SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Design Thinking Project (DDW)
Task 1

SC06 Group 09

07 November 2021

GROUP MEMBERS

Chua Zhuo Xuan - 1005304

Foo Chuan Shao - 1005549

Hoo Jun Jie - 1005017

Darius Lum - 1005451

Vernon Toh - 1005143

Introduction

The whole world grapples with the disruption that the appearance of the SARS-Cov-2 virus has brought us and we continue to feel this pandemic's effects strongly in its exceptional death toll. In this study, we will build a Multiple Linear Regression (MLR) model that predicts the **absolute number of deaths** due to COVID-19.

Analyzing potential predictor variables

To start , we determine the useful potential predictors to include in our model by conducting online research.

- **Number of cases**
  - The first intuitive and straightforward predictor is the number of cases. More cases would indicate more deaths. Therefore, one of the key predictor would be number of cases.
- **Health of the people & age**
  - Based on a study in Rio de Janerio, it is suggested that an individual's immunity affects the susceptibility to suffer more severe forms of infection. It was also found that older people have lower immunity, thus, mortality and lethality increase with age. [1]
- **Skin colour & gender**
  - Based on same study, skin colour and gender have also been found to be associated with the probability of death. [1]
- **Vaccination Status**
  - Based on a study by the Centers of Disease Control and Prevention (CDC), an individual that has been vaccinated has lower mortality rate. [2]
- **Population, Income, Occupation, & Transport**
  - Based on a study in the United States, it is suggested that the population, income, occupation, and transport have a high association with the spread of the virus. As the virus spreads, there would be more cases, and consequently more deaths. [3]
- **Healthcare Systems & Policies**
  - Based on a study in the European Union, the COVID-19 mortality rate depends on the healthcare system a country has, as well as the policies that they possess. [4]

Finding and reading data set

On the internet, we found our comprehensive dataset from https://ourworldindata.org/covid-deaths. [5] In our excel file, this sheet is called “Original”. To begin, we can quickly glance and look through our dataset.

Understanding our data set

Preliminary cleaning up of data set

Analyzing the dataset, we can begin screening the data that we do not want such as “World”, “Asia”, etc, under “location”. Next, we can also remove non-numerical or irrelevant data such as “iso_code”, “continent”, and “date”. We would also like to remove non-smoothed data as through online research, we have found that smoothed data has been modified to remove noise which helps us better identify patterns. [6] Lastly, we can remove relative data as we are going to work with absolute data. This new dataset has been created and is titled “preliminary clean”.

Visualizing relationship

To visualize our relationship between deaths and other variables, we can do a pair plot between our potential dependent variables "new_death_smoothed" and "total_deaths" and the other variables. The full visualization of each relationship can be found in the appendix (Fig 1) below.

Extracting predictor variables & cleaning up data set

Looking through the data set, we can locate some of the dependent and potential predictor variables as forementioned. Before analysis, we remove NAN data so that there are no lapses in the data.

- **new_deaths_smoothed (dependent 1)**
- **total_deaths (dependent 2)**
- **new_cases_smoothed / total_cases**
  - This is equivalent to number of cases and the spread of the virus.
- **people_fully_vaccinated_percent**
  - This is equivalent to vaccination status. This is modified to reflect the percentage of vaccinated individuals in the country. We assume that each person in the country has equal probability to be vaccinated. This was calculated using "people_fully_vaccinated" and "population".
- **life_expectancy**
  - This is used as a health status indicator for the country.
- **median_age**
  - This is used as an indicator of the age of the people of the country.
- **hospital_beds_per_thousand**
  - This is an indicator of the capabilities of the healthcare system that a country has.

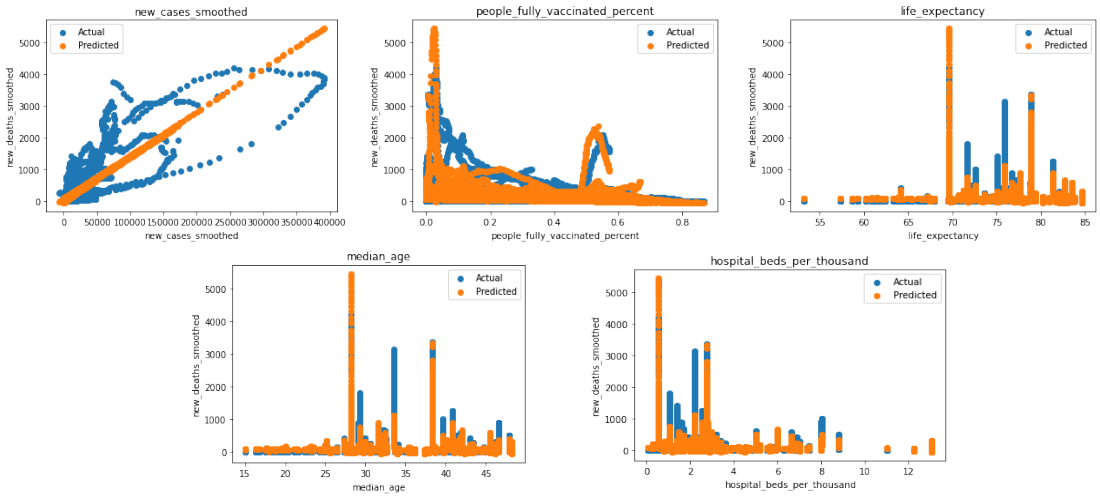
MLR Iteration 1:

