



IURI DOS SANTOS MANOEL

**USO DE MÉTODOS ROBUSTOS PARA TRATAMENTO DE
OUTLIERS NA ESTIMAÇÃO DE PARÂMETROS DO
MODELO LOGÍSTICO**

LAVRAS – MG

2025

IURI DOS SANTOS MANOEL

**USO DE MÉTODOS ROBUSTOS PARA TRATAMENTO DE OUTLIERS NA
ESTIMAÇÃO DE PARÂMETROS DO MODELO LOGÍSTICO**

Tese apresentada à Universidade Federal de Lavras,
como parte das exigências do Programa de Pós-
Graduação em Estatística e Experimentação
Agropecuária, na área de Ciências Agrárias, para a
obtenção do título de Doutor.

Prof. Dr. Tales Jesus Fernandes
Orientador

LAVRAS – MG

2025

**Ficha Catalográfica elaborada pelo Sistema de Geração
de Ficha Catalográfica da Biblioteca Universitária da UFLA, com
dados informados pelo(a) próprio(a) autor(a).**

Manoel, Iuri dos Santos.

Uso de métodos robustos para tratamento de outliers na estimação de parâmetros
do modelo logístico / Iuri dos Santos Manoel. - 2025.

98 p. : il.

Orientador: Tales Jesus Fernandes

Tese (Doutorado) - Universidade Federal de Lavras, 2025.

Bibliografia.

1. Métodos robustos. 2. Jackknife. 3. Least trimmed squares. 4. Curvas de
crescimento. 5. Cunicultura. I. Fernandes, Tales Jesus . II. Universidade Federal de
Lavras. III. Título.

IURI DOS SANTOS MANOEL

**USO DE MÉTODOS ROBUSTOS PARA TRATAMENTO DE OUTLIERS NA
ESTIMAÇÃO DE PARÂMETROS DO MODELO LOGÍSTICO**

**USE OF ROBUST METHODS FOR OUTLIER TREATMENT IN THE ESTIMATION
OF PARAMETERS OF THE LOGISTIC MODEL**

Tese apresentada à Universidade Federal de Lavras,
como parte das exigências do Programa de Pós-
Graduação em Estatística e Experimentação
Agropecuária, na área de Ciências Agrárias, para a
obtenção do título de Doutor.

Aprovada em 6 de Março de 2025.

Prof. Dr. Marcelo Ângelo Cirillo UFLA

Prof. Dr. Joel Augusto Muniz UFLA

Prof. Dr. Tadeu Vilela de Souza IFSUL de Minas

Prof. Dr. Rodrigo Ferreira de Abreu IFSUL de Minas

Prof. Dr. Tales Jesus Fernandes
Orientador

**LAVRAS – MG
2025**

*Dedico essa tese a todos familiares, companheira, amigos, orientador e todos que fizeram
parte na trajetória desse doutorado.*

AGRADECIMENTOS

Agradeço aos meus pais, Francisco e Gemini, ao meu irmão, Jonathan, à minha companheira, Daniela, por todo apoio, compreensão e amor durante essa jornada.

Agradeço ao meu orientador, Tales, por toda sabedoria, orientação e paciência.

Agradeço aos membros da banca por fazerem parte dessa etapa especial da minha carreira acadêmica.

Agradeço à Universidade Federal de Lavras (UFLA), em especial ao Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, por todo acolhimento e sabedoria.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo fomento e incentivo, os quais me possibilitaram concluir meu doutorado e esta tese.

"Sábio é aquele que conhece os limites da própria ignorância." (Sócrates de Atena)

RESUMO

Em análise de dados é comum utilizar modelos estatísticos para descrever a relação entre as variáveis em estudo. Entretanto, quando há outliers no conjunto amostral, as estimativas dos parâmetros do modelo podem ser tendenciosas. Desse modo, a detecção e tratamento adequado de outliers é imprescindível para garantir resultados representativos. Neste contexto, métodos robustos podem ser uma solução para lidar com outliers, pois podem mitigar os efeitos que esses valores têm em estimativas de interesse. Entre a diversidade de metodologias robustas difundidas na literatura, destaca-se os métodos robustos Jackknife e Least Trimmed Squares (LTS). Esses métodos são eficientes e versáteis na seleção de subconjuntos amostrais com menos outliers que o conjunto amostral. Isso permite selecionar uma proporção da amostra que a melhor representa. Diante do exposto, este trabalho tem como objetivo propor metodologias robustas inovadoras aplicadas na descrição de curvas de crescimento. Esses métodos estão apresentados em três capítulos em que a proposta é obter um ajuste do modelo não linear Logístico que melhor representa o conjunto de dados. No primeiro capítulo, o método de reamostragem Jackknife é implementado ao considerar estimativas para o coeficiente de curtose como critério de seleção. Já no segundo capítulo, o método LTS é implementado ao considerar partições da amostra para selecionar valores de acordo com a soma de quadrados dos desvios (SQD). Ressalta-se que nesses dois primeiros capítulos as metodologias propostas são aplicadas na descrição de um conjunto de dados simulados via Monte Carlo. Por fim, no terceiro capítulo ambos os métodos são aplicados na descrição de um conjunto de dados reais. Os dados são de cunicultura oriundos do Departamento de Zootecnia da Universidade Federal de Lavras. As medidas observadas são o peso e o comprimento da pata dianteira dos coelhos. Para a implementação das metodologias, é considerado um ponto de ruptura como critério para determinar o tamanho dos subconjuntos selecionados para o ajuste. Além disso, a comparação entre ajustes é feita pelo desvio médio absoluto (DMA). As metodologias propostas neste trabalho são promissoras em ajustes do modelo Logístico na descrição de curvas crescimento, sendo que o método LTS apresentou melhores resultados do que o método Jackknife. O ponto de ruptura e o DMA são apropriados como critérios para determinar o tamanho do subconjunto amostral e comparação entre os ajustes, respectivamente.

Palavras-chave: métodos robustos; jackknife; least trimmed squares; curvas de crescimento; cunicultura.

ABSTRACT

In data analysis, it is common to use statistical models to describe the relationship between the variables under study. However, when there are outliers in the sample set, the estimates of the model parameters may be biased. Therefore, the detection and appropriate treatment of outliers is essential to ensure representative results. In this context, robust methods can be a solution for dealing with outliers, as they can mitigate the effects that these values have on estimates of interest. Among the diversity of robust methodologies disseminated in the literature, the robust methods Jackknife and Least Trimmed Squares (LTS) stand out. These methods are efficient and versatile in selecting sample subsets with fewer outliers than the sample set. This allows selecting a proportion of the sample that best represents it. In view of the above, this work aims to propose innovative robust methodologies applied to the description of growth curves. These methods are presented in three chapters in which the proposal is to obtain an adjustment of the nonlinear logistic model that best represents the data set. In the first chapter, the Jackknife resampling method is implemented by considering estimates for the kurtosis coefficient as a selection criterion. In the second chapter, the LTS method is implemented by considering sample partitions to select values according to the sum of squared deviations (SSD). It is worth noting that in these first two chapters the proposed methodologies are applied to the description of a set of simulated data via Monte Carlo. Finally, in the third chapter both methods are applied to the description of a set of real data. The data are from rabbit farming and come from the Department of Animal Science of the Federal University of Lavras. The observed measurements are the weight and length of the front paw of the rabbits. For the implementation of the methodologies, a breaking point is considered as a criterion to determine the size of the subsets selected for the adjustment. In addition, the comparison between adjustments is made by the mean absolute deviation (MAD). The methodologies proposed in this work are promising in adjusting the Logistic model in the description of growth curves, with the LTS method presenting better results than the Jackknife method. The break point and the MAD are appropriate as criteria to determine the size of the sample subset and comparison between the adjustments, respectively.

Keywords: robust methods; jackknife; least trimmed squares; growth curves; rabbit farming.

LISTA DE FIGURAS

Figura 2.1 – Exemplo de curva do modelo Logístico.....	22
Figura 2.1 – Gráfico de dispersão do conjunto de dados simulados.....	40
Figura 2.2 – Processo de seleção dos subconjuntos reamostrados pelo método de reamostragem Jackknife, de acordo com o critério da menor curtose, para estimação dos parâmetros do modelo não linear Logístico.....	43
Figura 3.1 – Gráfico do modelo não linear Logístico ajustado sobre o conjunto de dados com outliers.....	45
Figura 3.2 – Gráfico dos modelos ajustados sobre a totalidade dos dados e sobre a mediana das estimativas das reamostragens e a dispersão do conjunto de dados.....	47
Figura 3.3 – Dispersão do conjunto de dados reamostrados pelo método de reamostragem Jackknife com menor curtose.....	48
Figura 3.4 – Gráficos dos modelos Logísticos ajustados sobre a totalidade dos dados e sobre o conjunto reamostrado pelo método de reamostragem Jackknife.....	49
Figura 2.1 – Processo de simulação do conjunto de dados com valores discrepantes.....	59
Figura 2.2 – Dispersão do conjunto de dados simulados.....	60
Figura 2.3 – Processo de seleção dos arranjos pelo método LTS para estimação dos parâmetros do modelo Logístico.....	61
Figura 3.1 – Plots das curvas dos ajustes do modelo não linear Logístico convencional e robusto na descrição do conjunto de dados.....	64
Figura 3.2 – Intervalos de confiança normalizados para as estimativas dos parâmetros do modelo não linear Logístico.....	64
Figura 2.1 – Gráfico de dispersão da relação do comprimento da pata dianteira (centímetros) e a idade (dias) de coelhos mestiços.....	72
Figura 3.1 – Gráfico da curva do modelo Logístico ajustado pelo método convencional.....	76
Figura 3.2 – Gráfico de dispersão do conjunto central obtido ao implementar o método Jackknife em dados de cunicultura.....	77

Figura 3.3 – Gráfico das curvas ajustadas pelo ajuste convencional e ajuste robusto Jackknife na descrição dos dados de cunicultura.....	78
Figura 3.4 – Gráfico de dispersão do conjunto central obtido ao implementar o método Least Trimmed Squares aos dados de cunicultura.....	79
Figura 3.5 – Gráfico das curvas ajustadas pelo ajuste convencional e ajuste robusto LTS na descrição dos dados de cunicultura.....	80
Figura 3.6 – Gráfico das curvas ajustadas pelo ajuste convencional e ajustes robustos Jackknife e LTS na descrição do CPD em função da idade de coelhos mestiços.....	81

LISTA DE TABELAS

Tabela 2.1 – Exemplo de implementação do método Jackknife.....	26
Tabela 2.2 – Exemplo de implementação do método Least Trimmed Squares.....	28
Tabela 2.1 – Subconjuntos Jackknife do exemplo didático.....	41
Tabela 2.2 – Representatividade de 1 valor diante de conjuntos de dados com tamanhos diferentes.....	41
Tabela 3.1 – Estimativas dos parâmetros para o ajuste do modelo não linear Logístico sobre a totalidade dos dados.....	44
Tabela 3.2 – Estimativas dos parâmetros do modelo não linear Logístico obtidas pela mediana das estimativas dos parâmetros de reamostragens como 151 valores.....	46
Tabela 3.3 – Estimativas para os parâmetros do modelo não linear Logístico ajustado ao subconjunto reamostrado pelo método Jackknife.....	48
Tabela 3.1 – Estimativas para os parâmetros dos ajustes do modelo não linear Logístico convencional e robusto.....	64
Tabela 3.1 – Estimativas para os parâmetros do modelo Logístico e DMA associado ao ajuste convencional.....	76
Tabela 3.2 – Estimativas para os parâmetros do modelo Logístico convencional e os DMA associados aos ajustes convencional e robusto Jackknife.....	78
Tabela 3.3 – Estimativas para os parâmetros do modelo Logístico e os DMA associados aos ajustes convencional e robusto LTS.....	79
Tabela 3.4 – Estimativas para os parâmetros do modelo Logístico e os DMA associados aos ajustes convencional, Jackknife e LTS.....	80

SUMÁRIO

	PRIMEIRA PARTE	16
1	INTRODUÇÃO	17
2	REFERENCIAL TEÓRICO	19
2.1	Ajustes de modelos	19
2.1.1	Método de Mínimos Quadrados Ordinários (MQO)	19
2.1.2	Linearidade de um modelo	20
2.2	Modelos não lineares	21
2.3	Modelo Logístico	21
2.4	Outliers	23
2.5	Estimadores robustos	23
2.5.1	Ponto de ruptura	24
2.5.2	Jackknife	25
2.5.3	Least Trimmed Squares (LTS)	27
2.6	Simulação de Monte Carlo	29
2.7	Desvio médio absoluto (DMA)	30
	REFERÊNCIAS	31
	SEGUNDA PARTE	34
	ARTIGO 1 – MÉTODO JACKKNIFE EM AJUSTES DE MODELOS NÃO LINEARES EM PRESENÇA DE OUTLIERS NO CONJUNTO DE DADOS	35
1	INTRODUÇÃO	38
2	MATERIAL E MÉTODOS	39
2.1	Conjunto de dados	39
2.2	Método de reamostragem Jackknife	40

2.2.1	Abordagem 1: Mediana sobre as reamostragens	42
2.2.1	Abordagem 2: Curtose sobre as reamostragens	43
2.3	Desvio médio absoluto (DMA)	43
3	RESULTADOS	44
3.1	Ajuste convencional	44
3.2	Ajustes robustos	45
3.2.1	Abordagem 1	46
3.2.2	Abordagem 2	47
4	DISCUSSÕES	49
5	CONCLUSÃO	50
	REFERÊNCIAS	52
	 ARTIGO 2 - MÉTODO ROBUSTO LEAST TRIMMED SQUARES IMPLEMENTADO EM AJUSTES DO MODELO NÃO LINEAR LOGÍSTICO	54
1	INTRODUÇÃO	57
2	MATERIAL E MÉTODOS	58
2.1	Conjunto de dados	58
2.2	Ponto de ruptura	60
2.3	Least Trimmed Squares (LTS)	60
2.4	Desvio médio absoluto (DMA)	62
3	RESULTADOS	63
4	DISCUSSÕES	65
5	CONCLUSÃO	65
	REFERÊNCIAS	66

	ARTIGO 3 - MÉTODOS ROBUSTOS IMPLEMENTADOS EM AJUSTES DO MODELO NÃO LINEAR LOGÍSTICO NA DESCRIÇÃO DE CURVAS DE CRESCIMENTO	68
1	INTRODUÇÃO	71
2	MATERIAL E MÉTODOS	72
2.1	Conjunto de dados	72
2.2	Método de reamostragem Jackknife	73
2.3	Método robusto Least Trimmed Squares	74
2.4	Desvio médio absoluto (DMA)	75
3	RESULTADOS	75
3.1	Algumas considerações	75
3.2	Ajuste convencional	76
3.3	Ajuste robusto Jackknife	77
3.4	Ajuste robusto LTS	78
3.5	Ajuste robustos	80
4	DISCUSSÃO	82
5	CONCLUSÃO	83
	REFERÊNCIAS	85
	TERCEIRA PARTE	86
	CONSIDERAÇÕES FINAIS	87
	APÊNCICE A — Script utilizado no Artigo 1	88
	APÊNCICE B — Script utilizado no Artigo 2	92
	APÊNCICE C — Script utilizado no Artigo 3	96

PRIMEIRA PARTE

1 INTRODUÇÃO

Outliers são valores que se desviam significativamente do padrão geral das demais observações em um conjunto de dados. Esses valores podem surgir por diversas razões, como erros de medição, erros na entrada de dados, comportamentos extremos ou eventos raros (ROUSSEEUW; HUBERT, 2011; AGGARWAL; AGGARWAL, 2015). A identificação e o tratamento adequado desses valores são essenciais em análises de dados, pois outliers podem distorcer estimativas e comprometer a precisão de resultados (HEKIMOGLU et al., 2015).

Neste contexto, uma opção válida para lidar com outliers são os métodos robustos (ROUSSEEUW; HUBERT, 2018). Essas metodologias estatísticas são menos sensíveis a distorções que outliers podem causar nas estimativas de interesse. Dessa forma, os efeitos que outliers podem ter nos resultados são mitigados, proporcionando estimativas robustas e representativas (MANOEL et al., 2024a). Entre as diversas metodologias robustas, destacam-se por sua versatilidade e simplicidade de implementação os métodos de reamostragem Jackknife e o Least Trimmed Squares (LTS) (ROUSSEEUW; DRIESSEN, 2006; MANOEL et al., 2024a).

O método Jackknife considera reamostragens do conjunto amostral para obter estimativas de parâmetros (MILLER, 1974; FOX; HINKLEY; LARNTZ, 1980; YANG; KOLASSA, 2021). Isso possibilita que o método seja versátil e eficiente, sendo implementado em diversos contextos. Beyaztas e Alin (2013) detectam outliers em conjuntos de dados ao implementar o método de reamostragem Jackknife de modo a melhorar o ajuste de modelos de regressão linear. Wang e Yu (2020) introduzem o método como alternativa viável para melhorar os resultados de estimações do método de mínimos quadrados totais ponderados. Manoel et al. (2024a) implementam o método Jackknife em ajustes do modelo não linear Logístico na descrição de curvas de crescimento.

Por outro lado, o método robusto LTS seleciona um subconjunto da amostra que apresenta menor soma de quadrados dos desvios (SQD) (ROUSSEEUW; DRIESSEN, 2006; MANOEL et al., 2024b). Desse modo, valores que aumentam a SQD em relação a média dos dados não são selecionados. Isso possibilita que as estimativas obtidas a partir desse subconjunto sejam menos influenciadas por outliers. Assim, o método LTS pode ser implementado em diversos contextos. Gu et al. (2020) implementam o método LTS com o método de mínimos quadrados totais em ajustes de medição. O LTS também é utilizado por Xie et al. (2022) em regressões de mínimos quadrados parciais. Manoel et al. (2024b)

implementam o LTS em regressão linear adaptativa na estimação de parâmetros para um modelo multinível.

Diante do exposto, este trabalho tem como objetivo apresentar metodologias robustas aplicadas à descrição de curvas de crescimento. O modelo considerado é o não linear Logístico, pois possui bons ajustes e parâmetros interpretáveis (FERNANDES et al., 2017). Além disso, o ponto de ruptura (ROUSSEEUW; DRIESSEN, 2006) é utilizado como critério de determinação para o tamanho dos subconjuntos considerados para os ajustes do modelo. Como critério de comparação entre os ajustes, é considerado o desvio médio absoluto (DMA) (GORARD, 2015). Dessa forma, esta tese é composta por três partes que apresentam essas metodologias.

O primeiro capítulo apresenta uma nova implementação para o método de reamostragem Jackknife. Ressalta-se que é considerada uma adaptação desse método, sendo considerado exclusivamente para reamostragem, sem a inclusão de pseudo-valores. É utilizado o coeficiente de curtose como critério de seleção para os subconjuntos. Ao considerar o subconjunto que apresenta menor estimativa para o coeficiente de curtose, outliers não são selecionados. Desse modo, obtém-se um ajuste robusto e representativo para o modelo não linear Logístico. Os dados utilizados são simulados via Monte Carlo.

No segundo capítulo, o método robusto LTS é implementado ao considerar partições no conjunto de dados. Assim, a seleção que o método executa em relação a SQD, é aplicada em partes do conjunto de dados. Isso garante que outliers sejam identificados e excluídos. Dessa forma, o ajuste do modelo não linear Logístico apresenta robustez e representatividade na descrição da curva de crescimento. Os dados utilizados também são simulados via Monte Carlo.

Por fim, no terceiro capítulo os métodos apresentados nos dois capítulos precedentes são aplicados na descrição de um conjunto de dados reais. Os dados são de cunicultura, oriundos do Departamento de Zootecnia da Universidade Federal de Lavras. As variáveis analisadas, visando a implementação das metodologias propostas neste trabalho, são o peso e o comprimento da pata dianteira de coelhos mestiços. Ambas as metodologias robustas propostas nesta tese são promissoras no ajuste do modelo Logístico na descrição de curvas de crescimento.

2 REFERENCIAL TEÓRICO

Em análise de dados, quaisquer características observáveis de uma amostra são denominadas como variáveis (FERREIRA, 2009). Desse modo, para explicar um evento de interesse, pode-se estabelecer relações entre variáveis de causa e efeito. A causa, pode ser interpretada como variáveis independentes e o efeito como variáveis dependentes. Assim, a partir de uma amostra do evento, o ajuste de um modelo estabelece uma relação entre variáveis dependentes e independentes (BUSSAB; MORETTIN, 2010). Isto pode ser feito de muitas formas, a depender da natureza dos dados.

2.1 Ajustes de Modelos

De modo geral, uma amostra de tamanho n oriunda do evento de interesse pode ser explicada por um modelo de regressão da forma

$$Y = F(X, \beta) + \epsilon, \quad (2.1)$$

em que Y_i é a variável dependente, X_i a variável independente, β se refere aos parâmetros do modelo, $f(\cdot)$ a função matemática que relaciona as variáveis e ϵ_i o desvio associado a i -ésima observação, com $i \in \mathbb{N} \mid 1 \leq i \leq n$.

Neste sentido, tem-se como objetivo obter o vetor de estimativas dos parâmetros $\vec{\beta}$. Geralmente, utiliza-se o método de Mínimos Quadrados Ordinários (MQO) com o propósito de minimizar a soma de quadrados dos desvios (ROUSSEEUW; LEROY, 2005).

2.1.1 Método de Mínimos Quadrados Ordinários (MQO)

Considere a equação matricial de uma regressão descrita de acordo com a Equação 2.1, em que $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $F(X, \beta) = \begin{pmatrix} f(X_1, \beta) \\ \vdots \\ f(X_n, \beta) \end{pmatrix}$ e $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$. De 2.1, tem-se que o erro pode ser escrito como $\epsilon = Y - F(X, \beta)$. Assim, o quadrado dos desvios pode ser obtido por

$$\begin{aligned} \epsilon^T \epsilon &= [Y - F(X, \beta)]^T [Y - F(X, \beta)] \\ \Rightarrow \epsilon^T \epsilon &= Y^T Y - 2Y^T F(X, \beta) + F(X, \beta) F(X, \beta)^T. \end{aligned} \quad (2.2)$$

Assim, como o objetivo do método de MQO é minimizar a soma de quadrados dos desvios, tem-se que esta minimização pode ser feita pela derivada parcial de 2.2 em relação a β . Dessa forma, tem-se

$$\begin{aligned}\frac{\partial(\epsilon^T \epsilon)}{\partial \beta} &= \frac{\partial[Y^T Y - 2Y^T F(X, \beta) + F(X, \beta)F(X, \beta)^T]}{\partial \beta} \\ \Rightarrow \frac{\partial(\epsilon^T \epsilon)}{\partial \beta} &= -2Y^T \times \frac{\partial F(X, \beta)}{\partial \beta} + \frac{\partial F^T(X, \beta)}{\partial \beta} \times F(X, \beta) + F^T(X, \beta) \times \frac{\partial F(X, \beta)}{\partial \beta}. \\ \Rightarrow \frac{\partial(\epsilon^T \epsilon)}{\partial \beta} &= -2Y^T \times \frac{\partial F(X, \beta)}{\partial \beta} + \frac{\partial F^T(X, \beta)}{\partial \beta} \times F(X, \beta) + F^T(X, \beta) \times \frac{\partial F(X, \beta)}{\partial \beta}. \quad (2.3)\end{aligned}$$

Da Equação 2.3, ressalta-se que $\frac{\partial F(X, \beta)}{\partial \beta}$ é a matriz jacobiana cuja ordem é $n \times p$, em que p é a quantidade de parâmetros do modelo e n o tamanho amostral. E ainda, fazendo $J = \frac{\partial F(X, \beta)}{\partial \beta}$, tem-se

$$\begin{aligned}\frac{\partial(\epsilon^T \epsilon)}{\partial \beta} &= -2Y^T \times J + J^T \times F(X, \beta) + F^T(X, \beta) \times J = -2Y^T J + J^T F(X, \beta) + [J^T F(X, \beta)]^T \\ \Rightarrow \frac{\partial(\epsilon^T \epsilon)}{\partial \beta} &= -2Y^T J + 2J^T F^T(X, \beta). \quad (2.4)\end{aligned}$$

Por fim, igualando 2.4 a zero, obtém-se um ponto crítico, isto é, a minimização dos quadrados dos desvios (RUGGIERO; LOPES, 1996)

$$\begin{aligned}-2Y^T J + 2J^T F^T(X, \beta) &= 0 \Rightarrow F^T(X, \beta)J = Y^T J \Rightarrow [J^T F(X, \beta)]^T = [J^T Y]^T \\ \therefore J^T F(X, \beta) &= J^T Y. \quad (2.5)\end{aligned}$$

A equação (2.5) denota o sistema de equações normais (SEN).

