
SÉRIE TEMPORAL APLICADA AO COVID-19 - BRASIL

Autor: Manuel Ferreira Junior

Disciplina: Série Temporal

Professora: Tatiene Correia de Souza

ferreira.jr.ufpb@gmail.com

<https://manuelfjr.github.io/>



Sumário

Análise Descritiva

Série Temporal

Função de Autocorrelação e

Autocorrelação Parcial

Teste de Estacionariedade

Estratégia de previsão

Métricas

ME

RMSE

MAE

Modelo obtido pelo critério AIC

Treinamento

Análise de Resíduos

Teste de normalidade

ACF

Teste de Ijung-box

Previsão

Modelo obtido pelo critério BIC

Treinamento

Análise de Resíduos

Teste de normalidade

ACF

Teste de Ijung-box

Previsão

Comparando modelos: AIC e BIC

Treinamento

Previsão

Alisamento Exponencial

Treinamento

Previsão

Conclusão

Referências



Análise Descritiva

Estatística	Valores
Minímo	0
1º Quartil	17294
Mediana ($\tilde{\mu}$)	33704
Mean (μ)	37795
3º Quartil	55460
Máximo	154664
Desvio Padrão (σ)	26334.06

Tabela: Algumas estatísticas



Análise Descritiva

Para o seguinte estudo, iremos considerar os casos diários de COVID-19 para o Brasil como um todo, realizando as técnicas estudadas durante o decorrer da disciplina de Séries Temporais. Além disso, o banco de dados foi filtrado para obtermos apenas os resultados referente ao número de casos diários confirmados no BRASIL, segue o link¹ referente em anexo.



¹<https://github.com/elhenrico/covid19-Brazil-timeseries>

Análise Descritiva

Com relação aos valores da série, ela não apresenta valores negativos e não foi preciso retirar nenhuma observação do conjunto de dados, ou seja, possuímos os valores observados dos casos confirmados de COVID-19 no Brasil desde o dia 26 de fevereiro de 2020 até 30 de junho de 2021.



Série Temporal



Série Temporal

A série utilizada contempla o período de 26 de fevereiro até o último dia atualizado sobre o conjunto de dados, retirado do repositório do elhenrico, filtrado e atualizado do Ministério da Saúde (*Ministerio da Saúde* 2021), sendo atualizado diariamente. Iremos contemplar dos dia 26 de fevereiro de 2020 até 30 de junho de 2021. Ao analisarmos a série, podemos observar uma queda no número de confirmados entre os períodos de agosto a outubro de 2020, evidenciado pelo possível resultado da criação da lei N^o 14.046, de 24 de agosto de 2020, onde fica explícito o adiamento e cancelamento de serviços, de reservas e de eventos dos setores de turismo e de cultura, em função do estado de calamidade pública, reconhecida pelo Decreto Legislativo n^o 6, de 20 de março de 2020, e da emergência de saúde pública de importância internacional decorrente da pandemia da COVID-19.



Série Temporal

Após a diminuição das medidas de afastamento, como lockdown e quarentena, em algumas regiões do país, mantendo apenas medidas de afastamentos previstas pela lei da quarentena, alguns estados começaram a “relaxar”, e houve outro aumento no número de contaminados entre os meses de novembro e dezembro, tendo uma diminuição no aumento de contaminados no final de dezembro. Contudo, é fácil notar que para o ano de 2021, já para o mês de fevereiro, podemos encontrar altos números de contaminados, dado pelo relaxamento das medidas de enfrentamento ao covid, por parte de alguns estados, sendo evidenciado uma possível terceira onda de ataque do vírus.



Série Temporal

Por fim, podemos notar que até então, o Brasil encontra-se na sua terceira onda, a primeira sendo a mais duradoura, dando início seus altos números por volta do final de abril, e tendo uma queda no mês de agosto; a segunda onde é possível notar em um intervalo menor sendo entre o final de outubro e, mantendo uma constância no final de dezembro; e, atualmente, encontramos em uma terceira onda, vinda do início do ano de 2021.