A naive first iteration used all the data points. In our first iteration, we begin with "new_cases_smoothed" and "new_deaths_smoothed" to analyze the relationship between the daily deaths and daily cases. In our dataset, we extracted our data and applied z-normalization before applying the regression toolpak. The edited data set and analysis can be found in sheets “V1 Raw” and “V1Z” respectively. Shown below are the coefficients of betas and the metrics used.

Constant	New Cases Smoothed	People Fully Vaccinated %	Life Expectancy	Median age	Hospital Bed Per Thousand
114.548744	272.401057	-32.140266	-7.075465	8.728219	-6.920659
R^2	R^2 Adjusted	RMSE	Mean	Relative RMSE	
0.684132	0.684055	187.851398	114.548744	1.639925	

Visualizing data & relationship (Iteration 1)

We can visualize our relationship by plotting each variable and seeing how the predicted data fits compared to the actual data. Full sized relationship graphs can be found in the appendix (Fig 2)



Discussion of results from Iteration 1

The actual values and predicted values are considerably similar with a high percentage of predicted values matching the actual values. However, from the plots, there are many data points that are concentrated in certain regions, indicating a possible misbalance in the weightage of each country's contribution.

Numerically, improvements can be made as the R^2 adjusted values are only 0.68. Furthermore, the relative RMSE is considerably large. This can be explained as when describing the data, the standard deviation is significantly large as compared to the mean.

Another numerical value to note is β. A strongly correlated variable as predicted is "new_cases_smoothed" which has a significant impact on the model generated. Furthermore, by looking at the other coefficients, we can primarily deduce that the variables support the hypothesis forementioned.

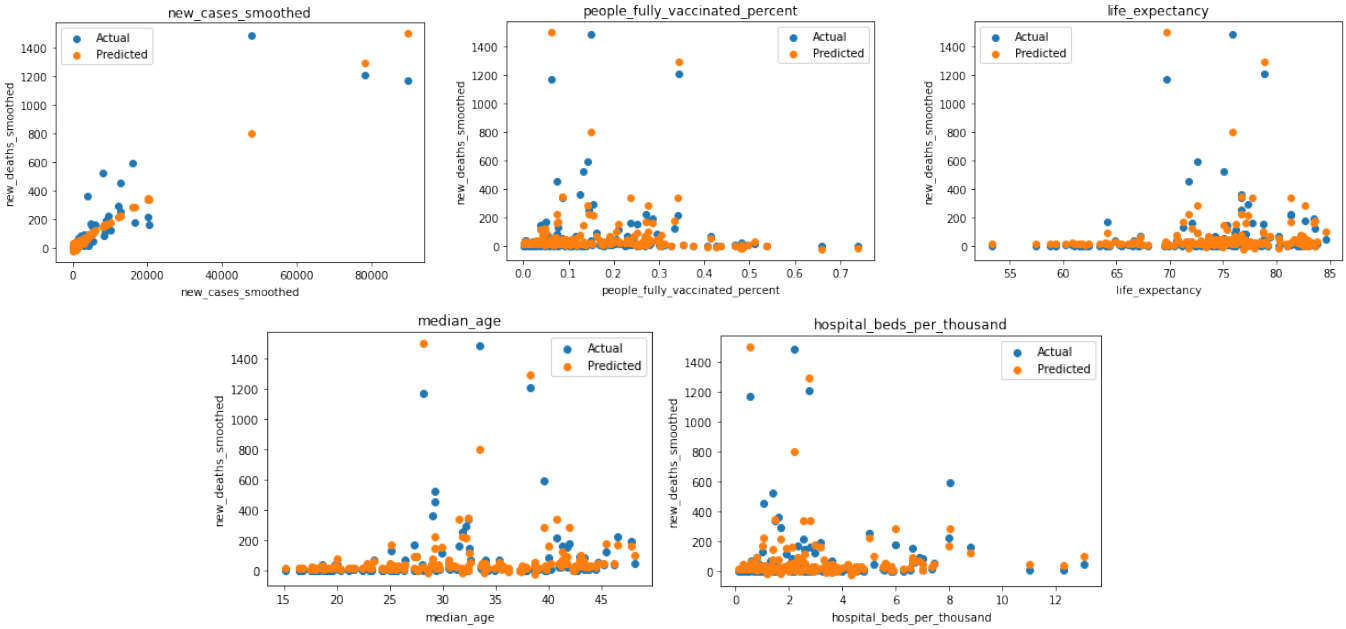
MLR Iteration 2:

From the first iteration, we identified a few issues. Some countries may have more data than others, which will skew the data and cause analysis to be imbalanced. Thus, we take the mean data for each country, and calculate the expected average death per day. Shown below are the coefficients of betas and the metrics used. The edited data set and analysis can be found in sheets “V2 Raw” and “V2Z” respectively.

Constant	New Cases Smoothed	People Fully Vaccinated %	Life Expectancy	Median age	Hospital Bed Per Thousand
68.997980	175.611552	-11.583231	-0.599621	6.507458	0.414734
R^2	R^2 Adjusted	RMSE	Mean	Relative RMSE	
0.817826	0.811794	84.694667	68.998004	1.227494	

Visualizing data & relationship (Iteration 2)

We can visualize our relationship by plotting each variable and seeing how the predicted data fits compared to the actual data. Full sized relationship graphs can be found in the appendix (Fig 3)



Discussion of results from Iteration 2

Compared to Iteration 1, visually, there are no obvious changes to the relationship and most predicted data still aligns with the actual data. However, the weights of the data are more evenly spread out which is supported by the lower standard deviation value.

Analyzing it numerically, we can see that the R^2 adjusted has increased to 0.81, indicating a better model with a balanced weight from each country. The relative RMSE also decreased significantly, indicating a better prediction.

However, a thing to note for β is that the coefficient for "life_expectancy" has a lesser impact now. This is unexpected as health has been cited in multiple studies to be an important factor in determining deaths. [7] [8] One possible explanation for this is that "life_expectancy" is also an indicator of age, which is possibly correlated with number of deaths. Furthermore, "life_expectancy" has been affected by the pandemic. [9] Thus, this indicator might not be suitable in indicating health.

Another thing to note for β is that "hospital_beds_per_thousand" has changed from negative correlation to positive. This is unexpected as better healthcare systems should yield lower deaths. One possible explanation would be that a country may increase their hospital beds to respond to higher demand. However, as the coefficient is very low, we used this variable again to verify.

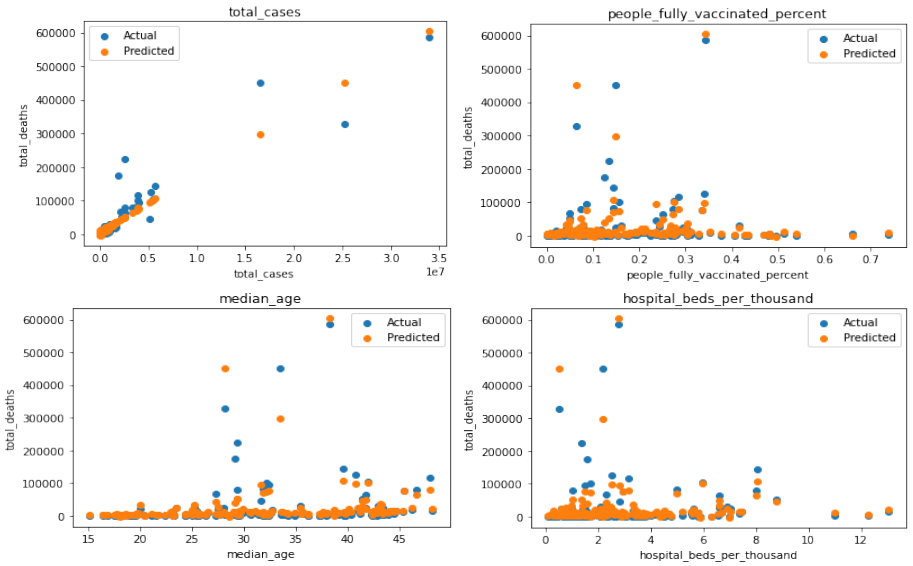
MLR Iteration 3:

We explored the use of tallied data instead of daily data. We also excluded "life_expectancy" as explained above. In Iteration 3, we tested the use of "hospital_beds_per_thousand" to verify if capability of healthcare system determines number of COVID-19 deaths. Shown below are the coefficients of betas and the metrics used. The edited data set and analysis can be found in sheets “V3 Raw” and “V3Z” respectively.

