UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

LAÍS ELIAS GIMENES

MODELOS NÃO LINEARES APLICADOS ÀS CURVAS DE CASOS ACUMULADOS E ÓBITOS DA COVID-19 A NÍVEL BRASIL E SÃO PAULO

PRESIDENTE PRUDENTE
2023

LAÍS ELIAS GIMENES

MODELOS NÃO LINEARES APLICADOS ÀS CURVAS DE CASOS ACUMULADOS E ÓBITOS DA COVID-19 A NÍVEL BRASIL E SÃO PAULO

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso II.

Orientadora: Profa. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE

2023

G491m

Gimenes, Laís Elias
Modelos não lineares aplicados às curvas de casos acumulados e óbitos da COVID-19 a nível Brasil e São Paulo / Laís Elias Gimenes. -- Presidente Prudente, 2023
85 f. : tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Miriam Rodrigues Silvestre

1. COVID-19. 2. Modelos não lineares. 3. Medidas de não linearidade. 4. Métodos iterativos. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

LAÍS ELIAS GIMENES

MODELOS NÃO LINEARES APLICADOS ÀS CURVAS DE CASOS ACUMULADOS E ÓBITOS DA COVID-19 A NÍVEL BRASIL E SÃO PAULO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores
Departamento de Estatística

Presidente Prudente, 08 de fevereiro de 2023.

RESUMO

A pandemia da COVID-19 tornou-se um grave problema de saúde pública no país e no mundo nos anos de 2020 e 2021. Até o dia 26 de fevereiro de 2022 o Brasil ocupava a segunda posição, em relação ao total de mortes acometidas pela doença, ao todo foram 648.913 óbitos, já para os casos confirmados o país ocupava a terceira posição com um total de 28.744.050 pessoas infectadas. Diante dessa problemática o principal objetivo foi o ajuste de modelos não lineares de crescimento, com a proposta de utilização de modelos como: Exponencial, Logístico e Gompertz. Alguns métodos de estimação dos parâmetros foram abordados, destacando os métodos iterativos de Gauss-Newton e Newton-Raphson bem como meios de escolha de bons valores iniciais para eles. Um aspecto fundamental dos modelos não lineares é que as técnicas de diagnóstico que são aplicadas em modelos lineares são inadequadas caso o grau de não linearidade seja alto, portanto, as medidas de não linearidade foram aplicadas. Análise de diagnóstico, algumas métricas para a seleção do modelo foram abordadas a fim de obter um melhor resultado. Para a análise, foram utilizados os dados acumulados para casos confirmados e mortes pela COVID-19 do Brasil e do estado de São Paulo. Dos modelos ajustados, o melhor modelo foi o Gompertz para a série de casos confirmados do período de 26/02/2020 a 28/07/2020 e também de 29/07/2020 a 13/10/2020 tanto do Brasil como de São Paulo. Já em relação a de mortes confirmadas o modelo de Gompertz também foi o melhor para ambas as séries de dados do Brasil e de São Paulo do período de 17/03/2020 a 07/08/2020 e também de 08/08/2020 a 18/10/2020.

Palavras-chave: COVID-19; Modelos não lineares; Medidas de não linearidade; Métodos iterativos.

ABSTRACT

The COVID-19 pandemic has become a serious public health problem in the country and in the world in 2020 and 2021. Until February 26, 2022, Brazil ranked second in terms of total deaths from the disease. disease, in all there were 648,913 deaths, as for confirmed cases the country occupied the third position with a total of 28,744,050 infected people. Faced with this problem, the main objective was the adjustment of non-linear growth models, with the proposal of using models such as: Exponential, Logistic and Gompertz. Some parameter estimation methods were discussed, highlighting the iterative methods of Gauss-Newton and Newton-Raphson as well as means of choosing good initial values for them. A key aspect of nonlinear models is that the diagnostic techniques that are applied to linear models are inadequate if the degree of nonlinearity is high, so nonlinearity measures were applied. Diagnostic analysis, some metrics for model selection were addressed in order to obtain a better result. For the analysis, accumulated data for confirmed cases and deaths from COVID-19 in Brazil and the state of São Paulo were used. Of the adjusted models, the best model was Gompertz for the series of confirmed cases from 02/26/2020 to 07/28/2020 and also from 07/29/2020 to 10/13/2020 in both Brazil and São Paulo. Paul. Regarding confirmed deaths, the Gompertz model was also the best for both data series from Brazil and São Paulo from 03/17/2020 to 08/07/2020 and also from 08/08/2020 to 10/18/2020.

Keywords: COVID-19; Non-linear models; Non-linearity measures; Iterative methods.

LISTA DE FIGURAS

Figura 1 – Modelo de crescimento Exponencial.....23
Figura 2 – Modelo de crescimento Logístico.....24
Figura 3 – Modelo de crescimento Gompertz.....24
Figura 4 – Gráfico dos casos acumulados da COVID-19 do ano de 2020 e 2021 para São Paulo.....46
Figura 5 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 em São Paulo.....49
Figura 6 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 em São Paulo do dia 26/02/2020 a 28/07/2020.....50
Figura 7 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 em São Paulo do dia 29/07/2020 a 13/10/2020.....51
Figura 8 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (26/02/2020 a 28/07/2020) de casos acumulados em São Paulo.....53
Figura 9 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (29/07/2020 a 13/10/2020) de casos acumulados em São Paulo.....54
Figura 10 – Ajuste do modelo Gompertz por partes aos dados de casos acumulados da COVID-19 para São Paulo e previsões.....55
Figura 11 – Gráfico das mortes acumuladas da COVID-19 do ano de 2020 e 2021 para São Paulo.....56
Figura 12 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo.....58
Figura 13 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo do dia 17/03/2020 a 07/08/2020.....59
Figura 14 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo do dia 08/08/2020 a 18/10/2020.....60
Figura 15 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (17/03/2020 a 07/08/2020) de mortes acumuladas em São Paulo.....62

Figura 16 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (08/08/2020 a 18/10/2020) de mortes acumuladas em São Paulo.....62
Figura 17 – Ajuste do modelo Gompertz por partes aos dados de mortes acumuladas de COVID-19 para São Paulo e previsões.....64
Figura 18 – Gráfico dos casos acumulados da COVID-19 do ano de 2020 e 2021 para o Brasil.....65
Figura 19 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados de COVID-19 no Brasil.....67
Figura 20 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 no Brasil de 26/02/2020 a 28/07/2020....68
Figura 21 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 no Brasil de 29/07/2020 a 13/10/2020....69
Figura 22 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (26/02/2020 a 28/07/2020) de casos acumulados no Brasil.....70
Figura 23 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (29/07/2020 a 13/10/2020) de casos acumulados no Brasil.....71
Figura 24 – Ajuste do modelo Gompertz por partes aos dados de casos acumulados de COVID-19 para o Brasil e previsões.....72
Figura 25 – Gráfico das mortes acumuladas da COVID-19 do ano de 2020 e 2021 para o Brasil.....73
Figura 26 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas de COVID-19 no Brasil.....75
Figura 27 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 no Brasil de 17/03/2020 a 07/08/2020...76
Figura 28 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 no Brasil de 08/08/2020 a 18/10/2020...77
Figura 29 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (17/03/2020 a 07/08/2020) de mortes acumuladas no Brasil.....79
Figura 30 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (08/08/2020 a 18/10/2020) de mortes acumuladas no Brasil.....79
Figura 31 – Ajuste do modelo Gompertz por partes aos dados de mortes acumuladas de COVID-19 para o Brasil e previsões.....80

LISTA DE TABELAS

Tabela 1 – Estudos estatísticos sobre a Covid-19.....14
Tabela 2 – Modelos não lineares utilizados para descrever curvas de crescimento....22
Tabela 3 – Medidas estatísticas referentes aos anos de 2020 e 2021 para casos acumulados da COVID-19 em São Paulo.....47
Tabela 4 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para os casos de COVID-19 em São Paulo.....48
Tabela 5 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.....49
Tabela 6 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados aos dados de 26/02/2020 a 28/07/2020 (D1) e para 29/07/2020 a 13/10/2020 (D2) 52
Tabela 7 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) de casos acumulados em São Paulo.....54
Tabela 8 – Medidas estatísticas referentes aos anos de 2020 e 2021 para mortes acumuladas da COVID-19 em São Paulo.....56
Tabela 9 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para as mortes acumuladas de COVID-19 em São Paulo.....57
Tabela 10 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.....58
Tabela 11 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) 61
Tabela 12 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) de mortes acumuladas em São Paulo.....63

Tabela 13 – Medidas estatísticas referentes aos anos de 2020 e 2021 para casos acumulados da COVID-19 no Brasil.....65
Tabela 14 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para os casos de COVID-19 no Brasil.....66
Tabela 15 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.....67
Tabela 16 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) 69
Tabela 17 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) de casos acumulados no Brasil.....71
Tabela 18 – Medidas estatísticas referentes aos anos de 2020 e 2021 para mortes acumuladas da COVID-19 no Brasil.....74
Tabela 19 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para as mortes de COVID-19 no Brasil.....74
Tabela 20 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.....75
Tabela 21 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) 78
Tabela 22 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) de mortes acumuladas no Brasil.....80

SUMÁRIO

1 INTRODUÇÃO 12

2 METODOLOGIA 14

2.1 LEVANTAMENTO BIBLIOGRÁFICO 14
2.2 REGRESSÃO NÃO LINEAR 18
2.2.1 Definição 19
2.3 VANTAGENS DOS MODELOS NÃO LINEARES 20
2.4 DESVANTAGENS DOS MODELOS NÃO LINEARES 21
2.5 CATEGORIAS DE MODELOS NÃO LINEARES 21
2.5.1 Modelos de crescimento 21
2.5.2 Modelos de rendimento 25
2.6 MÉTODOS DE ESTIMAÇÃO 25
2.6.1 Método da máxima verossimilhança 25
2.6.2 Método dos mínimos quadrados 26
2.6.3 Método dos mínimos quadrados ponderados 27
2.7 MÉTODOS ITERATIVOS 28
2.7.1 Método de Gauss-Newton 28
2.7.2 Método de Newton-Raphson 30
2.8 VALORES INICIAIS PARA O MODELO 31
2.9 MEDIDAS DE NÃO LINEARIDADE 32
2.9.1 Medidas de curvatura de Bates e Watts 33
2.9.2 Medida de vício de Box 34
2.10 ANÁLISES DE DIAGNÓSTICO 35
2.10.1 Matriz de projeção 35
2.10.2 Resíduo projetado 36
2.10.3 Medidas de influência 38
2.11 MÉTRICAS PARA SELEÇÃO DO MODELO 39
2.11.1 Coeficiente de determinação 39
2.11.2 Coeficiente de determinação ajustado 39
2.11.3 Erro quadrático médio (MSE) 40
2.11.4 Critérios de informação 40
2.11.5 Teste de razão de verossimilhanças 42
2.12 RESULTADOS ASSINTÓTICOS 42

2.13 TESTE DE PETTITT PARA RUPTURAS 43

3 APLICAÇÃO 45

3.1 ANÁLISE PARA O ESTADO DE SÃO PAULO 45

3.1.1 Casos confirmados acumulados pela doença COVID-19 45

3.1.1.1 Análise Exploratória para os dados de São Paulo 45

3.1.1.2 Ajuste de modelos para os dados do estado de São Paulo 47

3.1.1.3 Ajuste dos modelos de crescimento por partes 49

3.1.1.4 Análise dos resíduos e qualidade do ajuste 52

3.1.2 Mortes acumuladas pela doença COVID-19 55

3.1.2.1 Análise Exploratória para os dados de São Paulo 55

3.1.2.2 Ajuste de modelos para os dados do estado de São Paulo 57

3.1.2.3 Ajuste dos modelos de crescimento por partes 58

3.1.2.4 Análise dos resíduos e qualidade do ajuste 61

3.2 ANÁLISE PARA O BRASIL 64

3.2.1 Casos confirmados acumulados pela doença COVID-19 64

3.2.1.1 Análise Exploratória para os dados do Brasil 64

3.2.1.2 Ajuste de modelos para os dados do Brasil 66

3.2.1.3 Ajuste dos modelos de crescimento por partes 67

3.2.1.4 Análise dos resíduos e qualidade do ajuste 70

3.2.2 Mortes acumuladas pela doença COVID-19 72

3.2.2.1 Análise Exploratória para os dados do Brasil 72

3.2.2.2 Ajuste de modelos para os dados do Brasil 74

3.2.2.3 Ajuste dos modelos de crescimento por partes 76

3.2.2.4 Análise dos resíduos e qualidade do ajuste 78

4 CONSIDERAÇÕES FINAIS 81

REFERÊNCIAS 82

1 INTRODUÇÃO

No dia 26 de fevereiro de 2020, o Ministério da Saúde confirmou o primeiro caso de Covid-19, em um homem de 61 anos que residia em São Paulo, ele havia feito uma viagem para a Itália, região da Lombardia, entre os dias 9 e 21 de fevereiro, ficou doente e em 25 de fevereiro de 2020 deu entrada no Hospital Israelita Albert Einstein.

O Covid-19 foi descoberto em Wuhan, na China, no primeiro dia do mês de dezembro de 2019, um contagioso vírus, que é semelhante ao coronavírus SARS e coronavírus MERS (Liu *et al.* 2020). Desde então o número de casos de infecção pulmonar associada a este agente etiológico vem aumentando, disseminando-se em outras cidades do país e espalhando-se para diversas regiões do mundo (Li *et al.,* 2020), inclusive no Brasil.

De acordo com Winichakoon *et al.* (2020) a Covid-19 tornou-se emergência em saúde pública de preocupação internacional, dado que a infecção vem representando um enorme desafio à saúde e também a economia mundial (Reinholz e French, 2020). Conforme a doença foi se espalhando mundialmente de forma descontrolada a Organização Mundial da Saúde (OMS) declarou-a pandemia em 11 de março de 2020.

O Covid-19 é uma doença que afeta o sistema respiratório, causada pelo coronavírus da síndrome respiratória aguda grave 2 (Sars-CoV-2), potencialmente grave, tem elevada transmissibilidade e se espalha facilmente pelo ar quando as pessoas estão em contato direto, dessa forma, foram tomadas medidas preventivas para conter o avanço do vírus, sendo elas: o distanciamento social, uso de máscaras e a higienização das mãos constantemente com álcool em gel.

Conforme os dados da OMS desde o início da pandemia até o dia 26 de fevereiro de 2022 o Brasil ocupava a segunda posição no que diz respeito ao total de mortes acometidas pela Covid-19, ao todo são 648.913 óbitos. À frente, estava os Estados Unidos, que somava 948.215. Em terceiro, segue a Índia com 513.724; seguido da Rússia, com 343.178 e México, com 318.014. Já para os casos acumulados, foram confirmadas 434.247.399 pessoas positivadas para a Covid-19 no mundo, os Estados Unidos possuem o maior número (78.931.739), logo em seguida

vem Índia (42.916.117), Brasil (28.744.050), França (22.699.443) e Reino Unido (18.854.972).

Considerando que a maioria da população se mostrou vulnerável a essa doença e tentando entender a evolução do Coronavírus faz-se necessário estudos estatísticos, como modelos de regressão, que darão base a prevenção, ao controle e também a avaliação do atual cenário epidemiológico.

Neste sentido, existem inúmeras possibilidades de estudo, dentre elas a utilização de modelos de regressão que tem origem no século XIX com Francis Galton, quando ele procurava entender a relação entre as alturas dos pais e dos filhos, porém ele descobriu que não importava se essa estatura do pai era alta ou baixa, pois a altura do filho sempre tendia a média da população, daí surge o nome regressão. Os modelos de regressão são amplamente utilizados para descrever a relação entre uma ou múltiplas covariáveis e uma variável resposta. Dentre os modelos de regressão tem-se os lineares e os não lineares, Saraiva *et al.* (2021) utilizaram os modelos de crescimentos por partes para modelar o número de casos acumulados de Covid-19 em Campo Grande.

Do exposto, o objetivo deste trabalho é ajustar diferentes modelos de regressão não linear aos dados de mortes acumuladas e casos acumulados referentes ao Brasil e o estado de São Paulo, além de realizar as previsões a curto prazo bem como construir intervalos de confiança.

Diante do cenário atual de pandemia causado pela COVID-19 faz-se necessário medidas que auxiliem na tomada de decisão, com o intuito de causar menor impacto possível, tanto no âmbito social e também econômico, com isso análises quantitativas e o ajuste de modelos estatísticos aos dados são de grande relevância, pois ajudam a entender a dinâmica de propagação da doença, previsão e projeção da mesma.

Com o avanço rápido do coronavírus, foi preciso a criação do decreto nº 64.881, de 22 de março de 2020 (SÃO PAULO, 2020), implementando a quarentena no estado de São Paulo, para assim poder controlar a difusão do novo vírus. Uma providência para dar um parecer a respeito de que essas medidas e tantas outras, que foram implementadas, surtiram efeito é o uso de técnicas estatísticas, como por exemplo, os modelos de regressão não linear, com eles podemos analisar o quanto as ações tomadas contribuíram de forma eficaz para a contenção do vírus.

2 METODOLOGIA

2.1 Levantamento bibliográfico

Na Tabela 1, são apresentados alguns estudos feitos utilizando técnicas estatísticas aplicadas aos dados de Covid-19 em diversos lugares.

Tabela 1 – Estudos estatísticos sobre a Covid-19

Artigo	Autores	Modelo utilizado
Um modelo de crescimento por partes para modelar o número acumulado de casos de Covid-19 na cidade de Campo Grande.	SARAIVA *et al.* (2021)	- Modelo de crescimento exponencial. - Modelo de crescimento Logístico. - Modelo de crescimento Gompertz.
Modelos não lineares aplicados a mortalidade e casos da Covid-19 no Brasil, Itália e mundo.	SANTIAGO *et al.* (2020)	-Modelo de declínio hiperbólico. - Modelo de Hoerl. - Modelo Gaussiano. - Modelo poder logístico.
Interiorização do Covid-19: Uma análise da evolução dos casos/10 mil habitantes em municípios da microrregião de Garanhuns no estado de Pernambuco, através de modelos de regressão não linear.	AMARAL *et al.* (2020)	- Modelo Gompertz. - Modelo Logístico. - Modelo Von Bertalanffy.
Predição de casos e óbitos de Covid-19 em Mato Grosso e no Brasil.	ESPINOSA *et al.* (2020)	- Modelo exponencial duplo de séries temporais.

Fonte: Elaborado pela autora (2022).

Saraiva *et al.* (2021) foram solicitados para analisar a evolução do novo coronavírus em Campo Grande, sendo assim, aplicaram procedimentos de modelagem usados para ajustar um modelo de crescimento para os casos acumulados registrados na cidade. Foram ajustados o modelo de crescimento exponencial, o modelo de crescimento logístico e o modelo de crescimento Gompertz; trabalhou-se com a transformação logarítmica para cada um deles a fim de linearizar a função. A curva de crescimento Gompertz foi a que melhor se ajustou aos dados e essa escolha se deu através dos critérios do *mean square error* (MSE), ou erro quadrático médio e os métodos de seleção *Akaike information criterion* (AIC) e o *Bayesian information criterion* (BIC).

Conforme foram atualizando a base de dados perceberam que o modelo escolhido não se encaixava mais, logo a solução foi os modelos de crescimento por partes. Para separar os dados da melhor maneira possível, operou-se da seguinte forma, escolheu-se 10 pontos de corte para os dados, que são eles: $\boldsymbol{G} = \{\mathbf{25}, \mathbf{26}, \mathbf{27}, \mathbf{28}, \mathbf{29}, \mathbf{30}, \mathbf{31}, \mathbf{32}, \mathbf{33}, \mathbf{34}\}$, então o primeiro subconjunto era $\boldsymbol{D_1} = \{\boldsymbol{y_0}, \dots, \boldsymbol{y_d}\}$ e o segundo $\boldsymbol{D_2} = \{\boldsymbol{y_{d+1}}, \dots, \boldsymbol{y_{59}}\}$ para $\boldsymbol{d} \in \boldsymbol{G}$ o que formou 10 cenários, onde para cada um deles aplicou-se os três modelos e verificou-se a melhor curva através do AIC e BIC, o ponto $\boldsymbol{d} = \mathbf{28}$ levou a uma função ajustada com o menor MSE, logo ele foi utilizado na separação de $\boldsymbol{D_1}$ e $\boldsymbol{D_2}$.

Santiago *et al.* (2020) ajustaram modelos de regressão não linear aos dados de mortalidade e casos confirmados da COVID-19 no Brasil, Itália e mundo até 31/03/2020 bem como modelos de regressão linear aplicados a variável taxa de letalidade. Os parâmetros foram estimados utilizando o algoritmo Levenberg-Marquardt.

A respeito dos modelos aplicados foi feita a escolha de dois melhores modelos de um rol de mais de 70 modelos candidatos, eles foram selecionados com base na raiz quadrada do erro quadrático médio, coeficiente de determinação e determinação ajustado.

Para o número de mortes ocorridas no Mundo foram ajustados os seguintes modelos:

$\boldsymbol{f(x)} = \boldsymbol{c}\left(\mathbf{1} + \frac{\boldsymbol{bx}}{\boldsymbol{a}}\right)^{-\frac{\mathbf{1}}{\boldsymbol{b}}}$, modelo de declínio hiperbólico (HD)

$f(x) = (a + bx)^{-\frac{1}{c}}$, modelo Bleasdale (B)

Já para o Brasil foram os seguintes:

$f(x) = ab^x x^c$, modelo de Hoerl (H)

$f(x) = c\left(1 + \frac{bx}{a}\right)^{-\frac{1}{b}}$, modelo de declínio hiperbólico (HD)

E por fim para a Itália foram:

$f(x) = ae^{\left\{-\frac{(x-b)^2}{2c^x}\right\}}$, modelo gaussiano (G)

$f(x) = \frac{a}{1+\left(\frac{x}{b}\right)^c}$, modelo poder logístico (PG)

Em todos os modelos o x representa o número de dias a partir da primeira morte confirmada no respectivo local.

Sobre o número de casos confirmados de COVID-19 foram ajustados os seguintes modelos:

Para o mundo:

$f(x) = \frac{(a+bx)}{1+cx+dx^2}$, modelo racional (R)

$f(x) = c\left(1 + \frac{bx}{a}\right)^{-\frac{1}{b}}$, modelo de declínio hiperbólico (HD)

Para o Brasil:

$f(x) = ae^{\left(\frac{b}{x}\right)}$, modelo exponencial modificado (EM)

$f(x) = e^{\left\{a+\frac{b}{x}+\ln(x)c\right\}}$, modelo pressão de vapor (PV)

Para a Itália:

$f(x) = ae^{\left\{-\frac{(x-b)^2}{2c^x}\right\}}$, modelo gaussiano (G)

$f(x) = a - be^{\{-cx^{3,81}\}}$, modelo Weibull (W)

Nos modelos o x representa o número de dias a partir do primeiro caso confirmado no lugar analisado.

Já o estudo feito para verificar a taxa de letalidade da doença foi a regressão linear simples, em que a taxa foi calculada da seguinte forma:

$$\boldsymbol{Taxa\ de\ letalidade} = \frac{\boldsymbol{número\ de\ mortes\ confirmados\ da\ doença}}{\boldsymbol{número\ de\ casos\ confirmadas\ pela\ doença}}$$

Os dois modelos selecionados que foram descritos anteriormente para o Brasil, Itália e Mundo foram comparados utilizando o AIC e o BIC. Foram obtidos também, o coeficiente de determinação e determinação ajustado, a raiz quadrada do erro quadrático médio (RMSE) de cada modelo.

