

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

Resumo—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 [1], dinamizada pela 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, 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” [2], 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 respectivas conclusões presentes neste artigo nas secções IV, V, VI e VII.

III. EXPLORAÇÃO E PREPARAÇÃO DOS DADOS

No início do script em R estão presentes as variáveis globais que permitem facilitar a obtenção de dados necessários às diversas análises realizadas. Estas tiveram como função guardar os nomes, códigos e cores dos continentes e países necessários à exploração de dados pretendida. Para além disso, está também presente a importação de duas bibliotecas: *nortest*

[3] e *car* [4], que permitem a realização de diversos testes presentes nas análises efetuadas.

O script está organizado pelas análises, delimitadas de forma evidente, sendo que cada uma destas contém várias alíneas, listadas alfabeticamente. As letras indicadoras das alíneas são correspondentes às letras que delimitam as respetivas secções de análise neste artigo. Em cada alínea existe uma secção para a criação de variáveis referentes à própria alínea (ou mais do que uma alínea, caso os dados se repitam), seguidas da seleção dos dados conforme o pretendido, assim como modificações aos mesmos, caso necessário (por exemplo, a remoção de NAs ou organização dos dados).

Em seguida, é visível o código que permite a realização da respetiva análise, o qual inclui a criação de gráficos ou testes estatísticos. Nos gráficos que possuem diferentes amostras de populações, deu-se uso a paletas de cor [5], de modo a facilitar a visualização dos mesmos. Já os testes estatísticos, são visíveis na consola do script e referidos ao longo do artigo.

IV. ANÁLISE EXPLORATÓRIA DE DADOS

Na análise exploratória de dados pretendeu-se analisar diversas características referentes aos continentes e determinados países, como o número de infetados, número de mortes e taxa de transmissibilidade do vírus.

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

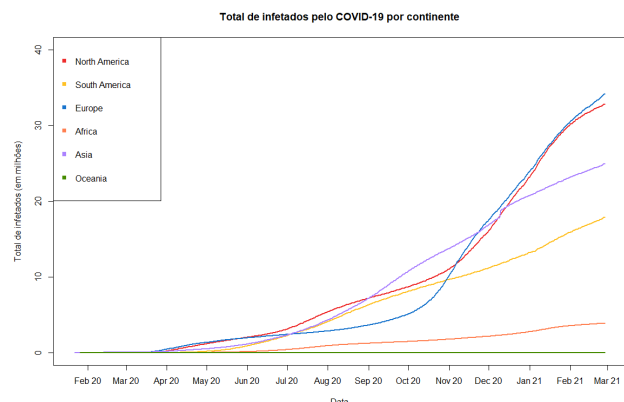


Figura 1. Gráfico de linhas representativo do total de infetados pelo COVID-19, por continente, ao longo do período de tempo estabelecido

No gráfico de linhas da Fig.1 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 distanciava 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

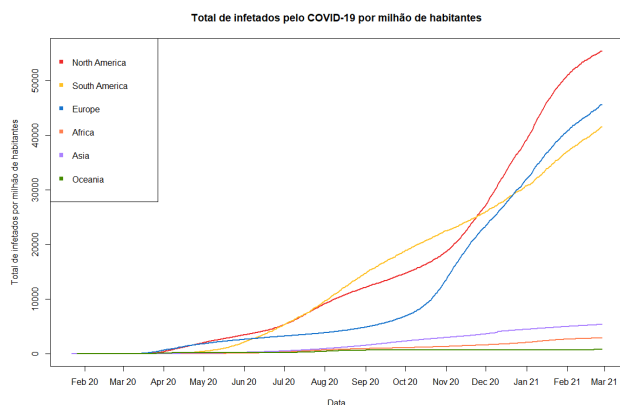


Figura 2. Gráfico de linhas representativo do total de infetados pelo COVID-19 por milhão de habitantes, por continente, ao longo do período de tempo estabelecido

Neste gráfico de linhas da Fig.2 conclui-se que a América do Norte é o continente que possui maior número total de infetados 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

O diagrama de caixa presente na Fig.3, 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.