Constant	Total Cases	People Fully Vaccinated %	Median Age	Hospital Bed Per Thousand
23909.780857	65139.113532	-2657.834601	4217.3519380	-1840.363371
R^2	R^2Adjusted	RMSE	Mean	Relative RMSE
0.869654	0.866224	25781.290932	23909.79223	1.078273

Visualizing data & relationship (Iteration 3)

We can visualize our relationship by plotting each variable and seeing how the predicted data fits compared to the actual data. Full sized relationship graphs can be found in the appendix (Fig 4)



Discussion of results from iteration 3

Compared to Iteration 2, visually, there are no obvious changes to the relationship and most predicted data still aligns with the actual data. However, the spread is still present as there is still significantly higher deaths. This is supported by the high standard deviation value.

Analyzing it numerically, we can see that the R^2adjusted has increased to 0.87, indicating that the model is a better fit using tallied data. Additionally, the relative RMSE decreased to 1.06, indicating a more accurate prediction.

For the coefficients of β, without "life_expectancy" in consideration, all variables support the hypothesis.

Conclusion

From all 3 iterations, we conclude that the number of COVID-19 related deaths is highly correlated with the number of cases. This is intuitive more cases mean a greater spread in the virus and this would lead to implications such as more variants that can be deadlier, and also a higher amount of people susceptible to COVID-19 deaths.

We also conclude that the percentage of people vaccinated in a country and average age of the population also affects their mortality. [1] [2] This is shown in all 3 models as "people_fully_vaccinated_percent" always has a negative correlation and "median_age" always has a positive correlation. From these correlations, we can deduce that higher vaccination count yields higher immunity for the people, suggesting herd immunity. [10] As the "median_age" increases, we also infer that there would be more seniors having higher mortality rate.

However, two fluctuating variables "life_expectancy" and "hospital_beds_per_thousand" has considerably lower coefficients and changes their relationship with the number of deaths. Thus, the suitability of these 2 variables is inconclusive and requires more testing to verify their relationship.

Through 3 iterations, we tested various methods of modelling the number of deaths. Overall, with balanced input from each country and tallied data, we can obtain better predictions. This effect is explained by a surplus of data from wealthier countries, thus skewing the dataset. Daily cases can fluctuate and may not be a good indicator for the number of deaths.

Comparing Excel and Python Code

Comparing the Excel data and Python code, even though the computed values are slightly different, they are not significantly differently and have the same conclusion. One possible reason could be the iterations or rounding off difference that can happen in the system.

Overall, the conclusion and discussion on data for both the Excel and Python code applies to both dataset and both information provided are the same.

Excel Interaction

In the excel sheets “V1Z”, “V2Z”, “V3Z”, there are interactive cells that allows for inputs to calculate the predicted deaths in a country for each iteration. Under variable, we once again apply the z-normalization using our data set mean and standard deviation to calculate the normalized input.

Here, we are assuming that the new input data would not skew the mean or standard deviation greatly such that the predicted values would turn out to be inaccurate.

There is a cell called “True Predicted Total Deaths” which uses the “if” function to ensure that the deaths do not exceed the number of cases, and that it is not less than 0.

References

[1] Cini Oliveira, M., de Araujo Eleuterio, T., de Andrade Corrêa, A.B. et al. Factors associated with death in confirmed cases of COVID-19 in the state of Rio de Janeiro. BMC Infect Dis 21, 687 (2021). https://doi.org/10.1186/s12879-021-06384-1

[2] Xu S, Huang R, Sy LS, et al. COVID-19 Vaccination and Non–COVID-19 Mortality Risk — Seven Integrated Health Care Organizations, United States, December 14, 2020–July 31, 2021. MMWR Morb Mortal Wkly Rep 2021;70:1520–1524. DOI: http://dx.doi.org/10.15585/mmwr.mm7043e2

[3] Cifuentes-Faura, J. (2021). Factors influencing the COVID-19 mortality rate in the European Union: importance of medical professionals. Public Health, 200, 1–3. https://doi.org/https://doi.org/10.1016/j.puhe.2021.09.003

[4] Pasha, D. F., Lundeen, A., Yeasmin, D., & Pasha, M. F. K. (2021). An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science. Case Studies in Chemical and Environmental Engineering, 3, 100067. https://doi.org/https://doi.org/10.1016/j.cscee.2020.100067

[5] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus'

[6] FDhir, Rajeev. "How Data Smoothing Works". Investopedia, 2021, https://www.investopedia.com/terms/d/data-smoothing.asp.

