COVID-19 na População Mundial

Bruno Carvalho
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200145@isep.ipp.pt

Sofia Canelas
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200185@isep.ipp.pt

***Abstract*—Através de dados retirados de uma base de dados internacional, pretende-se analisar o impacto da pandemia na população mundial seguindo processos de análise exploratória de dados, análise inferencial e análise de correlação e regressão.**

***Index Terms*—análise, dados, estatística, COVID-19, pandemia, exploração, inferência, correlação, regressão**

I. Introdução

No âmbito da pandemia atual, foram extraídos da base de dados internacional "Our World in Data" [**?**], dinamizada pela universidade Johns Hopkins University (JHU), dados reais incidentes em casos confirmados de COVID-19, taxa de transmissibilidade do vírus, mortes, pacientes nos cuidados intensivos, testagem, vacinação e dados acerca da população. Estes dados são referentes ao período entre 01 de janeiro de 2020 a 27 de fevereiro de 2021.

Pretende-se analisar o impacto da pandemia na população mundial, com o objetivo de compreender a expansão e tratamento do vírus em diferentes partes do mundo. Irá destacar-se a análise do número de mortes ocorridas, o número total de infetados e a taxa de transmissibilidade do vírus, entre outros fatores exploratórios. Estes dados serão discutidos através de processos de análise exploratória de dados, análise inferencial e análise de correlação e regressão.

II. Metodologia de Trabalho

Tendo por base o ficheiro "owid-covid-data.csv" [**?**], foi criado um script em R com quatro diferentes tipos de análise: Análise Exploratória de Dados, Análise Inferencial, Análise de Correlação e Análise de Regressão. Cada uma destas análises possui alíneas independentes que pretendem tratar de dados específicos referentes ao ficheiro de dados. Após a conclusão das diferentes alíneas, foram analisados os dados e tiradas as respetivas conclusões presentes neste artigo nas secções IV, V, VI e VII.

III. Exploração e Preparação dos Dados

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

IV. Análise Exploratória de Dados

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

A. Número total de infetados ao longo do período de tempo estabelecido, por continente

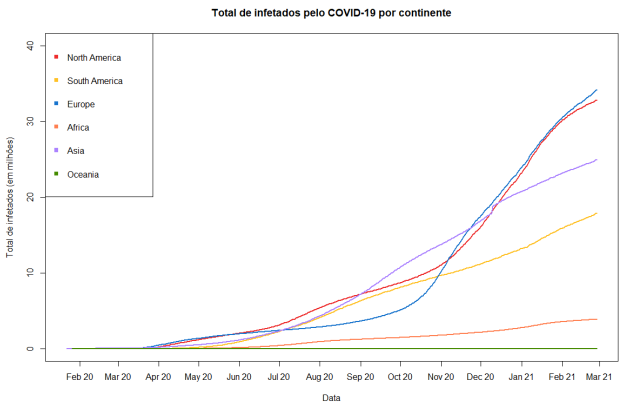

Fig. 1. Gráfico de linhas representativo do total de infetados pelo COVID-19, por continente, ao longo do período de tempo estabelecido

No gráfico de linhas apresentado acima observa-se que a América do Norte e Europa são os continentes que se destacam com os valores mais elevados de infetados, tendo ambos registado uma subida acentuada a partir dos meses de outubro e novembro, onde a Europa registou mais de 34 milhões de infetados.

Relativamente à Ásia e América do Sul, estes apresentam uma subida regular a partir de maio, sendo que a Ásia se distanciou com mais casos registados desde setembro.

Em contrapartida, África apresenta valores relativamente baixos em relação aos continentes referidos e a Oceânia destaca-se com um número reduzido de casos detetados durante todo o período apresentado no gráfico, com um máximo registado de quase 33.000 casos de infeção por COVID-19.

B. Número total de infetados por milhão de habitantes, ao longo do período de tempo, por continente

Neste gráfico de linhas conclui-se que a América do Norte é o continente que possui maior número total de infetados

