.28, n.4, p.85-101, 2010.

ROSS, G. J. S. **Nonlinear Estimation**. New York: Springer-Verlag. 1990.

RUBIN, D. B. **The Bayesian Bootstrap**. Ann. Statistic. 1981. 504p.

SCHABENBERGER, O.; PIERCE, F.I. **Contemporary statistical models for the plant and soil sciences**. Boca Raton: CRC, 2002. 753p.

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

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete sample). **Biometrika**, Great Britain, v. 52, 1965.

SHAO, J.; TU, D. **The Jackknife and Bootstrap**, Springer-Verlag, New York, 1996.

STONE, R. Sigmoids. **Bulletin of Applied Statistics**. 1980. v. 7, p. 59-119.

TIBSHIRANI, R. Correction to Discussion of: "Jackknife, *Bootstrap* and Resampling Methods in Regression Analysis". **Ann. Statist.** v. 16, n. 1, 1988.

TIETJEN, G. L.; MOORE, R. H.; BECKMAN, R. J. Testing for a single *outlier* in simple linear regression. **Technometrics**. 1973. p. 717-721.

TSE, Y. K. Assessing Lund's critical values for testing for *outliers* in linear regression models. **J. Appl. Statistic**.1988. p. 363-366.

WU, C. F. J. Jackknife, *Bootstrap* and other Resampling Methods in Regression Analysis. **Ann. Statist.** v. 14, n. 4. 1986. p. 1261-1350.

ZEVIANI, W. M. **Avaliação de modelos de regressão não linear na cinética de liberação de potássio de resíduos orgânicos**. 2009. 87p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) – Universidade Federal de Lavras.

APÊNDICE

Script do R utilizado na simulação Monte Carlo dos modelos Logístico e Gompertz.

```

options(show.error.messages = F)
nls.control(maxiter = 100, tol = 1e-10, minFactor=1/1024)

# ##### Função para calcular as probabilidades boot ##### #
n1=0 ; conta=0
# ##### Calculo dos pesos ##### #
calcpeso=function(vfi,ci,n)
{
  # ci:vetor de constantes de afinação
  # vfi: valores da função influencia
  # n: tamanho amostral
w=c(rep(0,n))
for (j in 1:n)
  {
aux=abs(vfi[j])
if (aux>=abs(ci[j])) w[j]=abs(aux)
if (aux<abs(ci[j])) w[j]=abs(ci[j])
  }
  p=w/sum(w)
return(p)
}

```

```

##### Função para selecionar o conjunto de dados reamostrados##### #
gera_amo=function(n,delta,alfa,beta,gama,a,b)
{
y=matrix(0,n,2)
x=seq(rep(1,n))
for (i in 1:n)
{
u=runif(1)
if (u>=delta) res=rnorm(1,0)
if (u<delta) res=rbeta(1,a,b)
res=rnorm(1)

y[i,1]=alfa/(1+exp(beta-(gama*x[i]))) + res      # logístico
##### y[1,1]=alfa*(exp(-exp(-beta*(x-gama)))) + res  # Gompertz ##### #

y[i,2]=x[i]
}
return(am=y)
}
dad_ream=function(naux,indboot,dados)
{
dr=matrix(0,1,2)
for (k in 1:n)
{
ind=indboot[k]

```

```
for (u in 1:naux)
{
if (ind==dados[u,1]) dr=rbind(dr,dados[u,])
}
}
return(dr)
}

# ##### INICIO DO PROGRAMA DE SIMULAÇÃO MONTE CARLO ##### #
# ##### Valores paramétricos ##### #

pdelta=0.05
palfa=1
pbeta=1
pgama=1.5
a=6 # ##### Par. da beta ### #
  b=6 # ##### Par. da Beta ### #
  n=50
sim=5000
nb=500
contal=0
contab=0

# ##### #

mres=matrix(0,sim,6)
```

```

for (s in 1:sim)
{
  amostra=gera_amo(n,pdelta,palfa,pbeta,pgama,a,b)
  y1= amostra[,1]
  x=amostra[,2]

  ##### Ajuste do modelo logístico ##### #
  mlog<-
  try(nls(y1~alfal/(1+exp(betal-gamal*x)),start=list(alfal=0.5,betal=1,gamal=0.2),
  trace=FALSE))

  ##### Ajuste do modelo gompertz ##### #
  #mlog<-
  try(nls(y1~alfal*(exp(-exp(-betal*(x-gamal))))),
  start=list(alfal=0.5,betal=1,gamal=0.2),trace=FALSE))

  if (!is.null(names(mlog)))
  {
    mcl=summary(mlog)
    resul_log=as.matrix(mcl$coefficients)
    contal=contal+1
    Xlog<- attr(mlog$m$fitted(), "gradient")
    linlog<- lm(y1~-1+Xlog)

    rstl=rstudent(linlog)      # ### residuos est. ##### #
    covl=covratio(linlog)    # ### cov. dos par. ##### #
  }
}