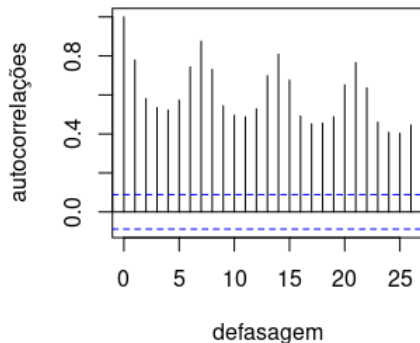
Série Temporal

Além disso, ao analisarmos a série temporal como um todo, podemos notar um comportamento estacionário, onde a série aparenta oscilar para baixo e para cima em torno de um valor esperado, ao desconsiderarmos o início dos registros de casos, por apresentar grande número de zeros inicialmente; ademais, também é possível notar uma variância constante sobre as mesmas circunstâncias citadas anteriormente.

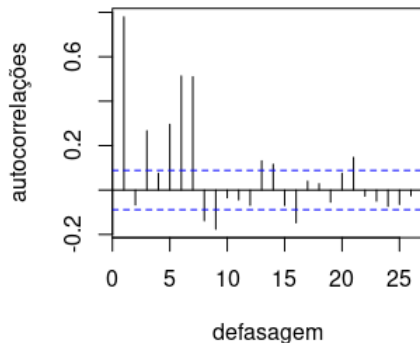


Função de Autocorrelação e Autocorrelação Parcial

Função de autocorrelação



Função de autocorrelação parcial



Função de Autocorrelação e Autocorrelação Parcial

Considerando o gráfico de autocorrelação, podemos notar um decaimento das autocorrelações ao longo das defasagens lento, ou seja, uma diminuição de uma autocorrelação entre os valores observados ao longo dos períodos, decaindo de forma lenta. Da mesma forma, podemos notar que a função de autocorrelação parcial encontra-se em torno de 0, apresentando apenas algumas defasagens superior ao intervalo calculado para as suas autocorrelações, porém algumas das suas autocorrelações são superiores a 0.4. Além disso, ao repararmos que para ambos os gráficos, função de autocorrelação e função de autocorrelação parcial, algumas autocorrelações extrapolam os limites calculados, evidenciando uma violação da suposição que $\rho = 0$, ou seja, que as correlações ao longo das defasagens é diferente de 0.



Teste de Estacionariedade

Ao considerarmos tanto o teste de de Phillips e Perron (Phillips e Perron 1988) quanto o de Dickey e Fuller (Dickey e Fuller 1979), tomando um nível de significância de 5%, temos que para ambos é possível notar um p-valor menor que o nível de significância, respectivamente equivalentes a 0.01 e 0.000032, ou seja, rejeitamos a hipótese nula de que a série é não estacionária (Testamos se $\phi = 1$, sobre h_0), ou seja, há evidências suficientes para afirmar que a série temporal do número de casos de COVID-19 confirmados diariamente é estacionária.



Estratégia de previsão

Para estratégia de previsão a ser adotada, será retirada as últimas 7 observações do banco de dados, assim reajustando os modelos obtidos, ARIMA(2,1,2) com o critério BIC (Schwarz 1978) e o ARIMA(3,1,3) pelo critério AIC (Akaike 1976), resultando em uma base de dados das datas de 26 de fevereiro de 2020 até 23 de junho de 2021. Iremos realizar previsões a **um**, **cinco** e **sete** dias a frente. As seguintes métricas serão adotadas:



ME

O **ME** ou **Erro médio** é a diferença do valor observado para o valor previsto pelo modelo, dado pela seguinte expressão:

$$ME = \frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \hat{Y}_i) \quad (1)$$

sendo n o tamanho amostral.



RMSE

O **RMSE** ou **Raiz do erro quadrático médio** trata-se da raiz da soma do quadrado das distâncias do valor observado para o previsto pelo modelo, dado por:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (2)$$



MAE

O **MAE** ou **Erro absoluto médio** é a diferença, **em valor absoluto**, do valor observado para o valor previsto pelo modelo, dado pela seguinte expressão:

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (3)$$

sendo n o tamanho amostral.