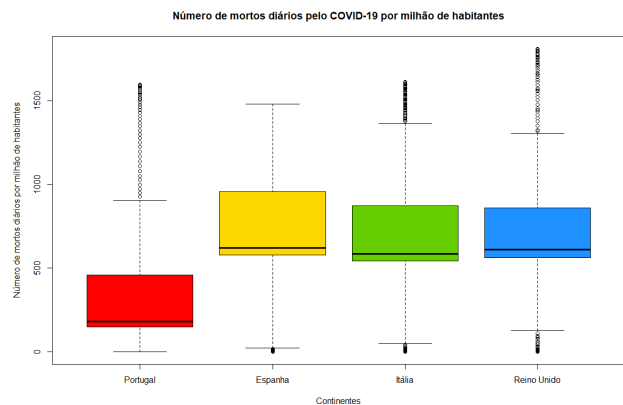


Figura 3. Diagrama de Caixa representativo do número de mortos diários pelo COVID-19 por milhão de habitantes, em Portugal, Espanha, Itália e Reino Unido

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

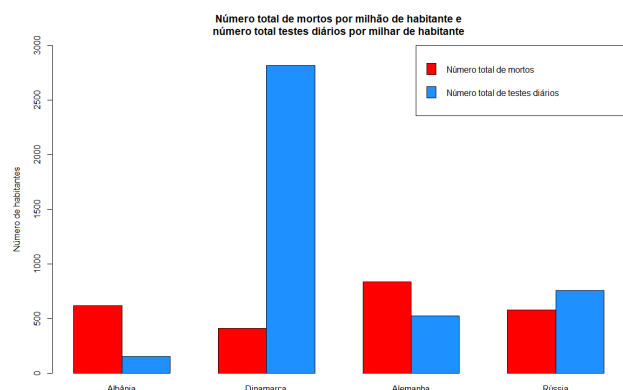


Figura 4. Gráfico de Barras representativo do número total de mortos pelo COVID-19 por milhão de habitantes, e número de testes diários por milhar de habitantes, na Albânia, Dinamarca, Alemanha e Rússia

Como se observa no gráfico de barras da Fig.4, 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 2.800 pessoas testadas por cada milhar de habitante. Contrastando, a Albânia possui o menor número de testes, tendo atingido pouco mais de 150 pessoas testadas, por milhar de habitantes, num único dia. Relativamente à Alemanha e Rússia, estes encontram-se entre os 500 e 1.000 testes diários, nos seus melhores dias de testagem.

A nível do número total de mortos, a Dinamarca volta a destacar-se como o país com o número mais reduzido, tendo aproximadamente um total de 400 mortos. O número mais notório encontra-se na Alemanha, 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

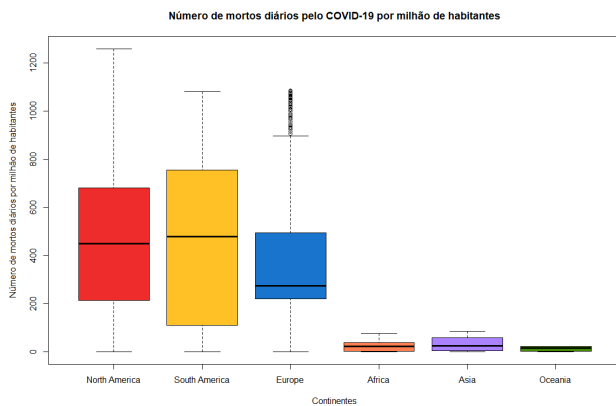


Figura 5. Diagrama de Caixa representativo do número de mortos diários pelo COVID-19 por milhão de habitantes, por continente

O diagrama de caixa representado na Fig.5 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.

$$x \in [Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR] \quad (1)$$

Por último, a Europa, mesmo depois de serem removidos alguns *outliers* pelo critério x é *outliers* se e só se satisfizer o critério indicado na equação (1), é o continente que ainda apresenta bastantes valores discrepantes, estando estes acima do limite máximo.

V. ANÁLISE INFERENCIAL

A análise inferencial incidiu na geração de amostras pseudoaleatórias, com diferentes *seeds* indicadas em cada alínea, para análise e verificação de diferenças significativas da taxa de transmissibilidade em Portugal e no Reino Unido, no número de mortes diárias em quatro países europeus e no número médio diário de mortes por continente. Estes dados são referentes ao período de 01 de abril de 2020 a 27 de fevereiro de 2021.

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

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

O teste de hipóteses segue a definição em (2). Como os dados são relativos a 30 dias, e utilizada uma *seed* de 118, 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), Lilliefors (p-value = 0.1533) e, também, pelo método gráfico (representado na Fig.6), 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.