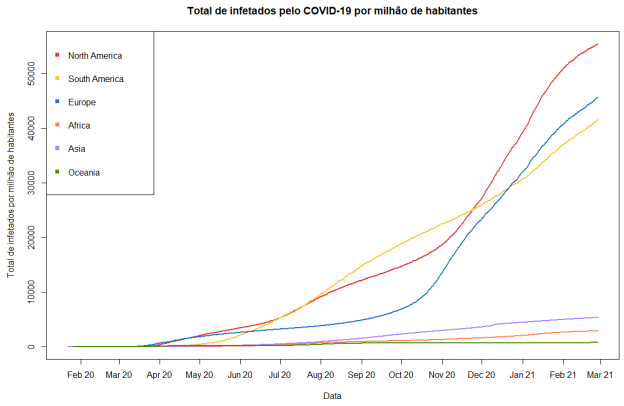

Fig. 2. Gráfico de linhas representativo do total de infetados pelo COVID-19 por milhão de habitantes, por continente, ao longo do período de tempo estabelecido

por milhão de habitantes, ultrapassando os 50.000 milhões de habitantes. Por outro lado, a Oceânia é, não só o continente mais estável uma vez que não se verificam oscilações significativas no número de infetados, como também é o que está mais próximo do 0, ou seja, é o continente com menor número de infetados por milhão de habitante.

É de notar que a América e a Europa são os únicos a ultrapassar o valor de 10.000 milhões de habitantes e cujas linhas tendem a subir de forma exponencial, ao contrário dos restantes continentes que se mantêm abaixo desse patamar e cujo crescimento é mais linear.

C. Número de mortos diários, por milhão de habitantes, dos seguintes países: Portugal, Espanha, Itália e Reino Unido

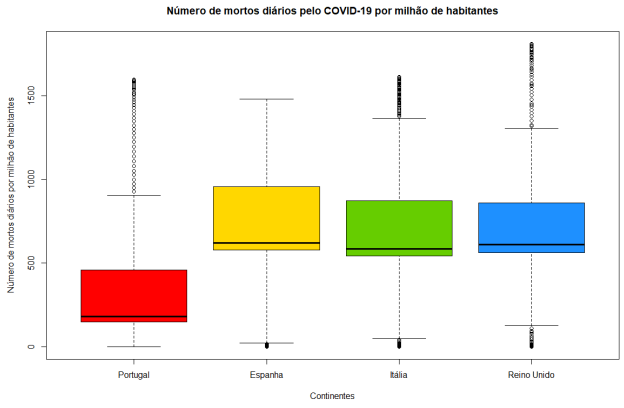

Fig. 3. Diagrama de Caixa representativo do número de mortos diário pelo COVID-19 por milhão de habitantes, em Portugal, Espanha, Itália e Reino Unido

O diagrama de caixa presente acima, demonstra, no geral, que Portugal, Itália e o Reino Unido têm amplitudes interquartis semelhantes, com Espanha um pouco acima destes três países. Também é possível observar que todos os países presentes nesta análise apresentam uma assimetria negativa, onde a mediana se encontra na parte inferior das caixas.

Espanha é o país com maior número de mortes diárias, por milhão de habitantes, e, também, o que apresenta maior discrepância dos dados, no entanto, sem apresentar outliers acima do terceiro quartil, ou seja, não registou nenhum dia com mortes acima da normalidade dos dados.

Em contrapartida, Portugal apresenta números mais reduzidos em relação aos restantes, com a particularidade de ter outliers acima do terceiro quartil, o que representa que, em vários dias, Portugal teve um número de mortes diárias significativamente acima da maioria do período temporal dos dados registados.

Em relação a Itália e Inglaterra, estes apresentam dados bastante semelhantes, onde a maior diferença encontra-se no elevado número de outliers acima do terceiro quartil.

D. Número total de mortos, por milhão de habitantes, e o número de testes diários, por milhar de habitantes, dos países: Albânia, Dinamarca, Alemanha e Rússia

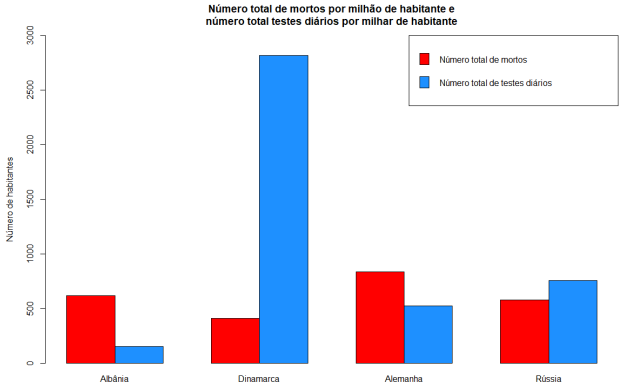

Fig. 4. Gráfico de Barras representativo do número total de mortos pelo COVID-19 por milhão de habitantes, e número de testes diários por milhar de habitantes, na Albânia, Dinamarca, Alemanha e Rússia

Como é possível observar no gráfico de barras, a Dinamarca destaca-se como o país com maior número total de testes diários em relação aos outros presentes nesta análise, com um máximo superior a 2800 pessoas testadas por cada milhão de habitantes. Contrastando, a Albânia é o país com menor número de testes, por milhar de habitante, tendo atingido pouco mais de 150 pessoas testadas, por milhão de habitantes, num único dia. Relativamente à Alemanha e Rússia, estes encontram-se entre os 500 e 1.000 testes diários, no melhor dia de testagem de ambos.