Em relação as mortes por COVID-19, o melhor modelo ajustado de acordo com os critérios de seleção foi o modelo de Hoerl e poder logístico, para o Brasil e a Itália, respectivamente. Para o mundo os dois modelos testados apresentaram os mesmos valores paras estatísticas AIC, BIC e RMSE, logo são equivalentes para modelar a variável em questão, porém o escolhido foi o modelo Bleasdale. Os parâmetros foram todos significativos, com exceção do modelo HD para o Brasil, pois ao testar a significância obteve-se p-valor de 0,098 maior do que o nível de significância adotado de 5%, mas esse não foi utilizado.

Sobre os casos confirmados da doença, todos os modelos tiveram seus coeficientes de determinação e determinação ajustado bem alto, maior do que 0,99, os melhores indicados através dos critérios de seleção foram o R, EM, e G, respectivamente, para o mundo, Brasil e Itália. Demais resultados são encontrados no artigo.

Amaral *et al.* (2020) trabalharam com dados referentes aos casos de COVID-19 nos seguintes municípios do agreste meridional pernambucano: Bom Conselho, Garanhuns e Lajedo. As informações coletadas foram convertidas em casos/10 mil habitantes. Para os dados citados foram ajustados os modelos não lineares Gompertz, Logístico e Von Bertalanffy.

Para os modelos analisados foram utilizados o coeficiente de determinação ajustado, o critério de informação de Akaike e o desvio médio absoluto para avaliar a qualidade dos ajustes, bem como a análise dos pressupostos residuais, a normalidade, a independência e a homoscedasticidade e ambos os modelos não

violaram os pressupostos. Além disso, para os municípios de Bom Conselho e de Garanhuns o melhor modelo foi o de Von Bertalanffy; já para Lajedo foi o de Gompertz.

Outro estudo feito por Amaral *et al.* (2020) foi a extração da taxa de crescimento absoluto (TCA) devido a COVID-19, que é a primeira derivada da função ajustada em relação ao tempo, com ela foi possível verificar o ponto de inflexão das curvas.

Espinosa *et al.* (2020) teve como objetivo a previsão dos casos e óbitos da COVID-19 no Brasil e, especificamente, em Mato Grosso, por meio do modelo estatístico duplo exponencial de séries temporais. As duas variáveis dependentes consideradas foram os casos e óbitos no Brasil e a variável independente foi a data de divulgação da ocorrência do óbito. Já para o Mato Grosso considerou-se somente o número de casos, pois esse registro até aquele momento do estudo era muito baixo.

Anteriormente a modelagem foi construído um gráfico para visualizar o comportamento da série temporal, a fim de identificar tendências, sazonalidades. Como os dados analisados apresentavam tendência os autores aplicaram o método de suavização exponencial de Holt ou suavização dupla exponencial e para verificar a adequação do modelo ajustado utilizaram o erro de porcentagem absoluto médio.

Após a análise residual, foi necessária uma transformação nos dados tanto de óbitos e confirmados da doença nos dois lugares analisados, foi aplicada o método de Box e Cox, a fim de estabilizar a variância, devido à grande variabilidade desses dados para o modelo se tornar adequado.

O modelo adotado com os dados transformados foi, estatisticamente, adequado, para fazer previsões dos dois eventos em estudo, atendendo os critérios de validação com valores reais e a precisão requerida.

2.2 Regressão não linear

Segundo Bates e Watts (1988) a regressão linear é um método poderoso para analisar dados descritos por modelos que são lineares nos parâmetros. Porém, muitas vezes em diversos estudos a função matemática que relaciona a variável resposta às variáveis preditoras não é linear, sendo assim, uma outra abordagem é utilizada, a regressão não linear. Nesses casos, as técnicas de regressão linear devem ser estendidas, o que aumenta o grau de complexidade.

A regressão não linear é uma técnica estatística que engloba diversos modelos que abrangem uma grande parte de aplicações nas mais diversas áreas do

conhecimento como: agricultura, biologia, medicina, química, econometria, engenharia, entre outros. Particularmente, destaca-se o livro de Seber e Wild (2003), que descreve vários modelos não lineares segundo diversos aspectos, destacando-se os modelos sigmoidais e as curvas de crescimento assintóticas.

2.2.1 Definição

Um modelo de regressão não linear pode ser escrito por

$$y_i = f(\boldsymbol{X}_i, \boldsymbol{\theta}) + \varepsilon_i \ , i = 1, \dots , n \qquad (2.1)$$

onde y_i é a variável resposta, $f(\boldsymbol{X}_i, \boldsymbol{\theta})$ é a função resposta não linear com forma conhecida, $\boldsymbol{X}_i$ é o vetor de variáveis explicativas, $\boldsymbol{\theta}$ é o vetor de parâmetros desconhecidos que precisam ser estimados e ε_i são os erros aleatórios, sobre os quais são assumidos os pressupostos usuais de que o erro seja independente e identicamente distribuído, seguindo distribuição normal, com $E(\varepsilon_i) = 0$ e variância$(\varepsilon_i) = \sigma^2$.

A função de regressão ou função esperança é dada por $E(y_i|x_i) = f(\boldsymbol{X}_i, \boldsymbol{\theta})$. De acordo com Bates e Watts (1988) se pelo menos uma das derivadas parciais da função f em relação aos parâmetros depender de pelo menos um dos parâmetros do modelo, logo tem-se um modelo de regressão não linear. Por exemplo, o modelo:

$$y = f(\boldsymbol{X}_i, \boldsymbol{\theta}) = \theta_1 e^{\theta_2 x} + \varepsilon \qquad (2.2)$$

não é linear nos parâmetros desconhecidos θ_1 e θ_2, pois suas derivadas com respeito a θ_1 e θ_2 são funções dos mesmos, ou seja,

$$\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x} \text{ e } \frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}$$

Os modelos de regressão não lineares podem ser “intrinsecamente lineares” ou “intrinsecamente não lineares”, ou seja, se for possível reduzi-los a um modelo linear através de uma reparametrização adequada ou linearizados através de alguma transformação eles são considerados intrinsecamente lineares. Como exemplo tem-se a função definida em 2.2, pois aplicando a transformação logarítmica o modelo é linearizado, como segue:

$$\log\big(\, f(\boldsymbol{X}_i, \boldsymbol{\theta})\big) = \log\big(\theta_1 e^{\theta_2 x} + \varepsilon\big) =$$
$$= \log(\theta_1) + \log\big(e^{\theta_2 x}\big) + \log\,(\varepsilon) = \log(\theta_1) + \theta_2 x + \log\,(\varepsilon)$$

ou,

$$g(\boldsymbol{X}_i, \boldsymbol{\theta}) = \beta_1 + \beta_2 x + \varepsilon^*,$$

no qual, $g(\boldsymbol{X}_i, \boldsymbol{\theta}) = \log\big(\, f(\boldsymbol{X}_i, \boldsymbol{\theta})\big)$, $\beta_1 = \log(\theta_1)$, $\beta_2 = \theta_2$ e $\varepsilon^* = \log\,(\varepsilon)$.

Já os modelos que não podem ser linearizados como feito acima, são chamados de “intrinsecamente não linear”, segundo Bates e Watts (1988). Um problema ao utilizar uma transformação no modelo não linear é que o termo do erro ε, quando transformado, geralmente perde seus pressupostos usuais, ou seja, $\log\,(\varepsilon)$ não irá satisfazer as suposições de normalidade, independência e homogeneidade de variância.

Todavia, a abordagem de linearizar do modelo não linear é interessante para a obtenção de valores iniciais para o método iterativo de ajuste utilizado, como será descrito mais adiante.

2.3 Vantagens dos modelos não lineares

A seguir são apresentadas as vantagens da utilização de modelos não lineares sobre os lineares:

- A escolha de modelos não lineares é baseada em teoria ou princípios mecanísticos como físicos, químicos, biológicos ou qualquer outra informação prévia sobre os dados que serão trabalhados;
- Os modelos não lineares são mais parcimoniosos, ou seja, apresentam uma quantidade menor de parâmetros, que os lineares;
- Geralmente as predições tendem a ser mais robustas do que os modelos lineares, especialmente fora da faixa dos dados observados;
- Os parâmetros apresentam interpretação prática, o que é uma vantagem muito grande para o pesquisador porque são informações a mais que ele terá no seu objeto de estudo;
- Os conhecimentos do pesquisador são levados em conta, pois os modelos não lineares partem deste princípio.

2.4 Desvantagens dos modelos não lineares

Abaixo são citadas as desvantagens dos modelos não lineares:

- Para a estimação dos parâmetros desconhecidos do modelo é preciso a utilização de procedimentos iterativos de estimação baseados na provisão de valores iniciais para eles;
- Modelos não lineares com comportamento distante do linear podem fazer com que seus resultados assintóticos sejam invalidados, especialmente quando a amostra disponível é pequena;
- Os métodos inferenciais são aproximados;
- O pesquisador precisa necessariamente ter conhecimento sobre o fenômeno estudado.

2.5 Categorias de modelos não lineares

Nesta seção será abordada duas categorias de modelos não lineares, os de crescimento e de rendimento, que tem impulsionado significativamente a área Estatística, no que diz respeito a parte de regressão não linear.

2.5.1 Modelos de crescimento

De acordo com Seber e Wild (2003), analisar dados a respeito do crescimento de algum fenômeno é de grande importância em muitas áreas de estudo. Os biólogos estão interessados na descrição do crescimento de plantas e animais tentando entender os mecanismos essenciais. Os químicos tem o interesse na formulação da reação química de um produto ao longo do tempo. Na agricultura saber como anda o crescimento da produção é de grande valia para as questões econômicas e administrativas. O crescimento infantil, assim como o aumento de tumores e os efeitos de tratamento sobre tais crescimentos são conhecimentos de grande relevância na medicina. Já para os cientistas sociais o interesse está na expansão da população, no fornecimento de alimentos e demanda de energia.

Para vários tipos de dados a taxa de crescimento não diminui constantemente, mas em vez disso, aumenta até um máximo antes de diminuir constantemente até zero, esse comportamento sugere curvas em forma de S ou sigmoidais. Na modelagem em que envolve dados de crescimento, é almejado obter informações

sobre a interpretação física dos parâmetros, para construir um modelo padrão para as observações em estudo.

Na literatura existem diversos modelos de crescimento. Na Tabela 2, são apresentados alguns dos principais modelos desse tipo.

Tabela 2 – Modelos não lineares utilizados para descrever curvas de crescimento.

Modelo	Função	Modelo Estatístico
1	$y_i = \alpha e^{\beta t_i} + \varepsilon_i$	Exponencial
2	$y_i = \alpha\left(1 - e^{(\gamma\beta - \gamma t_i)}\right) + \varepsilon_i$	Mitscherlich
3	$y_i = \frac{\alpha}{1 + \beta e^{(-\gamma t_i)}} + \varepsilon_i$	Logístico
4	$y_i = \alpha e^{\left(-\beta e^{(-\gamma t_i)}\right)} + \varepsilon_i$	Gompertz
5	$y_i = \alpha\left(1 - \beta e^{(-\gamma t_i)}\right) + \varepsilon_i$	Brody
6	$y_i = \alpha\left(1 - \beta e^{(-\gamma t_i)}\right)^3 + \varepsilon_i$	Von Bertalanffy
7	$y_i = \alpha\left(1 - \beta e^{(-\gamma t_i)}\right)^m + \varepsilon_i$	Richards

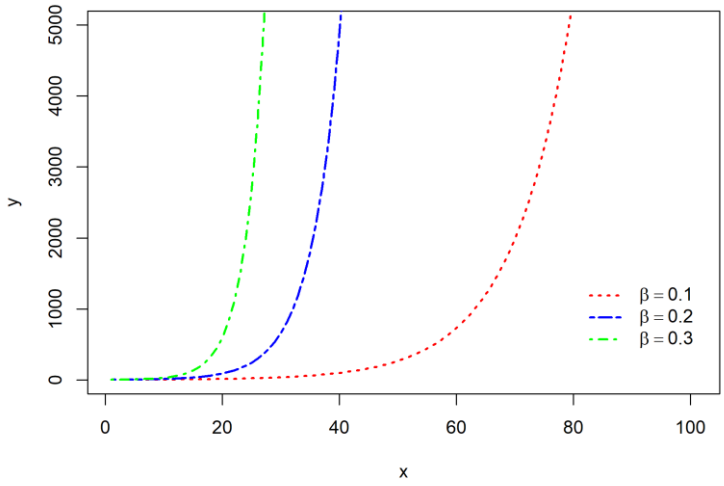
Fonte: Elaborado pela autora (2022).

Os modelos de crescimento aplicados neste trabalho serão os modelos 1, 3 e 4 expostos na Tabela 2 acima, que são eles:

- O modelo de crescimento Exponencial: é muito conhecido e utilizado quando se trata de doenças epidemiológicas, sua equação é dada na Tabela 2. Tendo em vista que as análises serão feitas em duas perspectivas, confirmados e óbitos da doença, logo y_i em um primeiro momento será o número acumulado de casos de COVID-19 no tempo t e em segundo a quantidade de mortes causadas pelo coronavírus no tempo t, onde $t \geq 0$. Além disso, α é o número de casos (ou mortes) no tempo inicial, $t = 0$, β é a taxa de crescimento. Na Figura 1 pode ser observado os gráficos das curvas do modelo Exponencial, foram gerados valores para a variável tempo de 0 a 100, o valor inicial é $\alpha = 2$ e para β utilizou-se três valores conforme a legenda do gráfico. Observa-se que à medida que o β aumenta mais inclinada é a curva, indicando

um rápido crescimento dos casos (ou mortes).

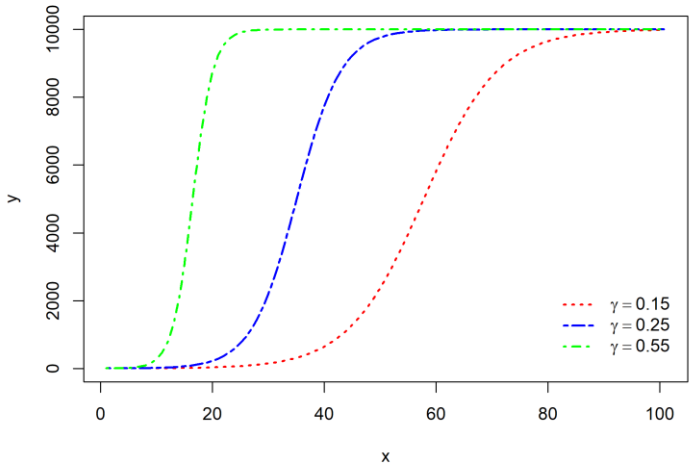
Figura 1 – Modelo de crescimento Exponencial



Fonte: Elaborado pela autora (2022).

- O modelo de crescimento Logístico: tem um formato de S (curva sigmoide), de modo que ele possui um limite de crescimento. O parâmetro α é a assíntota superior, ou seja, no contexto analisado seria o valor estimado para a quantidade máxima de casos, β corresponde as coordenadas (x, y) do ponto de inflexão onde $x = \frac{\log(\beta)}{\gamma}$ e $y = \frac{\alpha}{2}$, isto é, a partir daquele ponto a curvatura do modelo se altera, relacionando com o cenário de COVID significa que, se os casos (ou mortes) estão aumentando então a partir do ponto de inflexão ocorrerá uma queda, ou vice e versa. Já γ é a taxa de crescimento, quanto maior for esse valor menos tempo será necessário para o objeto de estudo atingir o seu valor máximo. Na Figura 2 pode ser observada três curvas do modelo citado, com $\alpha = 10000$, $\beta = 5000$ e γ variando conforme apresentada na legenda da Figura 2, aumentando esse valor mais inclinada é a curva.

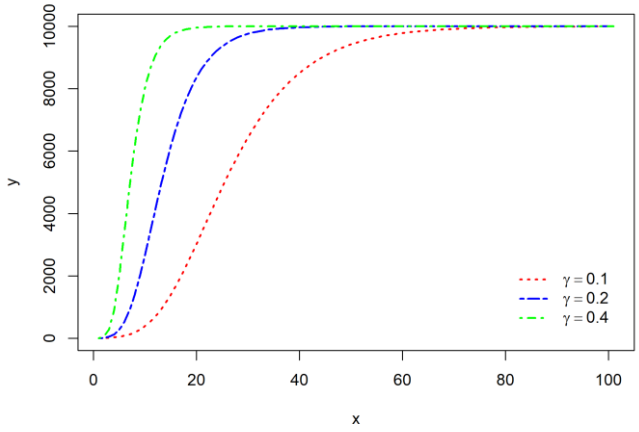
Figura 2 – Modelo de crescimento Logístico



Fonte: Elaborado pela autora (2022).

- O modelo de crescimento Gompertz: semelhante ao Logístico, tem o formato de S, porém a principal diferença é que o modelo Gompertz não é simétrico em relação ao ponto de inflexão, ou seja, ao invés da ordenada ser metade do valor do parâmetro α ela é aproximadamente 37% de α. As interpretações dos parâmetros são similares as descritas para o modelo Logístico. Na Figura 3 pode ser observada três curvas, com $\alpha = 10000$, $\beta = 8$ e γ variando conforme a legenda.

Figura 3 – Modelo de crescimento Gompertz



Fonte: Elaborado pela autora (2022).

2.5.2 Modelos de rendimento

Na literatura existe vários outros modelos que estão relacionados com as curvas de crescimento. Na agricultura são utilizados para quantificar a relação entre o rendimento total de uma cultura (y) e a densidade da cultura (x).

Os modelos de rendimento são importantes por exemplo na área agrícola, onde os agrônomos estão interessados em prever o rendimento de determinada plantação; conhecer a densidade de plantio onde obteve-se o rendimento máximo da lavoura; observar o quanto o rendimento diminuiu conforme o aumento da área de plantio, pois existe uma competição entre as plantas por recursos; bem como poder compará-la no que diz respeito as diferentes condições, seja de clima, quantidade de chuva, etc.

2.6 Métodos de estimação

A literatura apresenta diferentes métodos de estimação para a obtenção das estimativas dos parâmetros do modelo (SEBER e WILD (2003)), dentre eles temos o método de máxima verossimilhança e o método de mínimos quadrados. O obstáculo é que, para modelos não lineares, o sistema de equações normais não é facilmente resolvido, o que implica na utilização de métodos iterativos para obter as estimativas dos parâmetros. A seguir são expostos alguns métodos de estimação.

2.6.1 Método da máxima verossimilhança

Se a distribuição de ε_i na equação (2.1) é conhecida, com distribuição $N{\sim}(0, \sigma^2 I)$, então a estimativa de máxima verossimilhança é obtida maximizando a função de verossimilhança L que é dada por

$$L(\boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[yi - f(Xi,\boldsymbol{\theta})]^2}. \qquad (2.3)$$

Para facilitar nos cálculos da estimativa de máxima verossimilhança aplica-se o log em (2.3) e ignorando as constantes, a função log-verossimilhança de (2.3) é dada por

$$l(\boldsymbol{\theta},\sigma^2) = \log\left(L(\boldsymbol{\theta},\sigma^2) = -\frac{n}{2}\, log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[yi - f(Xi,\boldsymbol{\theta})]^2\right.$$

$$= -\tfrac{n}{2}\, log\sigma^2 - \tfrac{1}{2\sigma^2} SQRes(\boldsymbol{\theta}). \qquad (2.4)$$

Dado σ^2, (2.4) é maximizada com respeito a $\boldsymbol{\theta}$ quando $SQRes(\boldsymbol{\theta})$ é minimizada, ou seja, quando $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ (estimativa de mínimos quadrados), (SEBER e WILD, 2003). Além disso, fazendo a derivada de (2.4) em relação a σ^2 e igualando a zero, temos que:

$$\frac{\partial l(\boldsymbol{\theta},\sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\hat{\sigma}^2}\right) - \left(-\frac{1}{\hat{\sigma}^4}\right) SQRes(\boldsymbol{\theta}) = 0$$

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{\hat{\sigma}^4} SQRes(\boldsymbol{\theta}) = 0$$

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^2\hat{\sigma}^2} SQRes(\boldsymbol{\theta}) = 0$$

$$\frac{1}{2\hat{\sigma}^2\hat{\sigma}^2} SQRes(\boldsymbol{\theta}) = \frac{n}{2\hat{\sigma}^2}$$

$$SQRes(\boldsymbol{\theta}) = \frac{n}{2\hat{\sigma}^2}(2\hat{\sigma}^2\hat{\sigma}^2)$$

$$\hat{\sigma}^2 = \tfrac{SQRes(\boldsymbol{\theta})}{n}. \qquad (2.5)$$

Portanto, $\frac{\partial l}{\partial \sigma^2} = 0$, tem solução dada por (2.5), o que dá um valor máximo quando a segunda derivada é negativa. Isto sugere que $\widehat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2 = \frac{SQRes(\boldsymbol{\theta})}{n}$ são estimativas de máxima verossimilhança.

2.6.2 Método dos mínimos quadrados

Tomando-se o modelo descrito em (2.1) sabe-se da teoria de estimação por mínimos quadrados que a estimativa obtida por esse método para $\boldsymbol{\theta}$, indicado por $\widehat{\boldsymbol{\theta}}$, minimiza a soma de quadrados dos erros dada por:

$$SQRes(\boldsymbol{\theta}) = \sum_{i=1}^{n}[y_i - f(\boldsymbol{X}_i,\boldsymbol{\theta})]^2 \qquad (2.6)$$

sobre $\boldsymbol{\theta} \in \Theta$.

Para encontrar tais estimativas é preciso derivar (2.6) com relação a cada elemento do vetor $\boldsymbol{\theta}$ e igualar as equações a zero, isto é,

$$\left.\frac{\partial SQRes(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\widehat{\theta}} = 0 \quad (j = 1,2, \dots, p)$$

obtém-se o seguinte vetor de equações normais:

$$\sum_{i=1}^{n} [y_i - f(\boldsymbol{X}_i, \boldsymbol{\theta})] \left.\frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\theta})}{\partial \theta_j}\right|_{\theta=\widehat{\theta}} = 0 \quad (j = 1,2, \dots, p). \tag{2.7}$$

Em modelos lineares esse sistema possui um vetor único de soluções para os parâmetros θ e cada parâmetros depende somente dos dados observados. Já para modelos não lineares, os estimadores $\widehat{\boldsymbol{\theta}}$ obtidos de (2.7) não apresentam soluções de forma analíticas, sendo assim é preciso aplicar métodos iterativos.

2.6.3 Método dos mínimos quadrados ponderados

De acordo com Gallant (1987) nos modelos não lineares em que há heteroscedasticidade ou autocorrelação dos erros utiliza-se um método similar ao exposto acima, o dos mínimos quadrados ponderados, no qual a soma a ser minimizada é dada por

$$SQRes(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(\boldsymbol{X}_i, \boldsymbol{\theta})]' \, \boldsymbol{V}^{-1} \, [y_i - f(\boldsymbol{X}_i, \boldsymbol{\theta})] \tag{2.8}$$

onde $\boldsymbol{V}$ é a matriz de variâncias e covariâncias dos resíduos em que é conhecida e positiva definida. Quando $\boldsymbol{V} = \boldsymbol{I}_n$, tem-se novamente o método de mínimos quadrados exposto na seção 2.6.2.