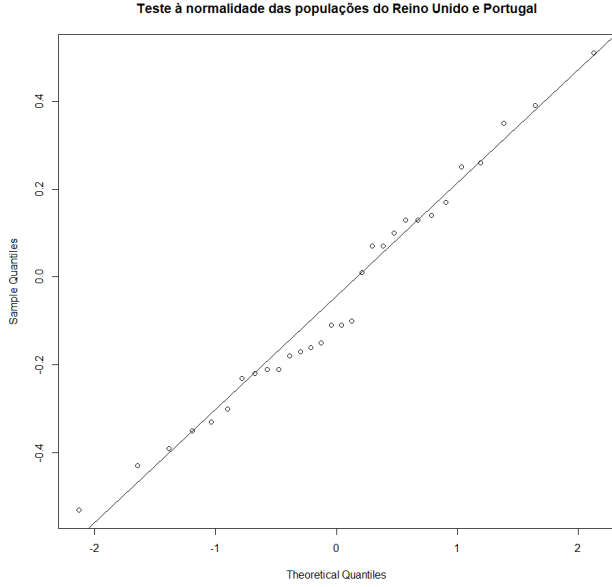


Figura 6. Diagrama de Dispersão representativo do teste à normalidade das populações do Reino Unido e Portugal

B. Verificação da existência de diferenças significativas no número de mortes diárias, por milhão de habitantes, em Espanha, França, Portugal e Itália

$$\begin{aligned} H_0 : No.Deaths_{Spain} &= No.Deaths_{France} = \\ &= No.Deaths_{Portugal} = No.Deaths_{Italy} \\ H_1 : No.Deaths_{Spain} &\neq No.Deaths_{France} \neq \\ &\neq No.Deaths_{Portugal} \neq No.Deaths_{Italy} \end{aligned} \quad (3)$$

O teste de hipóteses segue a definição em (3). Nesta análise foram utilizados 15 dias aleatórios e uma *seed* de 115. Procedeu-se à realização de um teste de Friedman, uma vez que dispomos de $n \times k$ observações (sendo n os dias e k os países), com o objetivo de perceber se as amostras apresentam diferenças significativas. A escolha deste teste deve-se a possuírmos quatro amostras emparelhadas, uma vez que os dados dos diversos países têm os dias em comum.

O p-value obtido foi de 0.06018, que, por ser superior a 0.05, permite-nos aceitar a hipótese nula e concluir que não há diferenças significativas no número de mortes diárias por milhão de habitantes nestes quatro países.

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

$$\begin{aligned} H_0 : \mu_{Asia} &= \mu_{Africa} = \mu_{Europe} = \mu_{NorthAmerica} = \\ &= \mu_{SouthAmerica} \\ H_1 : \mu_i &\neq \mu_j, i, j \in [Asia, Africa, Europe, \\ &NorthAmerica, SouthAmerica] \end{aligned} \quad (4)$$

$$\begin{aligned} Asia_{p-value} &= 0.07204, \\ Africa_{p-value} &= 0.001552, \\ Europe_{p-value} &= 0.001734, \\ NorthAmerica_{p-value} &= 0.08603, \\ SouthAmerica_{p-value} &= 2.932 \times 10^{-8} \end{aligned} \quad (5)$$

O teste de hipóteses segue a definição em (4). Nesta análise foram utilizadas *seeds* de 100 a 104 referentes aos quatro continentes. 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 (5) 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.

$$p - value = 1.437 \times 10^{-11} \quad (6)$$

Após estas confirmações realizou-se um teste Levene para verificar a homogeneidade dos dados, sendo que o valor obtido do p-value (6) permite concluir que as variâncias dos dados são significativamente diferentes.

$$p - value < 2 \times 10^{-16} \quad (7)$$

Com os resultados destes testes, realizou-se um teste ANOVA, ainda que o teste de Levene tenha indicado que as variâncias sejam diferentes, sendo que se obteve um p-value (7) reduzido e inferior a 0.05.

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.