2.1.2 Linearidade de um modelo

De acordo com Bates e Watts (1980), dado um modelo da forma $Y = F(X, \beta) + \epsilon$, tem-se que a linearidade deste pode ser definida pelas derivadas de $F(X, \beta)$ em relação ao parâmetro β . Um modelo é definido como não linear se pelo menos uma das derivadas parciais de $F(\cdot)$ em relação a β depender de algum parâmetro.

Assim, modelos estatísticos podem ser classificados em três categorias: linear, não linear e linearizável (DRAPER; SMITH, 1998). Logo, há três situações possíveis:

- a) se $\frac{\partial F(X, \beta)}{\partial \beta} = G(X)$, então o modelo é linear;
- b) se $\frac{\partial F(X, \beta)}{\partial \beta} = G(X, \beta)$, então o modelo é não linear;
- c) se o modelo é não linear, mas se torne linear por transformações algébricas em $f(\cdot)$, então o modelo é linearizável.

A linearidade se relaciona diretamente com o vetor de estimativas $\vec{\beta}$ pois, se linear, o SEN resultante da equação 2.5 tem solução analítica. Caso contrário, não há forma fechada para resolução do mesmo, tornando necessário o uso de algoritmos iterativos para se obter as estimativas, como o de Gauss-Newton, Marquardt e Gradiente (RUGGIERO; LOPES, 1996; FERNANDES et al., 2017).

2.2 Modelos Não Lineares

Se tratando de modelos com poucos parâmetros, os não lineares têm destaque na descrição de eventos nas mais diversas áreas de pesquisa. Seus parâmetros apresentam informações úteis para melhor interpretabilidade prática dos dados, salientando o uso destes em curvas de crescimento (FERNANDES et al., 2017).

Entre as famílias de modelos não lineares, a Richards é mais encontrada na literatura, abrangendo os modelos de Gompertz, Logístico, Brody e Von Bertalanffy (BATES; CHAMBERS, 2017). O modelo Logístico é o mais difundido na literatura, pois apresenta bons ajustes e parâmetros interpretáveis (MANOEL et al., 2024a). Desse modo, neste trabalho é considerado o modelo não linear Logístico para a implementação das metodologias robustas propostas.

2.3 Modelo Logístico

Na intenção de se estudar o crescimento de populações humanas, Verhulst propôs a função logística em 1838 (SEBER; WILD, 1989). Por ser muito utilizada, a função apresenta diferentes parametrizações, sendo a que proporciona melhor desempenho, de acordo com Fernandes et al. (2017), é apresentada no formato

$$y_i = \frac{\alpha}{1 + e^{\kappa(\beta - x_i)}} + \epsilon_i \quad (2.6)$$

em que y_i e x_i são as i -ésimas observações das variáveis dependente e independente, respectivamente, e ϵ_i o desvio associado a i -ésima observação no qual se pressupõe que $\epsilon \sim N(0, \sigma^2)$.

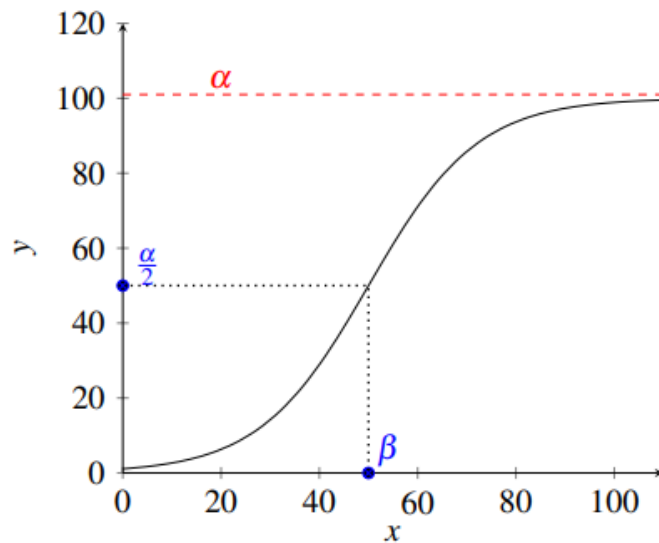
Na Equação 2.6, destacam-se os parâmetros do modelo logístico, sendo que suas interpretações práticas são: α é o valor da assíntota horizontal superior; β é a abscissa do ponto de inflexão, indicando que a característica em estudo passa de um crescimento acelerado para um crescimento desacelerado; κ é um índice associado ao crescimento da variável dependente.

Ressalta-se ainda que o modelo logístico tem algumas características importantes. Primeiramente, a função é limitada verticalmente, tendo $y = \alpha$ como assíntota horizontal superior, pois

$$\begin{aligned} \lim_{x \rightarrow \infty} \left[\frac{\alpha}{1 + e^{\kappa(\beta - x)}} \right] &= \lim_{x \rightarrow \infty} \left[\frac{\alpha}{1 + e^{-\kappa(x - \beta)}} \right] = \lim_{x \rightarrow \infty} \left[\frac{\alpha}{1 + \frac{1}{e^{\kappa(x - \beta)}}} \right] = \frac{\alpha}{1 + \frac{1}{e^{\kappa(\infty - \beta)}}} \\ &= \frac{\alpha}{1 + \frac{1}{e^\infty}} = \frac{\alpha}{1 + 0} = \alpha. \end{aligned}$$

Além do mais, a função é estritamente crescente e sem pontos críticos, além de apresentar simetria em relação ao ponto de inflexão. Graficamente, tem-se a representação de um modelo logístico e seus parâmetros α e β na Figura 2.1.

Figura 2.1 – Exemplo de curva do modelo Logístico.



Fonte: Do autor (2025).

2.4 Outliers

Outliers são frequentemente considerados como observações que são excepcionalmente altas ou extraordinariamente baixas em comparação com a maioria dos valores de um conjunto de dados (ROUSSEEUW; HUBERT, 2011). Essa discrepância sugere que a observação tenha sido gerada por outro evento (HAWKINS, 1980; ROUSSEEUW; HUBERT, 2018). Assim, outliers podem distorcer as estimativas obtidas em análises estatísticas e afetar a interpretação dos resultados (MANOEL et al., 2024a).

Em uma amostra é comum que ocorram outliers, pois experimentos de campo são suscetíveis a fatores aleatórios não controláveis, como chuvas, pragas, doenças, erros de mensuração, mal planejamento experimental, entre outros (ROUSSEEUW; HUBERT, 2018). No entanto, outliers não são sempre oriundos de erros, podendo assim ser a representatividade de um fenômeno no evento. Por exemplo, um animal prenho tende a apresentar maior peso, assim como um animal doente tende a apresentar menor peso.

Desse modo, antes de o pesquisador procurar uma solução para tratar uma amostra com outliers, deve entender a origem dos mesmos. Assim, depois de contextualizar esses valores, deve-se considerar alternativas que lidam com o problema de modo que mitigue a influência de outliers nas estimativas dos parâmetros de interesse. Ressalta-se que a decisão simplista de excluir esses valores da amostra sem nenhum critério formal de seleção pode comprometer a capacidade de generalização dos resultados (FARCOMENI; VENTURA, 2010). Diante do exposto, é recomendado o uso de metodologias robustas (ROUSSEEUW; HUBERT, 2018; MANOEL et al., 2024a).

2.5 Estimadores Robustos

Métodos robustos são técnicas estatísticas desenvolvidas para serem menos sensíveis a outliers no cálculo de estimativas (ROUSSEEUW; DRIESSEN, 2006). Essa robustez pode ser obtida ao reduzir a influência dos outliers em estimativas, ou ainda na eliminação desses valores no conjunto de dados. A mediana, por exemplo, é uma estimativa robusta para a posição central de uma amostra, pois ela não é afetada por valores extremos. Por outro lado, a adaptação do método de reamostragem Jackknife, apresentada por Manoel et al. (2024a), é robusta ao eliminar outliers do conjunto de dados. Ambos os métodos supracitados são robustos, pois mitigam os efeitos dos outliers em estimativas de interesse.

Diante do exposto, os métodos robustos são uma alternativa válida para analisar amostras possivelmente contaminadas por outliers, pois estes podem identificar, substituir ou eliminar esses valores (ROUSSEEUW; HUBERT, 2018). Deste modo, o conjunto de dados pode ser alterado tal que métodos convencionais de estimação forneçam boa predição ao modelo quando aplicados, denominando-os assim como estimadores robustos (MANOEL et al., 2024a).

Neste trabalho, são considerados os métodos robustos Jackknife e Least Trimmed Squares (LTS). Ambos os métodos eliminam outliers do conjunto de dados ao considerar uma subamostra. Isso permite o ajuste de modelos menos influenciados e, portanto, mais representativos. Nas seções a seguir, são apresentados alguns conceitos relevantes para a implementação das metodologias robustas supracitadas.

2.5.1 Ponto de ruptura

O ponto de ruptura é um método que mensura a robustez de um estimador robusto. A partir do tamanho amostral, é determinada a quantidade máxima de outliers que o estimador suporta tal que forneça informações confiáveis sobre o parâmetro em questão (ROUSSEEUW; DRIESSEN, 2006). Desse modo, ao considerar métodos robustos como Jackknife e LTS é possível definir quantas observações podem ser eliminadas da amostra (MANOEL et al., 2024a).

Seja $Z = (X, Y)$ um conjunto de dados e T um estimador. Quando T é aplicado a amostra Z , tem-se como resultado um vetor de parâmetros $\vec{\beta}$, isto é, $T(Z) = \vec{\beta}$, em que $\vec{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ para um modelo com p parâmetros. Neste sentido, definindo Z^* como um conjunto de tamanho n composto por $n - m$ elementos de Z e m valores possivelmente discrepantes, em que n é a cardinalidade de Z e $1 \leq m \leq n$, pode-se definir uma fração de contaminação que pode ser descrita por $\delta = \frac{m}{n}$.

A diferença entre o valor esperado e o verdadeiro, o vício, pode ser escrito por $b(m; T, Z)$. Logo, o vício máximo que uma contaminação pode causar é dada por pela Equação 2.7.

$$b(m; T, Z) = \sup ||T(Z^*) - T(Z)||. \quad (2.7)$$

Desta forma, se o vício for finito, então os valores discrepantes têm um efeito arbitrário no estimador T que perde a robustez a partir de uma certa quantidade desses valores, definindo

o ponto de ruptura de T . Na prática, o ponto de ruptura, definido por Rousseeuw e Driessen (2006) é descrito por

$$h \geq \frac{3n + p + 1}{4}, \quad (2.8)$$

em que h é a quantidade de observações necessárias para explicar o evento, n o tamanho amostral e p a quantidade de parâmetros do modelo considerado. Assim, o valor de h pode ser definido pelo pesquisador, desde que satisfaça a Equação 2.8. Isso permite preservar a amostra e reduzir a perda de informações. Desse modo, métodos robustos como Jackknife e LTS podem ser implementados de modo a não descaracterizar o conjunto de dados.

2.5.2 Jackknife

O método Jackknife é uma estatística não-paramétrica introduzido por Maurice Quenouille em 1949 e posteriormente desenvolvido por John Tukey em 1958 (MILLER, 1974; FOX; HINKLEY; LARNTZ, 1980). Essa metodologia se baseia em reamostragens feitas a partir da amostra (GOMES, 1994; BEYAZTAS; ALIN, 2013). Cada reamostragem é sem reposição, excluindo uma ou mais observações, de acordo com um critério pré-estabelecido. A seleção dos valores é feita aleatoriamente e de modo iterativo (WANG; YU, 2020; MANOEL et al., 2024a).

Seja θ o parâmetro de interesse de uma amostra da forma $Y = \{y_1, \dots, y_n\}$, com $i = 1, \dots, n$, e v a quantidade de valores que são excluídos a cada iteração. Assim, a cada reamostragem são introduzidos uma quantidade n^* de pseudo-valores de modo a preservar o tamanho amostral. Os pseudo-valores podem ser obtidos pela Equação 2.9 (FOX; HINKLEY; LARNTZ, 1980).

$$y^* = n\hat{\theta} - (n - n^*)\hat{\theta}_{-n^*}, \quad (2.9)$$

em que $\hat{\theta}$ é a estimativa para o parâmetro de interesse obtido ao considerar a amostra e $\hat{\theta}_{-n^*}$ a estimativa para o parâmetro ao considerar a reamostragem.

A princípio, o método Jackknife foi implementado para obter estimativas de um mesmo parâmetro para construir um intervalo de confiança do mesmo, permitindo o estudo do viés de uma estatística. No entanto, por sua versatilidade, esse método foi implementado em diversos contextos (MANOEL et al., 2024a). Nesta tese, é considerada uma versão adaptada dessa metodologia, em que considera reamostragens de tamanho $n - m$, com $0 < m < n - h$, h o

ponto de ruptura (Equação 2.8) e m a quantidade de valores excluídos em cada reamostragem. Além disso, não há a introdução de pseudo-valores. Isso possibilita uma abordagem robusta, como implementado por Manoel et al. (2024a), na redução dos efeitos de outliers nas estimativas dos parâmetros do modelo não linear Logístico.

No entanto, ressalta-se que para amostras relativamente grandes, o método Jackknife tende a exigir maior trabalho computacional (MANOEL et al., 2024a). Isso se dá devido ao grande número de possíveis combinações $C_{n,n-m} = \binom{n}{n-m} = \binom{n}{m}$ para as reamostragens. Assim, a escolha de um m relativamente próximo a n pode diminuir o custo computacional, tornando o método mais acessível. Porém, dessa forma o método é menos eficiente, pois elimina menos valores.

Para demonstrar a aplicação do método, considere o exemplo a seguir. Seja o conjunto amostral fictício $Y = \{1,2,5,6,9\}$ com $n = 5$ elementos. Suponha que o interesse seja obter um conjunto B de tamanho $n - m$ tal que $B \subset Y$ e B apresente menor média aritmética possível. Desse modo, o critério de seleção é a média aritmética. Definindo $m = 1$ e sabendo que $n = 5$, há $C_{5,4} = \binom{5}{4} = 5$ possíveis resultados para o conjunto B . As reamostragens e suas respectivas estimativas para a média amostral estão apresentadas na Tabela 2.1.

Tabela 2.1 – Exemplo de implementação do método Jackknife.

Iteração	Reamostragem	Estimativas
1	{1,5,6,9}	$\bar{y}_1 = 21$
2	{2,5,6,9}	$\bar{y}_2 = 22$
3	{1,2,5,9}	$\bar{y}_3 = 17$
4	{1,2,5,6}	$\bar{y}_4 = 14$
5	{1,2,6,9}	$\bar{y}_5 = 18$

Fonte: Do autor (2025).

Desse modo, de acordo com os resultados apresentados na Tabela 2.1, o conjunto B é dado por $B = \{1,2,5,6\}$. Assim, o método de reamostragem Jackknife determinou um subconjunto da amostra que apresenta menor estimativa para média. De forma análoga, essa metodologia pode ser implementada a qualquer conjunto de dados ao definir um critério de seleção, o ponto de ruptura e uma quantidade de elementos para serem eliminados (m).

Diante do exposto, entende-se que a implementação dessa metodologia de reamostragem como uma abordagem robusta é promissora. Trabalhos como os de Beyaztas e Alin (2013), Wang e Yu (2020) e Manoel et al. (2024a) são exemplos da eficiência do método Jackknife diante outliers. Em especial, destaca-se o trabalho de Manoel et al. (2024a) que implementam o método Jackknife em ajustes do modelo não linear Logístico na descrição de

curvas de crescimento. Nessa implementação, são implementadas e comparadas duas abordagens para o método (veja o capítulo 1 da segunda parte desta tese).

2.5.3 Least Trimmed Squares

O método robusto LTS consiste em determinar um subconjunto da amostra que apresenta menor estimativa para a soma de quadrados dos desvios (SQD) (ROUSSEEUW; DRIESSEN, 2006; GU et al., 2020). Para determinar esse subconjunto, o LTS compara todos os possíveis arranjos de modo iterativo. Assim, os outliers são eliminados nessa seleção, pois esses valores apresentam um maior desvio em relação a média e, conseqüentemente, aumenta a estimativa para a SQD (XIE et al., 2022). Dessa forma, o subconjunto selecionado é utilizado para o cálculo das estimativas de interesse (MANOEL et al., 2024b). A estimativa para a SQD é calculada de acordo com a Equação 2.10.

$$SQD = \sum_{i=1}^q (y_i - \bar{y})^2, \quad (2.10)$$

em que: SQD é o estimador para a soma de quadrados dos desvios, q é o tamanho do subconjunto, y_i a variável dependente selecionada para o subconjunto e $\bar{y}$ a média aritmética da amostra.

Para a implementação do método LTS, este trabalho considera o ponto de ruptura (Equação 2.8) como critério para determinar o tamanho dos subconjuntos. No entanto, assim como apresentado para o método Jackknife (subseção 2.5.2), o LTS pode exigir muito trabalho computacional quando a amostra é relativamente grande. Assim, a escolha de uma quantidade m de observações para serem eliminadas pode diminuir esse trabalho, desde que não seja superior a h , determinado pelo ponto de ruptura. Uma forma de tornar o método mais eficiente em uma perspectiva computacional é considerar partições na amostra.

Seja k a quantidade de partições a serem feitas. O valor de k é escolhido de modo que cada partição tenha uma mesma quantidade de observações tal que respeite a estrutura experimental da amostra. Logo, se A é o conjunto amostral, então $A = A_1 \cup A_2 \cup \dots \cup A_k$, em que A_j são as partições, com $j = 1, 2, \dots, k$. Desse modo, o método LTS é implementado em cada partição A_j tal que o subconjunto obtido esteja contido em cada partição. Assim, a quantidade de possíveis subconjuntos é reduzida, diminuindo o custo computacional.

Um possível problema é quando não se pode definir k tal que cada partição seja do mesmo tamanho. Nesse caso, pode-se completar a amostra com elementos neutros diante o LTS, ou seja, completar o conjunto amostral com pseudo-observações iguais a média aritmética da amostra. Desse modo, o método LTS pode ser implementado de acordo com o apresentado anteriormente.

Para ilustrar o método robusto LTS, considere o exemplo a seguir. Seja o conjunto de dados fictício $A = \{5, 8, 10, 13, 15, 16, 18, 22, 23, 28\}$ de tamanho $n = 10$. Ao definir duas partições, $k = 2$, o conjunto pode ser escrito por $A = A_1 \cup A_2$, em que $A_1 = \{5, 8, 10, 13, 15\}$ e $A_2 = \{16, 18, 22, 23, 28\}$ são as partições com cinco observações cada. Define-se ainda, $m = 1$ valores a serem eliminados de cada partição. Isso possibilita $C_{5,4} = 5$ possíveis subconjuntos em cada partição. Na Tabela 2.2 estão apresentados os resultados da implementação do método LTS.

Tabela 2.2 – Exemplo de implementação do método Jackknife.

Partição A_1		
Iteração	Reamostragem	SQD
1	$\{8, 10, 13, 15\}$	35,76
2	$\{5, 10, 13, 15\}$	57,96
3	$\{5, 8, 13, 15\}$	62,76
4	$\{5, 8, 10, 15\}$	54,96
5	$\{5, 8, 10, 13\}$	39,76
Partição A_2		
Iteração	Reamostragem	SQD
1	$\{1, 5, 6, 9\}$	58,04
2	$\{2, 5, 6, 9\}$	75,64
3	$\{1, 2, 5, 9\}$	86,84
4	$\{1, 2, 5, 6\}$	84,64
5	$\{1, 2, 6, 9\}$	43,64

Legenda: Soma de Quadrados dos Desvios (SQD).

Fonte: Do autor (2025).

Ao considerar os resultados apresentado na Tabela 2.2, os subconjuntos das partições A_1 e A_2 que apresentam menor estimativa para a SQD são, respectivamente, $\{8, 10, 13, 15\}$ e $\{16, 18, 22, 23\}$. Assim, o conjunto a ser considerado para se fazer as análises estatísticas de interesse, é dado por $B = \{8, 10, 13, 15, 16, 18, 22, 23\}$. Comparando os conjuntos A e B , percebe-se que o conjunto B não apresenta os valores das extremidades de A . Logo, as estimativas calculadas ao considerar B são menos influenciadas por valores extremos e ainda são capazes de descrever o conjunto amostral A .

Portanto, a metodologia robusta LTS é promissora para lidar com outliers. Além disso, é um método versátil, sendo implementado em diversos contextos. Trabalhos como os de Gu et al. (2020), Xie et al. (2022) e (MANOEL et al., 2024b) são exemplos de sucessos de aplicação

do LTS. Na parte dois desta tese, o capítulo dois é dedicado a implementação dessa metodologia em ajustes do modelo não linear Logístico na descrição de curvas de crescimento. No entanto, os dados utilizados são simulados. Dessa forma, de modo a complementar esse trabalho, no capítulo três há uma aplicação na descrição de dados reais.

2.6 Simulação de Monte Carlo

Essa técnica de amostragem foi desenvolvida na década de 40 pelos matemáticos John von Neumann e Stanislaw Ulam para resolver problemas estatísticos. Entretanto, dada a dificuldade de realizar simulações manualmente, a SMC só foi difundida quando os computadores foram desenvolvidos (SOBOL, 1994; KROESE; RUBINSTEIN, 2012). Basicamente, a partir de um conjunto, a SMC gera outros conjuntos de valores que apresentam a mesma distribuição dos dados originais. A SMC permite avaliar incertezas de medição, assim como possibilita investigar o impacto que variações nos valores têm sobre as estatísticas de interesse, já que se pode gerar tantas amostras quanto queira (YANG; TIAN, 2012).

Segundo Dunn e Bahadori (2024), na técnica SMC as variáveis de entrada devem ser caracterizadas por suas respectivas funções de probabilidade, além de suas médias e desvios padrão. Nesse sentido, a distribuição resultante é obtida pela combinação de amostras aleatórias das variáveis de entrada, o que produz uma propagação de suas funções de probabilidade. Assim, a função de probabilidade resultante descreve os valores mensurados. Por isso, a SMC também é conhecida como método de propagação de distribuições. Além disso, essa técnica proporciona condições para compreender conceitos e propriedades de amostragem repetida e de um estimador.

Os resultados obtidos pela técnica são estimativas do valor das grandezas de saída e a incerteza expandida. E mesmo que a SMC não apresente dificuldades quanto a aplicação, como em relação a distribuições assimétricas ou correlações entre as grandezas de influência, ressalta-se que a qualidade dos resultados obtidos, de acordo com Donatelli e Konrath (2005), dependem dos seguintes fatores:

- a) representatividade do modelo matemático;
- b) qualidade da caracterização das variáveis de entrada;
- c) características do gerador de números pseudo-aleatórios;
- d) quantidade de simulações;
- e) procedimento de definição do intervalo de abrangência.