Definindo $\widehat{\boldsymbol{\theta}}_G$ como sendo o vetor das estimativas de mínimos quadrados ponderados que minimiza (2.8). Seja $\boldsymbol{V} = \boldsymbol{U}'\boldsymbol{U}$ uma decomposição de Cholesky de $\boldsymbol{V}$ onde $\boldsymbol{U}$ é uma matriz triangular superior e $\boldsymbol{R} = (\boldsymbol{U}')^{-1}$; multiplica-se a equação (2.1) por $\boldsymbol{R}$ tem-se que:

$$\boldsymbol{z} = \boldsymbol{k}(\boldsymbol{\theta}) + \boldsymbol{\eta}$$

de modo que, $\boldsymbol{z} = \boldsymbol{Ry}$, $\boldsymbol{k}(\boldsymbol{\theta}) = \boldsymbol{Rf}(\boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\eta} = \boldsymbol{R\varepsilon}$. Portanto, a soma de quadrado dos resíduos descrita em (2.8) pode ser reescrita como

$$SQRes(\boldsymbol{\theta}) = [z - k(\boldsymbol{\theta})]'[z - k(\boldsymbol{\theta})].$$

2.7 Métodos iterativos

Os métodos iterativos são utilizados como alternativa para obter as estimativas dos parâmetros, pois as equações normais dependem dos mesmos, ou seja, são não lineares em relação aos estimadores dos parâmetros fazendo com que essas equações não apresentem formas explícitas.

Diferentes métodos iterativos são apresentados na literatura para a obtenção das estimativas de mínimos quadrados dos parâmetros de um modelo não linear. Dentre eles temos os seguintes métodos: do gradiente descendente, de Marquardt, de Newton-Raphson, de Gauss-Newton ou da linearização, entre outros. Ambos os métodos citados se diferenciam em como é calculado o termo $\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)$. De maneira geral o procedimento de Gauss-Newton e Marquardt efetuam a regressão dos resíduos em relação as primeiras derivadas do modelo não linear em relação aos parâmetros, até a convergência. Já o de Newton-Raphson faz a regressão desses resíduos em relação a uma função das segundas derivadas do modelo não linear com relação aos parâmetros. O método do gradiente, como o nome diz, se baseia no gradiente ou grau de variação dos resíduos. A seguir serão apresentados dois métodos em detalhes.

2.7.1 Método de Gauss-Newton

A $SQRes(\boldsymbol{\theta})$ descrita em (2.6), necessita de algum método numérico iterativo para poder minimiza-la. O método iterativo de Gauss-Newton se baseia em aproximações lineares para a função $f(\boldsymbol{X}_i, \boldsymbol{\theta})$ a cada passo.

Supondo que $\boldsymbol{\theta}^{(a)}$ é uma boa aproximação para a estimativa de mínimos quadrados $\boldsymbol{\theta}$ de $f(\boldsymbol{X}_i, \boldsymbol{\theta})$. Então, para $\boldsymbol{\theta}$ próximo de $\boldsymbol{\theta}^{(a)}$, tem-se que a expansão em série de Taylor de primeira ordem de $f(\boldsymbol{X}_i, \boldsymbol{\theta})$ para $\boldsymbol{\theta}$ em torno de $\boldsymbol{\theta}^{(a)}$ é dada por

$$f(\boldsymbol{X}_i, \boldsymbol{\theta}) \approx f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right) + \frac{\partial f(\boldsymbol{X}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right). \quad (2.9)$$

Fazendo $\boldsymbol{F}(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{X}_i,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$ a matriz com as derivadas da função não linear em relação a cada um dos parâmetros e definindo $\boldsymbol{r}(\boldsymbol{\theta}) = \boldsymbol{Y} - f(\boldsymbol{X}_i, \boldsymbol{\theta})$ o vetor de resíduos ordinários. Assim, da mesma forma os resíduos podem ser aproximados por

$$\boldsymbol{r}(\boldsymbol{\theta}) \approx \boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right).$$

O objetivo é minimizar a $SQRes = \boldsymbol{r}'(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta})$ pelo método de mínimos quadrados, onde $\boldsymbol{r}'(\boldsymbol{\theta})$ é o vetor transposto de $\boldsymbol{r}(\boldsymbol{\theta})$, logo tem-se que:

$$SQRes = \boldsymbol{r}'(\boldsymbol{\theta})\boldsymbol{r}(\boldsymbol{\theta}) \approx \left[\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)\right]'\left[\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)\right] =$$

$$= \left\{\boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right) - \left[\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)\right]'\right\}\left\{\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)\right\} =$$

$$= \left\{\boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right) - \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)'\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\right\}\left\{\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)\right\} =$$

$$= \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) - \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)'\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) +$$

$$+\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)'\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) =$$

$$= \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) - \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) +$$

$$+\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)'\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)$$

que pode ser reduzida, por:

$$SQRes = \boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) - \mathbf{2}\boldsymbol{r}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) +$$

$$+\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)'\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) \qquad (2.10)$$

(2.10) será minimizada quando

$$\frac{\partial SQRes}{\partial\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)} = 0$$

$$-2\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right) + 2\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) = 0$$

$$\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) = \boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right). \qquad (2.11)$$

Pré-multiplicando ambos os lados da equação (2.11) por $\left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}$, tem-se que,

$$\left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) = \left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right)$$

$$\boldsymbol{I}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) = \left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right)$$

$$\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}\right)=\left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right).$$

A cada iteração é gerada uma nova aproximação $\boldsymbol{\theta}^{(a+1)}$ para o vetor de parâmetros $\boldsymbol{\theta}^{(a)}$, através de

$$\boldsymbol{\theta}^{(a+1)}=\boldsymbol{\theta}^{(a)}+\left[\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right)\right]^{-1}\boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{r}\left(\boldsymbol{\theta}^{(a)}\right). \quad (2.12)$$

Repete-se o processo em (2.12) até que se tenha a convergência de acordo com algum critério, como por exemplo, $\boldsymbol{\theta}^{(a+1)}-\boldsymbol{\theta}^{(a)}<0{,}000001$.

2.7.2 Método de Newton-Raphson

Outro método iterativo muito utilizado é o de Newton-Raphson, que usa aproximação quadrática local para a função objetivo (função a ser minimizada), Bates e Watts (1988). Esse método é uma abordagem mais geral do que o citado acima, no qual $SQRes(\boldsymbol{\theta})$ é expandido diretamente usando uma expansão quadrática de Taylor.

Definindo,

$$\boldsymbol{g}(\boldsymbol{\theta})=\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

e

$$\boldsymbol{H}(\boldsymbol{\theta})=\frac{\partial^2 S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$$

denota-se, respectivamente, o vetor gradiente e a matriz Hessiana de $SQRes(\boldsymbol{\theta})$. Então temos a aproximação quadrática dada abaixo

$$\boldsymbol{S}(\boldsymbol{\theta})\approx \boldsymbol{S}\left(\boldsymbol{\theta}^{(a)}\right)+\boldsymbol{g}'\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}\right)+\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}\right)'\boldsymbol{H}\left(\boldsymbol{\theta}^{(a)}\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}\right) \quad (2.13)$$

(2.13) terá um ponto estacionário quando seu gradiente é zero e esse ponto será um mínimo se $\boldsymbol{H}\left(\boldsymbol{\theta}^{(a)}\right)$ é positiva definida, isto é, quando

$$\frac{\partial S(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)})} = \boldsymbol{g}(\boldsymbol{\theta}^{(a)}) + \frac{1}{2}2\boldsymbol{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = \boldsymbol{g}(\boldsymbol{\theta}^{(a)}) + \boldsymbol{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)})$$

$$\frac{\partial S(\boldsymbol{\theta})}{\partial(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)})} = 0 \;\rightarrow\; \boldsymbol{g}(\boldsymbol{\theta}^{(a)}) + \boldsymbol{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = 0$$

$$\boldsymbol{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = -\boldsymbol{g}(\boldsymbol{\theta}^{(a)}). \tag{2.14}$$

Pré-multiplicando ambos os lados da equação (2.14) por $\left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}$, tem-se que:

$$\left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}\boldsymbol{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = -\left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(a)})$$

$$\boldsymbol{I}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = -\left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(a)})$$

$$(\boldsymbol{\theta}-\boldsymbol{\theta}^{(a)}) = -\left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(a)}).$$

Se $\boldsymbol{H}(\boldsymbol{\theta}^{(a)})$ é positiva definida, o passo de Newton-Raphson é

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} - \left[\boldsymbol{H}(\boldsymbol{\theta}^{(a)})\right]^{-1}\boldsymbol{g}(\boldsymbol{\theta}^{(a)}). \tag{2.15}$$

O processo em (2.15) é repetido até a convergência como no método citado anteriormente.

2.8 Valores iniciais para o modelo

Para os métodos iterativos citados acima é necessário a escolha de bons valores iniciais. Decisões ruins a respeito dessas estimativas podem levar o algoritmo a convergir para um máximo local ao invés de um máximo global, além de se tornar mais demorado, ou seja, uma má escolha fará com que o método precise de muitas iterações, ou até mesmo poderá não atingir a convergência. Entre as técnicas mais conhecidas para escolha de estimativas iniciais de um modelo não linear estão as seguintes:

- Linearização: aplicar alguma transformação no modelo não linear para que a função $f(\boldsymbol{X}_i, \boldsymbol{\theta})$ se torne linear nos parâmetros $\boldsymbol{\theta}$. Se for possível linearizá-la pode-se realizar uma regressão linear da variável transformada sobre os novos parâmetros e usar as estimativas

resultantes (depois de aplicar a transformação inversa) dessa regressão como estimativas iniciais para o modelo não linear, ignorando o erro aleatório.

- Propriedades do modelo: estudar o comportamento limite da função esperança quando x_i vai para zero ou infinito e assim substituir os parâmetros por valores iniciais que representem nas condições de escala e contexto do problema.

- Análise gráfica: outra maneira bastante utilizada é a análise visual dos dados para “chutar” valores iniciais para os parâmetros, mas o conhecimento do pesquisador sobre o modelo estudado é fundamental.

2.9 Medidas de não linearidade

O objetivo principal das medidas de não linearidade é verificar o grau de não linearidade, no que diz respeito ao problema de estimação de um modelo não linear, ele precisa ser suficientemente pequeno para que as técnicas conhecidas de estimação, desenvolvidas para a regressão linear, se torne uma boa aproximação e só assim é possível utilizar as mesmas no modelo não linear.

> Dentre as muitas vantagens em se buscar modelos com comportamento próximo do comportamento linear, podem-se citar algumas:
> - as estimativas de mínimos quadrados podem ser facilmente obtidas;
> - os estimadores são aproximadamente não-viciados, normalmente distribuídos com variância mínima, mesmo em pequenas amostras;
> - os valores de previsões são mais precisos;
> - os métodos iterativo convergem mais rapidamente;
> - os estimadores têm propriedades similares às propriedades ótimas de modelos lineares. (MAZUCHELI; ACHCAR, 2002, p. 1766).

Beale (1960) foi o primeiro a tentar medir a não linearidade de um modelo, porém suas medidas previam um comportamento linear, mesmo que a não linearidade estivesse presente. Outra contribuição importante foi a de Box (1971), que apresentou cálculo associado aos parâmetros estimados, o viés de ordem n^{-1} do estimador de máxima verossimilhança. Já Gills e Ratkowsky (1978), constataram através do estudo de simulação que, a fórmula do vício de Box não só estima o viés de maneira aproximada como também traz uma boa indicação do comportamento não linear do modelo.

Bates e Watts (1980) desenvolveram novas medidas de não linearidade com base no conceito de curvatura geométrica, elas são conhecidas e muito utilizadas na prática.

2.9.1 Medidas de curvatura de Bates e Watts

Bates e Watts (1980) mostraram que a não linearidade de um modelo pode ser decomposta em duas componentes, e são elas: a curvatura intrínseca e a curvatura devida ao efeito dos parâmetros. A primeira está relacionada com a superfície esperada que é gerada pelo modelo, a segunda se refere a suposição de que o sistema de coordenadas que está sobre a superfície esperada possuem retas paralelas igualmente espaçadas.

Em modelos de regressão linear as duas medidas terão valor zero, pois a superfície gerada será um plano e as coordenadas sobre ela serão regulares.

Suponha que $\boldsymbol{\theta}$ esteja próximo de $\boldsymbol{\theta}^{(a)}$, então do mesmo modo feito em (2.9), porém fazendo a expansão de Taylor até o termo quadrático, tem-se que:

$$f(\boldsymbol{X}_i, \boldsymbol{\theta}) \approx f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right) + \frac{\partial f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right)}{\partial \boldsymbol{\theta}_r}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) +$$

$$+\frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right)' \frac{\partial^2 f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right)}{\partial \boldsymbol{\theta}_r \boldsymbol{\theta}_s}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right); \quad r, s = 1,2, \dots, p. \tag{2.16}$$

Fazendo $\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) = \boldsymbol{\delta}$, $\frac{\partial f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right)}{\partial \boldsymbol{\theta}_r} = \boldsymbol{F}$ a matriz de derivadas parciais de primeira ordem de dimensões $n \times p$ e $\frac{\partial^2 f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^{(a)}\right)}{\partial \boldsymbol{\theta}_r \boldsymbol{\theta}_s} = \boldsymbol{H}$ é a matriz $n \times p \times p$ em que cada face de $\boldsymbol{H}$ é uma matriz de derivadas parciais de segunda ordem de dimensões $p \times p$.

Reescrevendo a equação (2.16), segue que

$$f(\boldsymbol{\theta}) \approx f\left(\boldsymbol{\theta}^{(a)}\right) + \boldsymbol{F\delta} + \frac{1}{2}\boldsymbol{\delta' H \delta}.$$

De acordo com Bates e Watts (1988) o grau de linearização de $f(\boldsymbol{\theta})$ é medido comparando a magnitude do termo quadrático $\boldsymbol{\delta' H \delta}$ em relação ao termo linear $\boldsymbol{F\delta}$, assim $\boldsymbol{\delta' H \delta}$ foi dividido em duas componentes ortogonais, as projeções no plano tangente e normal ao plano tangente. Essa decomposição pode ser realizada utilizando-se a matriz de projeção

$$\widehat{\boldsymbol{P}} = \boldsymbol{F}(\widehat{\boldsymbol{\theta}})\left[\boldsymbol{F}'(\widehat{\boldsymbol{\theta}})\boldsymbol{F}(\widehat{\boldsymbol{\theta}})\right]^{-1}\boldsymbol{F}'(\widehat{\boldsymbol{\theta}}).$$

Portanto, as duas componentes são dadas por:

$$\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{H}}^{T} + \widehat{\boldsymbol{H}}^{N},$$

em que $\widehat{\boldsymbol{H}}^{T} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{H}}$ é a componente tangencial e $\widehat{\boldsymbol{H}}^{N} = \left(\boldsymbol{I} - \widehat{\boldsymbol{P}}\right)\widehat{\boldsymbol{H}}$ é a componente normal. Com base nessa decomposição Bates e Watts (1980) definiram as medidas como:

$$K^{T} = \frac{\left\|\boldsymbol{\delta}'\widehat{\boldsymbol{H}}^{T}\boldsymbol{\delta}\right\|}{\left\|\widehat{\boldsymbol{F}}\boldsymbol{\delta}\right\|^{2}} \tag{2.17}$$

e

$$K^{N} = \frac{\left\|\boldsymbol{\delta}'\widehat{\boldsymbol{H}}^{N}\boldsymbol{\delta}\right\|}{\left\|\widehat{\boldsymbol{F}}\boldsymbol{\delta}\right\|^{2}} \tag{2.18}$$

em que K^{T} é o efeito de curvatura de parâmetros e K^{N} é o efeito de curvatura intrínseca. Para obter uma medida invariante à escala da variável resposta, adotaram padronizar esses valores pelo raio padrão ρ, logo essas medidas foram definidas por:

$$\gamma^{T} = \rho K^{T} \quad \text{e} \quad \gamma^{N} = \rho K^{N}$$

em que $\rho = s\sqrt{p}$, com p o número de parâmetros e $s = \left\{\frac{\sum_{i=1}^{n}[y_i - f(\boldsymbol{X}_i,\boldsymbol{\theta})]^2}{n-p}\right\}^{1/2}$.

Bates e Watts (1980) sugerem para determinar o quanto um modelo ajustado possui curvatura significativa o critério $\gamma > 2\left(F_{p,n-p}^{\alpha}\right)^{-\frac{1}{2}}$, quanto mais próximo de zero os valores de γ^{T} e γ^{N} mais válidos serão os resultados inferenciais do modelo não linear utilizado.

2.9.2 Medida de vício de Box

Segundo Cook *et al*. (1986), os estimadores de máxima verossimilhança dos parâmetros em modelos não lineares geralmente são tendenciosos ou viesados. À medida que a amostra se torna menor ou a não linearidade se intensifica, o vício das estimativas passa a ser muito alto, logo é de grande valia a disponibilidade de fórmulas para poder avaliá-lo.

Box (1971), propôs a seguinte expressão, também conhecida como vício de Box:

$$B(\widehat{\boldsymbol{\theta}}) = -\frac{\sigma^2}{2}\left(\sum_{i=1}^{n}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{F}'(\boldsymbol{\theta})\right)^{-1}\sum_{i=1}^{n}\boldsymbol{F}(\boldsymbol{\theta})\, tr\left[\left(\sum_{i=1}^{n}\boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{F}'(\boldsymbol{\theta})\right)^{-1}\boldsymbol{H}(\boldsymbol{\theta})\right] \quad (2.19)$$

onde $\boldsymbol{F}(\boldsymbol{\theta})$ é o vetor $p \times 1$ de primeiras derivadas do modelo e $\boldsymbol{H}(\boldsymbol{\theta})$ é a matriz $p \times p$ de segundas derivadas, ambos com respeito aos elementos de $\boldsymbol{\theta}$. No lugar das quantidades desconhecidas são utilizadas $\widehat{\boldsymbol{\theta}}$ e $\hat{\sigma}^2$ e tr indica a operação traço.

Em geral o vício de Box é dado em porcentagem, o que é mais interessante pois é livre de escala e é dada por:

$$\%B(\widehat{\boldsymbol{\theta}}) = \frac{100.B(\widehat{\boldsymbol{\theta}})}{\widehat{\boldsymbol{\theta}}} \quad (2.20)$$

De acordo com Box (1971) a porcentagem de vício aceitável seria aquela em que $B(\theta_i) < 1\%$.

2.10 Análises de diagnóstico

A análise de diagnóstico é fundamental quando se trabalha com a modelagem estatística, pois é essa a etapa em que se avalia os modelos, em termos de qualidade e performance dos ajustes dos mesmos, bem como a verificação se as suposições estão sendo satisfeitas. A seguir será apresentado alguns métodos de diagnóstico.

2.10.1 Matriz de projeção

A matriz de projeção ou matriz chapéu é utilizada na detecção de pontos mais afastados dos demais, possivelmente, pontos influentes, que em geral exercem uma grande influência sobre a matriz de variâncias e covariâncias, ela é definida por

$$\widehat{\boldsymbol{P}} = \boldsymbol{F}(\widehat{\boldsymbol{\theta}})\left[\boldsymbol{F}'(\widehat{\boldsymbol{\theta}})\boldsymbol{F}(\widehat{\boldsymbol{\theta}})\right]^{-1}\boldsymbol{F}'(\widehat{\boldsymbol{\theta}}). \quad (2.21)$$

Diferentemente da regressão linear, essa é uma matriz de projeção local, pois depende de $\boldsymbol{\theta}$ e deve ser estimada em $\widehat{\boldsymbol{\theta}}$, porém o critério que serve como guia para detectar pontos suspeitos de serem influentes é o mesmo da regressão linear, calculado como: $\hat{p}_{ii} \geq \frac{2p}{n}$.

2.10.2 Resíduo projetado

Cook e Tsai (1985) empregaram uma expansão quadrática para poderem investigar o comportamento dos resíduos em modelos não lineares, pois eles podem produzir resultados enganosos quando utiliza-se os mesmos métodos de diagnósticos da regressão linear. Eles sugeriram um novo tipo de resíduo que supera as muitas falhas dos resíduos ordinários.

No modelo de regressão não linear os resíduos ordinários são definidos por $r_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\theta}),\ \ i = 1, \dots, n.$ Esses podem não produzir a forma correta da distribuição dos erros. Cook e Tsai (1985) apresentaram um novo tipo de resíduo que pode solucionar as muitas deficiências dos resíduos usuais.

Definindo $C(F)$ como sendo o subespaço gerado pelas colunas de F e $\bar{C}(F)$ o subespaço ortogonal a $C(F)$. Sejam as matrizes $p \times p$, $\boldsymbol{W}_i = \frac{\partial^2 f(x_i,\theta)}{\partial \theta_r \theta_s}$, $r, s = 1,2, \dots, p,$ de derivadas de segunda ordem para $i = 1, \dots, n$ e $\boldsymbol{W}$ como uma matriz $n \times (p \times p)$ com n faces, em que, n é o número de observações e p o de parâmetros cuja i-ésima face é igual a $\boldsymbol{W}_i$. Cook e Tsai (1985) propuseram a seguinte aproximação para o resíduo ordinário $\boldsymbol{r}$

$$\boldsymbol{r} \cong (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon} - \boldsymbol{F}\sum_{i=1}^{n} r_i \boldsymbol{W}_i\, \boldsymbol{\delta} - \frac{1}{2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\delta}'\boldsymbol{W}\boldsymbol{\delta}. \qquad (2.22)$$

Projeta-se cada coluna de derivadas de segunda ordem de $\boldsymbol{W}$ sobre $\bar{C}(F)$, tem-se $\boldsymbol{P_1} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$ o operador de projeção ortogonal do espaço gerado por sobre essa projeção. $\boldsymbol{S} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{T}$, em que $\boldsymbol{T}$ é uma matriz $n \times q$ obtida através dos vetores não nulos, de derivadas de segunda ordem, $\boldsymbol{P} = \boldsymbol{F}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{F}'$ e $\boldsymbol{I}$ é a matriz identidade de dimensão $n \times n$. Definindo-se $\boldsymbol{V}$ uma matriz $n \times (p + q)$ definida como $\boldsymbol{V} = (\boldsymbol{F}, \boldsymbol{S})$ e $\boldsymbol{P_2} = \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$ o operador de projeção ortogonal em $C(V)$. Cook e Tsai (1985) definiram o resíduo projetado utilizando as matrizes descritas anteriormente

$$(\boldsymbol{I} - \boldsymbol{P_2})\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\varepsilon} - (\boldsymbol{I} - \boldsymbol{P_1})\boldsymbol{\varepsilon}. \qquad (2.23)$$

O primeiro termo de (2.23) é a aproximação para o resíduo ordinário $\boldsymbol{r}$ e o segundo é a perda de informação necessária para remover os componentes não lineares de (2.22). Além disso, sabe-se que

$$E\{(\boldsymbol{I}-\boldsymbol{P_2})\boldsymbol{r}\}=\boldsymbol{0},\ var\{(\boldsymbol{I}-\boldsymbol{P_2})\boldsymbol{r}\}=\ \sigma^2(\boldsymbol{I}-\boldsymbol{P_2})$$

e

$$E\{\boldsymbol{r}'(\boldsymbol{I}-\boldsymbol{P_2})\boldsymbol{r}\}=\sigma^2 tr(\boldsymbol{I}-\boldsymbol{P_2})\ .$$

Então, a estimativa alternativa para σ^2 é dada por

$$\tilde{\sigma}^2=\frac{\boldsymbol{r}'(\boldsymbol{I}-\boldsymbol{P_2})\boldsymbol{r}}{tr(\boldsymbol{I}-\boldsymbol{P_2})}$$

De acordo com Prudente (2009) os resíduos projetados são melhores do que os ordinários e muitas das técnicas de diagnóstico aplicadas na regressão linear são também aplicáveis aos mesmos. Em regressão linear, muitas vezes os resíduos ordinários não são adequados, devido à heterogeneidade das variâncias e a presença de correlação entre eles, por isso definem-se os resíduos ordinários estudentizados que mesmo na presença de correlação apresentam média zero e variância constante igual a um. O i-ésimo resíduo ordinário estudentizados é expresso por

$$t_i=\frac{r_i}{S\sqrt{(1-p_{ii})}}\ ,i=1,\dots,n. \tag{2.24}$$

onde, r_i é o i-ésimo resíduo ordinário, S é o desvio padrão estimado para o modelo ajustado e p_{ii} é o i-ésimo elemento da diagonal da matriz de projeção $\boldsymbol{P}$.