A respeito do número total de mortos, a Dinamarca volta a destacar-se como o país com o número total de mortos, por milhão de habitantes, mais reduzido, tendo aproximadamente um total de 400 mortos. O número mais notório de total de mortos encontra-se na Alemanha, dentro deste conjunto, com um valor acima 800 mortos, o dobro dos registos da Dinamarca. Os valores obtidos da Albânia e Rússia rondam um total de 600 mortos, por milhão de habitantes.

Assim, percebe-se que a Dinamarca é o país com maior destaque destes quatro, sendo aquele com maior número de

testes realizados num dia e menor número total de mortos. Também é possível concluir que a Alemanha e Rússia apresentam valores relativamente próximos nestes dois parâmetros e a Albânia demonstra uma discrepância significativa.

E. País europeu com maior número de infetados, por milhão de habitantes, num único dia

O Vaticano registou um total de 8.652,658/1.000.000 de habitantes infetados no dia 12 de outubro de 2020, sendo, assim, o país europeu com maior número de infetados, por milhão de habitante, num único dia.

F. Dia e país onde se registou a maior taxa de transmissibilidade do vírus

A Coreia do Sul registou, no dia 22 de fevereiro de 2020, a maior taxa de transmissibilidade do vírus, sendo esse valor de 6,72.

G. Número de mortos diários por milhão de habitantes, em cada continente

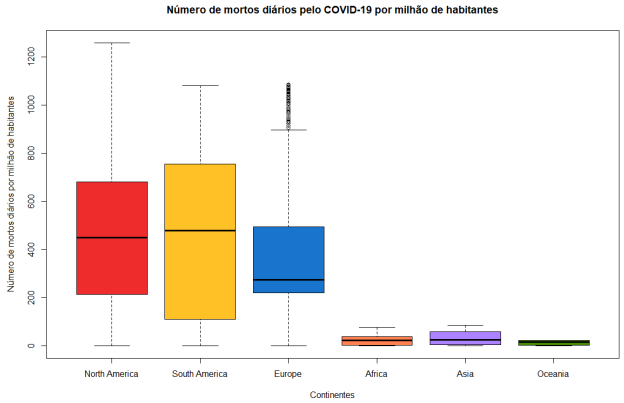

Fig. 5. Diagrama de Caixa representativo do número de mortos diários pelo COVID-19 por milhão de habitantes, por continente

O diagrama de caixa acima representado permite-nos afirmar que a América do Norte é o continente com maior número de mortos diários por milhão de habitantes.

A América do Sul é o continente mais disperso, visto que o tamanho da caixa (amplitude interquartil) é maior, comparativamente às restantes caixas, havendo, consequentemente, uma maior variação dos dados neste continente.

Analisando o fator de simetria, os dados da América do Norte e África são os que aparentam ter uma distribuição mais simétrica, dado que a linha da mediana se encontra posicionada no centro da caixa. Já a Europa e a Ásia denotam uma grande assimetria, sendo os seus dados assimetricamente negativos, uma vez que a linha da mediana se aproxima do primeiro quartil. A América do Sul e Oceânia também denotam uma assimetria, mas positiva, visto que a linha da mediana está mais próxima do terceiro quartil, contudo, a assimetria presente na América do Sul não é tão elevada.