Dunn e Bahadori (2024) apontam que a SMC fornece um detalhamento para as possibilidades de resultados a partir da distribuição de um conjunto, mas não uma sugestão explícita para tomadas de decisão. Por outro lado, essa técnica possibilita experimentos quanto aos efeitos de intervenções e planos de contingência, além de fornecer mais informações para as tomadas de decisão (YANG; TIAN, 2012; KROESE; RUBINSTEIN, 2012).

2.7 Desvio Médio Absoluto

Para mensurar a precisão dos ajustes do modelo não linear Logístico é considerado o desvio médio absoluto (DMA). Em uma perspectiva comparativa entre modelos ajustados, o DMA oferece uma interpretação intuitiva da variância residual (GORARD, 2015; LOPES et al., 2016). O DMA pode ser obtido pela média aritmética das diferenças absolutas entre os valores observados (y_i) e estimados ($\hat{y}_i$), de acordo com a Equação 2.11.

$$DMA = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (2.11)$$

Quanto menor o desvio, menor é o DMA. Logo, quando $y_i \rightarrow \hat{y}_i$, tem-se $DMA \rightarrow 0$. Assim, ao comparar dois ajustes, quanto menor o DMA, mais representativo é o modelo ajustado, pois apresenta desvios menores, ou ainda, menor variância residual (GORARD, 2015; MANOEL et al., 2024a). Portanto, diante de um conjunto de dados com outliers, é coerente definir o DMA como critério de seleção de ajustes do modelo não linear Logístico.

REFERÊNCIAS

- AGGARWAL, C. C.; AGGARWAL, C. C. **Data classification**. [S.l.]: Springer, 2015.
- BATES, D. M.; CHAMBERS, J. M. Nonlinear models. In: **Statistical models in S**. [S.l.]: Routledge, 2017. p. 421–454.
- BATES, D. M.; WATTS, D. G. Relative curvature measures of nonlinearity. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 42, n. 1, p. 1–16, 1980.
- BEYAZTAS, U.; ALIN, A. Jackknife-after-bootstrap method for detection of influential observations in linear regression models. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 42, n. 6, p. 1256–1267, 2013.
- BUSSAB, W. d. O.; MORETTIN, P. A. Estatística básica. In: **Estatística básica**. [S.l.: s.n.], 2010. p. xvi–540.
- DONATELLI, G. D.; KONRATH, A. C. Simulação de monte carlo na avaliação de incertezas de medição. 2005.
- DRAPER, N. R.; SMITH, H. **Applied regression analysis**. [S.l.]: John Wiley & Sons, 1998. v. 326.
- DUNN, W. L.; BAHADORI, A. A. Reflections on use of monte carlo methods. **Radiation Physics and Chemistry**, Elsevier, v. 218, p. 111634, 2024.
- FARCOMENI, A.; VENTURA, L. An overview of robust methods in medical research. **Statistical methods in medical research**, v. 21, p. 111–33, 10 2010.
- FERNANDES, T. J. et al. Double sigmoidal models describing the growth of coffee berries. **Ciência Rural**, SciELO Brasil, v. 47, p. e20160646, 2017.
- FERREIRA, D. F. **Estatística Básica**. 2nd. ed. Lavras: UFLA, 2009.
- FOX, T.; HINKLEY, D.; LARNTZ, K. Jackknifing in nonlinear regression. **Technometrics**, Taylor & Francis, v. 22, n. 1, p. 29–33, 1980.
- GOMES, M. I. Metodologias jackknife e bootstrap em estatística de extremos. In: **Actas do II Congresso Anual da Sociedade Portuguesa de Estatística**. [S.l.: s.n.], 1994. p. 31–46.
- GORARD, S. Introducing the mean absolute deviation ‘effect’ size. **International Journal of Research & Method in Education**, Taylor & Francis, v. 38, n. 2, p. 105–114, 2015.
- GU, T. et al. A robust moving total least-squares fitting method for measurement data. **IEEE Transactions on Instrumentation and Measurement**, IEEE, v. 69, n. 10, p. 7566–7573, 2020.
- HAWKINS, D. M. Multivariate outlier detection. In: **Identification of outliers**. [S.l.]: Springer, 1980. p. 104–114.

HEKIMOGLU, S. et al. A new outlier detection method considering outliers as model errors. **Experimental Techniques**, Springer, v. 39, p. 57–68, 2015.

KROESE, D. P.; RUBINSTEIN, R. Y. Monte carlo methods. Wiley Interdisciplinary Reviews: Computational Statistics, **Wiley Online Library**, v. 4, n. 1, p. 48–58, 2012.

LOPES, F. et al. Análises de dados longitudinais em bovinos nelore mocho por meio de modelos não lineares. **Archivos de zootecnia**, v. 65, n. 250, p. 123–129, 2016.

MANOEL, I. S. et al. Método jackknife aplicado em ajustes de modelos não lineares em presença de outliers no conjunto de dados. **Caderno Pedagógico**, v. 21, n. 10, p. e8938–e8938, 2024a.

MANOEL, I. S. et al. Simulation of robust adaptive regression multi-level models for quality analysis of special coffees in cold storage. **Acta Scientiarum. Technology**, Editora da Universidade Estadual de Maringá-EDUEM, v. 46, n. 1, 2024b.

MILLER, R. G. The jackknife-a review. **Biometrika**, Oxford University Press, v. 61, n. 1, p. 1–15, 1974.

ROUSSEEUW, P. J.; DRIESSEN, K. V. Computing lts regression for large data sets. **Data mining and knowledge discovery**, Springer, v. 12, p. 29–45, 2006.

ROUSSEEUW, P. J.; HUBERT, M. Robust statistics for outlier detection. Wiley interdisciplinary reviews: **Data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 1, p. 73–79, 2011.

ROUSSEEUW, P. J.; HUBERT, M. Anomaly detection by robust statistics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 8, n. 2, p. e1236, 2018.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust regression and outlier detection**. [S.l.]: John wiley & sons, 2005.

RUGGIERO, M. A. G.; LOPES, V. L. da R. **Cálculo numérico: aspectos teóricos e computacionais**. 2nd. ed. São Paulo: Pearson Makron Books, 1996. Edição traduzida.

SEBER, G. P.; WILD, C. J. Nonlinear regression. **New Jersey: J. Wiley**, p. 752, 1989.

SOBOL, I. M. A. **Primer for the Monte Carlo Method**. [S.l.]: CRC Press, 1994.

WANG, L.; YU, F. Jackknife resampling parameter estimation method for weighted total least squares. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 49, n. 23, p. 5810–5828, 2020.

XIE, Z. et al. Partial least trimmed squares regression. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 221, p. 104486, 2022.

YANG, J.; KOLASSA, J. E. The impact of application of the jackknife to the sample median. **The American Statistician**, Taylor & Francis, v. 75, n. 4, p. 445–449, 2021.

YANG, W.; TIAN, C. Monte-carlo simulation of information system project performance. **Systems Engineering Procedia**, Elsevier, v. 3, p. 340–345, 2012.

SEGUNDA PARTE

ARTIGO 1

Artigo publicado pela revista Caderno Pedagógico.

Redigido de acordo com as normas da Universidade Federal de Lavras.

MÉTODO JACKKNIFE EM AJUSTES DE MODELOS NÃO LINEARES EM PRESENÇA DE OUTLIERS NO CONJUNTO DE DADOS

Iuri dos Santos Manoel^a, Tales Jesus Fernandes^a, Daniela Aparecida Mafra^a, Carlos Antônio Zarzar^b

^a Universidade Federal de Lavras (UFLA), 3037 CEP: 37203-202, Lavras, Minas Gerais, Brasil

^b Universidade Federal do Oeste do Pará (UFOPA), Av. Maj. Francisco Mariano CEP: 68220-000, Cidade Alta, Monte Alegre, Pará, Brasil

RESUMO

A detecção e tratamento adequado de outliers é imprescindível em análises de dados, pois esses valores podem distorcer estimativas e prejudicar a precisão dos resultados. Quanto a ajustes de modelos, outliers têm um efeito negativo, distorcendo as estimativas de seus parâmetros, especialmente quando utilizado um estimador suscetível a outliers, como o Método de Mínimos Quadrados. Neste sentido, métodos robustos reduzem os efeitos que outliers têm em estimativas de interesse, proporcionando um resultado robusto e confiável. Dentre as diversas metodologias robustas, o método de reamostragem Jackknife apresenta versatilidade e eficiência quanto a minimização de vieses em estimadores. Diante do exposto, este artigo tem por objetivo apresentar o método Jackknife como uma metodologia robusta para ajustes do modelo não linear Logístico na descrição de um conjunto de dados simulados com outliers. A implementação da metodologia proposta apresenta resultados promissores quanto a redução da influência dos outliers no modelo ajustado.

Palavras-chave: outlier; método robusto; reamostragem; jackknife; modelo não linear.

ABSTRACT

Detecting and appropriately addressing outliers is crucial in data analysis because these values can distort estimates and compromise accuracy. Outliers have a negative effect on model fitting, distorting parameter estimation, particularly when using an estimator susceptible to outliers, such as the least squares method. In this sense, there are robust methods, which are techniques that mitigate the effects that outliers have on estimates of interest, providing a resilient, representative and reliable result. Among several robust methodologies, the jackknife resampling method exhibits versatility and efficiency in minimizing estimator bias. In view of the above, this article presents the jackknife method as a robust methodology for fitting a nonlinear logistic model using a simulated dataset with outliers. Promising results in terms of mitigating the influence of outliers on the adjusted model are demonstrated by implementing the proposed methodology shows promising results regarding mitigating the influence of outliers in the adjusted model.

Keywords: outlier; robust method; resampling; Jackknife; nonlinear model.

1 INTRODUÇÃO

Outliers são observações que diferem substancialmente do padrão geral dos demais valores em um conjunto de dados. Esses valores podem surgir devido a uma variedade de razões, como erros de medição, erros de entrada de dados, comportamentos extremos ou eventos raros (ROUSSEEUW; HUBERT, 2011; AGGARWAL; AGGARWAL, 2015). A detecção e tratamento adequado desses valores é imprescindível em análises de dados, pois outliers podem distorcer estimativas e prejudicar a precisão dos resultados (ROUSSEEUW; HUBERT, 2018).

Hekimoglu (2015) aponta duas abordagens para lidar com outliers: testes de detecção de outliers e métodos robustos. De modo geral, em testes de detecção, se o valor do resíduo de uma observação for maior que um valor crítico, então aquela observação é considerada um outlier (BASELGA, 2007; LIN, 2022). No entanto, Baselga (2007) demonstra que esses testes se tornam ineficientes ou pouco confiáveis para conjuntos com mais de um outlier.

Diante da limitação dos testes de detecção de outliers, é preferível os métodos robustos de estimação (HEKIMOGLU, 2005; HEKIMOGLU, 2015). Rousseeuw e Hubert (2018) apontam que esses métodos reduzem o efeito de outliers nas estimativas de interesse, proporcionando um resultado robusto e confiável. Os métodos robustos podem ser aplicados de diversas maneiras, sendo uma delas a reamostragem. Dentre os métodos de reamostragem, o método Jackknife apresenta versatilidade e eficiência quanto a minimização de vieses em estimativas (YANG; KOLASSA, 2021).

Beyaztas e Alin (2012) detectam observações influentes em conjuntos de dados ao implementar o método de reamostragem Jackknife, de modo que melhore o ajuste de modelos de regressão linear. Por outro lado, Wang e Yu (2019) introduzem o método como alternativa viável para melhorar os resultados de estimações do método de mínimos quadrados totais ponderados. O método Jackknife também é implementado para minimizar a heterogeneidade da variância em clusters (DU et al., 2022).

Neste sentido, este artigo tem por objetivo apresentar o método de reamostragem Jackknife como uma metodologia robusta para ajustes do modelo não linear Logístico na descrição de conjuntos de dados com outliers.

2 MATERIAL E MÉTODOS

2.1 Conjunto de dados

O conjunto de dados é gerado via simulação no software R utilizando o método de Simulação de Monte Carlo. São simulados $n = 200$ valores com base na aplicação da equação do modelo Logístico. Dentre as parametrizações do referido modelo, considera-se aquela mais eficiente, de acordo com Fernandes (2015), apresentada na Equação 2.1.

$$y_i = \frac{\alpha}{1 + e^{\kappa(\beta - x_i)}} + \epsilon_i \quad (2.1)$$

em que: $i = 1, \dots, n$ se refere a cada valor do conjunto de dados, y_i e x_i são os i -ésimos valores das variáveis dependente e independente, respectivamente, ϵ_i o desvio associado ao i -ésimo valor e α , β e κ os parâmetros do modelo. Os valores definidos para os parâmetros do modelo, são: $\alpha = 200$, $\beta = 50$ e $\kappa = 0.09$.

Quanto a variável independente x_i , pode ser interpretada como um vetor que contém 200 entradas, sendo 10 repetições de cada valor no intervalo de 5 a 100, com incrementos de 5. O vetor pode ser escrito de acordo com a Equação (2.2).

$$x = (\underbrace{5, \dots, 5}_{10}, \dots, \underbrace{100, \dots, 100}_{10}). \quad (2.2)$$

Ao definir a variável independente dessa forma, obtém-se um conjunto de dados longitudinais. Assim, interpreta-se o conjunto como valores oriundos de observações de um mesmo grupo de indivíduos em diferentes níveis da variável x_i . Logo, são considerados 10 ‘indivíduos’ no conjunto de dados simulados.

Desse modo, com o objetivo de aprimorar a representatividade, adiciona-se a cada valor y_i uma variável u_i , a qual segue uma distribuição uniforme. O valor da variável u_i aumenta de acordo com i , de modo a representar o aumento de dispersão nos dados, conforme definido a seguir.

- $-3 \leq u \leq 3$ para $i < 50$;
- $-7 \leq u \leq 7$ para $50 \leq i < 130$;
- $-14 \leq u \leq 14$ para $130 \leq i \leq 200$.

Nesse ponto, é adicionado 50 unidades a cada valor simulado, tal que o conjunto de dados sofra uma translação vertical. Assim, cada valor não discrepante y_i é definido pela Equação 2.3.

$$y_i = \frac{200}{1 + e^{0,09(50-x_i)}} + u_i + 50 \quad (2.3)$$

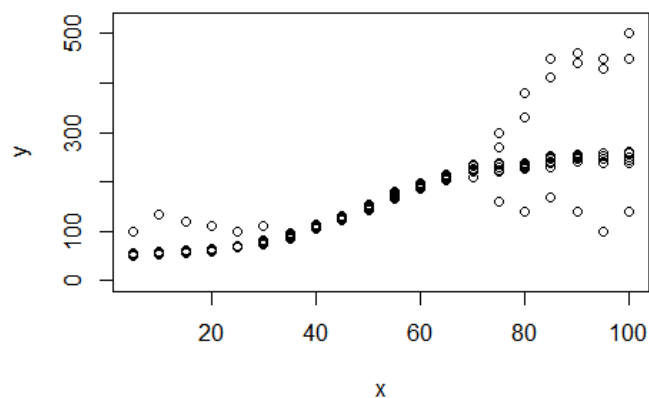
A adição de 50 unidades é necessária para prevenir valores negativos na parte inicial da curva, pois a variável u assume valores negativos. Contudo, essa adição implica no aumento do parâmetro α .

Diante do exposto, os valores reais dos parâmetros do modelo Logístico que descreve com maior precisão a relação entre as variáveis y e x , desconsiderando outliers, podem ser definidos por $\alpha = 250$, $\beta = 50$ e $\kappa = 0.09$. Esses valores são utilizados como referência em cada um dos ajustes apresentados neste artigo. Logo, o interesse é obter um ajuste cujos parâmetros estão mais próximos dos valores supracitados.

Salienta-se ainda que o ajuste do modelo não linear Logístico é implementado ao utilizar a função ‘nls()’ do software R (versão 4.2.2), em que o método iterativo utilizado para convergência é o algoritmo Gauss-Newton.

Por fim, 24 outliers são introduzidos no conjunto de dados, sendo 6 no início da curva e 18 no final. Os valores são substituídos de modo a representar um comportamento atípico de um indivíduo, o que justifica a abordagem robusta proposta (Figura 2.1).

Figura 2.1 — Gráfico de dispersão do conjunto de dados simulados.



Fonte: Do autor.

2.2 Método de reamostragem Jackknife

O método Jackknife é baseado em reamostragens que são feitas a partir do conjunto amostral. É um processo iterativo, em que a cada iteração um ou mais elementos são

selecionados de modo que o restante do conjunto seja considerado para o cálculo das estimativas de interesse. Para ilustrar o método em que a cada iteração é retirado um único elemento, considere o exemplo a seguir.

Exemplo: Seja um conjunto amostral com $n = 6$ elementos, sendo descrito por $\{x_1, x_2, x_3, x_4, x_5, x_6\}$. Deste modo, ao aplicar o método de reamostragem Jackknife, tem-se os subconjuntos apresentados na Tabela 2.1.

Tabela 2.1 — Subconjuntos Jackknife do exemplo didático.	
Iterações	Subconjuntos
1 ^a	$\{x_2, x_3, x_4, x_5, x_6\}$
2 ^a	$\{x_1, x_3, x_4, x_5, x_6\}$
$\vdots$	$\vdots$
6 ^a	$\{x_1, x_2, x_3, x_4, x_5\}$

Fonte: Do autor.

Uma vez obtidas as reamostragens, são estimadas as estimativas de interesse. Desse modo, se o interesse for a média aritmética dos elementos do conjunto $\{x_1, x_2, x_3, x_4, x_5, x_6\}$, então se tem 6 médias, uma para cada reamostragem Jackknife, de acordo com a seguir.

$$\begin{aligned}\bar{x}_1 &= \frac{x_2 + x_3 + x_4 + x_5 + x_6}{5}; \\ &\vdots \\ \bar{x}_6 &= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}.\end{aligned}$$

Geralmente o método é limitado a retirada de apenas 1 valor do conjunto, tornando-o menos eficiente em amostras grandes. Assim, um único valor representa cada vez menos de um conjunto, de acordo com seu tamanho, sendo sua representatividade inversamente proporcional ao tamanho amostral, isto é, quanto maior a amostra, menos representativo é um valor dela (veja a Tabela 2.2).

Tabela 2.2 — Representatividade de 1 valor diante de conjuntos de dados com tamanhos diferentes.

Tamanho Amostral (n)	Representatividade de 1 valor
30	3.3%
50	2.0%
100	1.0%
200	0.5%
500	0.2%
1000	0.1%

Fonte: Do autor.

Diante do exposto, propõe-se a retirada de quantidades maiores que 1, como a desigualdade do ponto de ruptura h , definida por Rousseeuw e Van Driessen (2006), em que $h \in \mathbb{N}$, tal que

$$h \geq \frac{3n + p + 1}{4}, \quad (2.4)$$

sendo p a quantidade de parâmetros do modelo estatístico utilizado e n o tamanho amostral.

O ponto de ruptura (Equação 2.4) pode ser interpretado pela proporção em que um subconjunto ainda representa o conjunto. Isso garante que não haja descaracterização dos dados em estudo. A partir das reamostragens, espera-se obter estimativas para os parâmetros do modelo Logístico que são menos influenciadas por valores discrepantes.

Diante do exposto, neste artigo são considerados três ajustes do modelo não linear Logístico. Para melhor referência, considera-se as definições apresentadas a seguir.

- Ajuste convencional: é o ajuste ao utilizar o modelo não linear Logístico sem nenhuma técnica robusta.
- Abordagem 1: considera-se as estimativas das medianas de cada reamostragem Jackknife para a estimação do parâmetro de interesse.
- Abordagem 2: considera-se o subconjunto reamostrado, via Jackknife, cujo coeficiente de curtose seja menor para a estimação do parâmetro de interesse.

2.2.1 Abordagem 1: Mediana sobre as reamostragens

A cada iteração são calculadas as estimativas para os parâmetros do modelo Logístico tal que ao final do processo, toma-se a mediana referente a cada parâmetro. Assim, a cada iteração, tem-se os vetores de estimativas dos parâmetros do modelo ajustado: $\hat{\theta}_1 = (\hat{\alpha}_1, \hat{\beta}_1, \hat{\kappa}_1)$, $\hat{\theta}_2 = (\hat{\alpha}_2, \hat{\beta}_2, \hat{\kappa}_2)$, $\dots$, $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\kappa}_n)$. De modo geral, a matriz $\hat{\theta}$ é composta pelos vetores $\hat{\theta}_j$, com $j = 1, \dots, n$, em que cada linha é um dos vetores, e pode ser escrita por

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_n \end{pmatrix} = \begin{pmatrix} \hat{\alpha}_1 & \hat{\beta}_1 & \hat{\kappa}_1 \\ \vdots & \vdots & \vdots \\ \hat{\alpha}_n & \hat{\beta}_n & \hat{\kappa}_n \end{pmatrix}.$$

A partir da matriz $\hat{\theta}$, estima-se a mediana de cada um dos parâmetros, de modo a se obter as estimativas robustas, de acordo com a Equação 2.5.

$$\begin{aligned} \hat{\theta}_{JKK} &= \left(\text{mediana}(\hat{\alpha}_1, \dots, \hat{\alpha}_n), \text{mediana}(\hat{\beta}_1, \dots, \hat{\beta}_n), \text{mediana}(\hat{\kappa}_1, \dots, \hat{\kappa}_n) \right) \\ &\Rightarrow \hat{\theta}_{JKK} = (\hat{\alpha}_{JKK}, \hat{\beta}_{JKK}, \hat{\kappa}_{JKK}), \end{aligned} \quad (2.5)$$

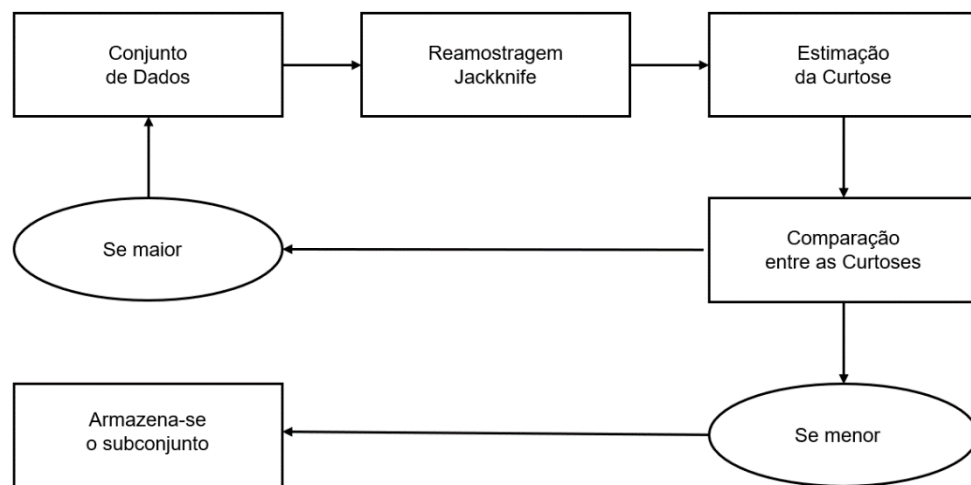
em que o índice JKK refere-se às estimativas obtidas por esta abordagem.

2.2.2 Abordagem 2: Curtose sobre as reamostragens

Em conjuntos de dados que apresentam valores discrepantes, a curva de distribuição apresenta caudas (extremidades) densas, pois há valores extremos. Desse modo, de acordo com a densidade das caudas, a estimativa de curtose aumenta. Logo, pode-se considerar que o conjunto com menor curtose contém menos valores nas extremidades, isto é, menos outliers.