Em modelos não lineares tem-se os resíduos projetados estudentizados que possui esperança nula e variância $\sigma^2(\boldsymbol{I}-\boldsymbol{P_2})$ e são definidos por

$$S_i=\frac{\left(\boldsymbol{I}-\widehat{\boldsymbol{P}}_{\boldsymbol{2}}\right)\boldsymbol{r_i}}{\tilde{\sigma}\sqrt{\left\{\left(\boldsymbol{I}-\widehat{\boldsymbol{P}}_{\boldsymbol{2}}\right)\right\}_{ii}}}\ ,i=1,2\dots,n. \tag{2.25}$$

> Para avaliar se os erros $\varepsilon_i's$ têm distribuição, aproximadamente, normal, assim como para detectar se há pontos aberrantes e/ou influentes, o gráfico de probabilidades dos resíduos projetados ordenados $S_{(i)}$ versus $\Phi^{-1}\left(\frac{i-\frac{3}{8}}{n+\frac{1}{4}}\right)$ pode ser útil, sendo $\Phi(.)$ a função de distribuição acumulada da normal padrão. (PRUDENTE, 2009, p. 49).

A partir de (2.25) a análise dos resíduos procede da mesma maneira do modelo linear.

2.10.3 Medidas de influência

Uma etapa importante na análise de diagnóstico é a detecção de observações influentes, elas podem alterar de forma considerável alguma propriedade do modelo ajustado, como por exemplo, as estimativas dos parâmetros e sua significância, seus erros padrões, valores ajustados, etc.

Uma medida de influência é uma quantidade que tem como objetivo medir o quanto uma particular observação afeta as propriedades do modelo. Uma muito conhecida é a distância de Cook, que foi proposta por Cook (1977) para modelos lineares e rapidamente estendida para os não lineares, sendo definida pela fórmula

$$D_i = \frac{\left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)'\left(\widehat{\boldsymbol{F}}'\widehat{\boldsymbol{F}}\right)\left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)}{p\hat{\sigma}^2} \tag{2.26}$$

em que $\widehat{\boldsymbol{\theta}}$ é o vetor das estimativas dos parâmetros, $\widehat{\boldsymbol{\theta}}_{(i)}$ é o vetor das estimativas de mínimos quadrados com a i-ésima observação excluída, $\widehat{\boldsymbol{F}}$ é o vetor gradiente, p é o número de parâmetros e $\hat{\sigma}^2 = \frac{SQRes(\hat{\theta})}{n-p}$.

A forma aproximada para expressão dada em (2.26) é definida por:

$$D_i^1 = \frac{t_i^2}{p}\frac{\hat{p}_{ii}}{(1-\hat{p}_{ii})}\ , i = 1,2,\dots,n \tag{2.27}$$

em que t_i são os resíduos ordinários estudentizados, p é o número de parâmetros e $\hat{p}_{ii}$ é o i-ésimo elemento da diagonal da matriz de projeção $\widehat{\boldsymbol{P}}$ definida em (2.21).

Graficamente pode-se avaliar a influência das observações nas estimativas dos parâmetros, plotando D_i^1 versus a ordem das observações, deve-se dar mais atenção para pontos muito afastados dos demais.

2.11 Métricas para seleção do modelo

A verificação da qualidade do ajuste de um modelo não linear é uma das partes importantes quando se trabalha com modelagem de dados, há diferentes aspectos a ser observados, busca-se aquele modelo que seja mais parcimonioso, ou seja, que envolve o número mínimo de parâmetros possíveis a serem estimados e que explique muito bem a variável resposta. Para esse fim existem vários avaliadores dessa qualidade, como teste de hipóteses e critérios de informação que podem ser aplicados, sendo fáceis de serem calculados e interpretados. A seguir são expostos alguns deles.

2.11.1 Coeficiente de determinação

O coeficiente de determinação é uma medida estatística que diz o quão próximo os dados estão do modelo ajustado, é a porcentagem da variabilidade da variável resposta y que é explicada pela variável independente x, quanto maior melhor. É também chamado de R-quadrado e é denotado por

$$R^2 = \frac{SQReg}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.28}$$

onde,

$0 < R^2 < 1$,

$SQReg$ é a soma de quadrados da regressão,

SQT é a soma de quadrados totais,

$\hat{y}_i$ são os valores estimados pelo modelo,

$\bar{y}$ é a média dos valores observados,

y_i são os valores observados.

2.11.2 Coeficiente de determinação ajustado

O coeficiente de determinação (R^2) possui limitações, uma delas é que ao acrescentar mais variáveis ao modelo analisado esse valor tende a aumentar, por mero acaso, ele nunca diminui, logo pode parecer que esse modelo com mais variáveis é melhor, somente por conter mais termos, portando o coeficiente de determinação ajustado (${R^2}_{aj}$) foi criado para eliminar esse problema, pois ele aumenta

somente se a nova variável acrescentada for significante para a melhoria do modelo avaliado.

O coeficiente de determinação ajustado é calculado da seguinte forma:

$$R^2{}_{aj} = 1 - \left[\frac{(1 - R^2)(n - i)}{(n - p)}\right] \tag{2.29}$$

em que,

$R^2 = \frac{SQReg}{SQT}$,

$SQReg$ = soma de quadrado da regressão,

SQT = soma de quadrado total,

n é o número de observações

p é a quantidade de parâmetros incluindo o intercepto,

i será 1 se tiver intercepto e 0 se não tiver.

Quanto maior for o $R^2{}_{aj}$, ou seja, mais próximo de um, melhor será o modelo ajustado.

2.11.3 Erro quadrático médio (MSE)

O Erro quadrático médio é uma métrica bastante utilizada, ela consiste na média do erro das previsões ao quadrado. Em outras palavras, é um valor que consiste na diferença entre o valor predito pelo modelo e o valor real, eleva-se o resultado ao quadrado, faz-se a mesma coisa para todas as observações, soma-os, e divide-se pelo número de observações preditas. Quanto menor esse valor, melhor é o modelo.

Essa métrica tem como mínimo o valor zero e sem valor máximo, e ela é dada pela seguinte fórmula:

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n} \tag{2.30}$$

em que,

n é o número de observações;

$\hat{y}_i$ é o valor estimado pelo modelo analisado;

y_i é o valor observado.

2.11.4 Critérios de informação

O critério de informação de Akaike (AIC), o critério de informação bayesiano (BIC) são outras estatísticas que servem para medir e comparar a qualidade de ajuste dos modelos, são baseadas no máximo da função de verossimilhança e dependem tanto do número de observações como dos parâmetros do modelo.

Quando o intuito é trabalhar com a modelagem, utiliza-se modelos para representar o processo que gerou aos dados, essa representação quase nunca é exata; logo, algumas informações serão perdidas. O AIC estima a quantidade relativa dessa perda ao aplicar o ajuste de certo modelo ao conjunto de dados e quanto menos informação ele perde, maior sua qualidade.

A estimativa do AIC é dada pela seguinte fórmula:

$$AIC = -2logL(\hat{\theta}) + 2p \tag{2.31}$$

onde,

$L(\hat{\theta})$ é o máximo da função de verossimilhança;

p é o número de parâmetros do modelo.

O BIC também mede a qualidade do ajuste, ele está intimamente relacionado ao critério de informação de Akaike.

Ao ajustar modelos, pode-se adicionar parâmetros, porém essa adição pode resultar em sobreajuste ou overfitting, que é quando um modelo se ajusta muito bem ao conjunto de dados, mas ele é ineficaz para prever novos valores. Sendo assim tanto o BIC como o AIC introduzem um termo de penalidade para o número de parâmetros.

A estimativa do BIC é dada pela fórmula abaixo:

$$BIC = -2logL(\hat{\theta}) + plog(n) \tag{2.32}$$

onde,

$L(\hat{\theta})$ é o máximo da função de verossimilhança;

p é o número de parâmetros do modelo;

n é o número de observações, ou seja, o tamanho da amostra.

Tanto o AIC como o BIC são utilizados para comparar modelos não aninhados e o melhor modelo, considerando o ajuste, será aquele que apresentar menor valor.

2.11.5 Teste de razão de verossimilhanças

Ao ajustar vários modelos aos dados trabalhados e ao final restar dois igualmente bem ajustados, é necessário fazer comparação entre eles e decidir qual deles é o preferido. Se estes forem aninhados, ou seja, um caso especial um do outro, pode-se utilizar um teste de razão de verossimilhanças (TRV), do mesmo modo feito em modelos lineares, baseado no ln da razão entre duas verossimilhanças, cujas hipóteses são:

$$H_0 : O\ modelo\ mais\ simples \text{ é } o\ correto;$$
$$H_a : O\ modelo\ mais\ parametrizado \text{ é } o\ correto.$$

O TRV utiliza a estatística $L = 2log\frac{L_1}{L_2} = 2[\log(L_2) - \log(L_1)\,]$ sendo L_2 o máximo do logaritmo natural da função de verossimilhança para o modelo com mais parâmetros e L_1 para o modelo mais simples. Sob a hipótese H_0 verdadeira $L \sim {\chi^2}_{(\delta,v)}$, com δ o parâmetro de não centralidade e v graus de liberdade, sendo v a diferença entre o número de parâmetros dos modelos que estão sendo comparados. Rejeita-se H_0 quando $L > {\chi^2}_{(\delta,v)}$.

2.12 Resultados Assintóticos

Em modelos de regressão linear, sabe-se que os estimadores de mínimos quadrados ordinários são não viesados e de variância uniformemente mínima para os parâmetros do modelo, ou seja, dentre todos os estimadores não viciados eles possuem a menor variância, podem ser chamados também de eficientes.

Já no caso não linear as propriedades expostas acima só são válidas assintoticamente, isso é, quando o tamanho da amostra for suficientemente grande. Geralmente para amostras pequenas esses resultados são desconhecidos. Entretanto, segundo Mazucheli e Achcar (2002) quanto maior o tamanho amostral os resultados assintóticos vão se tornando mais aplicáveis.

Assumidas as suposições usuais de que os erros, ε_i, são normalmente distribuídos, $\varepsilon_i \sim N(0, \sigma^2 \boldsymbol{I})$, também são iid, ou seja, independentes e identicamente distribuídos com variância σ^2, sob certas condições de regularidade, $\widehat{\boldsymbol{\theta}}$ e $S^2 = \frac{SQRes(\widehat{\boldsymbol{\theta}})}{n-p}$

são estimadores consistentes de $\boldsymbol{\theta}$ e σ^2 respectivamente; tem-se os seguintes resultados aproximados:

(i) $\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}\right) \sim N(\mathbf{0}, \sigma^2 \boldsymbol{C^{-1}})$, onde $\boldsymbol{C} = \boldsymbol{F}'\left(\boldsymbol{\theta}^{(a)}\right)\boldsymbol{F}\left(\boldsymbol{\theta}^{(a)}\right) = \boldsymbol{F}'\boldsymbol{F}$;

(ii) $\frac{(\mathrm{n-p})S^2}{\sigma^2} \sim \frac{\boldsymbol{\varepsilon}'(\boldsymbol{I_n} - \boldsymbol{P})\boldsymbol{\varepsilon}}{\sigma^2} \sim {\chi^2}_{n-p}$, onde $\boldsymbol{P} = \boldsymbol{F}[\boldsymbol{F}'\boldsymbol{F}]^{-1}\boldsymbol{F}'$;

(iii) $\widehat{\boldsymbol{\theta}}$ é estatisticamente independente de S^2;

(iv) $\frac{[SQRes\left(\boldsymbol{\theta}^{(a)}\right) - SQRes(\boldsymbol{\theta})]/p}{SQRes\left(\boldsymbol{\theta}^{(a)}\right)/(n-p)} \sim \frac{\boldsymbol{\varepsilon}'\boldsymbol{P}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'(\boldsymbol{I_n} - \boldsymbol{P})\boldsymbol{\varepsilon}}\left(\frac{n-p}{p}\right) \sim F_{p,n-p}$.

Para avaliar a precisão dos resultados assintóticos discutidos acima usa-se as medidas de não linearidade expostas na seção 2.9, através delas consegue-se avaliar o quão próximo um modelo não linear está do linear, e quanto mais aproximado for os resultados assintóticos serão mais precisos e, consequentemente, mais confiáveis serão as inferências.

Logo após obtida as estimativas dos parâmetros, é de interesse definir os intervalos de confiança $100(1-\alpha)\%$ tanto para os parâmetros bem como para a resposta esperada, eles são calculados pelas seguintes fórmulas, respectivamente

$$\hat{\theta}_i \pm t_{n-p}^{\alpha/2} S\sqrt{\hat{c}^{ii}}$$

e

$$\hat{y}_0 \pm t_{n-p}^{\alpha/2} S\sqrt{1 + \boldsymbol{f_0'}(\boldsymbol{F}'\boldsymbol{F})^{-1}\boldsymbol{f_0}}$$

em que, $\hat{c}^{ii}$ é o i-ésimo elemento da diagonal da matriz $\widehat{\boldsymbol{C}}$, $\hat{y}_0 = f\left(\boldsymbol{x_0}; \widehat{\boldsymbol{\theta}}\right)$ o valor previsto, $\boldsymbol{f_0'} = \left(\frac{\partial f(\boldsymbol{x_0};\boldsymbol{\theta})}{\partial \boldsymbol{\theta_1}}, \frac{\partial f(\boldsymbol{x_0};\boldsymbol{\theta})}{\partial \boldsymbol{\theta_2}}, \ldots, \frac{\partial f(\boldsymbol{x_0};\boldsymbol{\theta})}{\partial \boldsymbol{\theta_p}}\right)$.

Para testar as hipóteses H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ $versus$ H_1: $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, utiliza-se a estatística do teste dada por

$$F_1 = \frac{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'\widehat{\boldsymbol{C}}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}{pS^2}$$

que tem uma distribuição aproximada de $F_{p,n-p}$. Rejeita-se H_0 ao nível de significância α se, $F_1 > F_{p,n-p}$.

2.13 Teste de Pettitt para Rupturas

O teste de ruptura de Pettitt é um complemento do teste não-paramétrico de Mann-Kendall que é utilizado para determinar se uma série de dados possui uma

tendência temporal de alteração. O teste de Pettitt tem a finalidade de avaliar a existência de alguma ruptura (descontinuidade, quebra), ou seja, esta estatística localiza o ponto de mudança de tendência nos dados analisados. O objetivo é verificar se duas amostras pertencem a mesma população, a estatística utilizada faz uma contagem do número de vezes em que um elemento de uma amostra é maior do que na outra. A estatística considera os valores de $t = 2,\ldots,T$ e analisa a existência ou não de algum ponto de mudança na série de dados.

A estatística do teste é definida como: $K_T = \max_{1\leq t<T} |U_{t,T}|$, onde, $U_{t,T} = \sum_{i=1}^{t}\sum_{j=t+1}^{T} D_{ij}$ e $D_{ij} = sinal(X_i - X_j)$, considerando o sinal como sendo 1 se $x > 0$, 0 se $x = 0$ e -1 se $x < 0$.

As hipóteses para o teste são:

$$H_o: não\ há\ nenhuma\ mudança\ na\ série$$

$$H_A: existe\ alguma\ mudança\ na\ série$$

Para rejeição ou não da hipótese nula, utiliza-se o p-valor obtido no teste, rejeitando-se H_o para p-valor pequeno. Para esse trabalho adotou-se $p - valor \leq 0{,}05$. Ao rejeitar H_o pode-se concluir que existência um ponto de mudança na série no ponto t.

O pacote Trend do software R possui o teste de Pettitt, como resultado do mesmo tem-se a estatística do teste, p-valor e o ponto t de mudança de tendência.

3 APLICAÇÃO

Os dados utilizados nesta aplicação foram retirados de um aplicativo, resultado de um trabalho em conjunto de professores, alunas e alunos de pós-graduação em Estatística da UFMG e as análises foram feitas no software R. As bases de dados utilizadas foram para o Brasil e o estado de São Paulo, elas possuem informações sobre a quantidade acumulada de casos da COVID-19, óbitos acumulados, novos confirmados, novas mortes, bem como a data cujo período vai de 25/02/2020 a 14/10/2022. Para o estado de São Paulo e Brasil optou-se por trabalhar com os dados do ano de 2020 a partir de 26/02/2020, data que apresenta a primeira ocorrência da doença, até o dia 31/12/2020. Já para as mortes acumuladas tanto de São Paulo como do Brasil utilizou-se na modelagem o dia 17/03/2020 (dia do primeiro óbito) até 31/12/2020.

3.1 Análise para o estado de São Paulo

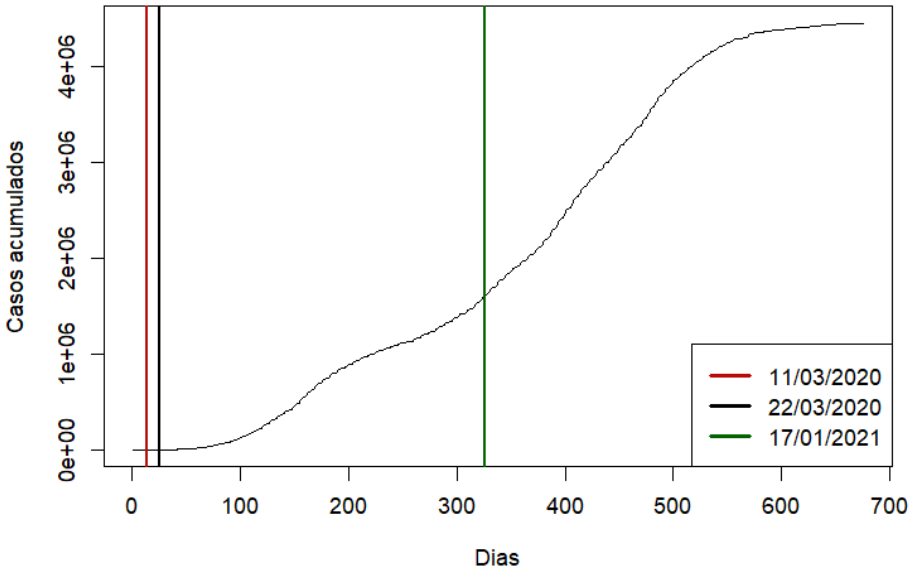
3.1.1 Casos confirmados acumulados pela doença COVID-19

3.1.1.1 Análise Exploratória para os dados de São Paulo

A seguir será apresentado o gráfico de casos acumulados do Estado de São Paulo referente aos anos de 2020 e 2021, cujo período vai de 26/02/2020 a 31/12/2021, a linha vermelha em destaque se refere a data que foi declarada pandemia em relação ao novo coronavírus pela Organização Mundial da Saúde, a linha preta marca o começo da quarentena no estado de São Paulo e a linha verde o início da vacinação contra doença após a aprovação do uso emergencial da vacina do Instituto Butantan pela Anvisa (Agência Nacional de Vigilância Sanitária), a primeira contemplada foi uma enfermeira de 54 anos da UTI do Instituto de Infectologia Emílio Ribas.

Na Figura 4 pode-se notar um aumento considerável nos casos acumulados do dia 306 ao dia 500 que corresponde respectivamente a 28/12/2020 a 10/07/2021, aproximadamente a partir do dia 29/08/2021 os casos começam se estabilizar.

Figura 4 – Gráfico dos casos acumulados da COVID-19 do ano de 2020 e 2021 para São Paulo.



Fonte: Elaborado pela autora (2022).

A Tabela 3 contém algumas informações estatísticas referentes aos casos acumulados da doença analisada, no estado de São Paulo. Nota-se através das médias que houve um aumento considerável dos casos acumulados passando de mais de meio milhão em 2020 para mais de 3 milhões em 2021. Pelo coeficiente de variação apresentado, observa-se uma alta dispersão nos dados de 2020, já para 2021 tem-se um valor mais baixo de 29,65%, geralmente considera-se um coeficiente de variação baixo quando este possuir valor menor ou igual a 25%. Pode-se verificar também, a partir da Tabela 3, que em 75% dos dias a quantidade de casos acumulados é menor ou igual a 1.049.975 para 2020 e para 2021 é 4.367.748. Já na metade dos dias esse valor é menor ou igual a 521.601 e na outra metade é maior ou igual a 521.601 em 2020. Já para 2021 tem-se uma quantidade de 3.762.758 ou menos na primeira metade dos dias e maior ou igual a esse valor para a outra metade.

Tabela 3 – Medidas estatísticas referentes aos anos de 2020 e 2021 para casos acumulados da COVID-19 em São Paulo.

	2020	2021
Quantidade de dias	310	365
Média	577.519	3.403.362
Desvio Padrão	495.105	1.009.143
Mínimo	1	1.466.191
Q1	51.894	2.513.178
Mediana	521.601	3.762.758
Q3	1.049.975	4.367.748
Máximo	1.462.297	4.456.108
Coeficiente de Variação	85,73%	29,65%

Fonte: Elaborado pela autora (2022).

3.1.1.2 Ajuste de modelos para os dados do estado de São Paulo

Ajustou-se aos dados de casos acumulados de 2020 os modelos 1, 3 e 4 expostos na Tabela 2 da seção 2.5.1, que são os modelos não lineares de crescimento Exponencial, Logístico e Gompertz respectivamente.

As estimativas encontradas para os parâmetros de cada modelo, bem como os valores do erro padrão das estimativas e o p-valor estão na Tabela 4. Observa-se que as estimativas dos parâmetros foram significativas ao nível de 0,05 de significância estatística.

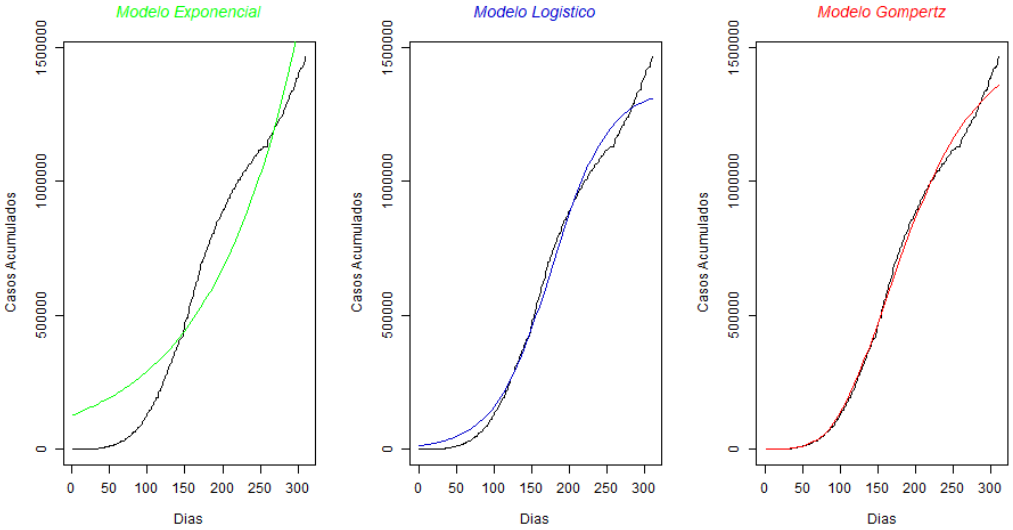
Tabela 4 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para os casos de COVID-19 em São Paulo.