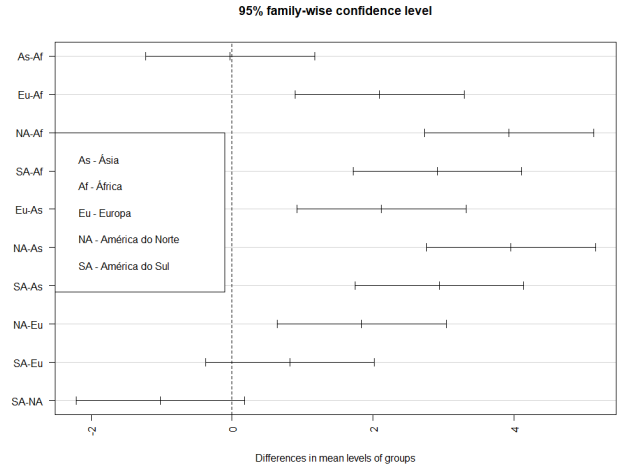


Figura 7. Gráfico representativo dos Intervalos de Confiança para a diferença das médias dos dados, entre os continentes

Após a realização e conclusões retiradas com o teste ANOVA, realizou-se um teste post-hoc [6] com o objetivo de perceber mais sobre as diferenças das médias dos dados entre os continentes. Assim sendo, realizou-se um teste Tukey's Honest Significant Difference [7], com o qual se obteve os dados apresentados na Fig.7.

Através do gráfico, é possível verificar que as médias da Ásia-África são estatisticamente semelhantes, uma vez que o intervalo de confiança contém o valor nulo. O mesmo acontece com as médias da América do Sul-Europa e América do Norte-América do Sul, no entanto carecem da mesma expressividade de igualdade relativamente à relação dos outros dois continentes, uma vez que, apesar do 0 se encontrar dentro dos intervalos, descai para os seus extremos.

Em relação às outras relações existentes, conclui-se que estas apresentam diferenças significativas nas médias, onde os continentes da Europa, América do Norte e América do Sul apresentam média superiores às de África e Ásia.

VI. ANÁLISE DE CORRELAÇÃO

Nesta análise pretende-se averiguar, em 2021, a existência de correlação entre dados referentes a países europeus com mais de 10 milhões de habitantes, nomeadamente o valor máximo da taxa de transmissibilidade, a densidade populacional, o total de mortos e a percentagem da população com 65 ou mais anos.

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 X1 a variável representativa do valor máximo da taxa diária de transmissibilidade de todos os países europeus com mais de 10 milhões de habitantes e Y1 da densidade populacional dos mesmos, foram verificados, em primeiro lugar, os seguintes pressupostos:

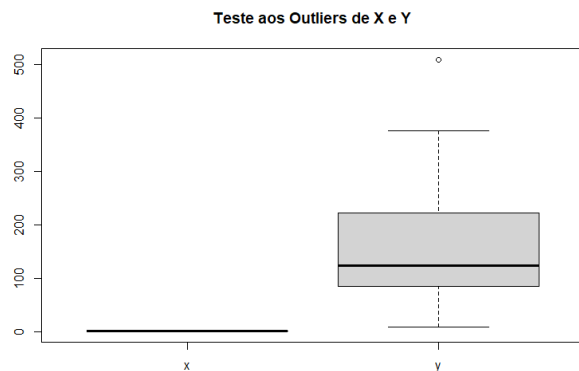


Figura 8. Diagrama de Caixa representativo das variáveis X1 e Y1

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 da Fig.8 que não o são.

2) *Deve existir uma relação linear entre as duas variáveis:* Para a confirmação desta condição realizou-se um diagrama de dispersão (Fig.9) 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.