A cada reamostragem Jackknife, é estimada o coeficiente de curtose de modo a comparar com a reamostragem anterior. O subconjunto que apresenta menor curtose, é armazenado e considerado para o ajuste do modelo Logístico. Todo o processo de seleção é ilustrado na Figura 2.2. Dessa forma, espera-se minimizar a influência de outliers sobre as estimativas dos parâmetros do modelo e, conseqüentemente, aumentar capacidade de predição.

Figura 2.2 — Processo de seleção dos subconjuntos reamostrados pelo método de reamostragem Jackknife, de acordo com o critério da menor curtose, para estimação dos parâmetros do modelo não linear Logístico.



Fonte: Do Autor.

2.3 Desvio médio absoluto (DMA)

Para mensurar a precisão dos ajustes do modelo não linear Logístico é considerado o desvio médio absoluto (DMA). Em uma perspectiva comparativa entre modelos ajustados, o DMA oferece uma interpretação intuitiva da variância residual, além de apresentar maior tolerância a outliers (GORARD, 2015; LOPES et al., 2016).

O DMA pode ser obtido pela média aritmética das diferenças absolutas entre os valores simulados (y_i) e estimados ($\hat{y}_i$), de acordo com a Equação 2.6.

$$DMA = \frac{1}{n} \cdot \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.6)$$

Quanto menor o desvio, menor é o *DMA*. Logo, quando $y_i - \hat{y}_i \rightarrow 0$, tem-se que $DMA \rightarrow 0$. Assim, ao comparar dois ajustes, quanto menor o *DMA*, mais representativo é o modelo ajustado, pois apresenta desvios menores, ou ainda, menor variância residual.

Portanto, como o conjunto de dados simulados apresenta outliers, é coerente definir o *DMA* como critério de seleção de ajustes do modelo não linear Logístico. Assim, o *DMA* é utilizado para comparações entre os ajustes convencional e robustos (abordagens 1 e 2).

3 RESULTADOS

A implementação do método de reamostragem Jackknife ao considerar as abordagens 1 e 2, apresentadas na subseção 2.2, é feita por script de autoria dos autores (Apêndice A).

3.1 Ajuste convencional

As estimativas dos parâmetros do modelo não linear Logístico sobre o conjunto de dados estão apresentadas na Tabela 3.1. Independentemente dos outliers no conjunto, todas as estimativas dos parâmetros são significativas para explicar a variável y em relação a x . O *DMA* apresentado é utilizado para comparação dos ajustes nas subseções seguintes.

Tabela 3.1 — Estimativas dos parâmetros para o ajuste do modelo não linear Logístico sobre a totalidade dos dados.

Parâmetros	Estimativas	$p - \text{valor}$	<i>DMA</i>
α	341.3000	$2 \cdot 10^{-16}$	23.6875
β	55.5400	$2 \cdot 10^{-16}$	
κ	0.0397	$3.16 \cdot 10^{-14}$	

Legenda: Desvio médio absoluto (DMA).

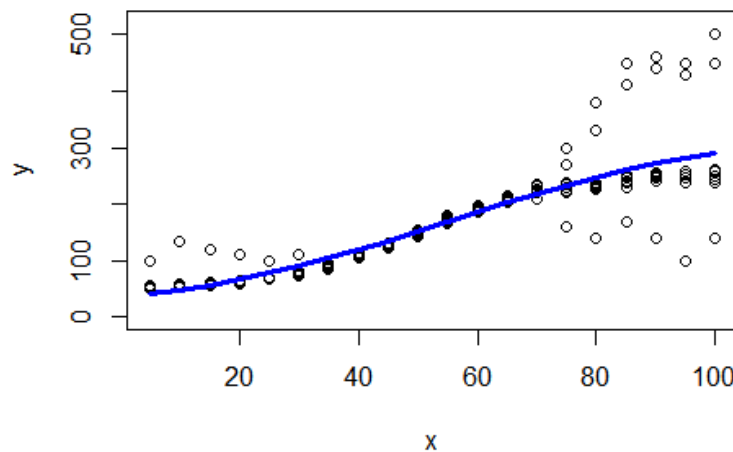
Fonte: Do autor.

As estimativas estão relativamente distantes de seus respectivos valores reais, apresentados na subseção 2.1. Essa diferença pode ser justificada pela presença dos outliers no conjunto, sendo que o valor estimado para o coeficiente de curtose é de 4.3001. Esse valor elevado, pode ser justificado pelos valores discrepantes na parte inicial e final da curva.

A presença dos outliers no conjunto de dados faz com que o modelo ajustado superestime as estimativas dos parâmetros α e β e subestime o parâmetro κ . A estimativa para o parâmetro α apresenta um aumento de 36.52%, enquanto β de 11.08%. Por outro lado, a estimativa para o parâmetro κ , apresenta um decaimento de 55.88%.

O gráfico do modelo Logístico ajustado está apresentado na Figura 3.1. Ressalta-se que visualmente o ajuste não está ruim, no entanto, requer especial atenção quanto às inferências realizadas se baseando nas interpretações práticas das estimativas (influenciadas) dos parâmetros.

Figura 3.1 — Gráfico do modelo não linear Logístico ajustado sobre o conjunto de dados com outliers.



Fonte: Do autor.

3.2 Ajustes robustos

Na implementação do método de reamostragem Jackknife são considerados subconjuntos formados por arranjos dos elementos do conjunto, de acordo com o ponto de ruptura h , sendo

$$h \geq \frac{3n + p + 1}{4} \Rightarrow h \geq \frac{3 \cdot 200 + 3 + 1}{4} \Rightarrow h \geq 151.$$

Assim, a cada iteração Jackknife, pode-se considerar no mínimo 151 valores para a estimação dos parâmetros do modelo não linear Logístico sem que descaracterize os dados. Visando uma maior eficiência do método, considera-se 151 elementos para compor um subconjunto para o ajuste do modelo.

Desse modo, é coerente tomar $C_{200,151} = \frac{200!}{151 \cdot (200-151)!}$ combinações tais que todos os possíveis subconjuntos sejam considerados. Contudo, tem-se que $C_{200,151} \approx 1,5 \cdot 10^{46}$, aumentando o custo computacional e inviabilizando sua implementação. Desse modo, neste

artigo é apresentada uma alternativa para o problema, definindo-se o número de 10^5 iterações, uma quantidade relativamente grande.

3.2.1 Abordagem 1

O método Jackknife geralmente é utilizado para estimar intervalos de confiança sobre um parâmetro de interesse a partir de n estimativas que são calculadas ao retirar 1 valor do conjunto. Neste manuscrito, visando o ajuste do modelo não linear Logístico, as estimativas dos parâmetros do modelo são estimadas de acordo com o apresentado na subseção 2.1, isto é, pela mediana das estimativas dos parâmetros dos modelos ajustados a cada reamostragem Jackknife. Matematicamente, escreve-se

$$\begin{aligned}\hat{\theta}_{JKK} &= \left(\text{mediana}(\hat{\alpha}_1, \dots, \hat{\alpha}_n), \text{mediana}(\hat{\beta}_1, \dots, \hat{\beta}_n), \text{mediana}(\hat{\kappa}_1, \dots, \hat{\kappa}_n) \right) \\ \Rightarrow \hat{\theta}_{JKK} &= (\hat{\alpha}_{JKK}, \hat{\beta}_{JKK}, \hat{\kappa}_{JKK}).\end{aligned}$$

Ao realizar 10^5 reamostragens, obtém-se um vetor de estimativas para os parâmetros do modelo não linear Logístico com 10^5 coordenadas. Assim, calcula-se a mediana das coordenadas de cada um dos vetores, obtendo as estimativas para α , β e κ , apresentadas na Tabela 3.2.

Tabela 3.2 — Estimativas dos parâmetros do modelo não linear Logístico obtidas pela mediana das estimativas dos parâmetros de reamostragens com 151 valores.

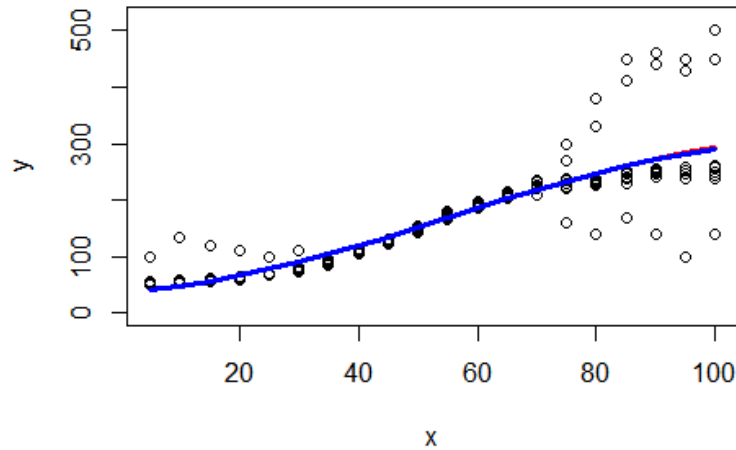
Parâmetros	Estimativas	DMA
α	342,5150	23,7288
β	55,7341	
κ	0,0396	

Legenda: Desvio médio absoluto (DMA).

Fonte: Do autor.

As estimativas obtidas por esta abordagem são semelhantes as estimativas do ajuste convencional, apresentadas na Tabela 3.2. Além disso, esta abordagem proporciona um ajuste com maior DMA ($DMA'_{JKK} > DMA_{MQO}$), em que DMA_{MQO} refere-se ao ajuste convencional e DMA'_{JKK} à abordagem 1. Isso indica que o modelo ajustado convencionalmente apresenta maior representatividade. Os gráficos dos modelos supracitados estão apresentados na Figura 3.2.

Figura 3.2 — Gráficos dos modelos ajustados sobre a totalidade dos dados e sobre a mediana das estimativas das reamostragens e a dispersão do conjunto de dados.



Fonte: Do autor.

3.2.2 Abordagem 2

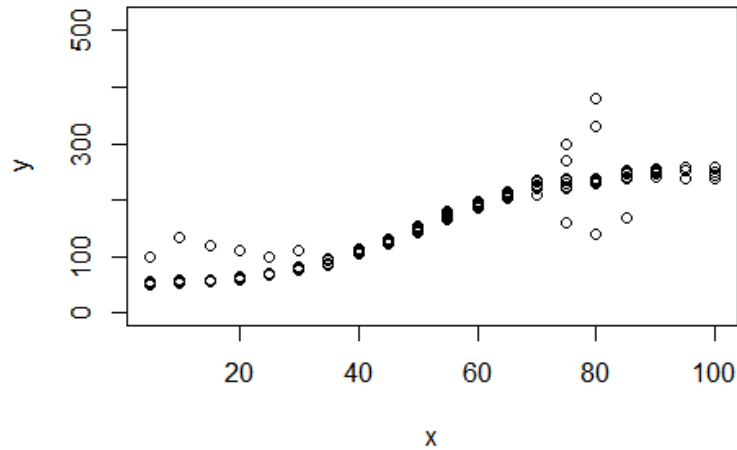
Ao implementar o método Jackknife com 10^5 iterações, tomando arranjos de $h = 151$ valores, é apresentado o gráfico de dispersão do subconjunto que obteve menor estimativa para o coeficiente de curtose (Figura 3.3). Observa-se que os maiores valores no final da curva não estão presentes no subconjunto por serem os maiores responsáveis pelo coeficiente de curtose elevada.

O subconjunto reamostrado apresenta uma estimativa para o coeficiente de curtose de 1.9877, indicando que houve redução na quantidade de outliers no conjunto. Isso sugere que o ajuste pela abordagem 2 será menos influenciado por outliers do que o ajuste convencional, em que é considerado a totalidade dos dados.

As estimativas dos parâmetros para o ajuste de acordo com a abordagem 2, estão apresentadas na Tabela 3.3. Todos os parâmetros são significativos para o ajuste do modelo, tornando-o apropriado para explicar o subconjunto.

No tocante a comparação entre as estimativas dos parâmetros obtidos neste ajuste e os valores reais dos parâmetros, a estimativa para o parâmetro α apresenta um aumento de 15.58%. Por outro lado, a estimativa para o parâmetro β apresenta um decaimento de 11.08%, enquanto κ de 51.44%.

Figura 3.3 — Dispersão do conjunto de dados reamostrados pelo método de reamostragem Jackknife com menor curtose.



Fonte: Do autor.

Tabela 3.3 — Estimativas para os parâmetros do modelo não linear Logístico ajustado ao subconjunto reamostrado pelo método Jackknife.

Parâmetros	Estimativas	$p - \text{valor}$	DMA
α	288,9604	$2 \cdot 10^{-16}$	22,1637
β	46,6922	$2 \cdot 10^{-16}$	
κ	0,0437	$2 \cdot 10^{-16}$	

Legenda: Desvio médio absoluto (DMA).

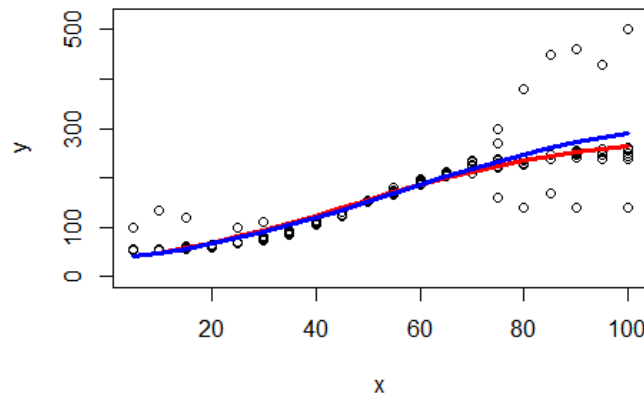
Fonte: Do autor.

Observa-se que as estimativas para os parâmetros α e β são menores em relação ao modelo ajustado convencionalmente. O mesmo resultado se dá ao comparar com o ajuste via abordagem 1. Isso pode ser justificado pela ausência dos maiores valores no subconjunto reamostrado.

Como consequência, tem-se que $DMA_{MQO} > DMA''_{JKK} \Rightarrow 23,6875 > 22,1637$, em que DMA''_{JKK} se refere ao ajuste segundo a abordagem 2. Isso indica que o modelo Logístico ajustado sobre o subconjunto é mais representativo para explicar o conjunto de dados do que o ajuste que considera a totalidade dos dados.

Os gráficos dos modelos ajustados estão apresentados na Figura 3.4. A curva ajustada pela abordagem 2 indica uma correção ao final curva, sugerindo uma atenuação sobre a influência dos outliers nos dados. Ademais, aponta-se que nas partes inicial e de crescimento das curvas, não são apresentadas diferenças expressivas.

Figura 3.4 — Gráficos dos modelos Logísticos ajustados sobre a totalidade dos dados e sobre o subconjunto reamostrado pelo método de reamostragem Jackknife.



Fonte: Do autor.

4 DISCUSSÕES

No tocante a descrição de conjuntos de dados com outliers, o método de reamostragem Jackknife apresenta resultados promissores quando implementado em ajustes do modelo não linear Logístico. Dentre as abordagens propostas, o critério de seleção da reamostragem pela menor curtose (abordagem 2) resulta em um ajuste com menor Desvio Médio Absoluto (DMA) que o ajuste convencional. Isto indica que a metodologia robusta é eficiente quanto a reduzir a influência de outliers nas estimativas dos parâmetros do modelo.

Contudo, a abordagem 1, que se baseia nas medianas das estimativas dos parâmetros de cada reamostragem, apresenta um ajuste com maior DMA que o ajuste convencional. Logo, não há evidência de que esta metodologia proporcione um ajuste menos influenciado por outliers. Isso pode ser justificado pela Simulação de Monte Carlo (MOONEY, 1997; ANDERSON, 2023).

A cada iteração são considerados h valores para a reamostragem. Se a maioria dos outliers são reamostrados, isso significa que os valores não reamostrados são valores não discrepantes, àqueles que contribuem positivamente para o ajuste de um modelo representativo. Como a proporção de outliers é relativamente pequena, para os dados simulados correspondem a 12%, a situação descrita anteriormente é comum, tornando essa abordagem pouco eficiente.

Portanto, o método de reamostragem Jackknife, ao utilizar a abordagem 2, apresenta resultados promissores quanto a redução dos efeitos de outliers em ajustes do modelo Logístico. Nesse sentido, tem-se a implementação do método Jackknife na melhoria de ajustes de

regressões não lineares, alinhando-se com o trabalho de Beyaztas e Alin (2012), que utiliza o método em regressões lineares.

Esse resultado ressalta que o método Jackknife é uma ferramenta versátil e eficiente para lidar com outliers em diversos contextos, como já afirmado por Chernick (2012). Ademais, Yang e Kolassa (2021) salienta a versatilidade e eficiência quanto a minimização de vieses em estimadores, como também é apresentado pelos trabalhos de Beyaztas e Alin (2012), Wang e Yu (2019) e Du et al. (2022)

Outros métodos robustos difundidos pela literatura e implementados à modelos não lineares são os M-estimator (COLOMBIER, 2009; ERTAŞ; TOKER; KAÇIRANLAR, 2015). No entanto, esses métodos requerem condições de aplicações. Lim, Sen e Peddada (2013) apresentam os estimadores robustos ordinary M-estimator e weighted M-estimator que requerem a homocedasticidade e heterocedasticidade do conjunto de dados, respectivamente. Além disso, há maior complexidade na implementação dessas metodologias, como testes preliminares de homocedasticidade (LIM; SEM; PEDDADA, 2012).

Diante de outras técnicas robustas, a implementação do método de reamostragem Jackknife proposta neste artigo é relativamente simples e não exige nenhuma condição do conjunto de dados. O custo computacional é relativo, podendo ser implementado em um computador pessoal.

Contudo, a exclusão indiscriminada de valores no conjunto de dados pode comprometer a capacidade de generalização dos resultados obtidos por análises estatísticas (AGGARWAL, 2015). Ademais, Farcomeni e Ventura (2014) apontam isso como um problema, pois os valores excluídos não são contextualizados. Neste sentido, o método de reamostragem Jackknife implementado em ajustes do modelo não linear Logístico é uma alternativa para lidar com outliers, sendo um critério para exclusão dos valores que prejudicam esses ajustes.

5 CONCLUSÃO

A implementação do método de reamostragem Jackknife em ajustes do modelo não linear Logístico foi bem-sucedida. A abordagem robusta apresentou um ajuste em que a influência dos outliers foi atenuada, sugerindo uma correção no ajuste em relação ao convencional.

O tamanho dos subconjuntos reamostrados pelo método foi definida por este artigo como o ponto de ruptura, escrito por $h \geq \frac{3n+p+1}{4}$. A quantidade h de valores garante a não descaracterização do conjunto de dados e maior impacto do método de reamostragem Jackknife.

Entre as abordagens consideradas, o critério de seleção da reamostragem pela menor curtose proporciona melhor ajuste. Por outro lado, a abordagem que se baseia nas medianas das estimativas dos parâmetros de cada reamostragem não se mostra eficiente.

REFERÊNCIAS

- AGGARWAL, C. C.; AGGARWAL, C. C. **Data classification**. Springer International Publishing, 2015.
- ANDERSON, B. et al. A Monte Carlo simulation framework for reject inference. **Journal of the Operational Research Society**, v. 74, n. 4, p. 1133-1149, 2023.
- BASELGA, S. Critical limitation in use of τ test for gross error detection. **Journal of Surveying Engineering**, v. 133, n. 2, p. 52-55, 2007.
- BEYAZTAS, U.; ALIN, A. Jackknife-after-bootstrap method for detection of influential observations in linear regression models. **Communications in Statistics-Simulation and Computation**, v. 42, n. 6, p. 1256-1267, 2013.
- COLOMBIER, C. Growth effects of fiscal policies: an application of robust modified M-estimator. **Applied Economics**, v. 41, n. 7, p. 899-912, 2009.
- DU, R. et al. A weight Jackknife approach utilizing linear model based-estimators for clustered data. **Communications in Statistics-Simulation and Computation**, v. 53, n. 2, p. 1048-1067, 2024.
- ERTAŞ, H.; TOKER, S.; KAÇIRANLAR, S. Robust two parameter ridge M-estimator for linear regression. **Journal of Applied Statistics**, v. 42, n. 7, p. 1490-1502, 2015.
- FARCOMENI, A.; VENTURA, L. An overview of robust methods in medical research. **Statistical Methods in Medical Research**, v. 21, n. 2, p. 111-133, 2012.
- FERNANDES, T. T. et al. Parameterization effects in nonlinear models to describe growth curves. **Acta Scientiarum. Technology**, v. 37, n. 4, p. 397-402, 2015.
- GORARD, S. Introducing the mean absolute deviation ‘effect’ size. **International Journal of Research; Method in Education**, v. 38, n. 2, p. 105-114, 2015.
- HEKIMOGLU, S. Do robust methods identify outliners more reliably than conventional tests for outliners?. **Zfv-Zeitschrift für Geodäsie, Geoinformation und Landmanagement**, n. zfv 3/2005, 2005.
- HEKIMOGLU, S.; ERDOGAN, B.; ERENOGLU, R. C. A new outlier detection method considering outliers as model errors. **Experimental Techniques**, v. 39, p. 57-68, 2015.
- HUBERT, M.; ROUSSEEUW, P. J.; SEGART, P. Multivariate functional outlier detection. **Statistical Methods; Applications**, v. 24, n. 2, p. 177-202, 2015.
- LIM, C.; SEN, P. K.; PEDDADA, S. D. Accounting for uncertainty in heteroscedasticity in nonlinear regression. **Journal of Statistical Planning and Inference**, v. 142, n. 5, p. 1047-1062, 2012.

LIM, C.; SEN, P. K.; PEDDADA, S. D. Robust nonlinear regression in applications. **Journal of the Indian Society of Agricultural Statistics. Indian Society of Agricultural Statistics**, v. 67, n. 2, p. 215, 2013.

LOPES, F. B. et al. Análises de dados longitudinais em bovinos Nelore Mocho por meio de modelos não lineares. **Archivos de zootecnia**, v. 65, n. 250, p. 123-129, 2016.

MOONEY, C. Z. **Monte carlo simulation**. Sage, 1997.

ROUSSEEUW, P. J.; HUBERT, M. Anomaly detection by robust statistics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 2, p. e1236, 2018.

ROUSSEEUW, P. J.; HUBERT, M. Robust statistics for outlier detection. **Wiley interdisciplinary reviews: Data mining and knowledge discovery**, v. 1, n. 1, p. 73-79, 2011.

ROUSSEEUW, P. J.; VAN DRIESSEN, K. Computing LTS regression for large data sets. **Data mining and knowledge discovery**, v. 12, p. 29-45, 2006.

WANG, L.; YU, F. Jackknife resampling parameter estimation method for weighted total least squares. **Communications in Statistics-Theory and Methods**, v. 49, n. 23, p. 5810-5828, 2020.

YANG, J.; KOLASSA, J. E. The Impact of Application of the Jackknife to the Sample Median. **The American Statistician**, v. 75, n. 4, p. 445-449, 2021.

ARTIGO 2

Artigo submetido e em avaliação pelos revisores da revista Scientia Agricola.

Redigido de acordo com as normas da Universidade Federal de Lavras.