AIC: Treinamento

Tabela: ARIMA(3,1,3) - Resultados (AIC=10501.68)

Parâmetro	ar_1	ar_2	ar_3	ma_1	ma_2	ma_3
Coeficientes	0.2324	0.1566	-0.5814	-0.7092	-0.6637	0.7029

$$\sigma^2 = 156691191$$



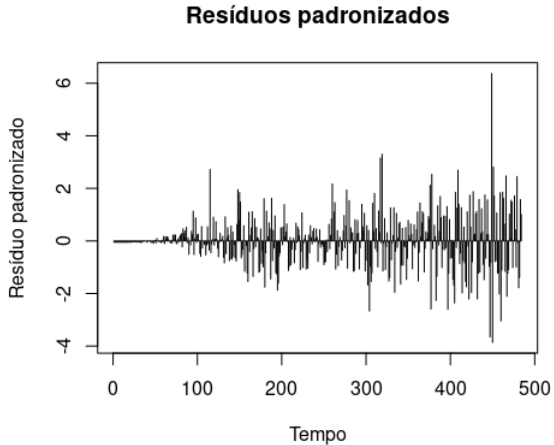
AIC: Treinamento

Tabela: Métricas obtidas durante o treinamento

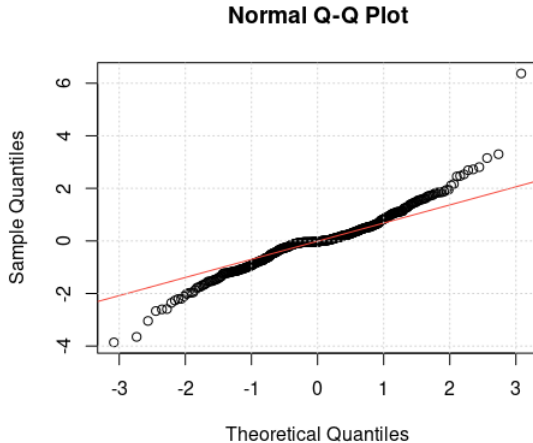
Métrica	ME	RMSE	MAE
Valor	592.08	12504.7	8566.21



AIC: Resíduos padronizados



AIC: Teste de normalidade

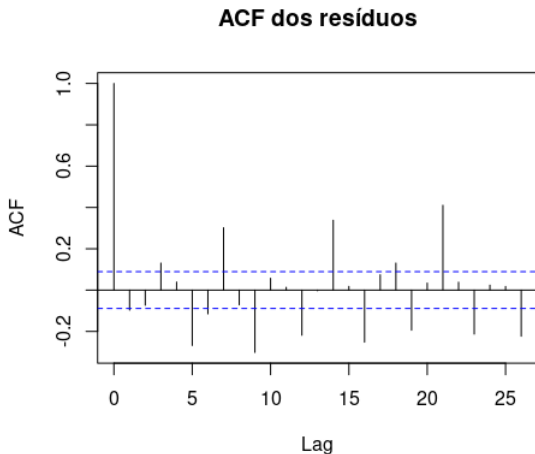


AIC: Teste de normalidade

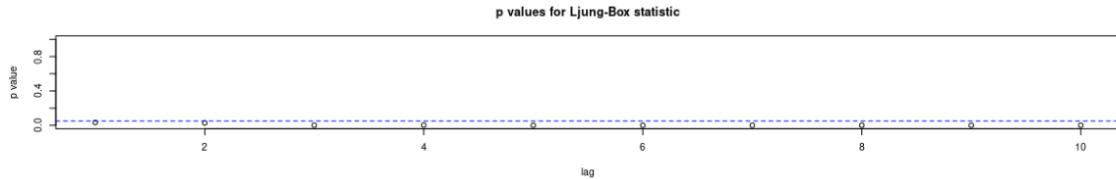
Ao realizar o teste de Lilliefors (Lilliefors 1967), considerando um nível de significância de 5%, obtemos um p-valor equivalente a $9.331 e^{-11}$, dessa forma, temos evidências suficientes para afirmar que a suposição de que os resíduos padronizados seguem uma normalidade está sendo violada.



AIC: Função de autocorrelação dos resíduos padronizados



AIC: Teste de Ljung-box



AIC: Teste de Ljung-box

Ao realizar o teste de Ljung-Box considerando um nível de significância de 5%, observamos que para todos os lags, os p-valores encontraram-se inferior ao nível de significância, dessa forma rejeitando a hipótese nula, logo podemos afirmar que os resíduos apresentam uma correlação estatisticamente significativa.