Por último, a Europa, mesmo depois de serem removidos alguns outliers pelo critério x é outlier se e só se satisfizer o critério indicado na equação abaixo (10), é o continente que ainda apresenta bastantes valores discrepantes, estando estes acima do limite máximo.

$$x \in [Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR] \tag{1}$$

V. Análise Inferencial

A. Verificação se a média da taxa transmissibilidade no Reino Unido é superior à média da taxa de transmissibilidade em Portugal, relativamente aos dados de 30 dias

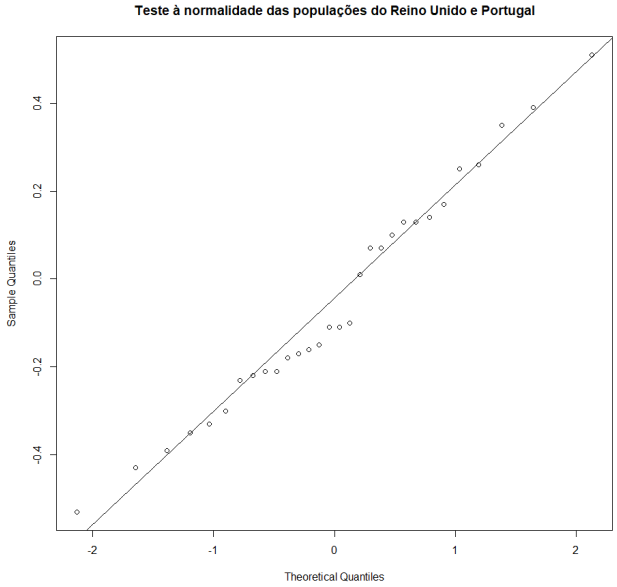

Fig. 6. Diagrama de Dispersão representativo do teste à normalidade das populações do Reino Unido e Portugal

O teste de hipóteses segue a seguinte definição:

$$\begin{aligned} H_0 &: \mu_{uk} \leq \mu_{pt} \\ H_1 &: \mu_{uk} > \mu_{pt} \end{aligned} \tag{2}$$

Como os dados são relativos a 30 dias, pode-se concluir que as amostra de ambos os países seguem uma distribuição aproximadamente normal, através do Teorema do Limite Central, sendo que esta conclusão foi verificada com os testes de Shapiro-Wilk (p-value = 0.686), Lillierfors (p-value = 0. 1533) e, também, pelo método gráfico (representado na figura acima), que demonstra que as variáveis têm uma relação linear.

Ao saber que as distribuições são normais e são emparelhadas, uma vez que partilham a variável dos dias, realizou-se um t-Test que obteve um p-value igual a 0.8652, que permite concluir que não se deve rejeitar a hipótese nula, ou seja, que as médias das taxas de transmissibilidade dos dois países são idênticas, a um nível de significância superior a 8.5%. Logo, a média da taxa de transmissibilidade do Reino unido é semelhante à média registada em Portugal.

B. X

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

C. Verificação da existência de diferenças significativas entre os números médios diários de mortes, por milhão de habitantes, entre os continentes

O teste de hipóteses segue a seguinte definição:

$$\begin{aligned} &H_0 : \mu_{Asia} = \mu_{Africa} = \mu_{Europe} = \mu_{NorthAmerica} = \\ &= \mu_{SouthAmerica} \\ &H_1 : \mu_i \neq \mu_j \\ &i, j \in [Asia, Africa, Europe, NorthAmerica, \\ &SouthAmerica] \end{aligned} \tag{3}$$

Com a obtenção dos dados pretendidos para cada continente, realizaram-se testes de Shapiro-Wilk para verificar a normalidade dos dados referentes a cada continente sendo que os resultados obtidos nos valores dos p-values (X) demonstram que Ásia, Europa e América do Sul não seguem distribuições normais uma vez que os seus p-values são inferiores a 0.05, rejeitando, assim, a hipótese nula.

$$\begin{aligned} &Asia_{p-value} = 0.07204, \\ &Africa_{p-value} = 0.001552, \\ &Europe_{p-value} = 0.001734, \\ &NorthAmerica_{p-value} = 0.08603, \\ &SouthAmerica_{p-value} = 2.932 \times 10^{-8} \end{aligned} \tag{4}$$

Após estas confirmações realizou-se um teste Levene para verificar a homogeneidade dos dados, sendo que o valor obtido do p-value indicado abaixo permite concluir que as variâncias dos dados são significativamente diferentes.

$$p - value = 1.437 \times 10^{-11} \tag{5}$$

Com os resultados destes testes, conclui-se que não é possível realizar um teste One-Way ANOVA para verificar a igualdade das médias dos dados de cada continente, pelo que se realizou um teste Kruskal-Wallis, onde se obteve o p-value indicado abaixo.

$$p - value = 2.2 \times 10^{-16} \tag{6}$$

Isto permite concluir que se deve rejeitar a hipótese nula, ou seja, que não existe igualdade entre os números médios de mortos diários, por milhão de habitantes, entre os continentes.