MÉTODO ROBUSTO LEAST TRIMMED SQUARES IMPLEMENTADO EM AJUSTES DO MODELO NÃO LINEAR LOGÍSTICO

Iuri dos Santos Manoel^a, Tales Jesus Fernandes^a, Daniela Aparecida Mafra^a, Carlos Antônio Zarzar^b

^a Universidade Federal de Lavras (UFLA), 3037 CEP: 37203-202, Lavras, Minas Gerais, Brasil

^b Universidade Federal do Oeste do Pará (UFOPA), Av. Maj. Francisco Mariano CEP: 68220-000, Cidade Alta, Monte Alegre, Pará, Brasil

RESUMO

O modelo não linear Logístico é amplamente implementado em pesquisas de campo, pois apresenta bons ajustes e parâmetros com interpretação prática. No entanto, este modelo é vulnerável a outliers, pois utiliza um estimador baseado na média, que pode ser significativamente afetada por esses valores extremos. Desse modo, este artigo tem por objetivo propor uma metodologia robusta para estimar os parâmetros do referido modelo na descrição de curvas de crescimento. Para isso, considera-se o método robusto Least Trimmed Squares (LTS), implementado em partições do conjunto de dados, que seleciona um subconjunto central que apresenta menor soma de quadrados dos desvios. Além disso, é utilizado o ponto de ruptura para determinar o tamanho dessas partições. O conjunto de dados é simulado via Simulação de Monte Carlo. A comparação entre o Desvio Médio Absoluto dos ajustes convencional e robusto, indica que a implementação do LTS ao ajuste do modelo não linear Logístico mitiga os efeitos que outliers têm nas estimativas de seus parâmetros. Portanto, a metodologia inovadora proposta neste artigo é promissora no aprimoramento de ajustes do modelo não linear Logístico na descrição de dados com outliers.

Palavras-chave: modelo não linear; curvas de crescimento; método robusto; least trimmed squares; ponto de ruptura.

ABSTRACT

The nonlinear logistic model is widely employed in field research due to its ability to provide accurate fits and parameters with practical interpretability. However, this model is sensitive to outliers, as its parameter estimation relies on mean-based estimators, which can be significantly influenced by extreme values. This article proposes a robust methodology for estimating the parameters of the nonlinear logistic model in the context of growth curve analysis. The robust Least Trimmed Squares (LTS) method is utilized, applied to partitions of the data set to select a central subset that minimizes the sum of squared deviations. Additionally, a breaking point is incorporated to determine the size of these partitions. Data sets are generated via Monte Carlo simulation. Comparisons based on the Mean Absolute Deviation between conventional and robust approaches demonstrate that implementing LTS mitigates the impact of outliers on parameter estimates. Consequently, the innovative methodology proposed in this study shows promise in enhancing the robustness of nonlinear logistic model adjustments for data containing outliers.

Keywords: nonlinear model; growth curves; robust method; least trimmed squares; breaking point.

1 INTRODUÇÃO

A utilização de modelos não lineares em pesquisas é comum, uma vez que esses modelos oferecem bons ajustes aos dados e possuem parâmetros interpretáveis (MISCHAN, DE PINHO; DE CARVALHO, 2011). No entanto, é essencial considerar que os conjuntos de dados podem conter informações que não representam adequadamente a população em estudo, especialmente em experimentos de campo, pois são sujeitos a fatores aleatórios não controláveis (ROUSSEEUW; VAN DRIESSEN, 2006).

Outliers são observações que diferem substancialmente do padrão geral dos demais valores em um conjunto de dados, podendo surgir devido a uma variedade de razões, como erros de medição, erros de entrada de dados, comportamentos extremos ou eventos raros (ROUSSEEUW; HUBERT, 2011; AGGARWAL; AGGARWAL, 2015). A detecção e tratamento adequado desses valores têm sido um desafio constante em análises de dados e nos ajustes de modelos, pois os outliers podem distorcer as estimativas de parâmetros e prejudicar a acurácia das análises estatísticas realizadas (ROUSSEEUW; HUBERT, 2018).

Nesse contexto, uma alternativa válida são os métodos robustos, pois a robustez destes métodos estatísticos permite que eles sejam menos sensíveis a outliers, evitando que distorçam significativamente os resultados (MARONNA et al., 2019). Uma abordagem robusta na análise de modelos não lineares pode proporcionar ajustes mais representativos e confiáveis diante de possíveis violações das suposições tradicionais.

No tocante à métodos robustos, o método Least Trimmed Squares (LTS) é uma opção para lidar com outliers no conjunto de dados (ROUSSEEUW; VAN DRIESSEN, (2006); GU et al. (2020)). O método robusto LTS seleciona um subconjunto da amostra que apresenta menor soma de quadrados dos desvios (XIE; CHEN, 2022). Desse modo, o ajuste do modelo considera apenas um subconjunto central, excluindo os extremos, mitigando as influências negativas que outliers podem incidir sobre as estimativas dos parâmetros do modelo ajustado (MANOEL et al., 2024).

O ponto de ruptura, proposto por Rousseeuw e Van Driessen (2006), é considerado como critério para delimitar o tamanho dos subconjuntos considerados pelo método LTS. Isso garante que os subconjuntos continuem a representar o conjunto de dados original. Assim, os modelos ajustados para o subconjunto, também explicam o conjunto.

Diante do exposto, este artigo tem como objetivo propor uma implementação inovadora do método LTS em partições da amostra, juntamente com o ponto de ruptura, como critério de seleção de valores que melhor representam o conjunto de dados para um ajuste do modelo não

linear Logístico. A partir disso, obtém-se um ajuste robusto para o referido modelo na descrição de curvas de crescimento.

2 MATERIAL E MÉTODOS

2.1 Conjunto de dados

Os dados longitudinais são gerados via software R, utilizando o método de Simulação de Monte Carlo (Apêndice B). Nesse procedimento, são gerados 200 valores simulados com base na aplicação da equação do modelo não linear Logístico. Dentre as parametrizações do referido modelo, considera-se aquela mais eficiente, de acordo com Fernandes (2015), apresentada na Equação 2.1.

$$y_i = \frac{\alpha}{1 + e^{\kappa(\beta - x_i)}} + \epsilon_i \quad (2.1)$$

em que: $i = 1, \dots, n$ se refere a cada valor dos dados, n é a quantidade de valores, y_i e x_i são os i -ésimos valores das variáveis dependente e independente, respectivamente, ϵ_i o desvio associado ao i -ésimo valor e α , β e κ os parâmetros do modelo. Para gerar os valores y , utiliza-se os seguintes valores para os parâmetros: $\alpha = 200$, $\beta = 50$ e $\kappa = 0,09$. Ressalta-se que a variável independente x é definida na Equação 2.2.

$$x = (\underbrace{5, \dots, 5}_{10}, \dots, \underbrace{100, \dots, 100}_{10}). \quad (2.2)$$

Com o objetivo de aprimorar a representatividade, acrescenta-se a cada valor y_i uma variável u , a qual segue uma distribuição normal, de modo que a dispersão dos dados aumente de acordo com a variável x . Os valores para os parâmetros de u , média e variância, são selecionados de modo a representar o incremento de variabilidade que ocorre naturalmente em curvas de crescimento. Especificamente:

- Se $i < 50$, então $u \sim N(6,1)$.
- Se $50 \leq i < 130$, então $u \sim N(9,4)$.
- Se $130 \leq i \leq 200$, então $u \sim N(12,9)$.

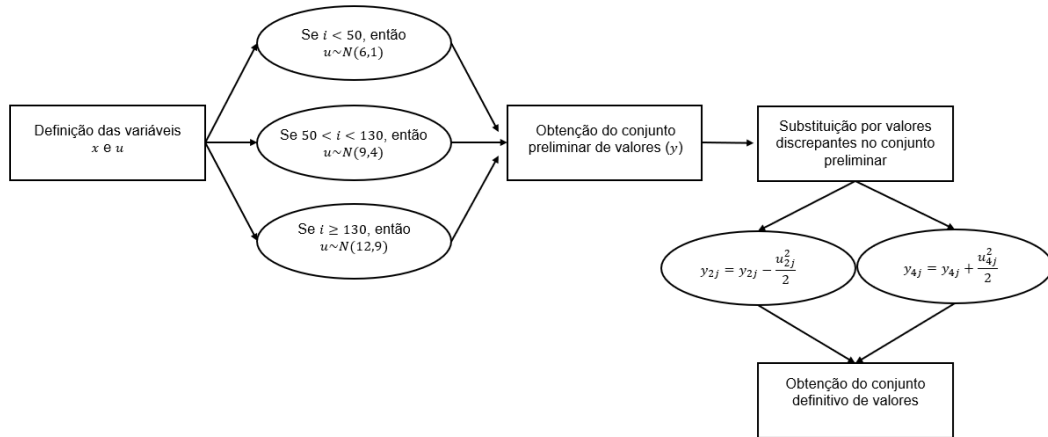
Para gerar outliers nesse conjunto, dado a variável u supracitada, define-se uma variável j , com $j \in \mathbb{N} \mid 1 \leq j \leq 50$, e um critério, escrito pela Equação 2.3.

$$\begin{cases} y_{2j}^* = y_{2j} - \frac{u_{2j}^2}{2} \\ y_{4j}^* = y_{4j} + \frac{u_{4j}^2}{2} \end{cases}, \quad (2.3)$$

em que y_{2j}^* e y_{4j}^* são os possíveis outliers inseridos nos dados simulados. Esse critério é definido tal que alguns dos valores gerados y sejam substituídos por possíveis outliers. Os índices $2j$ e $4j$ são definidos de modo a abranger valores por todo o conjunto de dados. Assim, adiciona-se $-\frac{u_{2j}^2}{2}$ e $\frac{u_{4j}^2}{2}$ a y_{2j} e y_{4j} , respectivamente, para diminuir os valores y_{2j} e aumentar os y_{4j} . Desse modo, outliers podem surgir em qualquer parte do conjunto, a depender da variável aleatória u .

Por fim, para prevenir valores negativos, considera-se a condição: se $y_i < 0$, então $y_i = |y_i|$. Vale ressaltar que valores negativos são evitados para melhor representatividade. O processo de simulação do conjunto de dados utilizado neste artigo está ilustrado na Figura 2.1.

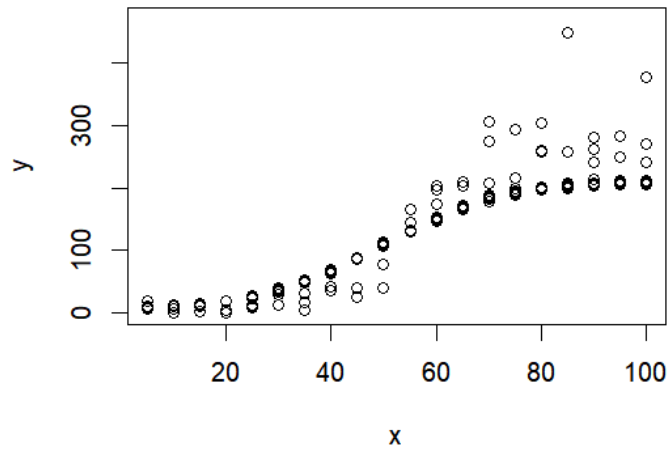
Figura 2.1 — Processo de simulação do conjunto de dados com valores discrepantes.



Fonte: Do autor.

O ajuste do modelo não linear Logístico é implementado ao utilizar a função ‘nls()’ do software R, em que o método iterativo utilizado para convergência é o algoritmo Gauss-Newton (Apêndice B). Os ajustes são estimados ao considerar os mesmos valores iniciais para gerar os dados (Figura 2.2): $\alpha = 200$, $\beta = 50$ e $\kappa = 0,09$.

Figura 2.2 — Dispersão do conjunto de dados simulados.



Fonte: Do autor.

2.2 Ponto de ruptura

O ponto de ruptura, proposto por Rousseeuw e Van Driessen (2006), pode ser interpretado pela proporção em que o subconjunto ainda representa o conjunto. Isso garante que não haja descaracterização dos dados em estudo. O ponto de ruptura h , com $h \in \mathbb{N}$, pode ser definido pela Equação 2.4.

$$h \geq \frac{3n + p + 1}{4}, \quad (2.4)$$

em que p é a quantidade de parâmetros do modelo a ser ajustado.

2.3 Least Trimmed Squares (LTS)

Sejam d os níveis da variável independente x , isto é, a quantidade de valores que esta assume, e q a quantidade de valores y por nível d , dado por $q = \frac{n}{p}$. Desse modo, define-se as partições do conjunto de dados, de acordo com a Equação 2.5.

$$A = A_1 \cup A_2 \cup \dots \cup A_p. \quad (2.5)$$

A r -ésima partição ($1 \leq r \leq d$), A_r , é composta por q valores de y . Assim, o s -ésimo elemento ($1 \leq s \leq q$) pertencente à A_r pode ser escrito por $y_{r,s}$. Por exemplo, o quarto elemento da quinta partição pode ser escrito por $y_{5,4}$, assim como o terceiro elemento da primeira partição é escrito por $y_{1,3}$.

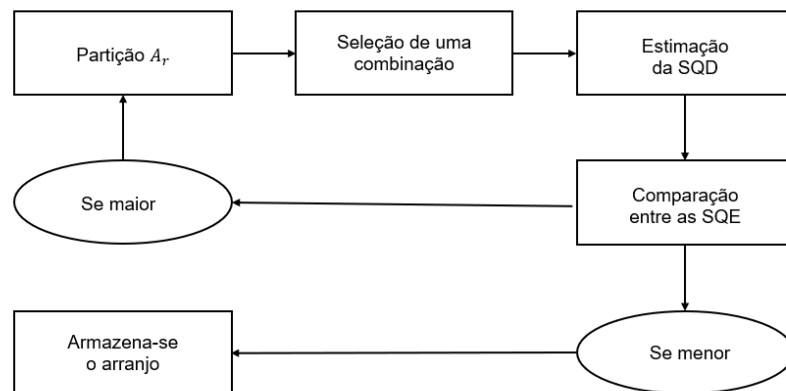
Nesse processo de simulação, tem-se que $n = 200$, $d = 20$ e $q = \frac{200}{20} = 10$. A quantidade de partições é definida de modo que as partições apresentem poucos pontos amostrais, diminuindo o custo computacional. Logo, há 20 partições com 10 elementos, $A = A_1 \cup A_2 \cup \dots \cup A_{20}$, escritas por $A_1 = \{y_{1,1}, y_{1,2}, \dots, y_{1,10}\}$, $A_2 = \{y_{2,1}, y_{2,2}, \dots, y_{2,10}\}$, $\dots$, $A_{20} = \{y_{20,1}, y_{20,2}, \dots, y_{20,10}\}$.

Diante do exposto, Least Trimmed Squares (LTS) é um método robusto e iterativo que seleciona um subconjunto dos dados de acordo com a soma de quadrados dos erros (SQE) (ROUSSEEUW; VAN DRIESSEN, 2006). Ao considerar cada uma das partições, combinações de tamanho h são tomados e suas respectivas SQE são estimadas e comparadas. As SQE são estimadas de acordo com a Equação 2.6, em que SQE_d e $\bar{y}_d$ são a SQE e média aritmética da combinação considerado na iteração da partição d , respectivamente.

$$SQE_d = \sum_{q=1}^h (y_{d,q} - \bar{y}_d)^2. \quad (2.6)$$

Por fim, cada partição é substituída por sua combinação que apresentou menor SQE (MANOEL et al., 2024). O procedimento é ilustrado na Figura 2.3.

Figura 2.3 — Processo de seleção das combinações pelo método LTS para estimação dos parâmetros do modelo Logístico.



Legenda: Soma de quadrados dos erros (SQE).

Fonte: Do autor.

Desse modo, após a implementação do LTS, o conjunto de dados é substituído por um subconjunto central que é formado pelas combinações supracitadas. Nesse processo, o método exclui os outliers, já que inflam a SQE. Portanto, o subconjunto central é o conjunto de dados a ser considerado para estimação dos parâmetros de interesse.

Para exemplificar, considere a primeira partição, $A_1 = \{y_{1,1}, y_{1,2}, \dots, y_{1,10}\}$. O ponto de ruptura, Equação 2.4, em que $m = 10$ e $p = 3$, é dado por

$$h \geq \frac{3m + p + 1}{4} \Rightarrow h \geq \frac{3 \cdot 10 + 3 + 1}{4} \Rightarrow h \geq \frac{34}{4} \Rightarrow h \geq 8,5.$$

Logo, combinações de $h = 9$ (pois $h \in \mathbb{N}$) são considerados para a estimação da SQE. Além disso, há $\binom{10}{9} = 10$ combinações possíveis. Assim, a SQE é estimada pela Equação 2.7, em que y_1, q é um elemento da partição A_1 e $\bar{y}_1$ a média aritmética de seus elementos.

$$SQE_1 = \sum_{q=1}^9 (y_{1,q} - \bar{y}_1)^2. \quad (2.7)$$

Ao final do procedimento, suponha que a combinação que apresenta menor SQE seja o seguinte $\{y_{1,1}, y_{1,2}, \dots, y_{1,9}\}$. Desse modo, o valor $y_{1,10}$ é desconsiderado para a estimação dos parâmetros de interesse, pois o mesmo aumenta a dispersão dos dados.

Portanto, o valor que tem maior incidência no acréscimo da SQE é excluído do conjunto. Assim, dos 10 valores por partição, 9 são considerados para o ajuste do modelo. Consequentemente, o subconjunto central utilizado para a estimação dos parâmetros do modelo não linear Logístico apresenta 180 valores, pois é formado pela união dos 20 subconjuntos que contêm 9 elementos cada.

2.4 Desvio médio absoluto (DMA)

Primeiramente, salienta-se que “ajuste convencional” é o ajuste do modelo não linear Logístico sem o uso de qualquer técnica robusta. Por outro lado, “ajuste robusto” é necessariamente o ajuste obtido a partir da metodologia proposta neste artigo. Nesse sentido, é coerente comparar os ajustes convencional e robusto. Para mensurar a precisão dos ajustes do modelo não linear Logístico é considerado o desvio médio absoluto (DMA). Em uma perspectiva comparativa entre modelos ajustados, o DMA oferece uma interpretação intuitiva da variância residual, além de apresentar maior tolerância a outliers (GORARD, 2015; LOPES et al., 2016).

O DMA pode ser obtido pela média aritmética das diferenças absolutas entre os valores simulados (y_i) e estimados ($\hat{y}_i$), de acordo com a Equação 2.8. Quanto menor o desvio, menor é o DMA. Logo, quando $y_i - \hat{y}_i \rightarrow 0$, tem-se que $DMA \rightarrow 0$. Assim, ao comparar dois ajustes, quanto menor o DMA, mais representativo é o modelo ajustado, pois apresenta desvios menores, ou ainda, menor variância residual.

$$DMA = \frac{1}{n} \cdot \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (2.8)$$

Ressalta-se que os ajustes convencional e robusto são ajustados em conjuntos de tamanhos diferentes, o que inviabiliza sua comparação direta pelo DMA. Nesse sentido, para o cálculo da estimativa do DMA para o ajuste robusto, que é ajustado em um conjunto central dos dados, considera-se como valor estimado ($\hat{y}$) os valores y_i preditos pelo modelo ajustado, com $n = 200$. Desse modo, é possível comparar os ajustes, pois as estimativas para o DMA são obtidas em um mesmo conjunto.

3 RESULTADOS

As estimativas para os parâmetros dos ajustes do modelo não linear Logístico estão apresentados na Tabela 3.1. Ressalta-se que os valores iniciais para gerar os dados são: $\alpha = 200$, $\beta = 50$ e $\kappa = 0,09$. Ambos os ajustes apresentam estimativas significativas, qualificando-os para representar o conjunto de dados. As estimativas para os parâmetros são mais superestimadas no ajuste convencional do que no ajuste robusto. Isso é justificado por maior presença de outliers nos dados em que se ajustou o ajuste convencional. Desse modo, há indícios de que a metodologia robusta foi eficiente na seleção de valores que melhor representa os dados.

Tabela 3.1 — Estimativas para os parâmetros dos ajustes do modelo não Linear Logístico convencional e robusto.

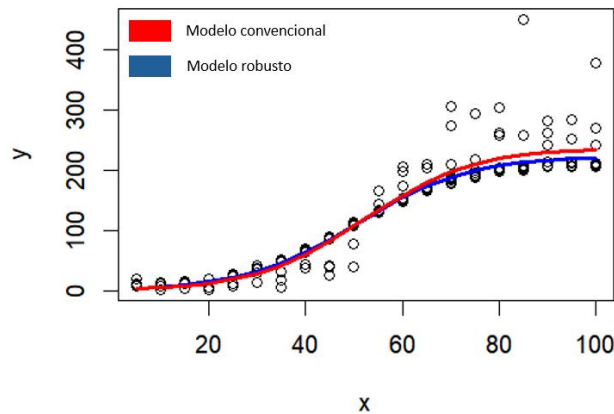
Ajuste convencional			
Parâmetros	Estimativas	$p - valor$	DMA
α	236,0000	$2 \cdot 10^{-16}$	17,1074
β	52,1800	$2 \cdot 10^{-16}$	
κ	0,0906	$2 \cdot 10^{-16}$	
Ajuste robusto (LTS)			
Parâmetros	Estimativas	$p - valor$	DMA
α	222,9000	$2 \cdot 10^{-16}$	14,2236
β	50,7200	$2 \cdot 10^{-16}$	
κ	0.0875	$2 \cdot 10^{-16}$	

Legenda: Desvio Médio Absoluto (DMA); Least Trimmed Squares (LTS).

Fonte: Do autor.

Na Tabela 3.1, tem-se que o DMA para o ajuste convencional é maior do que para o robusto ($17,1074 > 14,2236$). Logo, conclui-se que o ajuste robusto representa melhor o conjunto de dados do que o ajuste convencional, pois apresenta menores desvios. As curvas dos ajustes do modelo não linear Logístico estão apresentadas na Figura 3.1.

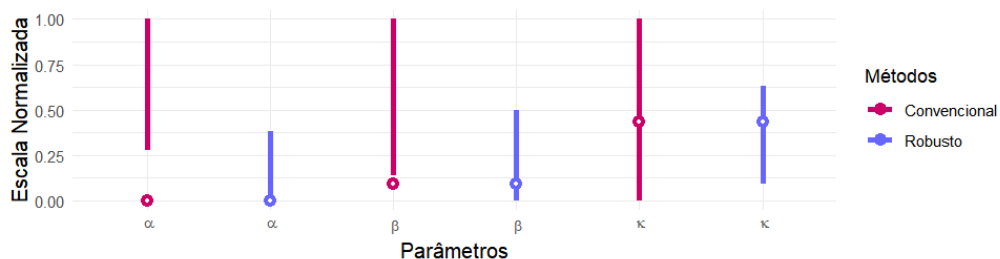
Figura 3.1 — Plots das curvas dos ajustes do modelo não linear Logístico convencional e robusto na descrição do conjunto de dados.



Fonte: Do autor.

Na Figura 3.2 são apresentados os intervalos de confiança normalizados considerando os valores iniciais para os parâmetros, $\alpha = 250$, $\beta = 50$ e $\kappa = 0,09$. Ressalta-se que as amplitudes dos IC para o ajuste robusto são menores do que para o ajuste convencional. Além disso, observa-se que os IC para o ajuste robusto, contemplam os valores iniciais dos parâmetros. Por outro lado, para o ajuste convencional, os IC não são representativos, já que não contêm seus respectivos valores iniciais para α e β . Esse resultado ressalta a mitigação dos efeitos que os outliers causam nas estimativas.

Figura 3.2 — Intervalos de confiança normalizados para as estimativas dos parâmetros do modelo não linear Logístico.



Fonte: Do autor.

Diante dos resultados, há indícios de que o método robusto proposto mitiga a influência que os outliers têm nas estimativas, pois apresentam menores diferenças em relação aos seus respectivos valores iniciais. Gu et al. (2020), Xie e Chen (2022) e Manoel et al. (2024) também

obtêm resultados promissores quanto à implementação do método LTS em suas aplicações particulares.

4 DISCUSSÕES

O desvio médio absoluto obtido ao implementar a metodologia robusta é menor do que o obtido no ajuste convencional. Isso se dá pela ausência de outliers, responsáveis pelo aumento da variabilidade no conjunto de dados. Assim, o método robusto LTS e o ponto de ruptura, como critério para determinar o tamanho das combinações, apresentam resultados promissores quando implementados em ajustes do modelo não linear logístico.