AIC: Previsão

Como dito na seção 5, realizamos as previsões a **um**, **cinco** e **sete** dias a frente, obtendo assim as seguintes métricas: (Considere h o número de passos preditos pelo modelo)

h / Métricas	ME-AIC	RMSE-AIC	MAE-AIC
1	-24109.41	24109.41	24109.41
5	-18632.96	27360.14	21866.52
7	-22989.73	30049.95	25299.42

Tabela: Métricas obtidas durante a previsão



BIC: Treinamento

Tabela: ARIMA(2,1,2) - Resultados (BIC=10505.99)

Parâmetro	ar_1	ar_2	ma_1	ma_2
Coeficientes	1.0658	-0.7284	-1.5949	0.7704

$$\sigma^2 = 159428697$$



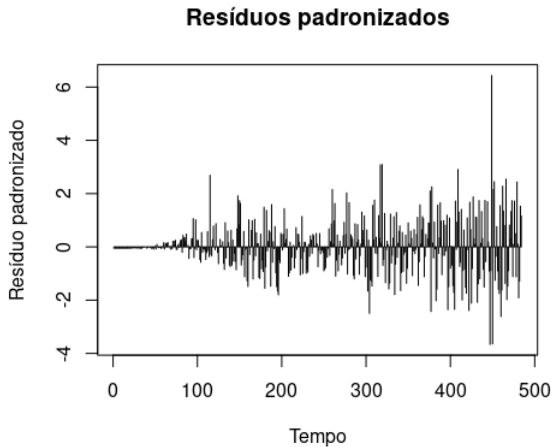
BIC: Treinamento

Tabela: Métricas obtidas durante o treinamento

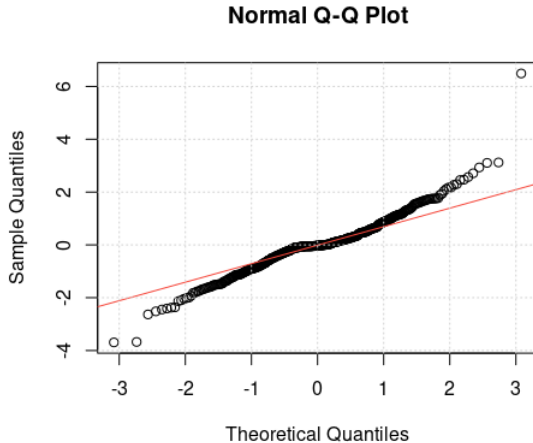
Métrica	ME	RMSE	MAE
Valor	621.07	12613.46	8697.075



BIC: Resíduos padronizados



BIC: Teste de normalidade

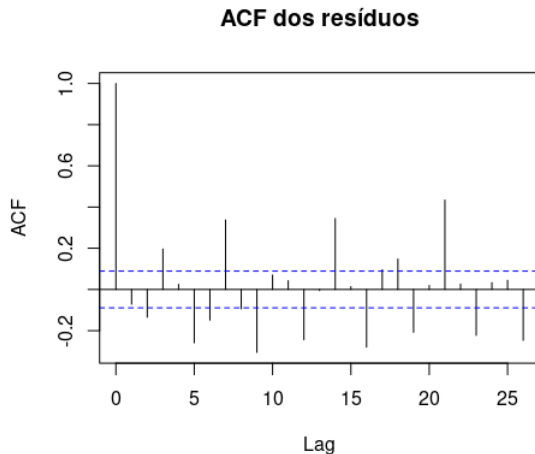


BIC: Teste de normalidade

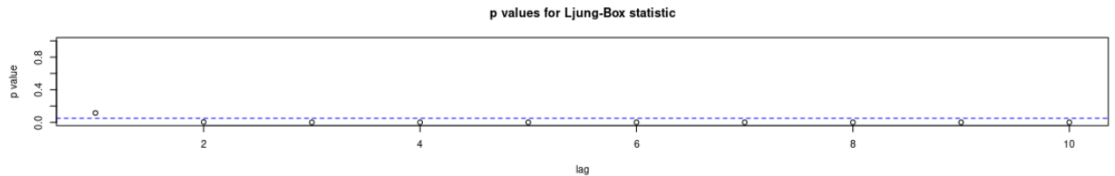
Ao realizar o teste de Lilliefors (Lilliefors 1967), considerando um nível de significância de 5%, obtemos um p-valor equivalente a $7.781e^{-11}$, dessa forma, temos evidências suficientes para afirmar que a suposição de que os resíduos padronizados seguem uma normalidade está sendo violada.