Modelo	Parâmetros	Estimativas	E.P.	p-valor
Exponencial	α	125.546,2	6.825,3010	<0,0001
	β	0,0084	0,0002	<0,0001
Logístico	α	1.346.761	9.274,5578	<0,0001
	β	106,3704	7,1555	<0,0001
	γ	0,0265	0,0004	<0,0001
Gompertz	α	1.522.285	11.167,7323	<0,0001
	β	10,2711	0,2677	<0,0001
	γ	0,0145	0,0002	<0,0001

Fonte: Elaborado pela autora (2022).

Na Figura 5 são apresentados os ajustes de cada modelo analisado em relação aos dados dos casos acumulados de Covid-19 para o estado de São Paulo, nota-se visualmente um melhor ajuste do modelo Gompertz, porém para os últimos dias do gráfico a curva não se adequa muito bem aos dados. Um modo de uma possível solução para resolver o ajuste não tão bom para os dados dos últimos dias foi a aplicação do teste de Pettitt para rupturas, explicado na seção 2.13, resumidamente será utilizado para verificar o dia onde houve mudança no comportamento dos dados e assim dividi-los e para cada parte será ajustado novamente os modelos não lineares de crescimento, a fim de obter melhores resultados.

Figura 5 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da Covid-19 em São Paulo.



Fonte: Elaborado pela autora (2022).

A partir das medidas de AIC, BIC, R^2 e o MSE expostos na Tabela 5 pode-se notar que, esses valores indicam que o modelo Gompertz é o mais adequado, pois o mesmo apresenta menores AIC, BIC e MSE, bem como um maior valor do coeficiente de determinação, em comparação com os modelos Exponencial e Logístico.

Tabela 5 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.

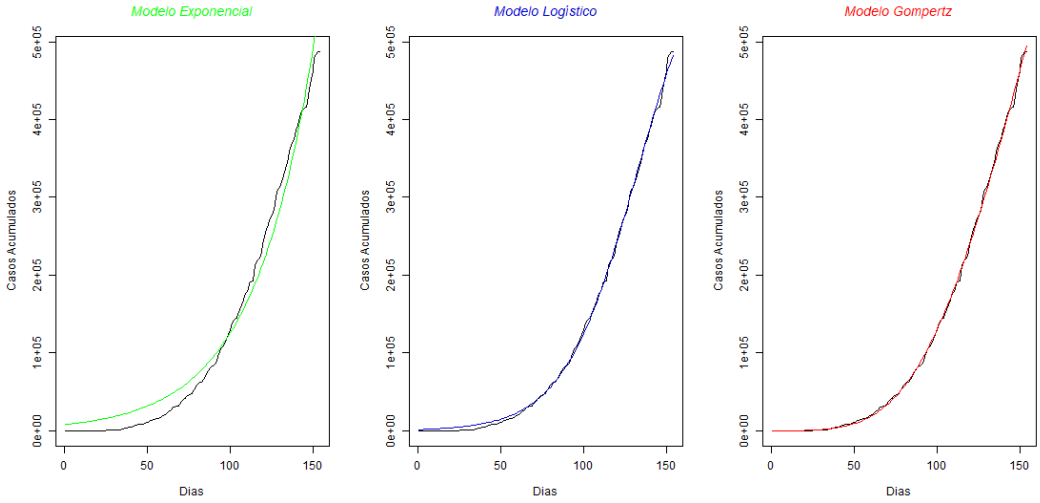
Modelo	AIC	BIC	R^2	MSE
Exponencial	8.278,588	8.289,797	0,916907	22.751.327.519
Logístico	7.502,012	7.516,958	0,9931319	1.846.123.912
Gompertz	7.206,207	7.221,153	0,9970977	710.973.755

Fonte: Elaborado pela autora (2022).

3.1.1.3 Ajuste dos modelos de crescimento por partes

O teste de Pettitt foi aplicado aos dados de casos acumulados de Covid-19 para o estado de São Paulo, resultando em uma mudança de tendência no dia 155, que corresponde ao dia 29/07/2020, logo, aplicou-se os modelos de crescimento para os dados do dia 26/02/2020 a 28/07/2020. Observa-se pela Figura 6 que o melhor ajuste foi utilizando o modelo de Gompertz.

Figura 6 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da Covid-19 em São Paulo do dia 26/02/2020 a 28/07/2020.

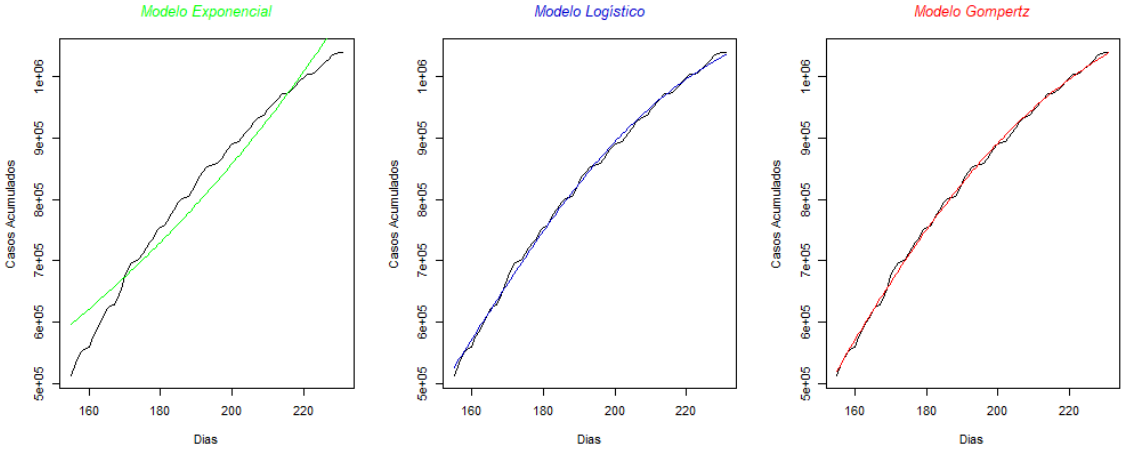


Fonte: Elaborado pela autora (2022).

Após a aplicação do teste para todos os dados, utilizou-se novamente, porém com os dados do dia 29/07/2020 a 31/12/2020, resultando em uma mudança de tendência no dia 78 que corresponde a 14/10/2020, com isso, aplicou-se os modelos para a segunda parte dos dados, do dia 29/07/2020 a 13/10/2020. Nota-se pela Figura 7 que, novamente o modelo Gompertz foi o que apresentou o melhor ajuste.

Ao calcular a curvatura de efeito do parâmetro e intrínseca do modelo para os dados D1, obteve-se, respectivamente $\gamma^T = 1{,}3026$ e $\gamma^N = 0{,}0134$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}225242$ calculado com 3 e 151 graus de liberdade. Logo, para o efeito do parâmetro foi significativo, então seria mais adequado utilizar uma função do modelo com outra parametrização para assim alcançar um resultado não significativo, ou seja, quanto mais próximo de zero mais válidos serão os resultados inferenciais do modelo não linear utilizado. Já para os dados D2 obteve-se a mesma conclusão, onde $\gamma^T = 2{,}2369$ e $\gamma^N = 0{,}0067$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}210837$ calculado com 3 e 74 graus de liberdade.

Figura 7 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da Covid-19 em São Paulo do dia 29/07/2020 a 13/10/2020.



Fonte: Elaborado pela autora (2022).

A partir das medidas de AIC, BIC, R^2 e o MSE expostos na Tabela 6 pode-se notar que, esses valores indicam que o modelo Gompertz é o mais adequado, pois o mesmo apresenta menores AIC, BIC e MSE, bem como um maior valor do coeficiente de determinação, em comparação com os modelos Exponencial e Logístico para ambas as partes D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) dos dados de casos acumulados de Covid-19 para o estado de São Paulo.

Portanto, o modelo ajustado é dado pelo seguinte modelo por partes:

$$\hat{Y}_t = \begin{cases} 1.526.134e^{(-10,5688e^{(-0,0145t)})}, para\ 0 \leq t < 155 \\ 1.202.097e^{(-28,9566e^{(-0,0229t)})}, para\ 155 \leq t < 232 \end{cases}$$

Tabela 6 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados aos dados de 26/02/2020 a 28/07/2020 (D1) e para 29/07/2020 a 13/10/2020 (D2).

Modelo	Dados D1		Dados D2	
	Métrica	Valor	Métrica	Valor
Exponencial	AIC	3.478,482	AIC	1.826,035
	BIC	3.487,593	BIC	1.833,066
	R^2	0,9866535	R^2	0,9540287
	MSE	363.289.199	MSE	1.078.608.711
Logístico	AIC	3.027,941	AIC	1.572,784
	BIC	3.040,089	BIC	1.582,16
	R^2	0,9992411	R^2	0,9983203
	MSE	19.232.797	MSE	39.192.612
Gompertz	AIC	2.939,401	AIC	1.540,73
	BIC	2.951,549	BIC	1.550,105
	R^2	0,9995007	R^2	0,9988916
	MSE	10.823.141	MSE	25.846.907

Fonte: Elaborado pela autora (2022).

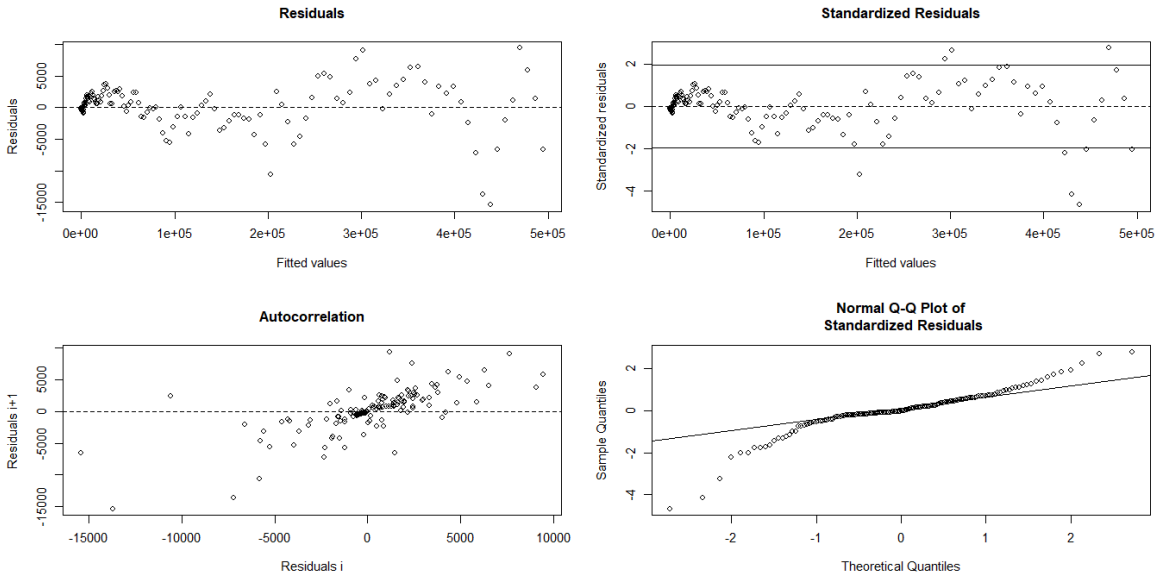
3.1.1.4 Análise dos resíduos e qualidade do ajuste

Após a etapa de ajuste dos modelos não lineares de crescimento aos dados de casos acumulados da doença em São Paulo foi feita a análise dos resíduos, porém, para o modelo Gompertz, pois este foi o melhor considerando os critérios de seleção.

Considerando-se as pressuposições necessárias de erros independentes e variância constante, pode-se observar pela Figura 8 que, para os dados de 26/02/2020 a 28/07/2020 no gráfico dos resíduos ou resíduos studentizados versus valores ajustados os pontos não estão distribuídos aleatoriamente em torno do zero, ou seja, a variância dos resíduos não é homogênea, o que pode ser confirmado pelo teste de Breusch-Pagan da Tabela 7, cujo p-valor é menor do que 0,05 de significância estatística adotado, logo, rejeita-se a hipótese nula de homogeneidade. Para a verificação de normalidade tem-se o gráfico QQ Plot da Figura 8, observa-se os pontos das extremidades distantes da reta, indicando a falta de normalidade dos resíduos, o

que se confirma pelo teste de Shapiro-Wilk da Tabela 7, um p-valor menor do que 0,05, rejeitando a hipótese nula de normalidade. Já para a pressuposição de independência dos erros, verifica-se por meio do gráfico de autocorrelação da Figura 8 que, os pontos não estão distribuídos aleatoriamente e sim agrupados, indicando que os resíduos não são independentes; pelo teste de Durbin-Watson pode-se notar a confirmação dessa afirmação, pois tem-se um p-valor menor do que 0,05, rejeitando assim a hipótese nula, de que não há autocorrelação entre os resíduos.

Figura 8 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (26/02/2020 a 28/07/2020) de casos acumulados em São Paulo.

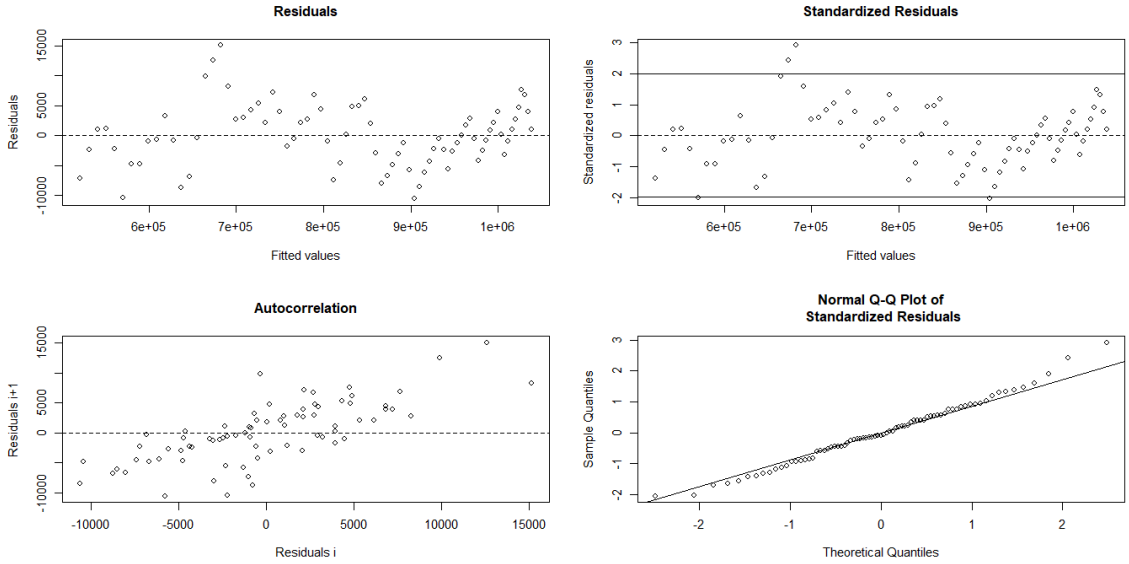


Fonte: Elaborado pela autora (2022).

A partir da Figura 9 que consta a análise gráfica dos resíduos para o período D2 (29/07/2020 a 13/10/2020) pode-se observar no gráfico dos resíduos versus valor ajustados que os pontos estão um pouco melhor distribuídos em torno do zero, logo, visualmente pode-se notar uma variância dos resíduos homogênea, o que pode ser confirmado pelo teste de Breusch-Pagan da Tabela 7, cujo p-valor é 0,1022, portanto como esse valor é maior do que o nível de significância adotado de 0,05 não rejeita-se a hipótese nula de homogeneidade, ou seja, não temos evidências suficientes para dizer que a heterogeneidade está presente nos resíduos do modelo não linear adotado. Observa-se também pela Figura 9 e pelos resultados expostos na Tabela 7

que, a suposição de normalidade foi satisfeita, porém a de erros independentes não foi confirmada.

Figura 9 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (29/07/2020 a 13/10/2020) de casos acumulados em São Paulo.



Fonte: Elaborado pela autora (2022).

Tabela 7 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) de casos acumulados em São Paulo.

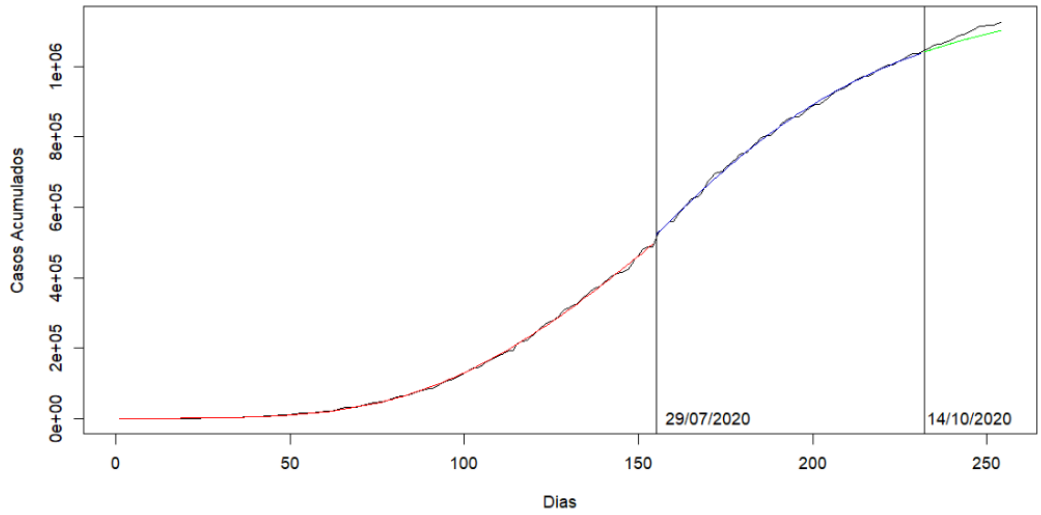
Teste	D1	D2
Shapiro-Wilk	0,88974	0,98945
p-valor	<0,0001	0,7808
Breusch-Pagan	31,682	4,5608
p-valor	<0,0001	0,1022
Durbin-Watson	0,536199	0,51001
p-valor	<0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Pela Figura 10 observa-se o modelo Gompertz ajustado aos dados de 26/02/2020 a 28/07/2020 sendo representado pela linha vermelha e também esse

modelo para 29/07/2020 a 13/10/2020, linha azul, bem como a data 29/07/2020 e 14/10/2020 de mudança de tendência, resultado do teste de Pettitt aplicado. A linha verde são as previsões para 14/10/2020 a 05/11/2020, nota-se que o modelo Gompertz não se ajustou tão bem como para os períodos anteriores, pois as previsões deram valores abaixo dos verdadeiros valores de casos acumulados da doença no estado de São Paulo, logo, é necessário utilizar outro modelo para esse período.

Figura 10 – Ajuste do modelo Gompertz por partes aos dados de casos acumulados da COVID-19 para São Paulo e previsões.



Fonte: Elaborado pela autora (2022).

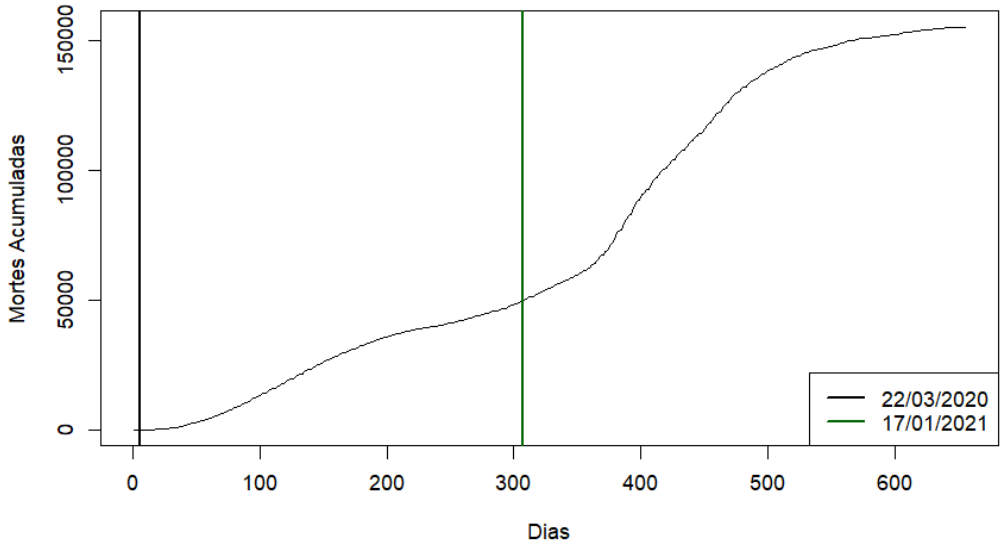
3.1.2 Mortes acumuladas pela doença COVID-19

3.1.2.1 Análise Exploratória para os dados de São Paulo

A seguir será apresentado o gráfico de mortes acumuladas no Estado de São Paulo referente aos anos de 2020 e 2021, cujo período vai de 17/03/2020, dia da primeira morte no estado de São Paulo, a 31/12/2021, a linha preta em destaque se refere ao começo da quarentena no estado de São Paulo e a verde é a data da primeira pessoa a ser vacinada no estado.

Na Figura 11 pode-se notar um aumento considerável nas mortes acumuladas do dia 370 ao dia 489 que corresponde respectivamente a 21/03/2020 a 18/07/2021, aproximadamente a partir do dia 19/07/2021, data que marca metade da população de São Paulo vacinada com a primeira dose, a quantidade de mortes acumuladas começa se estabilizar.

Figura 11 – Gráfico das mortes acumuladas da COVID-19 do ano de 2020 e 2021 para São Paulo.



Fonte: Elaborado pela autora (2022).

A partir da Tabela 8 nota-se que, em 2020 a média de mortes acumuladas foi de 23.070, já em 2021 foi de 114.100 para o estado de São Paulo. No ano de 2020 observa-se uma alta dispersão nos dados de 68,40%, em 2021 esse valor foi mais baixo de 33,64%.

Tabela 8 – Medidas estatísticas referentes aos anos de 2020 e 2021 para mortes acumuladas da COVID-19 em São Paulo.

	2020	2021
Quantidade de dias	290	365
Média	23.070,32	114.100,87
Desvio Padrão	15.780,96	38.385,33
Mínimo	1	46.775
Q1	7.053,75	76.552
Mediana	25.065	128.921
Q3	38.193,25	149.953
Máximo	46.717	155.205
Coeficiente de Variação	68,40%	33,64%

Fonte: Elaborado pela autora (2022).

3.1.2.2 Ajuste de modelos para os dados do estado de São Paulo

Ajustando os modelos não lineares de crescimento Exponencial, Logístico e Gompertz aos dados de mortes acumuladas em 2020 para o estado de São Paulo, tem-se as estimativas dos parâmetros expostas na Tabela 9, nota-se através do p-valor menor do que o nível de significância adotado de 0,05 que as estimativas foram significativas.

Tabela 9 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para as mortes acumuladas de COVID-19 em São Paulo.