Desse modo, esse resultado complementa outros trabalhos. Gu et al. (2020) utilizam subamostras, semelhantes às partições definidas neste manuscrito, para implementar o método LTS com o método de mínimos quadrados totais em ajustes de dados de medição. O método LTS também é implementado por Xie e Chen (2022) às regressões de mínimos quadrados parciais, uma técnica de regressão linear. Manoel et al. (2024) implementam o LTS à regressão linear adaptativa na estimação de parâmetros para um modelo multinível.

Há outros métodos robustos implementados em modelos não lineares, como os M-estimator (COLOMBIER, 2009; ERTAŞ et al., 2015). No entanto, Lim et al. (2013) apontam que esses estimadores requerem condições específicas de aplicação. Isso pode atribuir complexidade à implementação dessas metodologias (LIM et al., 2012) e ao ajustes de modelos não lineares (SALVADOR et al., 2025). Nessa perspectiva, a metodologia robusta proposta é de simples implementação, pois não há pré-requisitos para a aplicação do LTS ou do ponto de ruptura, dispensando testes preliminares.

5 CONCLUSÃO

A metodologia robusta implementada por este manuscrito é promissora na mitigação dos efeitos de outliers em estimativas de parâmetros do modelo não linear logístico. Os resultados obtidos pela interpretação prática das estimativas dos parâmetros supracitados são menos viesados do que as estimativas obtidas de forma convencional. Em trabalhos futuros, pode-se implementar essa metodologia robusta em ajustes de outros modelos não lineares, como Brody, Gompertz e Von Bertalanffy.

REFERÊNCIAS

- AGGARWAL, C. C.; AGGARWAL, C. C. **Data classification**. Springer International Publishing, 2015.
- COLOMBIER, C. Growth effects of fiscal policies: an application of robust modified M-estimator. **Applied Economics**, v. 41, n. 7, p. 899-912, 2009.
- ERTAŞ, H.; TOKER, S.; KAÇIRANLAR, S. Robust two parameter ridge M-estimator for linear regression. **Journal of Applied Statistics**, v. 42, n. 7, p. 1490-1502, 2015.
- FERNANDES, T. J. et al. Parameterization effects in nonlinear models to describe growth curves. **Acta Scientiarum. Technology**, v. 37, n. 4, p. 397-402, 2015.
- GORARD, S. Introducing the mean absolute deviation ‘effect’ size. **International Journal of Research ; Method in Education**, v. 38, n. 2, p. 105-114, 2015.
- GU, T. et al. A robust moving total least-squares fitting method for measurement data. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 10, p. 7566-7573, 2020.
- LIM, C.; SEN, P. K.; PEDDADA, S. D. Accounting for uncertainty in heteroscedasticity in nonlinear regression. **Journal of Statistical Planning and Inference**, v. 142, n. 5, p. 1047-1062, 2012.
- LIM, C.; SEN, P. K.; PEDDADA, S. D. Robust nonlinear regression in applications. **Journal of the Indian Society of Agricultural Statistics. Indian Society of Agricultural Statistics**, v. 67, n. 2, p. 215, 2013.
- MANOEL, I. S. et al. Simulation of robust adaptive regression multi-level models for quality analysis of special coffees in cold storage. **Acta Scientiarum. Technology**, v. 46, n. 1, 2024.
- MARONNA, R. A. et al. **Robust statistics: theory and methods (with R)**. John Wiley ; Sons, 2019.
- MISCHAN, M. M.; DE PINHO, S. Z.; DE CARVALHO, L. R. Determinação de um ponto suficientemente próximo à assíntota em funções de crescimento não lineares. **Scientia Agricola**, v. 68, n. 1, p. 109-114, 2011.
- ROUSSEEuw, P. J.; HUBERT, M. Anomaly detection by robust statistics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 2, p. e1236, 2018.
- ROUSSEEuw, P. J.; HUBERT, M. Robust statistics for outlier detection. **Wiley interdisciplinary reviews: Data mining and knowledge discovery**, v. 1, n. 1, p. 73-79, 2011.
- ROUSSEEuw, P. J.; VAN DRIESSEN, K. Computing LTS regression for large data sets. **Data mining and knowledge discovery**, v. 12, p. 29-45, 2006.
- SALVADOR, R. C. et al. Nonlinear models in the comparison of nitrogen fertilizers applied to coffee. **Scientia Agricola**, v. 82, p. e20230238, 2025.

XIE, Z. et al. Partial least trimmed squares regression. **Chemometrics and Intelligent Laboratory Systems**, v. 221, p. 104486, 2022.

ARTIGO 3

Redigido de acordo com as normas da Universidade Federal de Lavras.

MÉTODOS ROBUSTOS IMPLEMENTADOS NO AJUSTE DO MODELO LOGÍSTICO PARA A DESCRIÇÃO DE CURVAS DE CRESCIMENTO

Iuri dos Santos Manoel^a, Tales Jesus Fernandes^a

^a Universidade Federal de Lavras (UFLA), 3037 CEP: 37203-202, Lavras, Minas Gerais, Brasil

RESUMO

Na Estatística, modelos podem ser utilizados para descrever a relação entre as variáveis em estudo. No entanto, as estimativas dos parâmetros do modelo podem ser tendenciosas quando há outliers no conjunto de dados. Assim, os resultados obtidos podem ser distorcidos. Diante desse problema, métodos robustos são uma solução para lidar com outliers, proporcionando resultados robustos e representativos. Os métodos Jackknife e Least Trimmed Squares (LTS) são metodologias robustas e versáteis que podem ser implementadas para mitigar o efeito de outliers em estimativas. Neste contexto, este manuscrito tem por objetivo aplicar duas metodologias robustas na descrição de dados de cunicultura. A primeira metodologia é uma adaptação do método Jackknife, implementado ao considerar a estimativa de curtose de suas reamostragens como critério de seleção. A segunda é o método LTS, implementado em partições da amostra. Ambas as metodologias consideram o ponto de ruptura como critério de delimitação dos tamanhos dos conjuntos centrais a serem considerados para o ajuste da curva. O modelo não linear Logístico é considerado para descrever a curva de crescimento, pois apresenta bons ajustes e parâmetros com interpretações práticas. Os dados utilizados para as implementações das metodologias supracitadas são de cunicultura, oriundos do Departamento de Zootecnia da Universidade Federal de Lavras. As metodologias apresentadas neste manuscrito são promissoras em mitigar os efeitos de outliers em estimativas para os parâmetros do modelo não linear logístico. O ajuste robusto Jackknife apresenta menores desvios do que o ajuste convencional, mas interpretações redundantes. Já o ajuste robusto LTS apresenta menores desvios do que os demais, o que sugere que este proporciona interpretações práticas dos parâmetros menos tendenciosas.

Palavras-chave: least trimmed squares; jackknife; ponto de ruptura; modelo logístico; cunicultura.

ABSTRACT

In Statistics, models can be used to describe the relationship between the variables under study. However, estimates of model parameters may be biased when there are outliers in the data set. Thus, the results obtained may be distorted. Faced with this problem, robust methods are a solution to deal with outliers, providing robust and representative results. The Jackknife and Least Trimmed Squares (LTS) methods are robust and versatile methodologies that can be implemented to mitigate the effect of outliers in estimates. In this context, this manuscript aims to apply two robust methodologies in the description of rabbit farming data. The first methodology is an adaptation of the Jackknife method, implemented by considering the kurtosis estimate of its resamplings as a selection criterion. The second is the LTS method, implemented in sample partitions. Both methodologies consider the break point as a criterion for delimiting the sizes of the central sets to be considered for curve fitting. The nonlinear logistic model is considered to describe the growth curve, as it presents good adjustments and parameters with practical interpretations. The data used for the implementation of the aforementioned methodologies are from rabbit farming, originating from the Department of Animal Science of the Federal University of Lavras. The methodologies presented in this manuscript are promising in mitigating the effects of outliers in estimates for the parameters of the nonlinear logistic model. The robust Jackknife adjustment presents smaller deviations than the conventional adjustment, but redundant interpretations. The robust LTS adjustment presents smaller deviations than the others, which suggests that it provides less biased practical interpretations of the parameters.

Keywords: least trimmed squares; jackknife; breakpoint; logistic model; rabbit farming.

1 INTRODUÇÃO

Fatores como erros de medição, erros de entrada de dados, comportamentos extremos ou eventos raros, podem dar origem a outliers (ROUSSEEuw; HUBERT, 2011). Em um conjunto de dados, observações são denominadas outliers se seus valores diferem do padrão geral dos dados. Esses valores podem distorcer inferências estatísticas sobre os dados analisados (HUBERT; ROUSSEEuw; SEGAERT, 2015).

Desse modo, métodos robustos são uma alternativa para identificar, mitigar ou eliminar outliers de um conjunto de dados (ROUSSEEuw; HUBERT, 2018). Entre a variedade de metodologias robustas, o método de reamostragem Jackknife e o método robusto Least Trimmed Squares (LTS). Ambos os métodos são iterativos e relativamente simples, pois dispensam testes preliminares para sua aplicação (ROUSSEEuw; VAN DRIESSEN, 2006; MANOEL et al., 2024a).

O método Jackknife é versátil e eficiente para mitigar a influência de outliers em estimativas ao selecionar uma ou mais observações a serem eliminadas dos dados (YANG; KOLASSA, 2021). Manoel et al. (2024a) propuseram uma metodologia robusta constituída pelo método Jackknife aplicado na descrição de um conjunto de dados simulados. Os autores verificaram que a metodologia posposta foi eficiente ao lidar com outliers para estimar os parâmetros do modelo não linear Logístico.

Um outro método também utilizado para lidar com outliers no processo de estimação de parâmetros é o método LTS, que determina o subconjunto amostral que apresenta menor estimativa para a soma de quadrados dos desvio (SQD) (ROUSSEEuw; VAN DRIESSEN, 2006; XIE et al., 2022). Desse modo, todas as combinações possíveis são verificadas possibilitando sua implementação em ajustes de modelos. Manoel et al. (2024b) aplicam o método LTS para mitigar a distorção que outliers provocam nas estimativas dos parâmetros de um modelo multinível.

Para a aplicação das metodologias robustas Jackknife e LTS, é considerado um conjunto de dados de cunicultura. Cunicultura é o ramo da Zootecnia que trata da criação de coelhos. Essa é uma atividade que apresenta pouco impacto ambiental, pois coelhos consomem pouca água, ração e espaço em comparação com outros tipos de criação (GIDENNE et al., 2022). Além disso, a cunicultura pode ser integrada com outros tipos de atividades agrícolas, proporcionando diversificação da produção (FERNANDES et al., 2022). Da cunicultura, o produtor pode obter carne, pele e pelos dos coelhos. Esses animais também podem ser destinados a lares como animais de estimação ou a laboratórios para fins científicos.

Diante do exposto, este manuscrito tem por objetivo utilizar as metodologias robustas supracitadas para mitigar possíveis influências de outliers na descrição do crescimento de coelhos mestiços. A primeira metodologia é o método de reamostragem Jackknife implementado ao considerar o menor módulo da estimativa para o coeficiente de curtose de um subconjunto. A segunda, é o método robusto LTS ao considerar partições do conjunto de dados. Ambas as metodologias são implementadas ao considerar o ponto de ruptura para definir o tamanho de cada subconjunto.

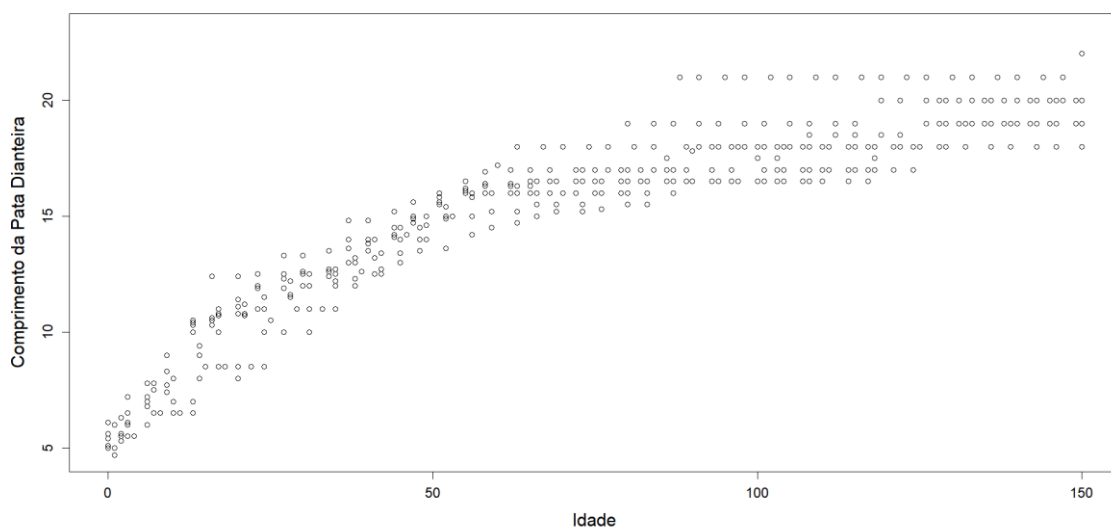
2 MATERIAL E MÉTODOS

2.1 Conjunto de dados

Os dados de cunicultura utilizados neste manuscrito são provenientes de um experimento realizado no setor de cunicultura do departamento de Zootecnia da Universidade Federal de Lavras. Os coelhos são mestiços das raças Gigante de Flandres (linha materna) e Nova Zelândia Branco (linha paterna). São observadas medidas corporais de 10 coelhos filhotes durante seus primeiros 150 dias, resultando em 440 pontos amostrais.

Para o interesse de implementar as metodologias robustas apresentadas neste manuscrito, foram utilizados dados do comprimento da pata dianteira (CPD) em função da idade dos animais. As unidades de medida são dias para a idade e centímetros para o CPD. O gráfico de dispersão do conjunto de dados está apresentado na Figura 2.1.

Figura 2.1 — Gráfico de dispersão da relação do comprimento da pata dianteira (centímetros) e a idade (dias) de coelhos mestiços.



Fonte: Do autor.

No conjunto de dados considerado, observa-se uma maior dispersão em coelhos para níveis maiores da variável idade. Assim, algum desses pontos pode prejudicar o ajuste de um modelo.

Para descrever a relação da idade (variável independente) e o CPD (variável dependente), considere o modelo não linear Logístico (Equação 2.1). Fernandes, Pereira e Muniz (2017) sugerem o modelo Logístico na descrição de curvas de crescimento, de acordo como é apresentado na Equação 2.1.

$$y_i = \frac{\alpha}{1 + e^{\kappa(\beta - x_i)}} + \epsilon_i \quad (2.1)$$

em que: $i = 1, \dots, n$ se refere a cada valor dos dados, n é o tamanho amostral, x_i e y_i são os i -ésimos valores das variáveis independente e dependente, respectivamente, ϵ_i o desvio associado ao i -ésimo valor observado, no qual se pressupõe que $\epsilon \sim N(0, \sigma^2)$,

Na Equação 2.1, destacam-se os parâmetros do modelo Logístico, sendo que suas interpretações práticas são: α é o valor da assíntota horizontal superior; β é a abscissa do ponto de inflexão, indicando que a característica em estudo passa de um crescimento acelerado para um crescimento desacelerado; κ é um índice associado ao crescimento da variável dependente.

2.2 Método de reamostragem Jackknife

O método Jackknife é um processo iterativo em que a cada iteração, uma ou mais observações são selecionadas e retiradas do conjunto amostral. As observações restantes, denominadas conjunto central, são utilizadas para o cálculo das estimativas de interesse. Desse modo, é possível restringir parte da amostra de acordo com algum critério.

Manoel et al. (2024) proporam o método Jackknife implementado ao considerar a estimativa do coeficiente de curtose como critério de seleção para obter um conjunto central. Um ponto de ruptura foi considerado de modo que o conjunto central ainda representasse o conjunto amostral (Equação 2.2).

$$h \geq \frac{3n + p + 1}{4}, \quad (2.2)$$

em que: h é o ponto de ruptura e p a quantidade de parâmetros do modelo a ser ajustado.

O conjunto amostral de cunicultura apresenta $n = 440$ observações. O modelo a ser ajustado é o Logístico, que apresenta $p = 3$ parâmetros (veja a Equação 2.1). Desse modo, o ponto de ruptura h é:

$$h \geq \frac{3 \cdot 440 + 3 + 1}{4} \Rightarrow h \geq 331.$$

Logo, pode-se considerar, no mínimo, 331 das 440 observações sem descaracterizar o conjunto de dados. Consequentemente, é possível implementar o método Jackknife para selecionar (ou eliminar) até $440 - 331 = 109$ observações de modo a obter um subconjunto com menor módulo da estimativa para o coeficiente de curtose.

O método Jackknife é relativamente simples e não exige testes preliminares. No entanto, para verificar todos os possíveis subconjuntos candidatos à conjunto central, é necessário um número grande de iterações, aumentando o custo computacional. Como são 440 valores, dos quais considera-se 331, há $\binom{440}{331} > 10^{100}$ combinações possíveis.

Diante desse grande número de combinações, o custo computacional aumenta consideravelmente. Para lidar com esse problema, essa metodologia é implementada ao considerar 10000 iterações. Desse modo, o conjunto central obtido pode não ser o que apresenta menor estimativa para o coeficiente de curtose entre as $\binom{440}{331}$ combinações possíveis, mas é um conjunto com menor estimativa do que a amostra.

2.3 Método robusto Least Trimmed Squares

O método LTS é um método iterativo que seleciona um conjunto central a partir da menor estimativa da soma de quadrados dos desvios (SQD). A estimativa para SQD é dada de acordo com a Equação 2.3.

$$SQD = \sum_{i=1}^n (y_i - \bar{y})^2, \quad (2.3)$$

em que: $\bar{y}$ é a média aritmética dos dados. Como outliers aumentam a dispersão dos dados, um conjunto que apresenta menor SQD tende a ter menor dispersão, isto é, menos (ou nenhum) outliers. Dessa forma, a seleção exclui os valores que são mais propícios a distorcer as estimativas de interesse.

No entanto, para implementar o LTS, propõe-se neste manuscrito considerar partições do conjuntos de dados. Desse modo, são definidos 44 partições com 10 observações cada ($44 \cdot 10 = 440$). Assim, a estrutura experimental é respeitada dado que são observados 10 coelhos ao longo de 150 dias.

O ponto de ruptura também é aplicado ao LTS, mas considerando as partições. Redefinindo o valor de n na Equação 2.2, com $m = n = 10$, tem-se:

$$h \geq \frac{3 \cdot 10 + 3 + 1}{4} \Rightarrow h \geq 8,5.$$

Entre 10 observações, pode-se considerar, no mínimo, 9 pontos para descrever os dados.

Diante do exposto, a cada iteração do método LTS, 9 de 10 valores são considerados para se estimar a SQD. Ressalta-se que no cálculo da estimativa da SQD (Equação 2.3), são considerados os valores e a média da partição considerada. A estimativa da SQD é comparada com a estimativa do próximo subconjunto de 9 valores. Aquele que apresentar menor estimativa é armazenado. Esse processo se estende até que todas as combinações possíveis sejam verificadas.

Felizmente, ao contrário do método Jackknife, a quantidade de combinações para o LTS é menor: $\binom{10}{9} = 10$. Isso se dá por considerar partições no processo, o que proporciona um baixo custo computacional. Desse modo, o conjunto central obtido pelo LTS sempre será aquele subconjunto que apresenta menor estimativa para a SQD.

2.4 Desvio médio absoluto (DMA)

O desvio médio absoluto (DMA) é utilizado em uma perspectiva comparativa entre modelos ajustados, pois oferece uma interpretação intuitiva da variância residual (GORARD, 2015; MANOEL et al., 2024a). O DMA pode ser obtido pela média aritmética das diferenças absolutas entre os valores simulados (y_i) e estimados ($\hat{y}_i$), de acordo com a Equação 2.4.

$$DMA = \frac{1}{n} \cdot \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (2.4)$$

Perceba que quanto menor o desvio, menor é o DMA. Desse modo, quando $y_i - \hat{y}_i \rightarrow 0$, tem-se que $DMA \rightarrow 0$. Portanto, o ajuste que apresenta menor DMA é considerado aquele que melhor descreve os dados, pois apresenta menores desvios.

3 RESULTADOS

3.1 Algumas considerações

Primeiramente, para simplificar o entendimento do leitor, considere que o ajuste do modelo Logístico obtido ao considerar a totalidade dos dados é denominado como ajuste

convencional. E ainda, os ajustes obtidos a partir dos subconjuntos obtidos pelas metodologias robustas apresentadas são denominados como ajuste Jackknife e ajuste LTS.

Os resultados apresentados nessa seção foram obtidos via software R (versão 4.2.2). O ajuste do modelo não linear Logístico foi implementado ao utilizar a função ‘nls()’, em que o método iterativo utilizado para convergência é o algoritmo Gauss-Newton (Apêndice C). Assumiu-se os seguintes valores iniciais para as estimativas dos parâmetros: $\alpha = 20$, $\beta = 50$ e $\kappa = 0,1$.

3.2 Ajuste convencional

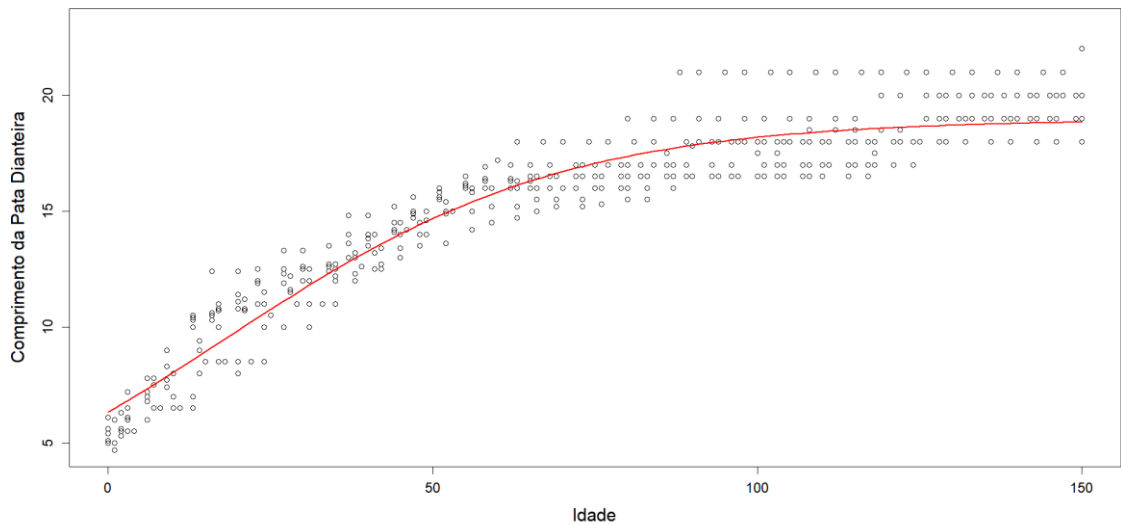
As estimativas obtidas para os parâmetros, assim como o DMA correspondente ao ajuste convencional, estão apresentadas na Tabela 3.1. O modelo foi bem ajustado, apresentando todas as estimativas para seus parâmetros significativas para a descrição dos dados. O gráfico da curva ajustada com a dispersão dos dados está apresentado na Figura 3.1.

Tabela 3.1 — Estimativas para os parâmetros do modelo Logístico e DMA associado ao ajuste convencional.

Parâmetros	Estimativas	p – valor	DMA
α	18,95	$< 10^{-16}$	0,8939
β	17,95	$< 10^{-16}$	
κ	0,0384	$< 10^{-16}$	

Legenda: Desvio médio absoluto (DMA).
Fonte: Do autor.