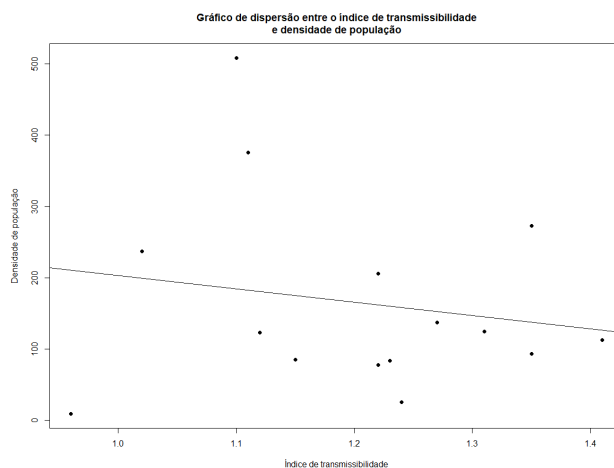


Figura 9. Diagrama de Dispersão representativo das variáveis X1 e Y1

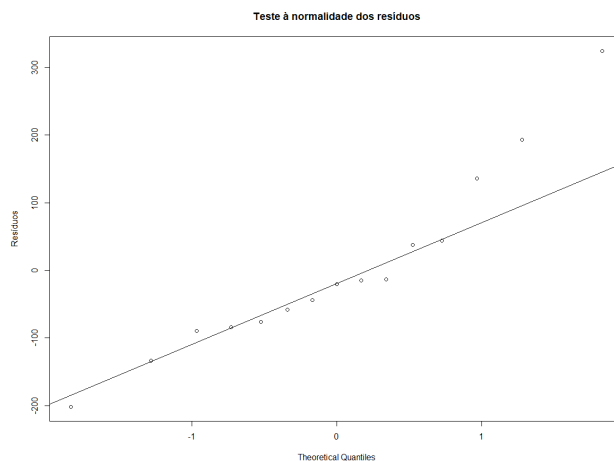


Figura 10. Gráfico de Dispersão Normal Q-Q representativo das variáveis X1 e Y1

3) *As variáveis devem ter, aproximadamente, uma distribuição normal:* Ao realizar um gráfico de dispersão Normal Q-Q (Fig.10), 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 da Fig.11 não se denota qualquer tendência no mesmo, 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 e $y = 15 < 30$), 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,

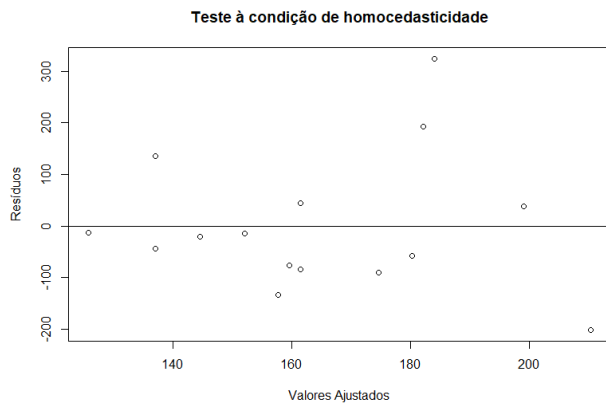


Figura 11. Diagrama de Dispersão ajustado representativo das variáveis X1 e Y1

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:

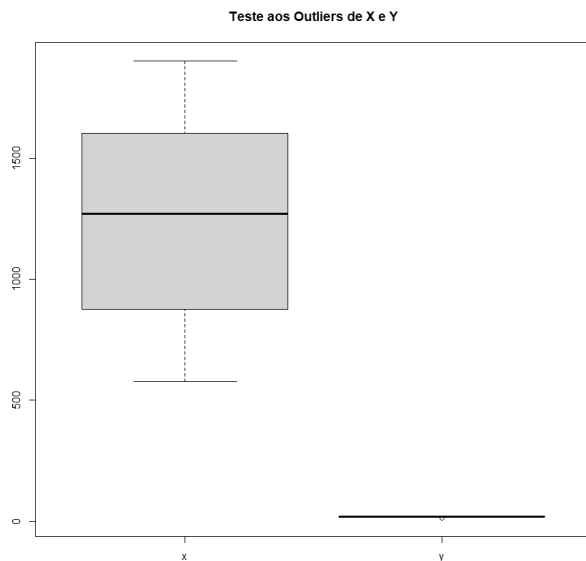


Figura 12. Diagrama de Caixa representativo das variáveis X2 e Y2

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 da Fig.12 que não o são.

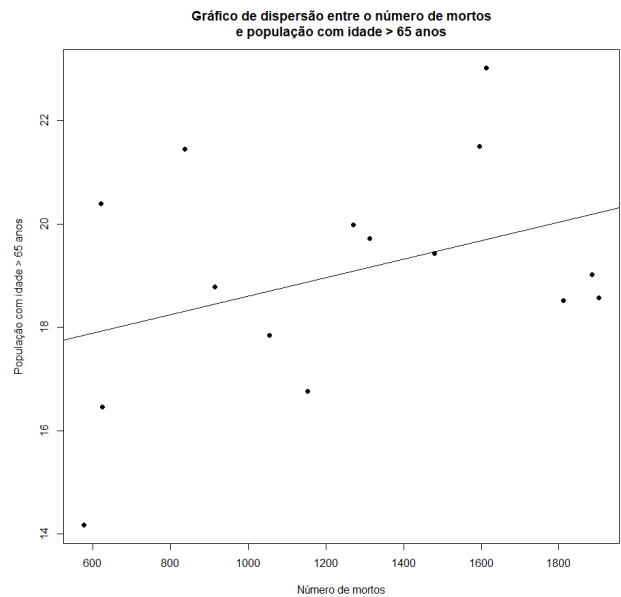


Figura 13. Diagrama de Dispersão representativo das variáveis X2 e Y2

2) *Deve existir uma relação linear entre as duas variáveis:* Através do diagrama de dispersão (Fig.13) 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.