BIC: Função de autocorrelação dos resíduos padronizados



BIC: Teste de Ljung-box



BIC: Teste de Ljung-box

Ao realizar o teste de Ljung-Box considerando um nível de significância de 5%, observamos que para todos os lags, exceto um, os p-valores encontraram-se inferior ao nível de significância, dessa forma rejeitando a hipótese nula, logo podemos afirmar que os resíduos apresentam uma correlação estatisticamente significativa.



BIC: Previsão

Assim como foi feito pelo para o modelo escolhido pelo critério B0IC, foi realizado previsões a **um**, **cinco** e **sete** dias a frente, obtendo as seguintes métricas:

h / Métricas	ME-BIC	RMSE-BIC	MAE-BIC
1	-26235.85	26235.85	26235.85
5	-18769.91	27422.68	21982.13
7	-23424.62	30538.25	25719.07

Tabela: Métricas obtidas durante a previsão



Comparando modelos: AIC e BIC (Treinamento)

Tabela: Métricas obtidas durante o treinamento

Métrica	ME	RMSE	MAE
AIC	592.08	12504.7	8566.21
BIC	621.07	12613.46	8697.08

$$\sigma_{AIC}^2 = 156691191$$

$$\sigma_{BIC}^2 = 159428697$$

Perceba que o modelo obtido pelo critério de AIC apresentou os menores erros para todas as métricas utilizadas, além disso apresentou uma variância também inferior, mas ainda sim, alta.



Comparando modelos: AIC e BIC (Previsão)

Tabela: Métricas obtidas durante a previsão

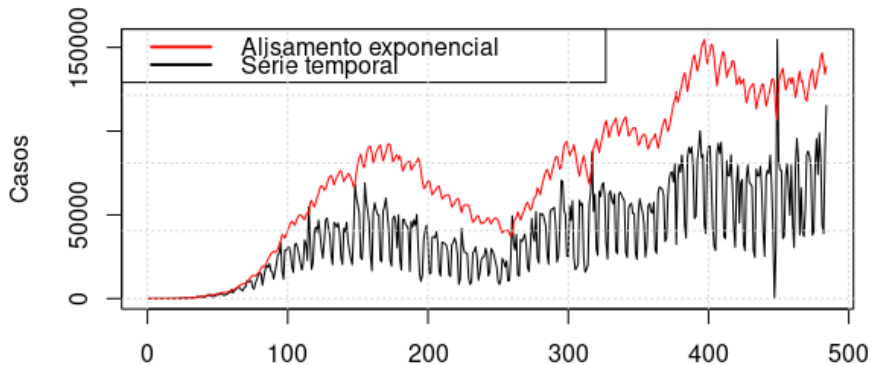
Critério	h	ME	RMSE	MAE
AIC	1	-24109.41	24109.41	24109.41
	5	-18632.96	27360.14	21866.52
	7	-22989.73	30049.95	25299.42
BIC	1	-26235.85	26235.85	26235.85
	5	-18769.91	27422.68	21982.13
	7	-23424.62	30538.25	25719.07

Note que o modelo selecionado pelo critério do AIC, podemos observar que para as previsões a **1**, **5** e **7** dias a frente, obtemos os menores erros, quando comparado ao modelo obtido pelo critério BIC.



Alisamento Exponencial

Figura: Série temporal vs Alisamento exponencial



Alisamento Exponencial

Graficamente, podemos observar que o algoritmo de alisamento exponencial conseguiu seguir o comportamento da série em momentos de aumento e de queda, porém com valores sempre superiores aos observados pela série original, contudo o máximo da série original é superior ao observado pela série, sendo um valor atípico.

Analogamente, como visto para os modelos selecionados pelos critérios AIC e BIC, será adotado o mesmo critério para estratégia de previsão, logo será retirado as 7 últimas observações e teremos como intuito estimar os valores a 1, 5 e 7 dias a frente.