Figura 3.1 — Gráfico da curva do modelo Logístico ajustado pelo método convencional na descrição do comprimento da pata dianteira em função da idade de coelhos mestiços.

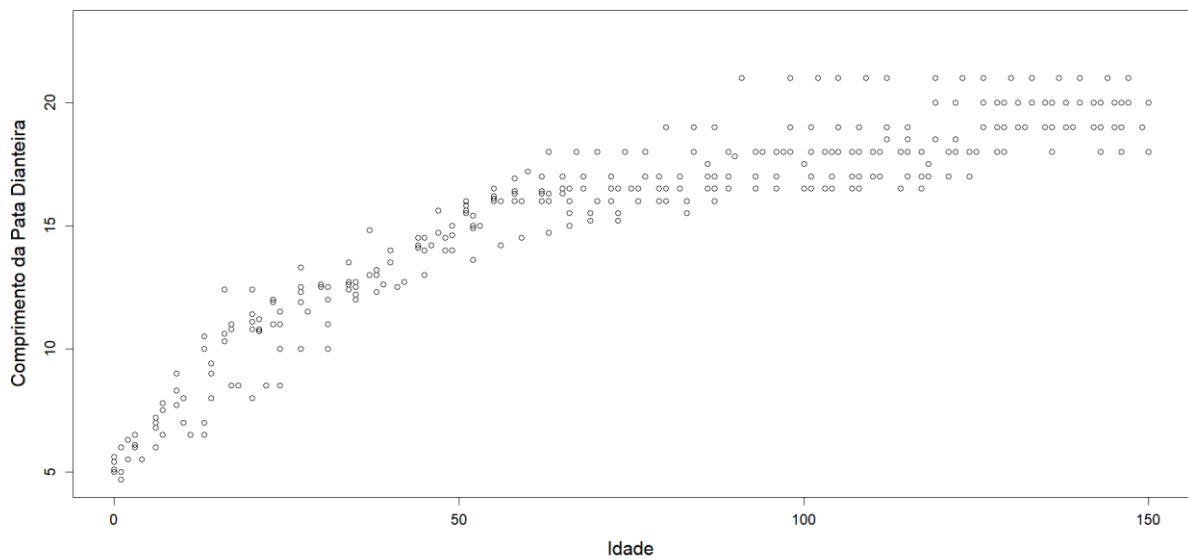


Fonte: Do autor.

3.3 Ajuste robusto Jackknife

A implementação do método de reamostragem Jackknife foi bem-sucedida. O gráfico de dispersão do conjunto central (331 observações) obtido por essa metodologia está apresentado na Figura 3.2. Pontos aleatórios foram removidos do conjunto de dados, de modo a formar um subconjunto (conjunto central) com menor módulo da estimativa para o coeficiente de curtose do que o conjunto original dos dados.

Figura 3.2 — Gráfico de dispersão do conjunto central obtido ao implementar o método Jackknife em dados de cunicultura.



Fonte: Do autor.

Ressalta-se que o conjunto central apresentado pode não ser o que apresenta menor módulo entre as $\binom{440}{331}$ combinações possíveis, pois não foi considerado todas as possibilidades, mas é o mais adequado entre as 10000 iterações consideradas. O módulo da estimativa para o coeficiente de curtose dos dados originais é de $3,3536 \times 10^{-2}$, enquanto para o conjunto central, obtém-se $1,1213 \times 10^{-4}$.

As estimativas para os parâmetros do modelo Logístico ao considerar o conjunto central estão apresentadas na Tabela 3.2. Para efeito de comparação, as estimativas obtidas pelo ajuste convencional (Tabela 3.1) estão também apresentadas na Tabela 3.2. O ajuste apresenta todas as estimativas significativas para a descrição dos dados. Além disso, o ajuste Jackknife apresenta menor DMA do o ajuste convencional. Isso sugere que esse ajuste é mais adequado para descrever a relação entre a idade e o CPD.

Tabela 3.2 — Estimativas para os parâmetros do modelo Logístico e os DMA associados aos ajustes convencional e robusto Jackknife.

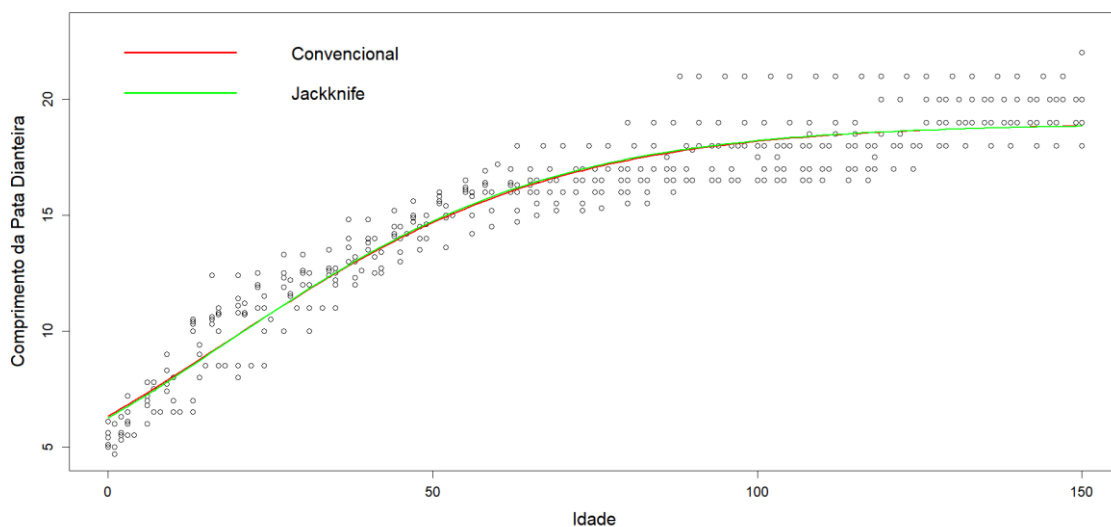
Ajuste convencional			
Parâmetros	Estimativas	$p - valor$	DMA
α	18,9558	$< 10^{-16}$	0,8939
β	17,9591	$< 10^{-16}$	
κ	0,0384	$< 10^{-16}$	
Ajuste Robusto Jackknife			
Parâmetros	Estimativas	$p - valor$	DMA
α	18,9336	$< 10^{-16}$	0,8867
β	18,0443	$< 10^{-16}$	
κ	0,0393	$< 10^{-16}$	

Legenda: Desvio médio absoluto (DMA).

Fonte: Do autor.

No entanto, os valores para as estimativas dos parâmetros para o ajuste robusto Jackknife são semelhantes das apresentadas para o ajuste convencional. Desse modo, os gráficos das curvas dos modelos supracitados se sobrepõem ao longo do eixo horizontal, tornando-os redundantes. As curvas estão apresentadas na Figura 3.3.

Figura 3.3 — Gráfico das curvas ajustadas pelo ajuste convencional e ajuste robusto Jackknife na descrição do comprimento da pata dianteira em função da idade de coelhos mestiços.

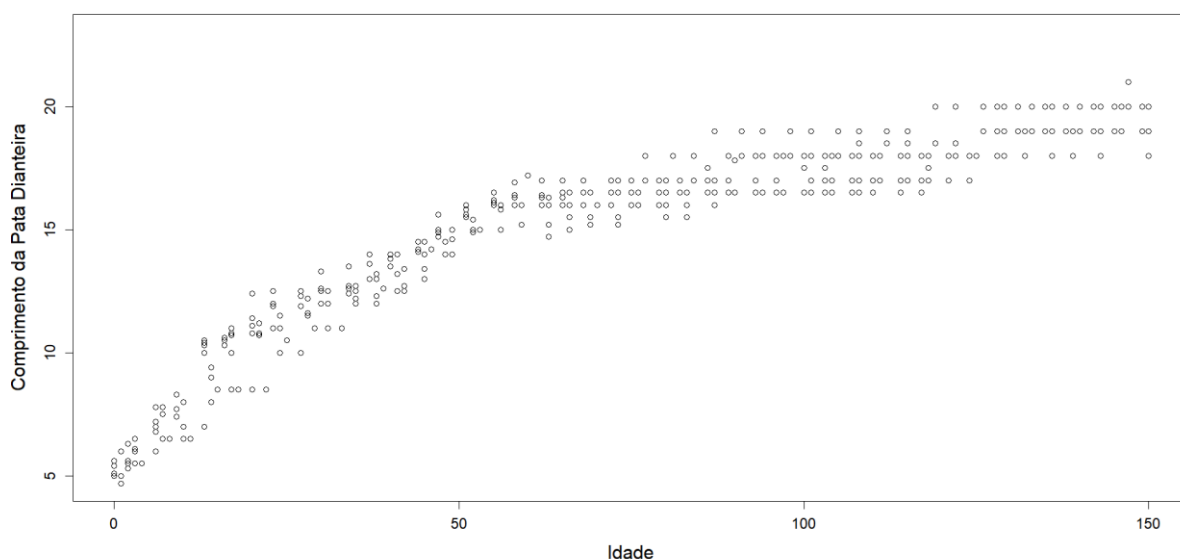


Fonte: Do autor.

3.4 Ajuste robusto LTS

Para implementar o método LTS o conjunto central é composto ao considerar 396 observações. Ressalta-se que esse valor é obtido ao selecionar 9 de 10 observações em cada uma das 44 partições, sendo essas definidas de acordo com os níveis, em ordem crescente, da variável independente. O conjunto central obtido ao implementar o método LTS está apresentado na Figura 3.4.

Figura 3.4 — Gráfico de dispersão do conjunto central obtido ao implementar o método Least Trimmed Squares aos dados de cunicultura.



Fonte: Do autor.

As estimativas para os parâmetros do modelo Logístico ao considerar o conjunto central para o ajuste, assim como seu DMA, estão apresentados na Tabela 3.3. Para efeito de comparação, as estimativas obtidas pelo ajuste convencional (Tabela 3.1) estão também apresentadas na Tabela 3.3. Todas as estimativas obtidas no ajuste LTS são significativas.

Tabela 3.3 — Estimativas para os parâmetros do modelo Logístico e os DMA associados aos ajustes convencional e Least Trimmed Squares na descrição do comprimento da pata dianteira em função da idade de coelhos mestiços.

em função da idade de colmos amostrados.			
Ajuste convencional			
Parâmetros	Estimativas	p – valor	DMA
α	18,9558	$< 10^{-16}$	0,8939
β	17,9591	$< 10^{-16}$	
κ	0,0384	$< 10^{-16}$	
Ajuste Robusto LTS			
Parâmetros	Estimativas	p – valor	DMA
α	18,6184	$< 10^{-16}$	0,7647
β	16,9195	$< 10^{-16}$	
κ	0,0397	$< 10^{-16}$	

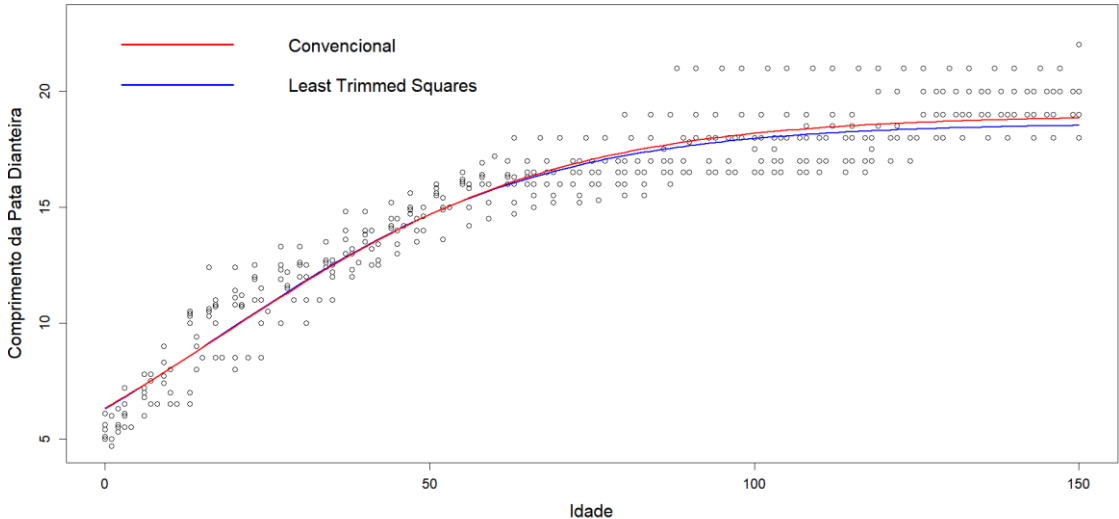
Legenda: Desvio médio absoluto (DMA).

Fonte: Do autor.

A diferença entre os modelos ajustados apresentados na Tabela 3.3 se dá pela presença de outliers. Observe que na Figura 3.4 há menos valores acima da maioria dos dados, diferente do gráfico de dispersão dos dados originais (Figura 2.1). Salienta-se ainda que o ajuste robusto LTS apresenta menor DMA que o ajuste convencional. Isso indica que o ajuste robusto é mais

adequado para descrever a relação entre as variáveis em estudo do que o ajuste convencional. O gráfico com as curvas dos ajustes supracitados está apresentado na Figura 3.5.

Figura 3.5 — Gráfico das curvas ajustadas pelo ajuste convencional e ajuste robusto Least Trimmed Squares na descrição do comprimento da pata dianteira em função da idade de coelhos mestiços.



Fonte: Do autor.

3.5 Ajustes Robustos

Na Tabela 3.4, estão apresentadas as estimativas para os parâmetros dos ajustes convencional, Jackknife e LTS, assim como seus respectivos DMA.

Tabela 3.4 — Estimativas para os parâmetros do modelo Logístico e os DMA associados aos ajustes convencional, Jackknife e Least Trimmed Squares.

Ajuste convencional			
Parâmetros	Estimativas	p – valor	DMA
α	18,9558	$< 10^{-16}$	0,8939
β	17,9591	$< 10^{-16}$	
κ	0,0384	$< 10^{-16}$	
Ajuste Robusto Jackknife			
Parâmetros	Estimativas	p – valor	DMA
α	18,9336	$< 10^{-16}$	0,8867
β	18,0443	$< 10^{-16}$	
κ	0,0393	$< 10^{-16}$	
Ajuste Robusto LTS			
Parâmetros	Estimativas	p – valor	DMA
α	18,6184	$< 10^{-16}$	0,7647
β	16,9195	$< 10^{-16}$	
κ	0,0397	$< 10^{-16}$	

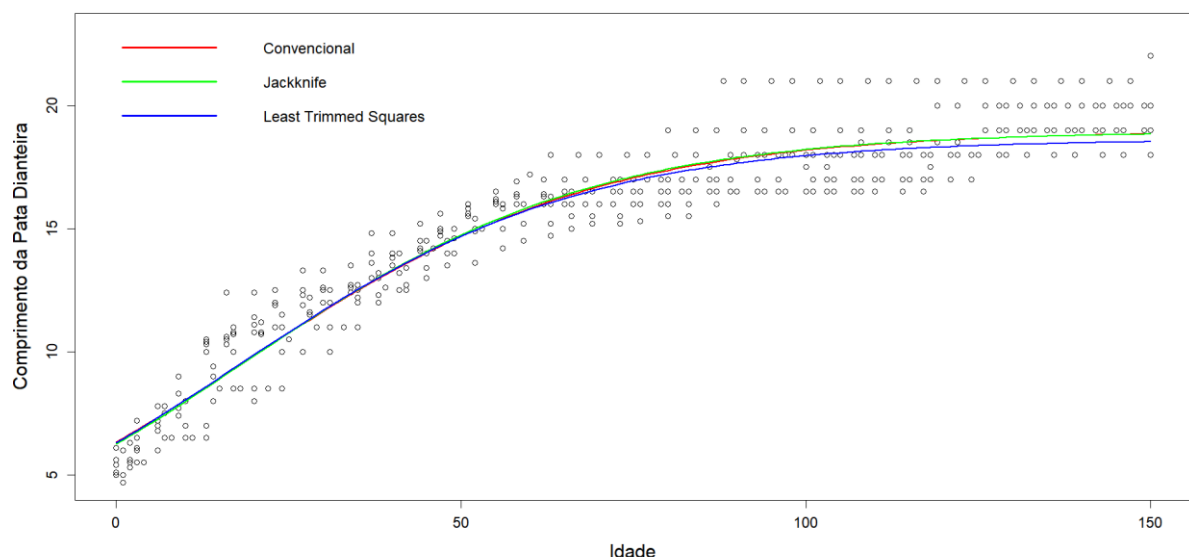
Legenda: Desvio médio absoluto (DMA).
Fonte: Do autor.

Os ajustes convencional e Jackknife proporcionam interpretações semelhantes quanto aos valores das estimativas para os parâmetros do modelo Logístico. Em 18 dias o comprimento das patas dianteiras dos coelhos apresenta metade do valor assintótico, $\frac{18,95}{2} = 9,475$ cm. Nos outros 132 dias, o crescimento é desacelerado gradualmente, até que o CPD atinja seu tamanho adulto, que é de 18,95 cm.

Para o ajuste LTS, o CPD máximo esperado é de 18,61 cm, enquanto os coelhos atingem metade dessa medida, 9,305 cm, em aproximadamente 17 dias. Desse modo, em contraste com os ajustes convencional e Jackknife, o ajuste LTS proporciona uma diferença de um dia para o início da redução da taxa de crescimento das patas dianteiras dos coelhos. Além disso, o DMA para o LTS é o menor (Tabela 3.4), o que sugere que esse ajuste é o mais adequado para explicar a relação entre as variáveis dos dados de cunicultura.

Na Figura 3.6 estão apresentados os gráficos das curvas ajustadas pelos métodos convencional, Jackknife e LTS. Na Tabela 3.4, observe que o ajuste LTS destoa dos outros dois ajustes quanto a estimativa para o parâmetro β . Como consequência, a curva ajustada pela metodologia LTS difere das demais quando a curva de crescimento tende ao valor adulto.

Figura 3.6 — Gráfico das curvas ajustadas pelo ajuste convencional e ajustes robustos Jackknife e Least Trimmed Squares na descrição do comprimento da pata dianteira em função da idade de coelhos mestiços.



Fonte: Do autor.

4 DISCUSSÃO

Na descrição do conjunto de dados de cunicultura, os métodos Jackknife e LTS são promissores na mitigação dos efeitos que outliers causam nas estimativas dos parâmetros do modelo não linear Logístico. Os ajustes obtidos apresentam um DMA menor do que o ajuste convencional (Tabela 3.4). Isso sugere que ambos os ajustes robustos são mais adequados do que o ajuste convencional para a descrição do conjunto de dados.

O ajuste obtido ao implementar o método Jackknife é semelhante ao ajuste convencional (Figura 3.3). Assim, mesmo que o método Jackknife apresente menores desvios, as interpretações de seus parâmetros são semelhantes as interpretações obtidas pelo ajuste convencional. Esse resultado pouco expressivo pode ter sido devido a diminuição no número de iterações realizadas.

Nesse contexto, aumentar a quantidade de iterações pode fazer com que o método apresente resultados ainda mais promissores. No entanto, uma abordagem mais eficiente requer maior custo computacional, como apontam Manoel et al. (2024a). Além disso, ressalta-se que a princípio o método de reamostragem Jackknife é indicado a amostras relativamente pequenas, o que não é o caso abordado neste trabalho (MILLER, 1974; WANG; YU, 2020).

De qualquer forma, esse é um resultado promissor para metodologias robustas implementadas a ajustes do modelo Logístico na descrição de curvas de crescimento. Isso complementa o trabalho de Manoel et al. (2024a) que propõem essa metodologia aplicada em ajuste do modelo Logístico na descrição de dados simulados. E ainda, esse resultado corrobora para com a ideia de que o método de reamostragem Jackknife é uma técnica robusta e versátil para a estimação de parâmetros de interesse, entrando em concordância com outros trabalhos, como Wang e Yu (2020) e Yang Kolassa (2021).

O ajuste robusto LTS sugere uma correção em relação ao ajuste convencional (Figura 3.5). As estimativas para o parâmetro β apresentam uma diferença de um dia para o ponto em que ocorre a redução da taxa de crescimento das patas dos coelhos. Além disso, esse ajuste apresenta o menor DMA em relação aos ajustes convencional e Jackknife. Desse modo, o ajuste robusto LTS é o mais adequado entre os considerados para descrever os dados considerados.

O resultado promissor referente ao ajuste LTS obtido neste trabalho complementa outros estudos, ressaltando a eficiência e versatilidade desse método. Manoel et al. (2024b) implementam o método LTS em ajuste de um modelo multinível. Gu et al. (2020) utilizam subamostras para implementar o método LTS com o método de mínimos quadrados totais na

descrição de dados de medição. Xie et al. (2022) implementam o LTS às regressões de mínimos quadrados parciais.

Salienta-se que as metodologias robustas apresentadas não apresentam pré-requisitos para suas aplicações. Isso não atribui complexidade em suas implementações, pois dispensam testes preliminares. O custo computacional é relativamente baixo, apesar de que o método Jackknife pode exigir maior trabalho computacional para se obter resultados mais expressivos. Entretanto, esse problema pode ser evitado ao limitar o número de iterações, como proposto neste manuscrito. O método LTS tende a exigir menor trabalho computacional de acordo com o tamanho de suas partições.

5 CONCLUSÃO

As metodologias robustas apresentadas neste manuscrito são promissoras na mitigação dos efeitos que outliers causam em estimativas dos parâmetros do modelo não linear Logístico. Apesar de apresentar menores desvios, o ajuste Jackknife apresenta pouca diferença em relação ao ajuste convencional. Por outro lado, o ajuste LTS apresenta estimativas diferentes em relação aos parâmetros do modelo Logístico. Além disso, as partições implementadas junto ao método LTS diminuem expressivamente seu custo computacional, tornando a metodologia mais eficiente.

O ponto de ruptura aplicado tanto no conjunto amostral, quanto nas partições para implementação do método LTS, é adequado para delimitar o tamanho do conjunto central de modo a não descaracterizar a amostra. Em trabalhos futuros, essas metodologias podem ser implementadas em ajustes de outros modelo não lineares, como os modelos Brody, Gompertz e Von Bertalanffy.

REFERÊNCIAS

- FERNANDES, F. A. et al. Cross-sectional and longitudinal method for describing growth curve of rabbits. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 74, p. 731-740, 2022.
- FERNANDES, T. J.; PEREIRA, A. A.; MUNIZ, J. A. Double sigmoidal models describing the growth of coffee berries. **Ciência Rural**, v. 47, p. e20160646, 2017.
- GIDENNE, T. et al. Cuniculture au pâturage et sous certification Agriculture Biologique en France: fonctionnement des systèmes, performances et règlementation. **INRAE Productions Animales**, v. 35, n. 3, p. 201-216, 2022.
- GORARD, S. Introducing the mean absolute deviation 'effect' size. **International Journal of Research & Method in Education**, v. 38, n. 2, p. 105-114, 2015.
- GU, T. et al. A robust moving total least-squares fitting method for measurement data. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 10, p. 7566-7573, 2020.
- HUBERT, M.; ROUSSEEUW, P. J.; SEGAERT, P. **Multivariate functional outlier detection. Statistical Methods & Applications**, v. 24, n. 2, p. 177-202, 2015.
- MANOEL, I. S. et al. Método Jackknife aplicado em ajustes de modelos não lineares em presença de outliers no conjunto de dados. **Caderno Pedagógico**, v. 21, n. 10, p. e8938-e8938, 2024a.
- MANOEL, I. S. et al. Simulation of robust adaptive regression multi-level models for quality analysis of special coffees in cold storage. **Acta Scientiarum. Technology**, v. 46, n. 1, 2024b.
- MILLER, R. G. The jackknife-a review. **Biometrika**, v. 61, n. 1, p. 1-15, 1974.
- ROUSSEEUW, P. J.; HUBERT, M. Anomaly detection by robust statistics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 2, p. e1236, 2018.
- ROUSSEEUW, P. J.; HUBERT, M. Robust statistics for outlier detection. **Wiley interdisciplinary reviews: Data mining and knowledge discovery**, v. 1, n. 1, p. 73-79, 2011.
- ROUSSEEUW, P. J.; VAN DRIESSEN, K. Computing LTS regression for large data sets. **Data mining and knowledge discovery**, v. 12, p. 29-45, 2006.
- WANG, L.; YU, F. Jackknife resampling parameter estimation method for weighted total least squares. **Communications in Statistics-Theory and Methods**, v. 49, n. 23, p. 5810-5828, 2020.
- XIE, Z. et al. Partial least trimmed squares regression. **Chemometrics and Intelligent Laboratory Systems**, v. 221, p. 104486, 2022.
- YANG, J.; KOLASSA, J. E. **The Impact of Application of the Jackknife to the Sample Median. The American Statistician**, v. 75, n. 4, p. 445-449, 2021.