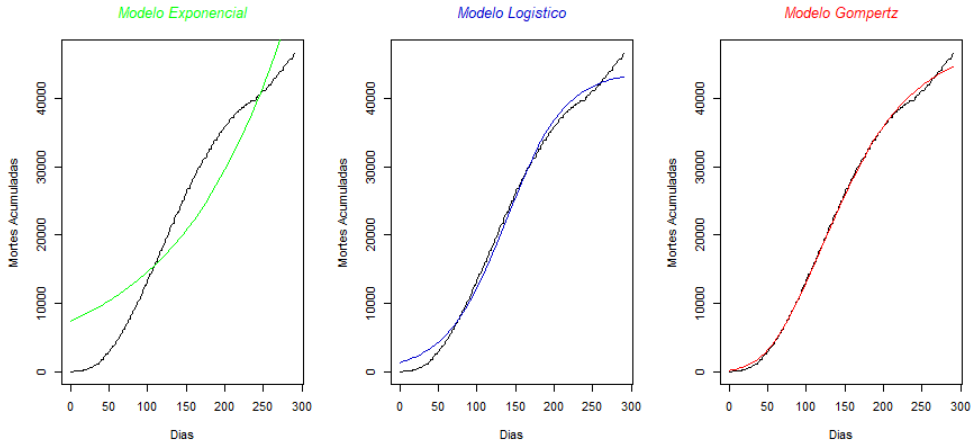
Modelo	Parâmetros	Estimativas	E.P.	p-valor
Exponencial	α	7.273,96	334,3417	<0,0001
	β	0,0070	0,0002	<0,0001
Logístico	α	44.107,84	212,4052	<0,0001
	β	33,5980	1,3437	<0,0001
	γ	0,0257	0,0003	<0,0001
Gompertz	α	48.302,64	145,6734	<0,0001
	β	5,7209	0,0538	<0,0001
	γ	0,0148	0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Na Figura 12 tem-se a curva dos dados de óbitos acumulados observados e os modelos Exponencial, Logístico e Gompertz ajustado a esses dados, nota-se um melhor ajuste do terceiro modelo, porém no final do período analisado observa-se que o mesmo não se adequada tão bem aos dados.

Pela Tabela 10 as considerações levantadas a partir da Figura 12 são confirmadas, pois o modelo Gompertz possui menores valores de AIC e BIC e um erro quadrático médio bem menor do que para o Exponencial e o Logístico.

Figura 12 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo.



Fonte: Elaborado pela autora (2022).

Tabela 10 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.

Modelo	AIC	BIC	R^2	MSE
Exponencial	5.826,29	5.837,3	0,8863691	30.465.028
Logístico	4.927,604	4.942,283	0,9952085	1.364.467
Gompertz	4.394,038	4.408,717	0,9991593	216.726,3

Fonte: Elaborado pela autora (2022).

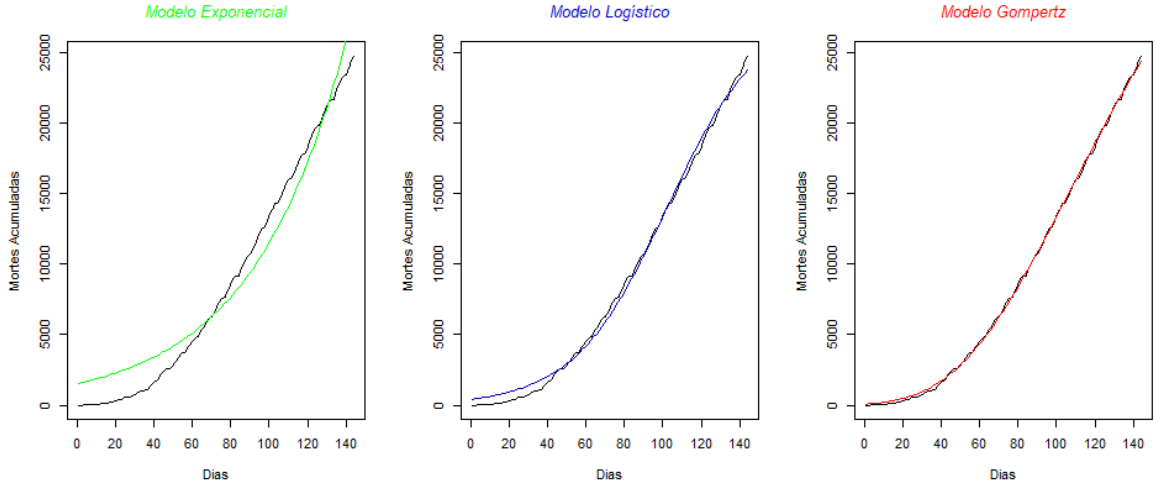
3.1.2.3 Ajuste dos modelos de crescimento por partes

Para tentar obter um melhor resultado dos modelos foi aplicado o teste de Pettitt aos dados de óbitos para 17/03/2020 a 31/12/2020, a fim de encontrar o ponto de mudança de tendência e assim ajustá-los a cada parte dos dados. Através do teste encontrou-se o dia 145, correspondente a 08/08/2020, logo a primeira parte dos dados (D1) utilizada para a modelagem foi de 17/03/2020 a 07/08/2020. Já para encontrar a segunda parte dos dados (D2) o processo foi a aplicação do teste novamente, porém para o período de 08/08/2020 a 31/12/2020, como resultado o dia da mudança de tendência na curva foi 19/10/2020, então utilizou-se de 08/08/2020 a 18/10/2020.

Observa-se tanto na Figura 13 como na 14 que o modelo de Gompertz melhor se ajustou aos dados de mortes acumuladas para o estado de São Paulo.

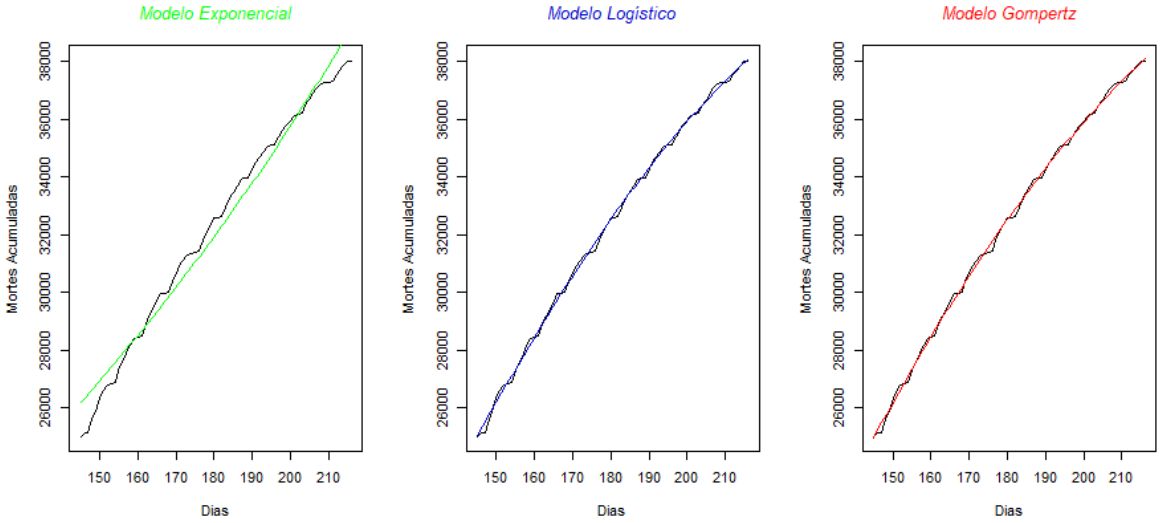
Ao calcular a curvatura de efeito do parâmetro e intrínseca do modelo para os dados D1, obteve-se, respectivamente $\gamma^T = 0{,}3309$ e $\gamma^N = 0{,}0097$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}224257$ calculado com 3 e 141 graus de liberdade. Logo, como ambos os valores foram não significativos os resultados inferenciais do modelo não linear utilizados são válidos. Já para os dados D2 obteve-se que $\gamma^T = 1{,}4555$ e $\gamma^N = 0{,}0062$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}210837$ calculado com 3 e 69 graus de liberdade, portanto para o efeito dos parâmetros foi significativo então o melhor seria utilizar um modelo com outra parametrização.

Figura 13 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo do dia 17/03/2020 a 07/08/2020 (D1).



Fonte: Elaborado pela autora (2022).

Figura 14 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 em São Paulo do dia 08/08/2020 a 18/10/2020 (D2).



Fonte: Elaborado pela autora (2022).

A partir das medidas de AIC, BIC, R^2 e o MSE expostos na Tabela 11 pode-se notar que, esses valores indicam que o modelo Gompertz é o mais adequado, pois o mesmo apresenta menores AIC, BIC e MSE, em comparação com os modelos Exponencial e Logístico para ambas as partes D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) dos dados de mortes acumuladas de Covid-19 para o estado de São Paulo.

Portanto, o modelo ajustado é dado pelo seguinte modelo por partes:

$$\hat{Y}_t = \begin{cases} 41.641{,}7e^{(-6{,}3051e^{(-0{,}0171t)})}, para\ 0 \leq t < 145 \\ 47.285{,}14e^{(-5{,}8527e^{(-0{,}0153t)})}, para\ 145 \leq t < 217 \end{cases}$$

Tabela 11 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020).

Modelo	Dados D1		Dados D2	
	Métrica	Valor	Métrica	Valor
Exponencial	AIC	2.525,818	AIC	1.118,436
	BIC	2.534,727	BIC	1.125,266
	R^2	0,9679007	R^2	0,9799225
	MSE	2.328.783	MSE	300.323,9
Logístico	AIC	2.167,61	AIC	903,1706
	BIC	2.179,49	BIC	912,2772
	R^2	0,9973713	R^2	0,9990152
	MSE	190.883,3	MSE	14.691,84
Gompertz	AIC	1.881,418	AIC	898,6213
	BIC	1.893,298	BIC	907,7279
	R^2	0,9996078	R^2	0,9990755
	MSE	26.159,62	MSE	13.792,26

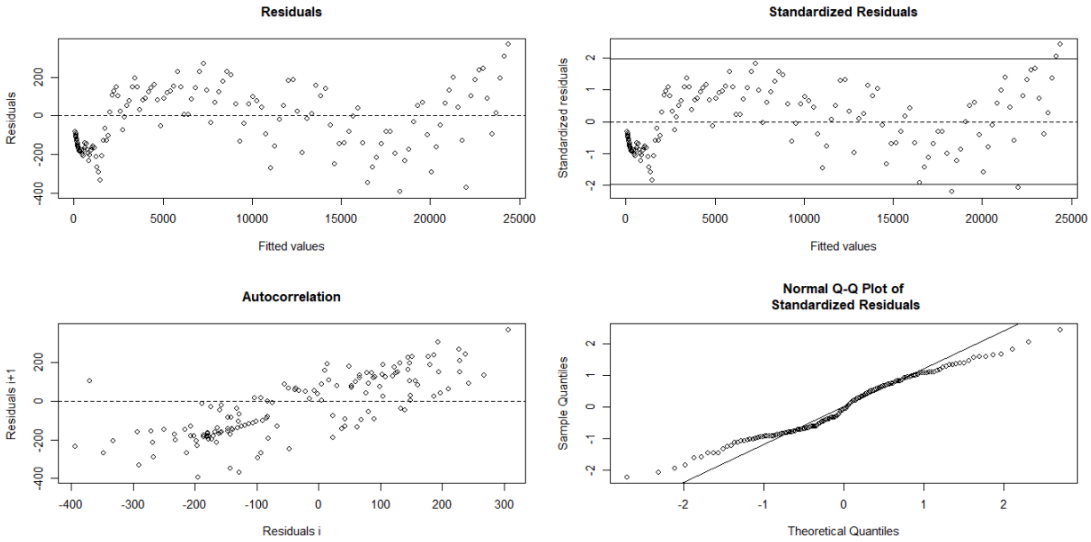
Fonte: Elaborado pela autora (2022).

3.1.2.4 Análise dos resíduos e qualidade do ajuste

A partir da Figura 15, que contém a análise gráfica dos resíduos da parte D1 dos dados, pode-se notar que os erros não estão distribuídos de forma aleatória nos gráficos dos resíduos versus valores ajustados, logo existe uma heterogeneidade da variância. Pelo gráfico de autocorrelação nota-se que os erros não são independentes e pelo QQ plot observa-se que eles não possuem distribuição normal. A partir dos resultados dos testes apresentados na Tabela 15 confirma-se as observações ditas.

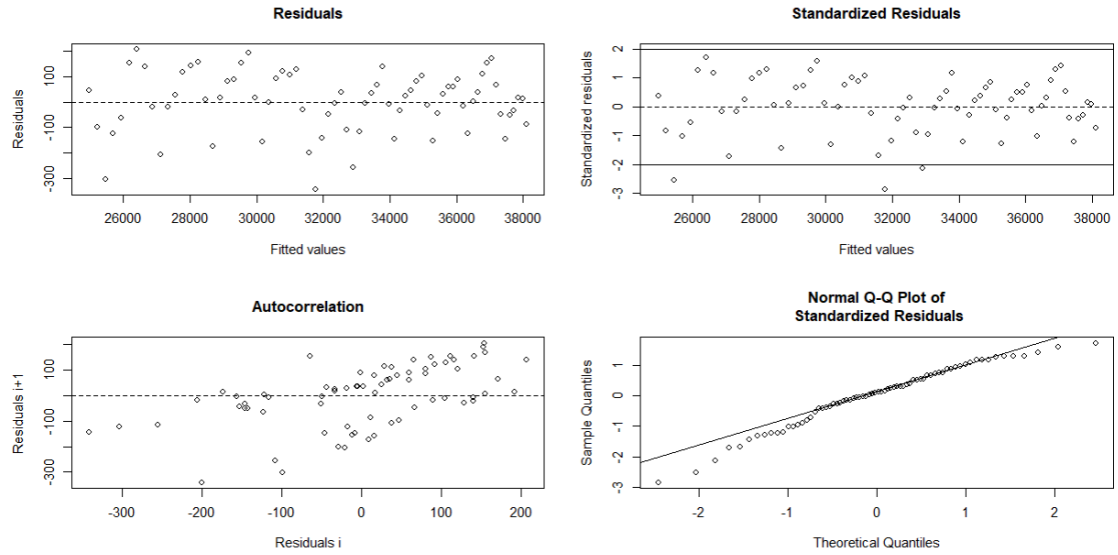
Para a análise dos resíduos da parte D2 dos dados, observa-se pela Figura 16 e pela Tabela 12 que, somente a hipótese de homogeneidade da variância é atendida.

Figura 15 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (17/03/2020 a 07/08/2020) de mortes acumuladas em São Paulo.



Fonte: Elaborado pela autora (2022).

Figura 16 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (08/08/2020 a 18/10/2020) de mortes acumuladas em São Paulo.



Fonte: Elaborado pela autora (2022).

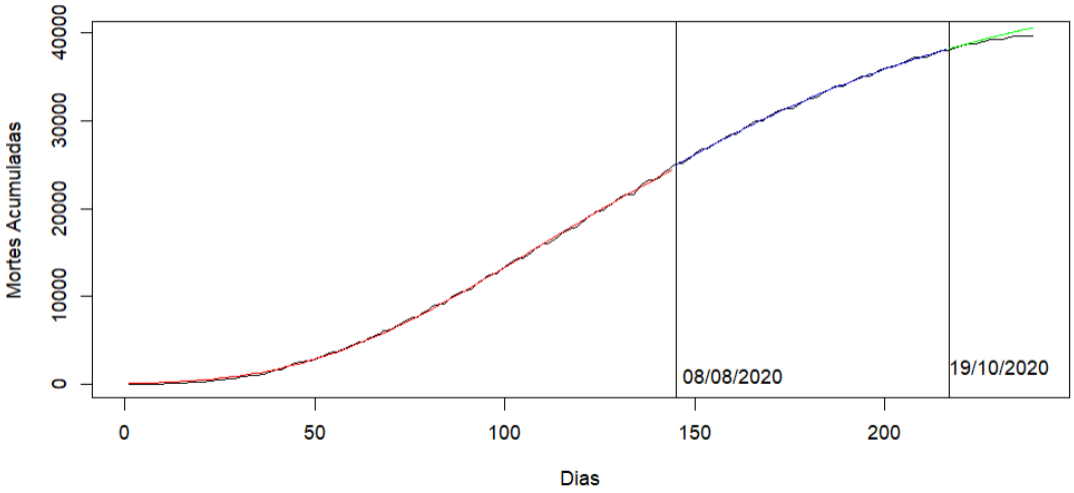
Tabela 12 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) de mortes acumuladas em São Paulo.

Teste	D1	D2
Shapiro-Wilk	0,97464	0,96621
p-valor	0,008925	0,04997
Breusch-Pagan	11,363	4,5613
p-valor	0,003408	0,1022
Durbin-Watson	0,39351	0,88264
p-valor	<0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Pela Figura 17 observa-se o modelo Gompertz ajustado aos dados de 17/03/2020 a 07/08/2020 sendo representado pela linha vermelha e também esse modelo para 08/08/2020 a 18/10/2020, linha azul, bem como a data 08/08/2020 e 19/10/2020 de mudança de tendência, resultado do teste de Pettitt aplicado. A linha verde são as previsões para 19/10/2020 a 10/11/2020, nota-se que o modelo Gompertz não se ajustou tão bem como para os períodos anteriores, pois as previsões deram valores acima dos verdadeiros valores de mortes acumuladas pela doença no estado de São Paulo, logo, é necessário utilizar outro modelo para esse período.

Figura 17 – Ajuste do modelo Gompertz por partes aos dados de mortes acumuladas de COVID-19 para São Paulo e previsões.



Fonte: Elaborado pela autora (2022).

3.2 Análise para o Brasil

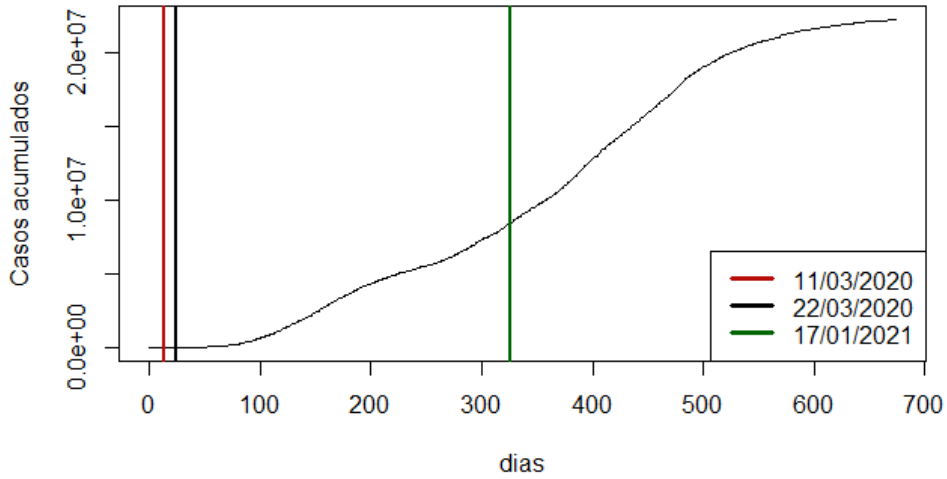
3.2.1 Casos confirmados acumulados pela doença COVID-19

3.2.1.1 Análise Exploratória para os dados do Brasil

A seguir será apresentado o gráfico de casos acumulados do Brasil referente aos anos de 2020 e 2021, cujo período vai de 26/02/2020 a 31/12/2021, a linha vermelha em destaque se refere a data que foi declarada pandemia em relação ao novo coronavírus pela Organização Mundial da Saúde, a linha preta marca o começo da quarentena no Brasil e a linha verde o início da vacinação contra a doença após a aprovação do uso emergencial da vacina do Instituto Butantan pela Anvisa (Agência Nacional de Vigilância Sanitária), a primeira contemplada foi uma enfermeira de 54 anos da UTI do Instituto de Infectologia Emílio Ribas.

Na Figura 18 pode-se notar um aumento considerável nos casos acumulados do dia 306 ao dia 500 que corresponde respectivamente a 28/12/2020 a 10/07/2021, aproximadamente a partir do dia 29/08/2021 os casos começam a se estabilizar, porém de maneira mais lenta.

Figura 18 – Gráfico dos casos acumulados da COVID-19 do ano de 2020 e 2021 para o Brasil.



Fonte: Elaborado pela autora (2022).

A Tabela 13 contém algumas informações estatísticas referentes aos casos acumulados da doença analisada, no Brasil. Nota-se através das médias que houve um aumento considerável dos casos acumulados passando de mais de 2 milhões em 2020 para mais de 17 milhões em 2021. Pelo coeficiente de variação apresentado, observa-se uma alta dispersão nos dados de 2020, já para 2021 tem-se um valor mais baixo de 28,10%.

Tabela 13 – Medidas estatísticas referentes aos anos de 2020 e 2021 para casos acumulados da COVID-19 no Brasil.

	2020	2021
Quantidade de dias	310	365
Média	2.874.056	17.000.495
Desvio Padrão	2.514.300	4.776.357
Mínimo	1	7.700.578
Q1	192.460	12.910.082
Mediana	2.581.184	18.687.469
Q3	5.162.255	21.445.651
Máximo	5.162.255	22.287.521
Coeficiente de Variação	87,48%	28,10%

Fonte: Elaborado pela autora (2022).

3.2.1.2 Ajuste de modelos para os dados do Brasil

Ajustando os modelos não lineares de crescimento Exponencial, Logístico e Gompertz aos dados de casos acumulados em 2020 para o Brasil, tem-se as estimativas dos parâmetros expostas na Tabela 14, nota-se através do p-valor menor do que o nível de significância adotado de 0,05 que as estimativas foram significativas.

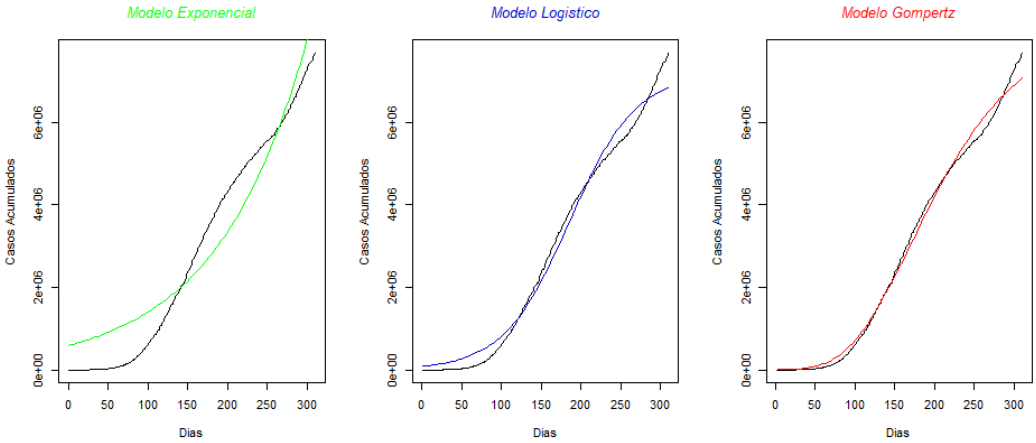
Tabela 14 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para os casos de COVID-19 no Brasil.

Modelo	Parâmetros	Estimativas	E.P.	p-valor
Exponencial	α	581.719,5	30.120,7944	<0,0001
	β	0,0087	0,0002	<0,0001
Logístico	α	7.175.396	70.819,1906	<0,0001
	β	86,034	6,2459	<0,0001
	γ	0,0240	0,0005	<0,0001
Gompertz	α	8.405.220	95.252,7978	<0,0001
	β	8,7391	0,2464	<0,0001
	γ	0,0127	0,0002	<0,0001

Fonte: Elaborado pela autora (2022).

Na Figura 19 tem-se a curva dos dados de casos acumulados observados e os modelos Exponencial, Logístico e Gompertz ajustado a esses dados, nota-se um melhor ajuste do terceiro modelo, porém no final do período analisado observa-se que o mesmo não se adequada tão bem aos dados.

Figura 19 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados de COVID-19 no Brasil.



Fonte: Elaborado pela autora (2022).

A partir das medidas de AIC, BIC, R^2 e o MSE expostos na Tabela 15 pode-se notar que, esses valores indicam que o modelo Gompertz é o mais adequado, pois o mesmo apresenta menores AIC, BIC e MSE, bem como um maior valor do coeficiente de determinação, em comparação com os modelos Exponencial e Logístico.