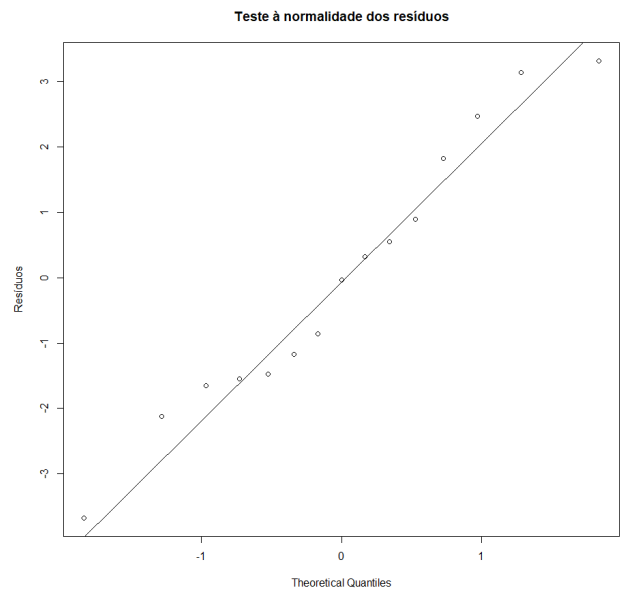


Figura 14. Gráfico de Dispersão Normal-QQ representativo das variáveis X2 e Y2

Ao realizar um gráfico de dispersão Normal Q-Q (Fig.14), verificam-se que os pontos estão ajustados à linha reta, ou seja, permite concluir que os resíduos seguem uma distribuição normal.

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.

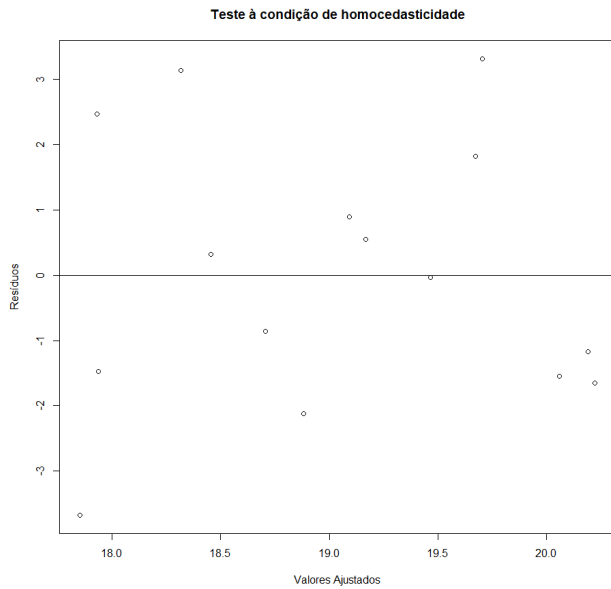


Figura 15. Diagrama de Dispersão ajustado representativo das variáveis X2 e Y2

4) As variâncias devem ser iguais (Homocedasticidade): Por fim, no gráfico da Fig.15 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

Na análise de regressão considerou-se o "Índice de rigor" médio mensal (Ir) em Portugal, a variável dependente e as variáveis independentes: média de mortes diárias por milhão de habitantes (Dm), média de casos diários por milhão de habitantes (Cm) e a média mensal da taxa de transmissibilidade (Rm). Os dados foram considerados no período entre 01 de abril de 2020 e 27 de fevereiro de 2021.

A. Construção do modelo de regressão linear múltipla

$$\begin{aligned}
 Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
 \beta_0 &= 81.827720 & \beta_{0p-value} &= 2 \times 10^{-16} \\
 \beta_1 &= 0.139378 & \beta_{1p-value} &= 0.141 \\
 \beta_2 &= 0.002183 & \beta_{2p-value} &= 0.259 \\
 \beta_3 &= -9.096406 & \beta_{3p-value} &= 8.55 \times 10^{-16}
 \end{aligned} \tag{8}$$

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 equação em (8) com os valores dos coeficientes obtidos.