TERCEIRA PARTE

CONSIDERAÇÕES FINAIS

Outliers podem ser um problema na análise de dados, especialmente em ajustes de modelos, já que esses valores podem distorcer as estimativas dos parâmetros. Desse modo, esta tese propõe metodologias robustas de relativamente simples implementação, pois os métodos não requerem testes preliminares ou condições específicas. Essas técnicas abrangem o ajuste de modelos não lineares, uma área que carece de metodologias robustas simples. Dessa forma, o conhecimento construído nesta tese contribui com a ciência no sentido de desenvolver, discutir e aplicar metodologias estatísticas para o ajuste de modelos não lineares.

As metodologias robustas utilizadas, o método Jackknife e o Least Trimmed Squares (LTS), são promissores quando implementadas para mitigar os efeitos de outliers em ajustes do modelo não linear Logístico na descrição de curvas de crescimento. O ponto de ruptura é um critério adequado para determinar o tamanho do subconjunto a ser considerado para descrever os dados. Além disso, o desvio médio absoluto (DMA) foi eficiente na comparação dos ajustes obtidos pelas metodologias propostas.

Entre os métodos Jackknife e LTS, o segundo apresentou melhores resultados. Apesar do ajuste robusto Jackknife apresentar menores desvios do que o ajuste convencional, ambos os modelos são semelhantes. Por outro lado, o método LTS apresenta resultados expressivos quanto a seleção de um subconjunto livre de outliers, além de proporcionar um ajuste robusto visualmente melhor do que o convencional.

APÊNDICE A

O script utilizado no Artigo 1 é detalhado a seguir.

PACOTES UTILIZADOS

```
library(qpcR)
```

```
library(fBasics)
```

```
set.seed(1)
```

```
n <- 200 # tamanho amostral
```

```
x <- c(rep(5, 10), rep(10, 10), rep(15, 10), rep(20, 10), rep(25, 10), rep(30, 10),  
      rep(35, 10), rep(40, 10), rep(45, 10), rep(50, 10), rep(55, 10), rep(60, 10),  
      rep(65, 10), rep(70, 10), rep(75, 10), rep(80, 10), rep(85, 10), rep(90, 10),  
      rep(95, 10), rep(100, 10)) # variável independente
```

```
u <- c() # variável de dispersão
```

```
for( i in 1:n ) {
```

```
  if( i < 50 ) {
```

```
    u[i] <- runif( 1, min = -3, max = 3 )
```

```
  } else{
```

```
    if( i < 130 ) {
```

```
      u[i] <- runif( 1, min = -7, max = 7)
```

```
    } else{
```

```
      u[i] <- runif( 1, min = -14, max = 14)
```

```
    }
```

```
  }
```

```
}
```

```
a <- 200; b <- 50; k <- 0.09 # parâmetros do modelo
```

```
# GERANDO O CONJUNTO DE DADOS
```

```
y <- c() # variável dependente
```

```
for( i in 1:n ) {
```

```
  y[i] <- a/(1+exp(k*(b-x[i]))) + u[i] + 50
```

```
  if( y[i] < 0){
```

```
    y[i] <- sqrt((y[i])^2)
```

```
  }
```

```
}
```

```

plot(x, y, ylim = c(0, 300)); max(y); min(y) # plot dos dados sem valores discrepantes
curtose.dados <- kurtosis(y) + 3; curtose.dados[[1]]; min(y); max(y)

#####

##### AJUSTE DO MODELO LOGÍSTICO PARA #####

#### OS DADOS SEM VALORES DISCREPANTES ####

#####

dados <- data.frame(x, y)
mod <- nls(y ~ a/(1+exp(k*(b-x))),
            start = c(a = 270, b = 50, k = 0.09), data = dados)
summary(mod)
Rsqr(mod)

# CONVERTENDO ALGUNS PONTOS EM VALORES DISCREPANTES

# baixo aumento da curtose
y[5] <- 100; y[15] <- 135; y[25] <- 120; y[35] <- 110; y[45] <- 100; y[55] <- 110

# baixo aumento da curtose
y[150] <- 160; y[160] <- 140; y[170] <- 170; y[180] <- 140; y[190] <- 100; y[200] <- 140

# alto aumento da curtose
y[141] <- 270; y[151] <- 330; y[161] <- 410; y[171] <- 460; y[181] <- 450; y[191] <- 500

# alto aumento da curtose
y[145] <- 300; y[155] <- 380; y[165] <- 450; y[175] <- 440; y[185] <- 430; y[195] <- 450
plot(x, y, ylim = c(0,520)); max(y); min(y) # plot dos dados com valores discrepantes
curtose.dados <- kurtosis(y) + 3; curtose.dados[[1]]; min(y); max(y)

y.dma <- y

#####

##### AJUSTE DO MODELO LOGÍSTICO PARA #####

#### OS DADOS COM VALORES DISCREPANTES ####

#####

dados <- data.frame(x, y)
mod.log <- nls(y ~ a/(1+exp(k*(b-x))),
               start = c(a = 270, b = 50, k = 0.09), data = dados)
summary(mod.log)
Rsqr(mod.log)

### DESVIO MÉDIO ABSOLUTO DO MODELO LOGÍSTICO

desvio0 <- c()

```

```

y.pred0 <- coef(mod.log)[1]/(1+exp(coef(mod.log)[3]*(coef(mod.log)[2]-x)))
desvio0 <- sqrt((y-y.pred0)^2)
(dma.log <- sum(desvio0)/n)
# ESTIMATIVAS JACKKNIFE - ABORDAGEM 1
alpha.jkk <- c(); beta.jkk <- c(); kappa.jkk <- c();
curtose <- c(); dados.jkk <- c()
set.seed(3)
for (i in 1:100000) {
  selecao <- sample(1:n, 151)
  dados.jkk <- dados1[selecao, ]
  log.jkk <- nls(y ~ a/(1+exp(k*(b-x))),
    start = c(a = 275, b = 50, k = 0.09), data = dados.jkk)
  alpha.jkk[i] <- coef(log.jkk)[1]
  beta.jkk[i] <- coef(log.jkk)[2]
  kappa.jkk[i] <- coef(log.jkk)[3]
  curtose[i] <- kurtosis(dados.jkk)[[1]] + 3
}
(a.jkk <- median(alpha.jkk)) # ESTIMATIVA DE ALPHA
(k.jkk <- median(kappa.jkk)) # ESTIMATIVA DE KAPPA
(b.jkk <- median(beta.jkk)) # ESTIMATIVA DE BETA
(curtose.jkk <- median(curtose))
# ESTIMATIVA DA CURTOSE DOS CONJUNTOS SELECIONADOS PELO JACKKNIFE
# II: DESVIO MÉDIO ABSOLUTO DO MODELO LOGÍSTICO PELO JACKKNIFE
y.pred <- c(); desvio <- c()
y.pred <- a.jkk/(1+exp(k.jkk*(b.jkk-x)))
desvio <- sqrt((y-y.pred)^2)
(dma.jkk <- sum(desvio)/n)
# ESTIMATIVAS JACKKNIFE
curtose <- c(); dados.jkk <- c()
curtose[0] <- 3
dados2 <- c()
a <- 0
b <- 5
set.seed(2)

```

```

for (i in 1:100000) {
  selecao <- sample(1:n, 151)
  dados2 <- dados[selecao, ]
  curtose[i] <- kurtosis(dados2$y)[[1]] + 3

  if(curtose[i] < b){
    b <- curtose[i]
    dados.jkk <- dados2
    a <- i
  }
}
a
b
kurtosis(dados.jkk$y)[[1]] + 3
plot(dados.jkk, ylim = c(0,520))
boxplot(dados.jkk$y, ylab = "y", ylim = c(0, 520), col = "white")
mod.rob <- nls(y ~ a/(1+exp(k*(b-x))),
               start = c(a = 275, b = 50, k = 0.09), data = dados.jkk)
summary(mod.rob)
# II: DESVIO MÉDIO ABSOLUTO DO MODELO LOGÍSTICO PELO JACKKNIFE
y.pred <- c(); desvio <- c()
y.pred <- coef(mod.rob)[1]/(1+exp(coef(mod.rob)[3]*(coef(mod.rob)[2]-x)))
desvio <- sqrt((y-y.pred)^2)
(dma.jkk <- sum(desvio)/n)
y.mqo <- coef(mod.log)[1]/(1+exp(coef(mod.log)[3]*(coef(mod.log)[2]-x)))
y.rob <- coef(mod.rob)[1]/(1+exp(coef(mod.rob)[3]*(coef(mod.rob)[2]-x)))
y.adp <- coef(mod.adpt)[1]/(1+exp(coef(mod.adpt)[3]*(coef(mod.adpt)[2]-x)))
plot(x,y.dma, ylim = c(0, 520), ylab = "y")
lines(x,y.mqo, col="blue",lwd=3)
lines(x,y.rob, col="red",lwd=3)
lines(x,y.adp, col="green",lwd=3)

```

APÊNDICE B

O script utilizado no Artigo 2 é detalhado a seguir.

PACOTES UTILIZADOS

```
library(qpcR)
```

```
library(fBasics)
```

```
set.seed(123)
```

```
n <- 200 # tamanho amostral
```

```
x <- c(rep(5, 10), rep(10, 10), rep(15, 10), rep(20, 10), rep(25, 10), rep(30, 10),  
      rep(35, 10), rep(40, 10), rep(45, 10), rep(50, 10), rep(55, 10), rep(60, 10),  
      rep(65, 10), rep(70, 10), rep(75, 10), rep(80, 10), rep(85, 10), rep(90, 10),  
      rep(95, 10), rep(100, 10)) # variável independente
```

```
u <- c() # variável de dispersão
```

```
for( i in 1:n ) {
```

```
  if( i < 50 ) {
```

```
    u[i] <- rnorm( 1, mean = 6, sd = 1 )
```

```
  } else{
```

```
    if( i < 130 ) {
```

```
      u[i] <- rnorm( 1, mean = 9, sd = 2 )
```

```
    } else{
```

```
      u[i] <- rnorm( 1, mean = 12, sd = 3 )
```

```
    }
```

```
  }
```

```
}
```

```
a <- 200; b <- 50; k <- 0.09 # parâmetros do modelo
```

```
# GERANDO O CONJUNTO DE DADOS
```

```
y <- c()
```

```
for( i in 1:n ) {
```

```
  y[i] <- a/(1+exp(k*(b-x[i]))) + u[i]
```

```
# prevenindo valores negativos
```

```
if( y[i] < 0){
```

```
  y[i] <- sqrt((y[i])^2)
```

```
}
```

```

}
# convertendo alguns valores em outliers
for( j in 1:50 ){
  y[4*j]<-y[4*j]+u[4*j]^2/2
  y[2*j]<-y[2*j]-u[2*j]^2/2
}
for( i in 1:n ) {
# prevenindo valores negativos
  if( y[i] < 0){
    y[i] <- sqrt((y[i])^2)
  }
}

plot(x, y, ylim = c(0, 470)); max(y); min(y) # plot dos dados sem valores discrepantes
curtose.dados <- kurtosis(y) + 3; curtose.dados[[1]]; min(y); max(y)

#####
##### AJUSTE DO MODELO LOGÍSTICO PARA #####
##### OS DADOS SEM VALORES DISCREPANTES #####
#####

dados <- data.frame(x, y)
mod <- nls(y ~ a/(1+exp(k*(b-x))),
           start = c(a = 270, b = 50, k = 0.09), data = dados)
summary(mod)
confint(mod)

#####
##### LEAST TRIMMED SQUARES #####
#####

library(gtools)      # pacote para a combinatória
p <- combinations(10, 9) # possíveis arranjos
sqd <- c() # somas de quadrados dos desvios
q <- matrix(nrow = 20, ncol = 9) # matriz para armazenar os 'melhores' valores
for(h in 1:20){
  z <- 10^10 # apenas um número grande
  t <- c(y[10*h-9], y[10*h-8], y[10*h-7], y[10*h-6], y[10*h-5],
        y[10*h-4], y[10*h-3], y[10*h-2], y[10*h-1], y[10*h-0])

```

```

for (i in 1:nrow(p)) {
  sqd[h] <- (t[p[i,]]-mean(t[p[i,]]))^2
  if(sqd[h] < z){
    q[h,] <- t[p[i,]]
    z <- sqd[h]
  }
}
}

q # subconjunto central
x.lts <- c(rep(5, 9), rep(10, 9), rep(15, 9), rep(20, 9), rep(25, 9), rep(30, 9),
  rep(35, 9), rep(40, 9), rep(45, 9), rep(50, 9), rep(55, 9), rep(60, 9),
  rep(65, 9), rep(70, 9), rep(75, 9), rep(80, 9), rep(85, 9), rep(90, 9),
  rep(95, 9), rep(100, 9)) # variável independente
y.lts <- c(q[1,],q[2,],q[3,],q[4,],q[5,],q[6,],q[7,],q[8,],q[9,],q[10,],
  q[11,],q[12,],q[13,],q[14,],q[15,],q[16,],q[17,],q[18,],q[19,],q[20,])
plot(x.lts, y.lts, ylim = c(0, 470))

#####
#### AJUSTE DO MODELO LOGÍSTICO ROBUSTO ####
#####

dados.lts <- data.frame(x.lts, y.lts)
mod.lts <- nls(y.lts ~ a/(1+exp(k*(b-x.lts))),
  start = c(a = 270, b = 50, k = 0.09), data = dados.lts)
summary(mod.lts)
Rsqr(mod.lts)

#####
##### DESVIO MÉDIO ABSOLUTO #####
#####

### LTS
y.pred <- c(); desvio <- c()
y.pred <- coef(mod.lts)[1]/(1+exp(coef(mod.lts)[3]*(coef(mod.lts)[2]-x)))
desvio <- sqrt((y-y.pred)^2)
(dma.lts <- sum(desvio)/n)

### AJUSTE CONVENCIONAL
y.pred0 <- c(); desvio0 <- c()

```

```
y.pred0 <- coef(mod)[1]/(1+exp(coef(mod)[3]*(coef(mod)[2]-x)))
desvio0 <- sqrt((y-y.pred0)^2)
(dma <- sum(desvio0)/n)
y.mqo <- coef(mod)[1]/(1+exp(coef(mod)[3]*(coef(mod)[2]-x)))
y.rob <- coef(mod.rob)[1]/(1+exp(coef(mod.rob)[3]*(coef(mod.rob)[2]-x)))
plot(x,y.dma, ylim = c(0, 520), ylab = "y")
lines(x,y.mqo, col="blue",lwd=3)
lines(x,y.rob, col="red",lwd=3)
```

APÊNDICE C

O script utilizado no Artigo 3 é detalhado a seguir.

```
### PACOTES
#library(dplyr)
#library(corrplot)
library(fBasics) # Coeficiente de curtose
library(qpcR)    # Coeficiente de determinação
library(gtools)  # Combinatória

### DADOS
dados <- read.table("dados_coelho.txt", h=T)
dados <- dados[order(dados[, 1]), ] # organizando os dados de acordo com os IDADEs
# Gráfico de dispersão
plot(dados$IDADE, dados$CPD, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",
      ylim = c(min(dados$CPD), max(dados$CPD)+1),
      cex.lab = 1.5, cex.axis = 1.2)
# Ajuste do modelo Logístico
mod <- nls(CPD ~ a/(1+exp(k*(b-IDADE))),
           start = c(a = 20, b = 50, k = 0.1), data = dados)
summary(mod)
Rsqr(mod)
# Desvio Médio Absoluto
y.mod <- coef(mod)[1]/(1+exp(coef(mod)[3]*(coef(mod)[2]-dados$IDADE)))#; y.mod <-
round(y.mod)
d_mod <- c()
d_mod <- sqrt((dados$CPD-y.mod)^2)
(dma_mod <- sum(d_mod)/nrow(dados)) # DMA do modelo ajustado
lines(dados$IDADE, y.mod, col="red",lwd=2)

### MÉTODO JACKKNIFE (CURTOSE)
# Ponto de Ruptura
pr <- (3*nrow(dados)+3+1)/4
# Seleção de valores
cts <- c() # Curtose
dados0 <- c() # Dados candidatos à seleção
```

```

dados1 <- c() # Dados selecionados
a <- 5 # Apenas um valor superior à curtose, usado para seleção dos dados
set.seed(12345)
for (i in 1:10^4) {
  selecao <- c()
  selecao <- sample(1:nrow(dados), pr)
  dados0 <- dados[selecao, ]
  cts[i] <- sqrt((kurtosis(dados0$CPD)[1])^2)
  if(cts[i] < a){
    a <- cts[i]
    dados1 <- dados0
  }
}

# Gráfico de dispersão para os dados selecionados via JACKKNIFE
plot(dados1$IDADE, dados1$CPD, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",
     ylim = c(min(dados$CPD), max(dados$CPD)+1),
     cex.lab = 1.5, cex.axis = 1.2)

# Ajuste do modelo Logístico Jackknife
mod_jkk <- nls(CPD ~ a/(1+exp(k*(b-IDADE))),
               start = c(a = 25, b = 50, k = 0.1), data = dados1)
summary(mod_jkk)

# Desvio Médio Absoluto (JACKKNIFE)
y.mod_jkk <- coef(mod_jkk)[1]/(1+exp(coef(mod_jkk)[3]*(coef(mod_jkk)[2]-
dados1$IDADE)))
d_mod_jkk <- sqrt((dados1$CPD-y.mod_jkk)^2)
(dma_mod_jkk <- mean(d_mod_jkk)) # DMA do modelo Jackknife ajustado

# Gráfico da curva convencional e jackknife
plot(dados$IDADE, dados$CPD, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",
     ylim = c(min(dados$CPD), max(dados$CPD)+1),
     cex.lab = 1.5, cex.axis = 1.2)
lines(dados$IDADE, y.mod, col="red",lwd=2)
lines(dados$IDADE, y.mod_jkk, col="green",lwd=2)
legend("topleft", legend=c("Convencional", "Jackknife"),
     col=c("red", "green"), lwd=3, cex = 1.5, bty = "n")

```

MÉTODO LEAST TRIMMED SQUARES

Ponto de Ruptura

```
pr1 <- (3*10+3+1)/4; pr1 <- round(pr1)+1
```

Possíveis arranjos

```
p <- combinations(10, pr1)
```

Seleção de valores

```
sqd <- matrix(nrow = 44, ncol = nrow(p)) # somas de quadrados dos desvios
```

```
q <- matrix(nrow = 44, ncol = 9) # matriz para armazenar os 'melhores' valores
```

```
r <- matrix(nrow = 44, ncol = 9)
```

```
for(h in 1:44){
```

```
  z <- 10^10 # apenas um número grande
```

```
  t <- c(dados$CPD[10*h-9], dados$CPD[10*h-8], dados$CPD[10*h-7], dados$CPD[10*h-6],  
        dados$CPD[10*h-5], dados$CPD[10*h-4], dados$CPD[10*h-3], dados$CPD[10*h-2],  
        dados$CPD[10*h-1], dados$CPD[10*h])
```

```
  s <- c(dados$IDADE[10*h-9], dados$IDADE[10*h-8], dados$IDADE[10*h-7],  
        dados$IDADE[10*h-6],  
        dados$IDADE[10*h-5], dados$IDADE[10*h-4], dados$IDADE[10*h-3],  
        dados$IDADE[10*h-2], dados$IDADE[10*h-1], dados$IDADE[10*h])
```

```
  for (i in 1:nrow(p)) {
```

```
    sqd[h,i] <- sum((t[p[i,]]-mean(t[p[i,]]))^2)
```

```
    if(sqd[h,i] < z){
```

```
      q[h,] <- t[p[i,]]
```

```
      r[h,] <- s[p[i,]]
```

```
      z <- sqd[h,i]
```

```
    }
```

```
  }
```

```
}
```

organizando os dados selecionados via lts

```
q1 <- as.vector(q); q1 <- matrix(q1, ncol = 1)
```

```
r1 <- as.vector(r); r1 <- matrix(r1, ncol = 1)
```

```
dados.lts <- data.frame(r1, q1)
```

```
plot(dados.lts$r1, dados.lts$q1, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",  
     ylim = c(min(dados$CPD), max(dados$CPD)+1),  
     cex.lab = 1.5, cex.axis = 1.2)
```

```

# Ajuste do modelo Logístico LTS
mod_lts <- nls(q1 ~ a/(1+exp(k*(b-r1))),
              start = c(a = 25, b = 50, k = 0.1), data = dados.lts)
summary(mod_lts)

# Desvio Médio Absoluto (LTS)
y.mod_lts <- coef(mod_lts)[1]/(1+exp(coef(mod_lts)[3]*(coef(mod_lts)[2]-dados.lts$r1)))
d_mod_lts <- sqrt((dados.lts$q1-y.mod_lts)^2) # Quadrado dos desvios
(dma_mod_lts <- sum(d_mod_lts)/nrow(dados.lts)) # DMA do modelo LTS ajustado

# Gráfico da curva convencional e LTS
plot(dados$IDADE, dados$CPD, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",
     ylim = c(min(dados$CPD), max(dados$CPD)+1),
     cex.lab = 1.5, cex.axis = 1.2)
lines(dados$IDADE, y.mod, col="red",lwd=2)
y.mod_lts <- coef(mod_lts)[1]/(1+exp(coef(mod_lts)[3]*(coef(mod_lts)[2]-dados$IDADE)))
lines(dados$IDADE, y.mod_lts, col="blue",lwd=2)
legend("topleft", legend=c("Convencional", "Least Trimmed Squares"),
     col=c("red", "blue"), lwd=3, cex = 1.5, bty = "n")

# Gráfico da curva convencional, Jackknife e LTS
plot(dados$IDADE, dados$CPD, xlab = "Idade", ylab = "Comprimento da Pata Dianteira",
     ylim = c(min(dados$CPD), max(dados$CPD)+1),
     cex.lab = 1.5, cex.axis = 1.2)
lines(dados$IDADE, y.mod, col="red",lwd=2)
y.mod_jkk <- coef(mod_jkk)[1]/(1+exp(coef(mod_jkk)[3]*(coef(mod_jkk)[2]-
dados$IDADE)))
lines(dados$IDADE, y.mod_jkk, col="green",lwd=2)
y.mod_lts <- coef(mod_lts)[1]/(1+exp(coef(mod_lts)[3]*(coef(mod_lts)[2]-dados$IDADE)))
lines(dados$IDADE, y.mod_lts, col="blue",lwd=2)
legend("topleft", legend=c("Convencional", "Jackknife", "Least Trimmed Squares"),
     col=c("red", "green", "blue"), lwd=3, cex = 1.2, bty = "n")

```