Treinamento

Tabela: Métricas obtidas durante o treinamento

Métrica	ME	RMSE	MAE
Valor	1413.70	16059.19	11968.13



Treinamento

Tabela: Métricas obtidas durante o treinamento

Métrica	ME	RMSE	MAE
AIC	592.08	12504.7	8566.21
BIC	621.07	12613.46	8697.075
AE	1413.70	16059.19	11968.13



Previsão

h / Métricas	ME	RMSE	MAE
1	97.77	97.77	97.77
5	-17800.03	27545.26	20148.25
7	-18181.38	26043.62	19858.67

Tabela: Métricas obtidas durante a previsão



Previsão

Tabela: Métricas obtidas durante a previsão

Critério	h	ME	RMSE	MAE
AIC	1	-24109.41	24109.41	24109.41
	5	-18632.96	27360.14	21866.52
	7	-22989.73	30049.95	25299.42
BIC	1	-26235.85	26235.85	26235.85
	5	-18769.91	27422.68	21982.13
	7	-23424.62	30538.25	25719.07
AE	1	97.77	97.77	97.77
	5	-17800.03	27545.26	20148.25
	7	-18181.38	26043.62	19858.67



Conclusão

Concluiu-se então que, inicialmente, ao realizar um estudo aprofundado da série de casos confirmados de COVID-19 no Brasil, é facilmente evidenciado que o Brasil, até a data de 30 de junho de 2021, passou por no mínimo 3 ondas de altas de casos, apresentando um pico máximo de 154664 casos observados no final do mês de abril e, aparentemente, desde o início do mês de junho, o Brasil apresentou um crescente aumento, de forma lenta quando comparado aos outros meses, evidenciando uma nova onda. Além disso, o discente implementou um site utilizando-se da tecnologia de framework *Shiny* (Chang 2021), utilizado dentro da linguagem *R* (R Core Team 2021).



Conclusão

Dessa forma, foi feita a implementação de um site² de monitoramento de COVID-19, completamente autônomo, para casos e mortes confirmadas, não apenas para o Brasil e sim por Estado e por grande Região (Norte, Nordeste, Sul, Sudeste e Centro-Oeste); o site contempla os dados puramente para consultas, além da possibilidade de gerar gráfico de uma série temporal em um intervalo determinado pelo usuário.



²<https://mfjr.shinyapps.io/timeseries/>

Conclusão

Por fim, ao compararmos os resultados obtidos pelos modelos encontrados pelos critérios de AIC e BIC, juntamente com o algoritmo de alisamento exponencial, podemos observar que o modelo obtido pelo AIC apresentou os menores erros, como visto na tabela 11, considerando o período de treinamento dos modelos e o processo realizado pelo algoritmo, não apresentando tanta diferença quando comparado apenas com o modelo obtido pelo BIC.



Conclusão

Entretanto, é possível notar que, considerando a etapa de previsão, comparando os dois modelos e o algoritmo de alisamento exponencial, pela tabela 13, observamos que o algoritmo apresentou o menor erro para todas as previsões, um, cinco e sete dias a frente, sendo mais preciso com a estimativa a um dia a frente, apresentando o menor erro, equivalente para as três métricas adotadas.








Conclusão

Concluindo então, podemos observar que existem diversos métodos para predição e estimação do comportamento da série temporal, porém necessitam uma não violação dos seus pressupostos, bem como o alisamento exponencial utilizando, sendo o método simples, pois os dados não apresentam sazonalidade. Além dos métodos utilizados, existem outros métodos como o alisamento exponencial de Holt e Winters, levando em consideração tanto tendência quanto sazonalidade presente em uma série.



Referências

-  Akaike, Hirotugu (1976). “Canonical correlation analysis of time series and the use of an”. Em:
-  Dickey, David A e Wayne A Fuller (1979). “Distribution of the estimators for autoregressive time series with a unit root”. Em: *Journal of the American statistical association* 74.366a, pp. 427–431.
-  Lilliefors, Hubert W (1967). “On the Kolmogorov-Smirnov test for normality with mean and variance unknown”. Em: *Journal of the American statistical Association* 62.318, pp. 399–402.
-  Phillips, Peter CB e Pierre Perron (1988). “Testing for a unit root in time series regression”. Em: *Biometrika* 75.2, pp. 335–346.
-  Schwarz, Gideon (1978). “Estimating the dimension of a model”. Em: *The annals of statistics*, pp. 461–464.



Referências



Chang, Winston (2021). *shiny: Web Application Framework for R*. URL:
<https://shiny.rstudio.com/>.



Ministerio da Saúde (2021). Ministerio da Saúde. URL:
<https://www.gov.br/saude/pt-br>.



R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL:
<http://www.R-project.org/>.