Tabela 15 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.

Modelo	AIC	BIC	R^2	MSE
Exponencial	9.229,621	9.240,831	0,7926119	489.054.408.293
Logístico	8.620,312	8.635,258	0,9407827	68.068.039.111
Gompertz	8.334,236	8.349,183	0,9814302	27.049.929.722

Fonte: Elaborado pela autora (2022).

3.2.1.3 Ajuste dos modelos de crescimento por partes

O teste de Pettitt aplicado aos dados de casos acumulados de Covid-19 para o Brasil, resultou em uma mudança de tendência no dia 155, que corresponde ao dia 29/07/2020, logo, aplicou-se os modelos de crescimento para os dados do dia 26/02/2020 a 28/07/2020. Observa-se pela Figura 20 que o melhor ajuste foi utilizando o modelo Gompertz. Já para o período de 29/07/2020 a 31/12/2020 a mudança de

tendência foi no dia 14/10/2020, logo ajustou-se os modelos para o dia 29/07/2020 a 13/10/2020. Nota-se pela Figura 21 que novamente o melhor modelo foi o Gompertz.

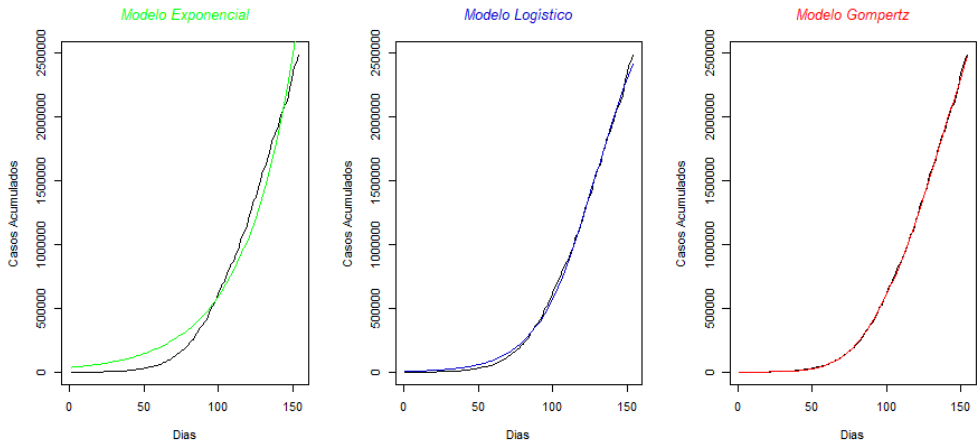
A partir dos resultados da Tabela 16 confirma-se as afirmações feitas acima, pois para o modelo de Gompertz os valores de AIC, BIC e MSE são menores.

Portanto, o modelo ajustado é dado pelo seguinte modelo por partes:

$$\hat{Y}_t = \begin{cases} 5.831.143e^{(-13,8062e^{(-0,0180t)})}, para\ 0 \leq t < 155 \\ 6.350.928e^{(-17,0221e^{(-0,0189t)})}, para\ 155 \leq t < 232 \end{cases}$$

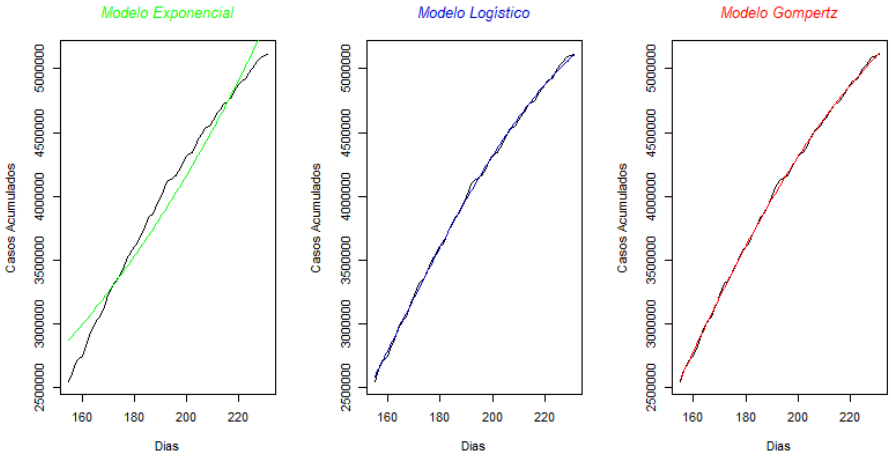
Ao calcular a curvatura de efeito do parâmetro e intrínseca do modelo para os dados D1, obteve-se, respectivamente $\gamma^T = 0{,}399$ e $\gamma^N = 0{,}0085$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}225242$ calculado com 3 e 151 graus de liberdade. Logo, como ambos os valores foram não significativos os resultados inferenciais do modelo não linear utilizados são válidos. Já para os dados D2 obteve-se que $\gamma^T = 1{,}318$ e $\gamma^N = 0{,}0042$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}210837$ calculado com 3 e 74 graus de liberdade, portanto para o efeito dos parâmetros foi significativo então o melhor seria utilizar um modelo com outra parametrização.

Figura 20 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 no Brasil de 26/02/2020 a 28/07/2020.



Fonte: Elaborado pela autora (2022).

Figura 21 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de casos acumulados da COVID-19 no Brasil de 29/07/2020 a 13/10/2020.



Fonte: Elaborado pela autora (2022).

Tabela 16 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020).

Modelo	Dados D1		Dados D2	
	Métrica	Valor	Métrica	Valor
Exponencial	AIC	4.011,067	AIC	2.045,101
	BIC	4.020,178	BIC	2.052,132
	R^2	0,9834	R^2	0,9678
	MSE	11.539.600.286	MSE	18.554.010.964
Logístico	AIC	3.569,987	AIC	1.749,596
	BIC	3.582,135	BIC	1.758,971
	R^2	0,9988	R^2	0,9993
	MSE	649.627.953	MSE	389.451.714
Gompertz	AIC	3.298,33	AIC	1.717,461
	BIC	3.310,478	BIC	1.726,836
	R^2	0,9998	R^2	0,9996
	MSE	111.317.943	MSE	256.570.213

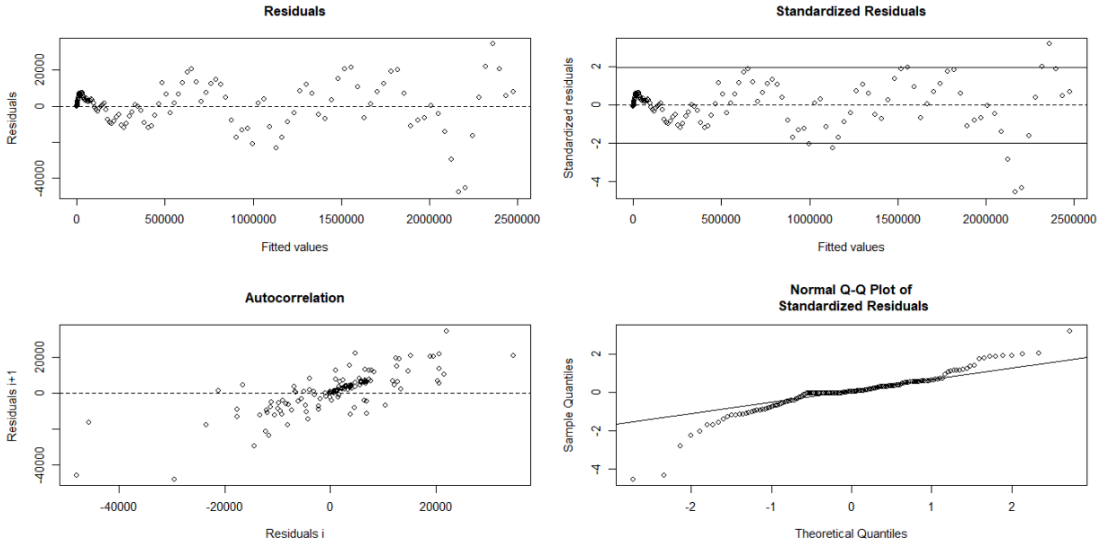
Fonte: Elaborado pela autora (2022).

3.2.1.4 Análise dos resíduos e qualidade do ajuste

Feito o ajuste dos modelos não lineares de crescimento Exponencial, Logístico e Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) constatou-se pelos critérios de seleção dos modelos expostos na Tabela 16 e pela observação gráfica dos dados que, o melhor modelo foi o Gompertz. Após essa constatação prévia, foi feita a análise dos resíduos considerando esse modelo.

Como pressupostos dos modelos tem-se erros independentes e variância constante, normalidade dos erros; pela Figura 22 observa-se no gráfico dos resíduos ou resíduos studentizados versus os valores ajustados que os pontos não estão distribuídos aleatoriamente, logo a variância não é homogênea; já no gráfico de autocorrelação observa-se que os pontos estão juntos na região central do gráfico, não possui uma aleatoriedade, portanto os erros não são independentes; por fim pelo QQ plot nota-se o distanciamento dos pontos em relação a reta, logo os resíduos não possuem distribuição normal. Todas as afirmações acima são confirmadas pelos resultados dos respectivos testes, apresentados na Tabela 17.

Figura 22 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (26/02/2020 a 28/07/2020) de casos acumulados no Brasil.

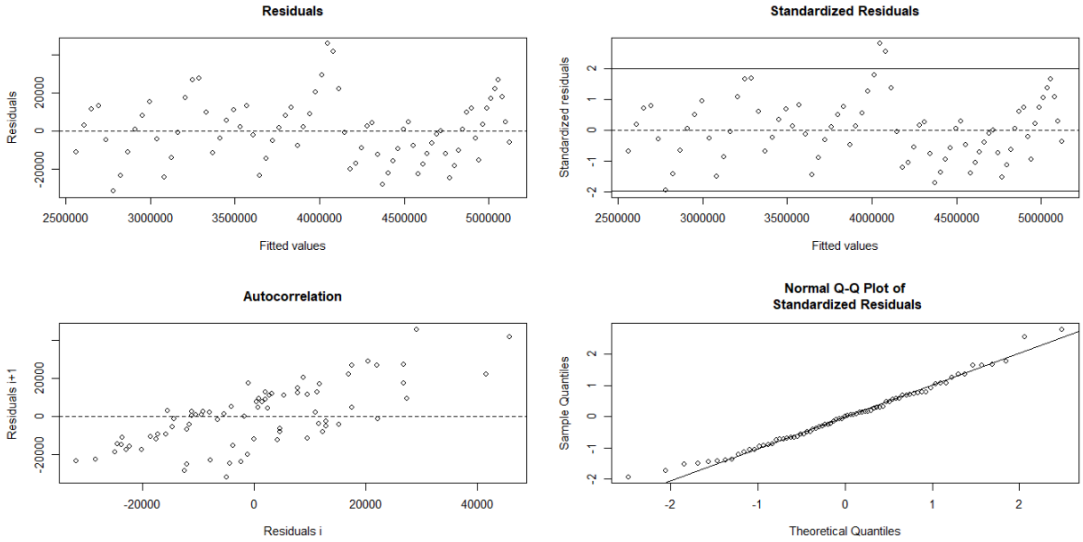


Fonte: Elaborado pela autora (2022).

Na Figura 23 tem-se os gráficos para a análise dos resíduos do modelo Gompertz para os dados D2 (29/07/2020 a 13/10/2020), nota-se pelos gráficos dos resíduos versus valores ajustados que os pontos estão melhores distribuídos, mais aleatórios,

logo contata-se a homogeneidade da variância, o que é confirmada pelo resultado do teste de Breusch-Pagan apresentado na Tabela 17, nota-se também a distribuição normal dos resíduos através do gráfico QQ plot da Figura 23 e pelo teste de Shapiro-Wilk da Tabela 17 e por fim a suposição de independência dos erros não é satisfeita, como pode-se observar pelo gráfico de autocorrelação e pelo teste de Durbin-Watson, apresentados na Figura 23 e Tabela 17 respectivamente.

Figura 23 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (29/07/2020 a 13/10/2020) de casos acumulados no Brasil.



Fonte: Elaborado pela autora (2022).

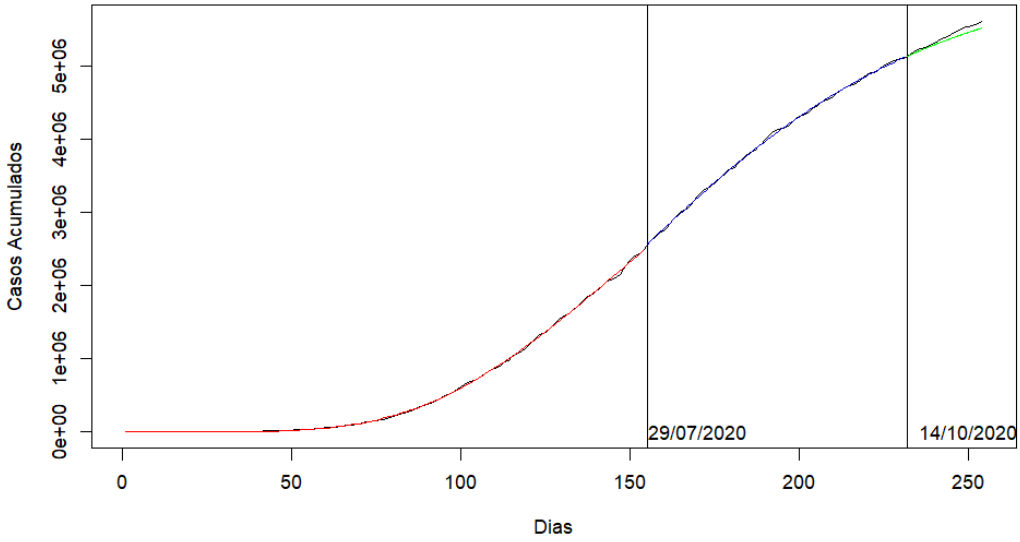
Tabela 17 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (26/02/2020 a 28/07/2020) e D2 (29/07/2020 a 13/10/2020) de casos acumulados no Brasil.

Teste	D1	D2
Shapiro-Wilk	0,89925	0,98421
p-valor	<0,0001	0,4571
Breusch-Pagan	32,897	1,3104
p-valor	<0,0001	0,5193
Durbin-Watson	0,41169	0,55956
p-valor	<0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Pela Figura 24 observa-se o modelo Gompertz ajustado aos dados de 26/02/2020 a 28/07/2020 sendo representado pela linha vermelha e também esse modelo para 29/07/2020 a 13/10/2020, linha azul, bem como a data 29/07/2020 e 14/10/2020 de mudança de tendência, resultado do teste de Pettitt aplicado. A linha verde são as previsões para 14/10/2020 a 05/11/2020, nota-se que o modelo Gompertz não se ajustou tão bem como para os períodos anteriores, pois as previsões deram valores abaixo dos verdadeiros valores de casos acumulados pela doença no Brasil, logo, é necessário utilizar outro modelo para esse período.

Figura 24 – Ajuste do modelo Gompertz por partes aos dados de casos acumulados de COVID-19 para o Brasil e previsões.



Fonte: Elaborado pela autora (2022).

3.2.2 Mortes acumuladas pela doença COVID-19

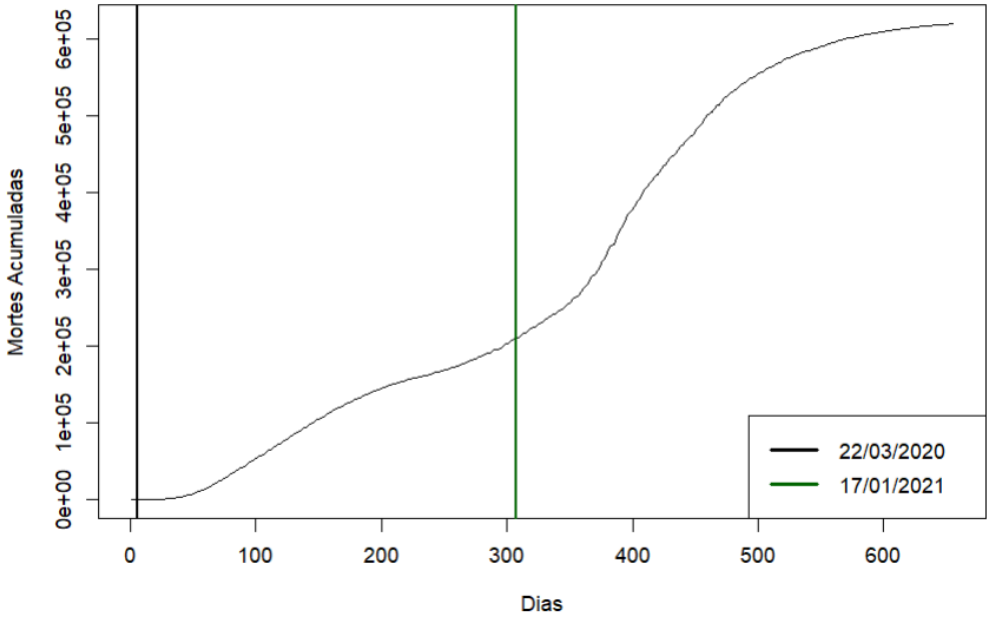
3.2.2.1 Análise Exploratória para os dados do Brasil

A seguir será apresentado o gráfico de mortes acumuladas no Brasil referente aos anos de 2020 e 2021, cujo período vai de 17/03/2020, dia da primeira morte no Brasil, a 31/12/2021, a linha preta em destaque se refere ao início da quarentena e a verde é a data da primeira pessoa a ser vacinada no país.

Na Figura 25 pode-se notar um aumento considerável nas mortes acumuladas do dia 370 ao dia 489 que corresponde respectivamente a 21/03/2020 a 18/07/2021, aproximadamente a partir do dia 508 (06/08/2021), data que marca metade da

população do Brasil vacinada com a primeira dose, a quantidade de mortes acumuladas começa se estabilizar.

Figura 25 – Gráfico das mortes acumuladas da COVID-19 do ano de 2020 e 2021 para o Brasil.



Fonte: Elaborado pela autora (2022).

A Tabela 18 contém algumas medidas estatísticas importantes, pode-se observar que a média de mortes por COVID-19 no Brasil em 2020 é de 93.355, já para 2021 esse valor foi de 464.251 óbitos. Em relação a dispersão dos dados há uma alta dispersão para o ano de 2020 de 69,95%, em 2021 esse valor foi mais baixo de 31,50%.

Na metade dos dias tem-se uma quantidade de mortes acumuladas acima de 100.763 e na outra metade dos dias estão abaixo desse valor.

Estão expostos na Tabela 18 também, os valores de desvio padrão, mínimo, máximo, primeiro e terceiro quartil.

Tabela 18 – Medidas estatísticas referentes aos anos de 2020 e 2021 para mortes acumuladas da COVID-19 no Brasil.

	2020	2021
Quantidade de dias	290	365
Média	93.355,34	464.251,6
Desvio Padrão	65.305,94	146.219,7
Mínimo	1	195.411
Q1	27.035	328.206
Mediana	100.763	521.952
Q3	154.671,75	597.255
Máximo	194.949	619.056
Coeficiente de Variação	69,95%	31,50%

Fonte: Elaborado pela autora (2022).

3.2.2.2 Ajuste de modelos para os dados do Brasil

A Tabela 19 contém as estimativas dos parâmetros para os modelos não lineares de crescimento analisado, nota-se pelos p-valores que eles foram significativos ao nível de 5% de significância estatística.

Tabela 19 – Estimativas dos parâmetros dos modelos Exponencial, Logístico e Gompertz, erro padrão (E.P.) e p-valores para as mortes de COVID-19 no Brasil.

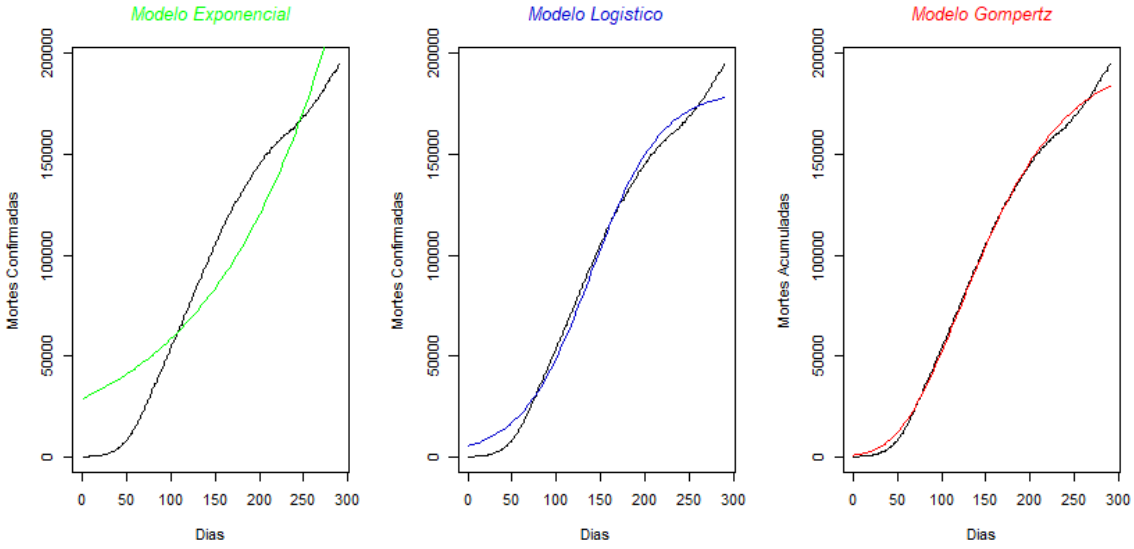
Modelo	Parâmetros	Estimativas	E.P.	p-valor
Exponencial	α	28.468,23	1.328,0899	<0,0001
	β	0,0072	0,0002	<0,0001
Logístico	α	181.958,2	1.096,0092	<0,0001
	β	34,9760	1,7088	<0,0001
	γ	0,0255	0,0004	<0,0001
Gompertz	α	199.905,8	894,6846	<0,0001
	β	5,8620	0,0801	<0,0001
	γ	0,0147	0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Na Figura 26 tem-se a curva dos dados de óbitos acumulados observados e os modelos Exponencial, Logístico e Gompertz ajustado a esses dados, nota-se um melhor ajuste do terceiro modelo, porém no final do período analisado observa-se que o mesmo não se adequada tão bem aos dados.

Pela Tabela 20 as considerações levantadas a partir da Figura 26 são confirmadas, pois o modelo Gompertz possui maiores valores de AIC e BIC e um erro quadrático médio bem menor do que para o Exponencial e o Logístico.

Figura 26 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas de COVID-19 no Brasil.



Fonte: Elaborado pela autora (2022).

Tabela 20 – Valores de AIC e BIC, o coeficiente de determinação (R^2) e o erro quadrático médio (MSE) para os três modelos ajustados.