VI. Análise de Correlação

A. Correlação, em 2021, entre o valor máximo da taxa diária de transmissibilidade e a densidade populacional de todos os países da Europa com mais de 10 milhões de habitantes

Sendo X a variável representativa do valor máximo da taxa diária de transmissibilidade de todos os países europeus com mais de 10 milhões de habitantes e Y da densidade populacional dos mesmos, foram verificados, em primeiro lugar, os seguintes pressupostos:

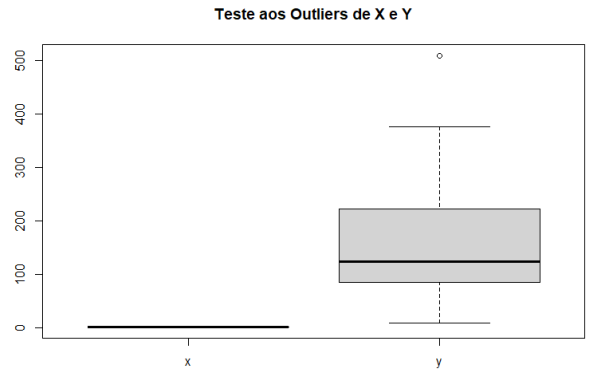

Fig. 7. Diagrama de Caixa representativo das variáveis X1 e Y1

1) As variáveis devem ser contínuas e não devem existir outliers significativos: Sendo as variáveis contínuas, pretendeu-se verificar, através de um diagrama de caixa, se os outliers seriam significativos, o que se conclui pelo gráfico acima que não o são.

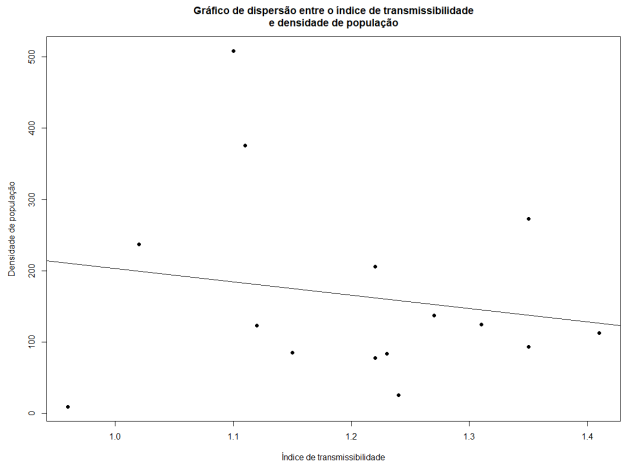

Fig. 8. Diagrama de Dispersão representativo das variáveis X1 e Y1

2) Deve existir uma relação linear entre as duas variáveis: Para a confirmação desta condição realizou-se um diagrama de dispersão onde se comprova que não existe uma relação linear entre as duas variáveis, estando os pontos demasiado dispersos em relação à linha traçada. Este pressuposto não é, então, verificado.

3) As variáveis devem ter, aproximadamente, uma distribuição normal: Ao realizar um gráfico de dispersão Normal Q-Q, verificam-se que os pontos estão razoavelmente ajustados à linha reta, o que nos permite a suspeita dos resíduos serem normais. Pelo teste de normalidade Shapiro-Wilk, obteve-se um p-value de 0,2014, pelo que se confirma a condição de normalidade, já que este é superior a 0,05.

4) As variâncias devem ser iguais (Homocedasticidade): Por fim, no gráfico de dispersão acima representado não se denota qualquer tendência no mesmo, estando os pontos