```

```

#fitsl=dffits(linlog)      # ### Inf. val. pred. ### #
#cookl=cooks.distance(linlog) # ### Inf. dist. Cook ## #

# ##### Calculo dos pesos ##### #
plog=calcpeso(covl,rstl,n)
plog=1-plog

# ##### Vieses Monte Carlo ##### #
mres[s,1]=(resul_log[1,1])
mres[s,2]=(resul_log[2,1])
mres[s,3]=(resul_log[3,1])

# ##### Inicio do bootstrap ##### #
vparb=matrix(0,nb,4)
for (b in 1:nb)
{
dadl=sample(yl, replace=TRUE, prob=plog)
indbl=sort(as.matrix(dadl))
reaml=dad_ream(n,indbl,amostra)
reaml=reaml[2:n,]

# ##### Estimar os parâmetros do modelo ##### #
nllb<-
try(nls(reaml[,1]~ab*(exp(-exp(-bb*(reaml[,2]-gb))))),
start=list(ab=0.5,bb=1,gb=0.2),trace=FALSE))
if (!is.null(names(nllb)))

```

```

{
mcbl=summary(nllb)
resbootl=as.matrix(mcbl$coefficients)
}
}
# ##### Média das estimativas bootstrap ##### #
mres[s,4]=resbootl[1,1]
mres[s,5]=resbootl[2,1]
mres[s,6]=resbootl[3,1]
}
}

# ##### Resultados que convergiram em boot e Monte Carlo ##### #
res=matrix(0,1,6)
res_f=matrix(0,1,6)
for (h in 1:nrow(mres))
{
if (mres[h,1]!=0.000) res=rbind(res,mres[h,])
}
res=res[2:nrow(res),]
for (h in 1:nrow(res))
{
ampmc=max(res[h,1:3])-min(res[h,1:3])
ampb=max(res[h,4:6])-min(res[h,4:6])

if (ampmc<=ampb) res_f=rbind(res_f,res[h,])
}
}

```

```

}
res_f[2:nrow(res_f),]

# ##### Vieses ##### #

vmca=(mean(res_f[,1])-palfa)/palfa# ## Vies Monte Carlo ## #
vm_cor_a=abs(((mean(res_f[,4])-vmca)-palfa)/palfa)

vmcb=(mean(res_f[,2])-pbeta)/pbeta# ## Vies Monte Carlo ## #
vm_cor_b=abs(((mean(res_f[,5])-vmcb)-pbeta)/pbeta)

vmcg=(mean(res_f[,3])-pgama)/pgama# ## Vies Monte Carlo ## #
vm_cor_g=abs(((mean(res_f[,6])-vmcg)-pgama)/pgama)

# ##### #
# #####Precisão ##### #

pr_mc_a=sd(res_f[,1])
pr_boot_a=sd(res_f[,4])

pr_mc_b=sd(res_f[,2])
pr_boot_b=sd(res_f[,5])

pr_mc_g=sd(res_f[,3])
pr_boot_g=sd(res_f[,6])
# ##### #

```

```
amos_val=nrow(res_f)
```

```
vmca
```

```
vm_cor_a
```

```
pr_mc_a
```

```
pr_boot_a
```

```
vmcb
```

```
vm_cor_b
```

```
pr_mc_b
```

```
pr_boot_b
```

```
vmcg
```

```
vm_cor_g
```

```
pr_mc_g
```

```
pr_boot_g
```

```
amos_val
```