Modelo	AIC	BIC	R^2	MSE
Exponencial	6.639,232	6.650,241	0,8909	502.615.393
Logístico	5.858,737	5.873,417	0,9930	33.837.478
Gompertz	5.420,74	5.435,42	0,9983	7.472.527

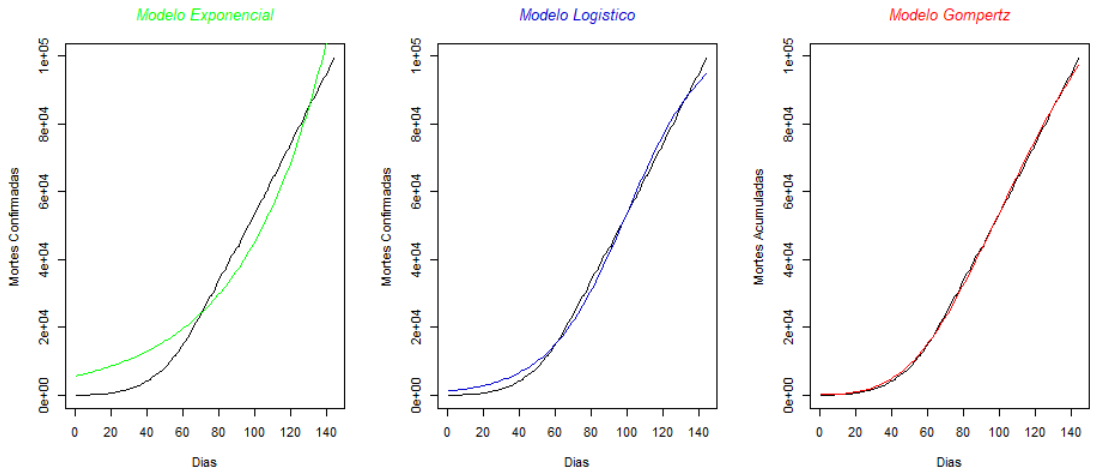
Fonte: Elaborado pela autora (2022).

3.2.2.3 Ajuste dos modelos de crescimento por partes

O teste de Pettitt foi aplicado aos dados de óbitos para 17/03/2020 a 31/12/2020, a fim de encontrar o ponto de mudança de tendência e assim ajustá-los a cada parte dos dados. Através do teste encontrou-se o dia 145, correspondente a 08/08/2020, logo a primeira parte dos dados (D1) utilizada para a modelagem foi de 17/03/2020 a 07/08/2020. Já para encontrar a segunda parte dos dados (D2) o processo foi a aplicação do teste novamente, porém para o período de 08/08/2020 a 31/12/2020, como resultado o dia da mudança de tendência na curva foi 19/10/2020, então utilizou-se de 08/08/2020 a 18/10/2020.

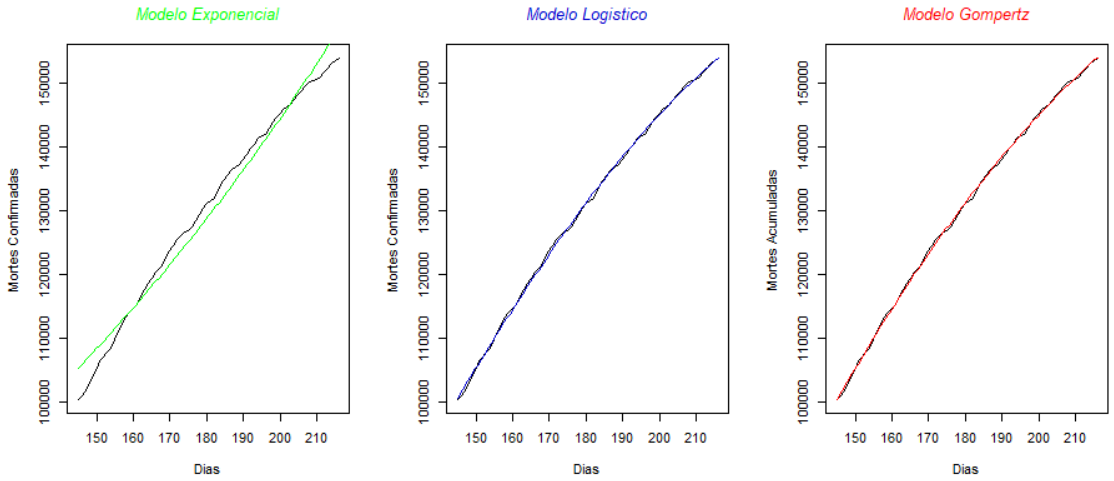
Observa-se tanto na Figura 27 como na 28 que o modelo de Gompertz melhor se ajustou aos dados de mortes acumuladas para o Brasil.

Figura 27 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 no Brasil de 17/03/2020 a 07/08/2020.



Fonte: Elaborado pela autora (2022).

Figura 28 – Curvas ajustadas dos modelos Exponencial, Logístico e Gompertz aos dados de mortes acumuladas da COVID-19 no Brasil de 08/08/2020 a 18/10/2020.



Fonte: Elaborado pela autora (2022).

A partir dos resultados da Tabela 21 confirma-se as afirmações feitas acima, pois para o modelo de Gompertz os valores de AIC, BIC e MSE são menores tanto para o período de 17/03/2020 a 07/08/2020 bem como para 08/08/2020 a 18/10/2020.

Portanto, o modelo ajustado é dado pelo seguinte modelo por partes:

$$\hat{Y}_t = \begin{cases} 145.410{,}1e^{(-7{,}8434e^{(-0{,}0207t)})}, para\ 0 \leq t < 145 \\ 191.551{,}8e^{(-5{,}9530e^{(-0{,}0153t)})}, para\ 145 \leq t < 217 \end{cases}$$

Ao calcular a curvatura de efeito do parâmetro e intrínseca do modelo para os dados D1, obteve-se, respectivamente $\gamma^T = 0{,}2334$ e $\gamma^N = 0{,}0123$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}224257$ calculado com 3 e 141 graus de liberdade. Logo, como ambos os valores foram não significativos os resultados inferenciais do modelo não linear utilizados são válidos. Já para os dados D2 obteve-se o mesmo resultado, onde $\gamma^T = 1{,}0068$ e $\gamma^N = 0{,}0043$ e o valor crítico para um nível de significância de 0,05 foi igual a $2F^{-\frac{1}{2}} = 1{,}208797$ calculado com 3 e 69 graus de liberdade, portanto os resultados inferenciais são válidos.

Tabela 21 – Valores para as métricas de seleção dos modelos Exponencial, Logístico e Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020).

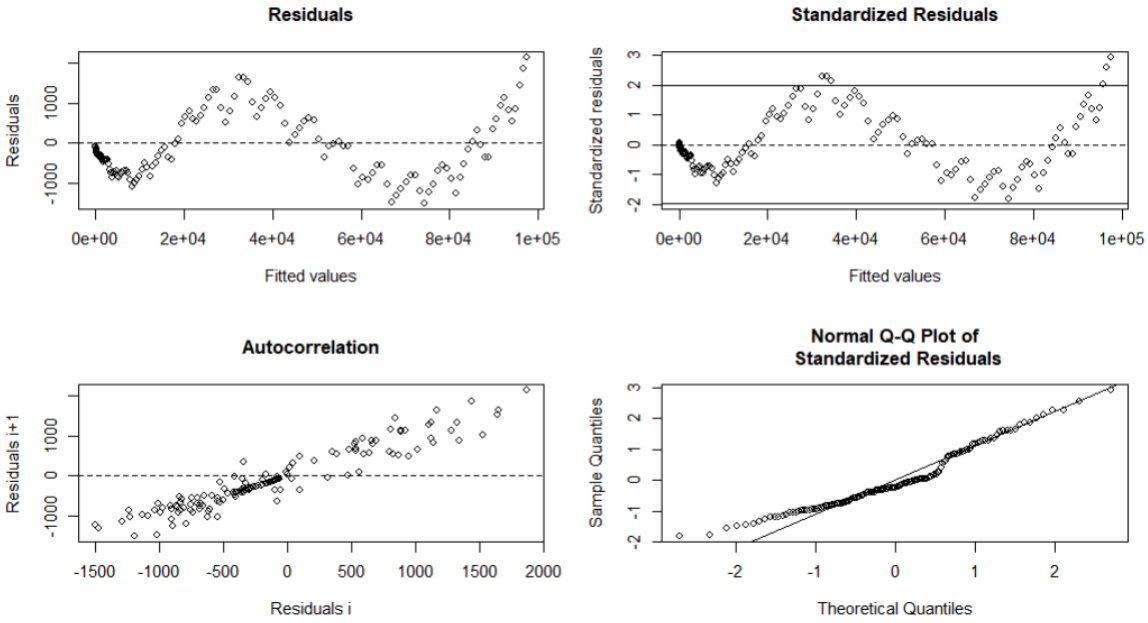
Modelo	Dados D1		Dados D2	
	Métrica	Valor	Métrica	Valor
Exponencial	AIC	2.961,095	AIC	1.319,386
	BIC	2.970,005	BIC	1.326,216
	R^2	0,9612	R^2	0,9804
	MSE	47.851.702	MSE	4.894.353
Logístico	AIC	2.607,826	AIC	1.060,937
	BIC	2.619,705	BIC	1.070,044
	R^2	0,9967	R^2	0,9995
	MSE	4.059.084	MSE	131.432,5
Gompertz	AIC	2.326,738	AIC	1.047,778
	BIC	2.338,617	BIC	1.056,884
	R^2	0,9995	R^2	0,9996
	MSE	576.348,3	MSE	109.478,3

Fonte: Elaborado pela autora (2022).

3.2.2.4 Análise dos resíduos e qualidade do ajuste

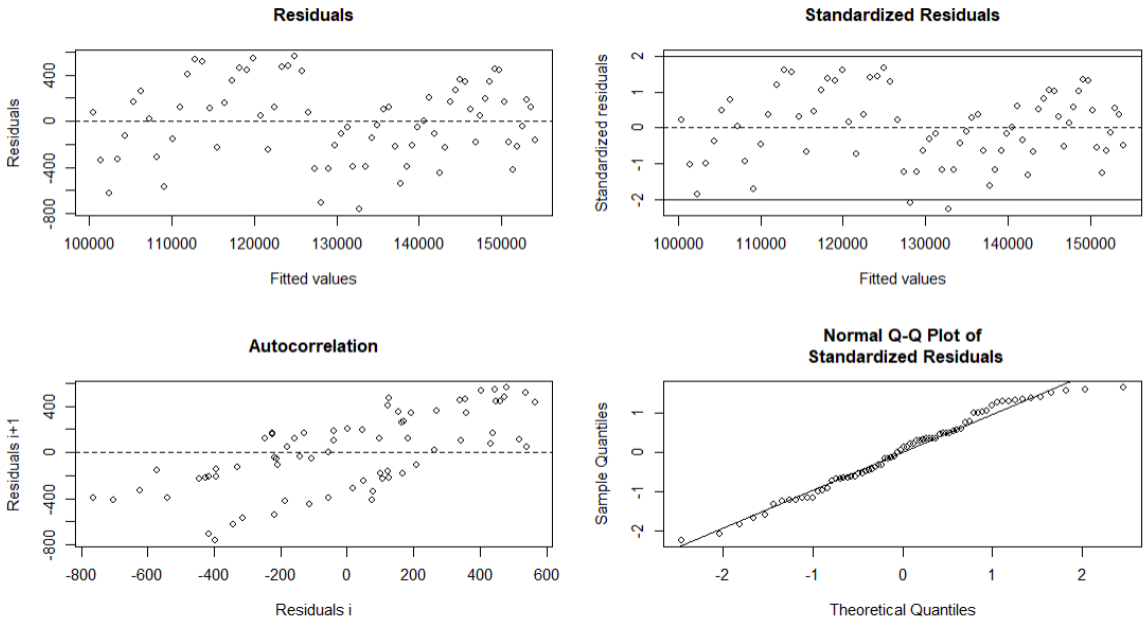
Para o período de 17/03/2020 a 07/08/2020 observa-se pela Figura 29 que a variância dos resíduos não é homogênea, a distribuição não é normal e os erros não são independentes, confirma-se tais interpretações gráficas por meio dos testes expostos na Tabela 22, cujos p-valores foram menores do que o nível de significância de 0,05 adotado. Já para o período de 08/08/2020 a 18/10/2020 nota-se pela Figura 30 que os erros possuem variância homogênea, distribuição normal e não são independentes. O que por meio dos resultados dos testes, Breusch-Pagan, Shapiro-Wilk e Durbin-Watson da Tabela 22 isso se confirma.

Figura 29 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D1 (17/03/2020 a 07/08/2020) de mortes acumuladas no Brasil.



Fonte: Elaborado pela autora (2022).

Figura 30 – Análise gráfica dos resíduos para o modelo Gompertz aplicado aos dados D2 (08/08/2020 a 18/10/2020) de mortes acumuladas no Brasil.



Fonte: Elaborado pela autora (2022).

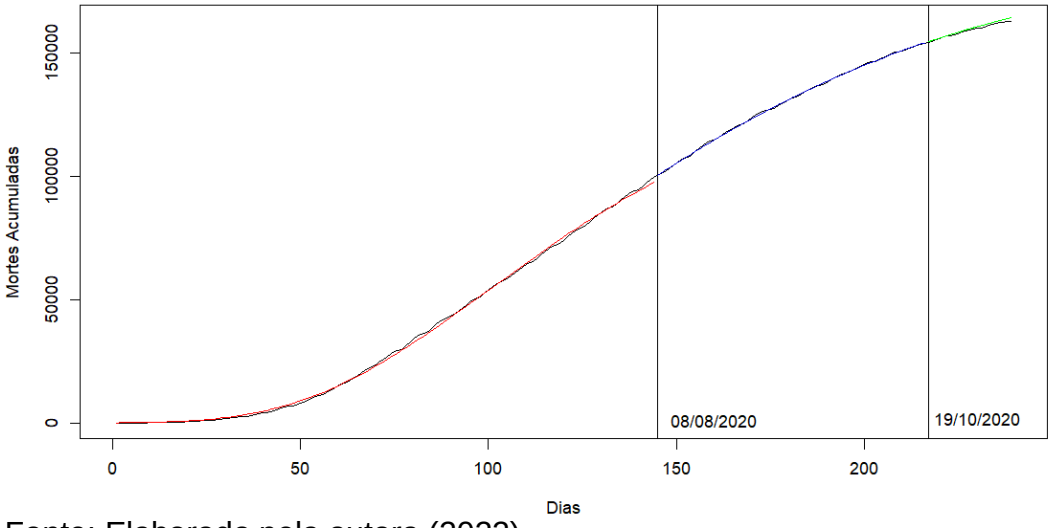
Tabela 22 – Estimativas das estatísticas dos testes de Shapiro-Wilk, Breusch-Pagan e Durbin-Watson e os respectivos p-valor dos resíduos do modelo Gompertz para os dados D1 (17/03/2020 a 07/08/2020) e D2 (08/08/2020 a 18/10/2020) de mortes acumuladas no Brasil.

Teste	D1	D2
Shapiro-Wilk	0,94964	0,97452
p-valor	<0,0001	0,1514
Breusch-Pagan	10,143	4,8978
p-valor	0,006272	0,08639
Durbin-Watson	0,092186	0,59972
p-valor	<0,0001	<0,0001

Fonte: Elaborado pela autora (2022).

Por fim, tem-se na Figura 31 o modelo de Gompertz por partes ajustado aos dados de óbitos acumulados para o Brasil. A data de 08/08/2020 e 19/10/2020 referem-se as mudanças de tendência, resultantes do teste de Pettitt. A linha vermelha é o modelo Gompertz ajustado aos dados de 17/03/2020 a 07/08/2020 e a azul ao período de 08/08/2020 a 18/10/2020, já linha verde são as previsões feitas para o dia 217 a 239 (19/10/2020 a 10/11/2020). Pode-se observar que os valores previstos estão acima dos observados, logo, o ajuste de outro modelo seria necessário.

Figura 31 – Ajuste do modelo Gompertz por partes aos dados de mortes acumuladas de COVID-19 para o Brasil e previsões.



Fonte: Elaborado pela autora (2022).

4 CONSIDERAÇÕES FINAIS

Com o avanço da pandemia do COVID-19 no Brasil no ano de 2020, medidas precisaram ser tomadas para conter a propagação do novo vírus, além do uso de máscaras, álcool em gel; o isolamento social, e o fechamento de estabelecimentos que não eram considerados essenciais (salão de beleza, comércio, academia, etc.), o que surpreendeu toda a população, pois a última pandemia registrada no mundo, dessa magnitude foi a gripe espanhola ocorrida no começo do século XX, medidas de isolamento também foram aplicadas no país e mesmo assim, cerca de 35 mil pessoas morreram no Brasil por conta da doença.

A vacinação contra a COVID-19 começou no dia 17/01/2021, com o avanço da mesma, as medidas se tornaram mais flexíveis, já no dia 19/07/21 a metade da população de São Paulo estava vacinada com a primeira dose. O dia 06/08/2020 foi a data onde 50% do país estava vacinado contra o coronavírus. Através dos gráficos apresentados nas seções 3.1.2.1 e 3.2.2.1 nota-se que a partir dessas datas as quantidades acumuladas de mortes pela doença começam a se estabilizar.

Ao utilizar os dados de todo o período de 2020 para ajustar os modelos não lineares de crescimento aos casos acumulados da doença tanto para São Paulo como para o Brasil notou-se um ajustamento ruim para o fim desse ano, logo optou-se por verificar o dia de mudança de tendência aplicando o teste de Pettitt, constatou-se que para os dias 29/07/2020 e 14/10/2020 houve a mudança de tendência, portando para cada parte ajustou-se novamente os modelos, sendo eles: Exponencial, Logístico e Gompertz; observou-se que os valores de AIC, BIC e MSE foram valores menores do que utilizando todo o período e as estimativas dos parâmetros foram significativas com 95% de confiança estatística. Já para os dados de mortes acumuladas em São Paulo e no Brasil a mudança de tendência ocorreu em 08/08/2020 e 19/10/2020, os valores para a seleção dos modelos também foram menores e as estimativas dos parâmetros foram significativas.

Os valores de R^2 foram altos, porém ao analisar os resíduos alguns casos não atenderam as suposições de normalidade, homogeneidade das variâncias e para a independência dos erros em nenhum ajuste essa suposição foi atendida, logo, uma melhor indicação é a aplicação de modelos não lineares para séries temporais ou até mesmo a modelagem da variância através do método dos mínimos quadrados ponderados, mas esse não era o objetivo do projeto em questão.

REFERÊNCIAS

AMARAL, L. S. et al. Interiorização do COVID-19: Uma análise da evolução dos casos/10 mil habitantes em municípios da Microrregião de Garanhuns no Estado de Pernambuco, através de modelos de Regressão não linear. **Research, Society and Development**, [s. l.], v. 9, n. 9, p. 1-19, ago. 2020. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/6582/6785. Acesso em: 15 mai. 2022.

BEALE, E.M. Confidence regions in non-linear estimation. **Journal of the royal statistical society**. Series B. Methodological, p.41-76, 1960. Disponível em: https://www.jstor.org/stable/2983877?seq=1#metadata_info_tab_contents. Acesso em: 16 jul. 2022.

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

BATES, D.M.; WATTS, D.G. Relative curvature measures of nonlinearity. **Journal of the royal statistical society**. Series B. Methodological, p.1-25, 1980. Disponível em: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1980.tb01094.x#:~:text=Simple%20relative%20curvature%20measures%20of,design%2Dparameterization%20combi%2D%20nation. Acesso em: 16 jul. 2022.

BOX, M. J. Bias in nonlinear estimation. **Journal of royal statistical society**. Série B. Methodological, p.171-201, 1971. Disponível em: https://www.jstor.org/stable/pdf/2985002.pdf. Acesso em: 20 jul. 2022.

COOK, R. D.; TSAI, C. L.; WEI, B. C. Bias in nonlinear regression. **Biometrika**, London, v. 73, n. 3, p. 615-623, Dec. 1986. Disponível em: https://www.jstor.org/stable/2336526#metadata_info_tab_contents. Acesso em: 23 jul. 2022.

COOK, R. D. Detection of influential observation in linear regression. **Technometrics**, v. 19, n. 1, p. 15-18, Feb. 1977.

COVIDLP. Disponível em: https://dest-ufmg.shinyapps.io/app_COVID19/. Acesso em: 18 out. 2022.

ESPINOSA, M. M. et al. Predição de casos e óbitos de COVID-19 em Mato Grosso e no Brasil. **Journal of Health and Biological Sciences**, Fortaleza, v. 8, n.1, p. 1-7, 2020. Disponível em: https://periodicos.unichristus.edu.br/jhbs/article/download/3224/1111. Acesso em: 15 mai. 2022.

GALANT, A. R. **Nonlinear statistical models.** John Wiley & Sons, 1987.

GILLS, P.R.; RATKOWSKY, D.A. The behavior of estimators of the parameters of various yield-density relationships. **Biometrics**, p.191-198, 1978. https://www.jstor.org/stable/2530009#metadata_info_tab_contents. Acesso em: 16 jul. 2022.

Li, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. **The New England Journal of medicine.** v. 382, n. 13, p. 1199-1207, Mar. 26, 2020. Disponível em: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2001316?articleTools=true. Acesso em: 11 jul. 2022.

LIU, Y. et al. The reproductive number of COVID19 is higher compared to SARS coronavirus. ***Journal of Travel Medicine***, Inglaterra, v. 27, p. 1-4, Feb. 13, 2020. Disponível em: https://doi:10.1093/jtm/taaa021. Acesso em: 11 jul. 2022.

MAZUCHELI, J.; ACHCAR, J. A. Algumas considerações em regressão não linear. **Acta Scientiarum**, Maringá, v. 24, n. 6, p. 1761-1770, 2002. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwi_saH-jLL5AhUxI7kGHc_uA7QQFnoECA4QAw&url=https%3A%2F%2Fperiodicos.uem.br%2Fojs%2Findex.php%2FActaSciTechnol%2Farticle%2Fdownload%2F2551%2F1574%2F&usg=AOvVaw2KC6Io3dtg4LPPJYgYHMjD. Acesso em: 16 mai. 2022.

PRUDENTE, Andrea Andrade. **Modelos não-lineares de regressão:** alguns aspectos da teoria assintótica. 2009. Dissertação (Pós-Graduação em Biometria e Estatística aplicada) – Universidade Federal Rural de Pernambuco, Recife, 2009.

REINHOLZ, M.; FRENCH, L. Medical education and care in dermatology during the SARS-CoV2 pandemia: challenges and chances. **Journal of the European Academy of Dermatology and Venereology**, v. 34, p. e214- e216, Mar. 27, 2020. Disponível em: https://doi:10.1111/jdv.16391. Acesso em: 11 jul. 2022.

SANTIAGO, E. J. P. et al. Modelos não lineares aplicados a mortalidade e casos da COVID-19 no Brasil, Itália e mundo. **Research, Society and Development**, [s. l.], v. 9, n. 6, p. 1-16, abr. 2020. Disponível em: https://rsdjournal.org/index.php/rsd/article/view/3561/3842. Acesso em: 15 mai. 2022.

SÃO PAULO. **Decreto nº 64.881, de 22 de março de 2020**. Decreta quarentena no Estado de São Paulo, no contexto da pandemia do COVID-19 (Novo Coronavírus), e dá providências complementares. São Paulo: Assembleia Legislativa do Estado de São Paulo, [2020]. Disponível em: https://www.al.sp.gov.br/repositorio/legislacao/decreto/2020/decreto-64881-22.03.2020.html. Acesso em: 27 abr. 2022.

SARAIVA, E. F. et al. A piecewise growth model for modeling the accumulated number of COVID-19 cases in the city of Campo Grande. **Brazilian Journal of Biometrics**, v. 39, n. 1, p. 240-265, Mar. 31, 2021. Disponível em: https://biometria.ufla.br/index.php/BBJ/article/view/542/301. Acesso em: 13 jul. 2022.

SEBER, G.; WILD, C. **Nonlinear Regression**. John Wiley & Sons, 2003.

WINICHAKOON, P. et al. Negative nasopharyngeal and oropharyngeal swabs do not rule out COVID-19. **Journal of Clinical Microbiology**, Boston, vol. 58, n. 5, p. 1-2, Apr. 23, 2020. Disponível em: https://doi:10.1128/jcm.00297-20. Acesso em: 11 jul. 2022.