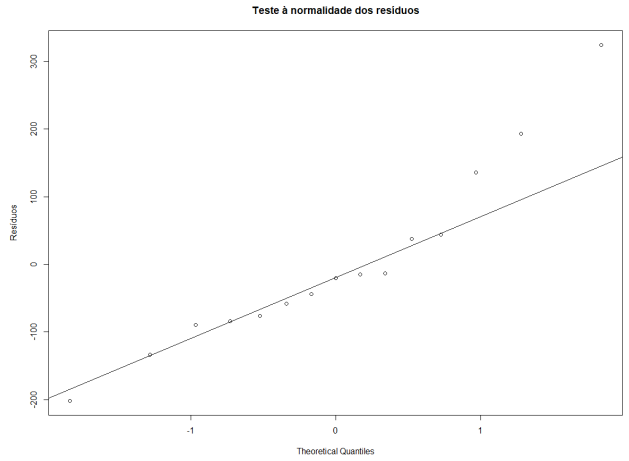

Fig. 9. Gráfico de Dispersão Normal Q-Q representativo das variáveis X1 e Y1

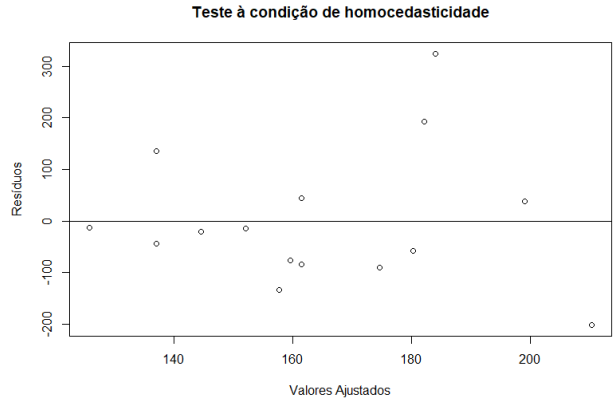

Fig. 10. Diagrama de Dispersão ajustado representativo das variáveis X1 e Y1

aleatoriamente distribuídos em torno do valor 0, onde se traçou uma linha. Assim, temos indícios de que a variância dos resíduos é homocedástica, contudo, sendo a amostra pequena (x e y = 15), testou-se também a variância. No teste à variância o p-value obtido é ligeiramente acima de 0,05 (0,05919), o que nos permite aceitar a análise anterior e concluir que as variâncias são iguais.

Verificados três dos pressupostos anteriores, procedeu-se à realização de um Teste de Correlação Linear de Pearson, onde se concluiu que, de facto, as variáveis estão fracamente correlacionadas, uma vez que o valor de r obtido é próximo de 0 (-0,175886) e o valor de p-value não é significativo, estando bastante acima de 0,05 (0,5306).

B. Correlação, em 2021, entre o total de mortos por milhão de habitantes e a percentagem da população com 65 anos ou mais em todos os países da Europa com mais de 10 milhões de habitantes

Primeiramente foram verificados os seguintes pressupostos:

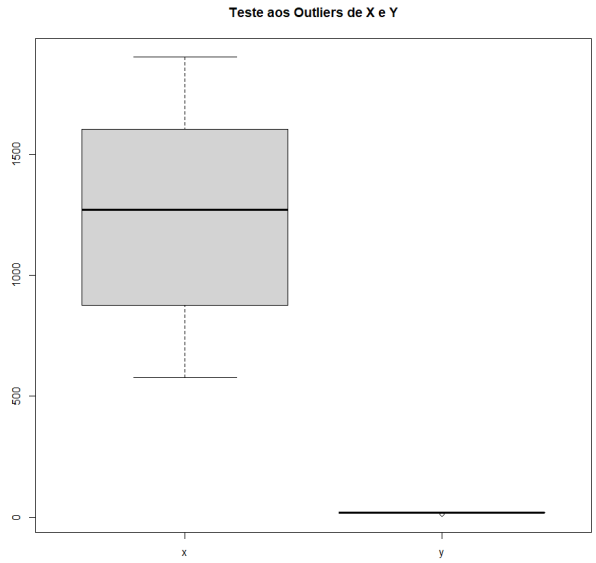

Fig. 11. Diagrama de Caixa representativo das variáveis X2 e Y2

1) As variáveis devem ser contínuas e não devem existir outliers significativos: Como as variáveis são contínuas, verificou-se, através de um diagrama de caixa, se os outliers seriam significativos, o que se conclui pelo gráfico acima que não o são.

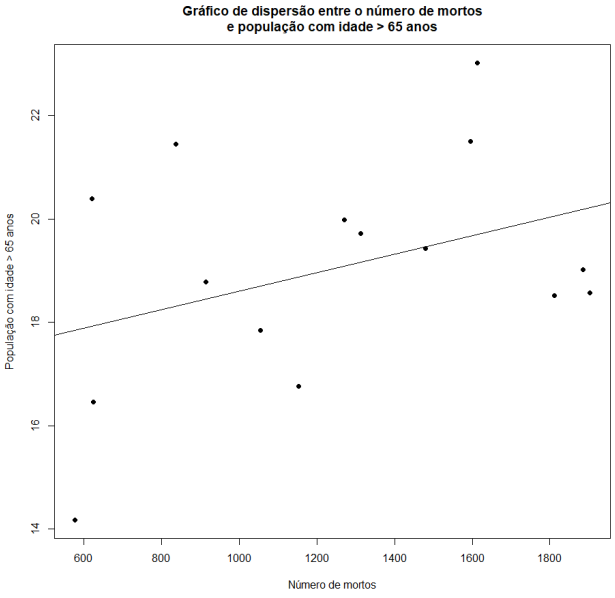

Fig. 12. Diagrama de Dispersão representativo das variáveis X2 e Y2

2) Deve existir uma relação linear entre as duas variáveis: Através do diagrama de dispersão acima comprova-se que não existe uma relação linear entre as duas variáveis, uma vez que os pontos se encontram distantes da linha de regressão traçada.