Relativamente à significância dos coeficientes obtidos, a interseção e a taxa de transmissibilidade são aqueles que apresentam p-values extremamente 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.

$$R^2_{ajustado} = 0.2275 \tag{9}$$

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

O valor do R-quadrado ajustado (9) indica que apenas 22.75% dos valores do Índice de Rigor são explicados pelos outros parâmetros. Como o valor do p-value da estatística de teste (10) é bastante reduzido, o valor do R-quadrado ajustado é estatisticamente significativo.

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

1) Homocedasticidade:

$$p-value = 6.232 \times 10^{-13} \tag{11}$$

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 (11).

$$p-value = 2.2 \times 10^{-16} \tag{12}$$

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 (12).

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:

$$\begin{aligned}D_m &= 5.744092 \\C_m &= 5.665572 \\R_m &= 1.715799\end{aligned}\quad (13)$$

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 em (13), verificou-se que D_m e C_m são multicolineares, contrariamente a R_m .

C. Estimação do Ir para os valores $D_m = 10$, $C_m = 460$, $R_m = 1.1$

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

VIII. CONCLUSÕES

Através da análise exploratória constata-se que a América do Norte e a Europa são os continentes que mais se evidenciam, quer no número de infetados, quer no número de mortes, registando os valores mais elevados. Em contrapartida, a Oceânia demonstra ser o continente menos afetado pelo COVID-19, assinalando os menores valores.

Da inferência estatística conclui-se que a média da taxa de transmissibilidade do Reino Unido é semelhante à média registada em Portugal. De igual modo, também se verificou não existirem diferenças significativas no número de mortes diárias, por milhão de habitantes, em Espanha, França, Portugal e Itália. Por fim, constatou-se a existência de diferenças significativas no número médio de mortos diários, por milhão de habitantes, entre os vários continentes, sendo a Ásia e África aqueles com menores médias, excluindo a Oceânia, uma vez que registou os números mais baixos comparativamente ao resto do mundo.

Relativamente à análise de correlação, observou-se que ambas as análises estão fracamente correlacionadas, obtendo valores de relação próximos do 0 e de p-value não significativos. Denotou-se também que ambas falharam no mesmo pressuposto, sendo este o critério da existência de uma relação linear entre as duas variáveis, o que previu a falta de correlação entre os dados de cada análise.

Finalmente, na análise de regressão demonstrou-se que a relação linear múltipla não é muito forte, sendo que apenas a variável independente R_m possui uma relação linear com a variável dependente (Ir), contrariamente às restantes. Este fator foi comprovado quer na construção do modelo de regressão linear múltipla, quer na verificação da condição de multicolinearidade. Consequentemente, existe uma falta de precisão das previsões de valores do Ir através dos valores das variáveis independentes.

REFERÊNCIAS

- [1] Ritchie, H. (2021, 3 de abril). *Coronavirus Source Data*. Our World in Data. <https://ourworldindata.org/coronavirus-source-data>
- [2] Our World in Data (2021, 3 de abril). [Ficheiro Csv]
- [3] Gross, J. & Ligges, U. (2015, 30 de julho). *nortest: Tests for Normality*. R Project. <https://cran.r-project.org/web/packages/nortest/index.html>
- [4] Fox, J. & Outros (2020, 29 de setembro). *car: Companion to Applied Regression*. R Project. <https://cran.r-project.org/web/packages/car/index.html>
- [5] Department of Statistics (2021, 3 de abril). *Colors in R*. Columbia University. http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf?utm_source=twitterfeed&utm_medium=twitter
- [6] Autor desconhecido (2020, 17 de agosto). *Post Hoc Tests*. LibreTexts Statistics. [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_An_Introduction_to_Psychological_Statistics_\(Foster_et_al.\)/11%3A_Analysis_of_Variance/11.08%3A_Post_Hoc_Tests](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_An_Introduction_to_Psychological_Statistics_(Foster_et_al.)/11%3A_Analysis_of_Variance/11.08%3A_Post_Hoc_Tests)
- [7] Frost, J. (2021, 22 de abril). *Using Post Hoc Tests with ANOVA*. Statistics By Jim. <https://statisticsbyjim.com/anova/post-hoc-tests-anova/>