Ao realizar um gráfico de dispersão Normal Q-Q, verificam-

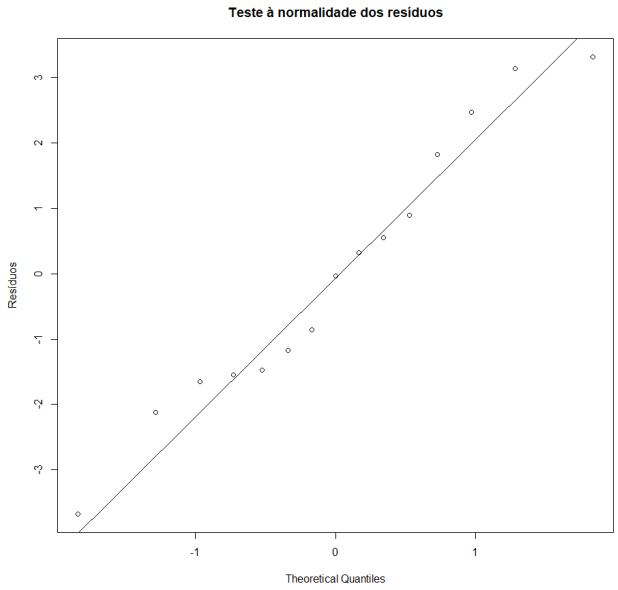

Fig. 13. Gráfico de Dispersão Normal-QQ representativo das variáveis X2 e Y2

se que os pontos estão ajustados à linha reta, ou seja, permite concluir que os resíduos seguem uma distribuição normal.

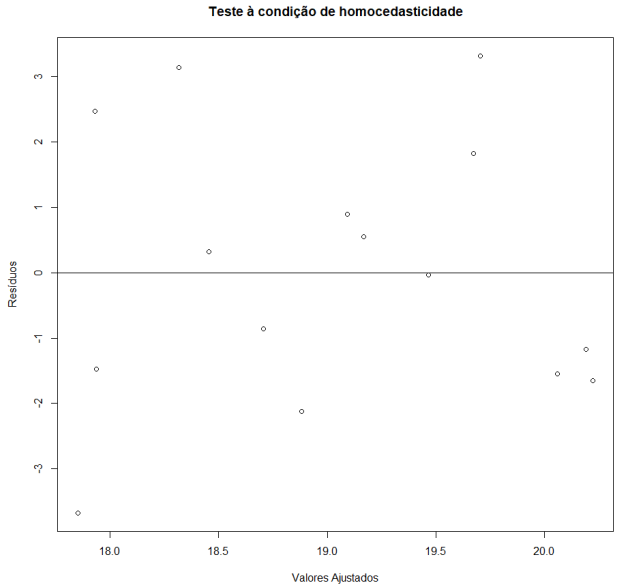

Fig. 14. Diagrama de Dispersão ajustado representativo das variáveis X2 e Y2

3) As variáveis devem ter, aproximadamente, uma distribuição normal: Pelo teste de Shapiro-Wilk, obteve-se um p-value de 0,7183, que confirma a condição de normalidade, obtendo a mesma conclusão do método gráfico.

4) As variâncias devem ser iguais (Homocedasticidade): Por fim, no gráfico acima representado não se denota qualquer tendência no gráfico, estando os pontos aleatoriamente distribuídos em torno do valor 0, onde se traçou uma linha. Assim, temos indícios de que a variância dos resíduos é homocedástica, contudo, sendo a amostra pequena (x = y = 15 ¡ 30), testou-se também a variância.

No teste à variância o p-value obtido é ligeiramente acima de 0,05 (0,05919), o que nos permite aceitar a análise anterior e concluir que as variâncias são iguais.

Verificados três dos pressupostos anteriores, procedeu-se à realização de um Teste de Correlação Linear de Pearson.

De facto, as variáveis estão fracamente correlacionadas, uma vez que o valor de r obtido é próximo de 0 (-0,175886) e o valor de p-value não é significativo, estando bastante acima de 0,05 (0,5306).

VII. Análise de Regressão

A. Construção do modelo de regressão linear múltipla

O modelo de regressão linear obtido através das variáveis independentes: mortes diárias, casos diários e taxa de transmissibilidade; para a variável dependente índice de rigor resulta na seguinte equação com os valores dos coeficientes obtidos:

$$\begin{aligned} &Y = \beta_0 + \beta_1 X_1 + \beta_1 X_3 + \beta_3 X_3 \\ &\beta_0 = 81.827720 \qquad \beta_{0p-value} = 2 \times 10^{-16} \\ &\beta_1 = 0.139378 \qquad \beta_{1p-value} = 0.141 \\ &\beta_2 = 0.002183 \qquad \beta_{2p-value} = 0.259 \\ &\beta_3 = -9.096406 \qquad \beta_{3p-value} = 8.55 \times 10^{-16} \end{aligned} \tag{7}$$

Relativamente à significância dos coeficientes obtidos, a interceção e a taxa de transmissibilidade são aqueles que apresentam p-values bastante baixos, sendo estatisticamente significantes, enquanto os coeficientes das mortes diárias e casos diários apresentam um p-value alto, de 0.141 e 0.259, respetivamente, ou seja, são valores com baixo nível de significância estatisticamente.

$$\begin{aligned} &R^2_{ajustado} = 0.2275 \\ &Estatistica_{p-value} = 2.2 \times 10^{-16} \end{aligned} \tag{8}$$

O valor do R-quadrado ajustado indica que apenas 22.75% dos valores Índice de Rigor é explicado pelos outros parâmetros. Como o valor do p-value da estatística de teste é bastante reduzido, o valor do R-quadrado ajustado é estatisticamente significativo com um grau de significância superior a 99%.

B. Verificação da satisfação das condições de Homocedasticidade, Autocorrelação nula e de Multicolinearidade

1) Homocedasticidade: Ao realizar um gráfico Normal Q-Q denota-se uma certa tendência linear nos pontos, levantando a suspeita de que a distribuição dos resíduos não é normal, o que se confirma pelo teste de Shapiro que adquire um p-value inferior a 0,05, como indicado em baixo.

$$p - value = 6.232 \times 10^{-13} \tag{9}$$

Para além de não se confirmar a normalidade dos resíduos, confirmou-se ainda através de um t.test que a média dos erros

não é zero e, através de um teste à variância, que comprovou também que as variâncias não são constantes, já que o p-value é inferior a 0,05, como indicado em baixo.

$$p - value = 2.2 \times 10^{-16} \tag{10}$$

2) Autocorrelação Nula: Para verificar esta condição recorreu-se a um teste de Durbin Watson, sendo as hipóteses testadas as seguintes:

H0: Os resíduos são independentes.

H1: Os resíduos não são independentes.

Obtendo um p-value de 0, sendo este inferior a 0.05, rejeita-se a hipótese nula, o que nos permite afirmar que a condição de independência não se verifica.

3) Multicolinearidade: Por último, foi testada a multicolinearidade dos resíduos pelo Fator de Inflação de Variância (VIF), sendo que se considera a ausência de multicolinearidade quando o VIF é inferior a 3. Através deste teste, cujos valores obtidos estão apresentados abaixo, verificou-se que Dm e Cm são multicolineares, contrariamente a Rm.

$$\begin{aligned} D_m &= 5.744092 \\ C_m &= 5.665572 \\ R_m &= 1.715799 \end{aligned} \tag{11}$$

C. Estimação do Ir para os valores Dm = 10, Cm = 460, Rm = 1.1

A estimação do valor do Índice de Rigor médio mensal para os valores de Dm = 10, Cm = 460, Rm = 1.1 resulta num índice de 74.21956, com um nível de confiança de 95%, pertencente ao intervalo de confiança [73.52288, 74.91623].

VIII. Discussão de Resultados

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

IX. Conclusões

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Referências Bibliográficas

Please number citations consecutively within brackets [**?**]. The sentence punctuation follows the bracket [**?**]. Refer simply to the reference number, as in [**?**]—do not use "Ref. [**?**]" or "reference [**?**]" except at the beginning of a sentence: "Reference [**?**] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [**?**]. Papers that have been accepted for publication should be cited as "in press" [**?**]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].