[7] Williamson, E.J., Walker, A.J., Bhaskaran, K. et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 584, 430–436 (2020). https://doi.org/10.1038/s41586-020-2521-4

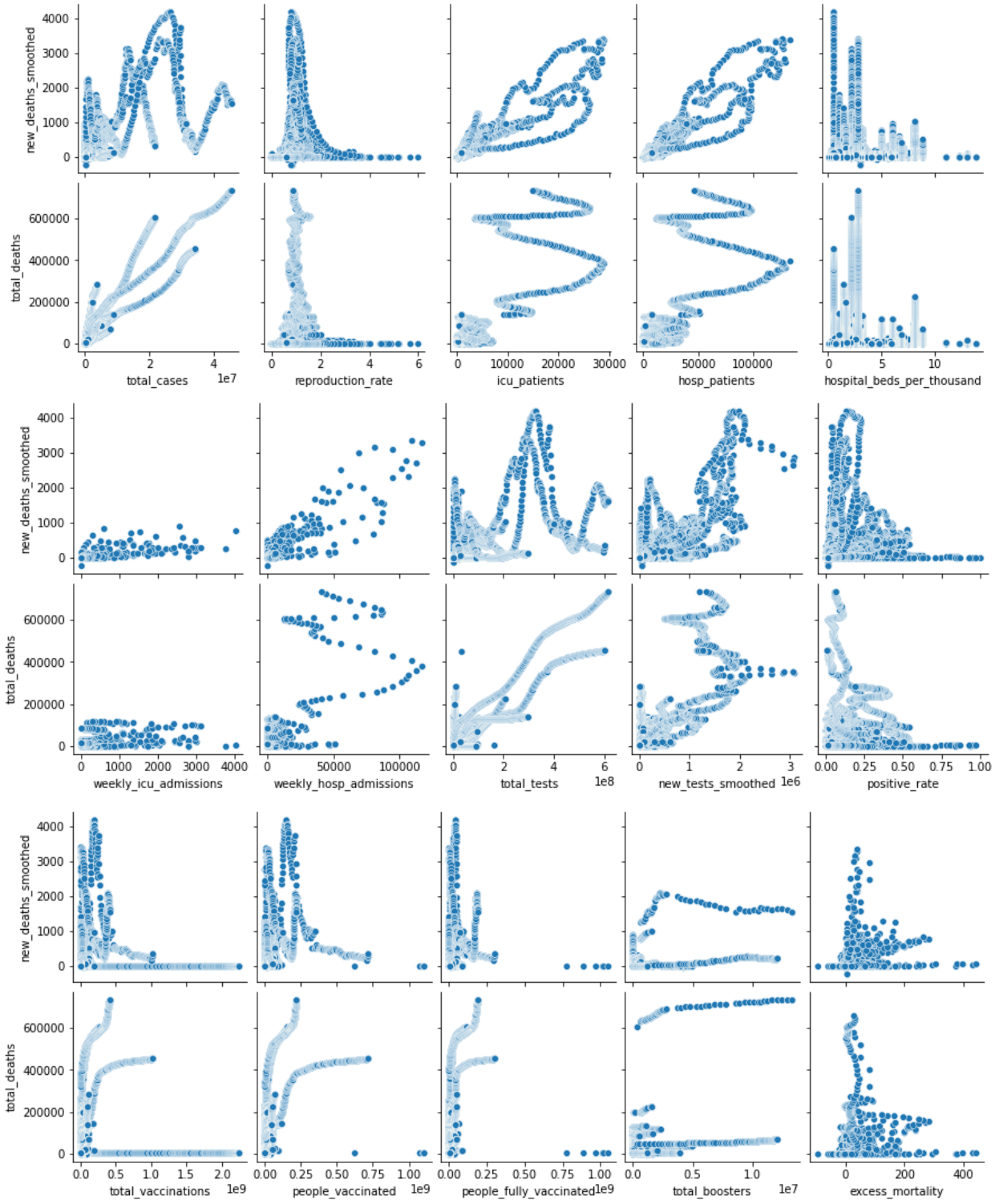
[8] Upadhyaya, A., Koirala, S., Ressler, R. and Upadhyaya, K. (2020), "Factors affecting COVID-19 mortality: an exploratory study", Journal of Health Research, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JHR-09-2020-0448

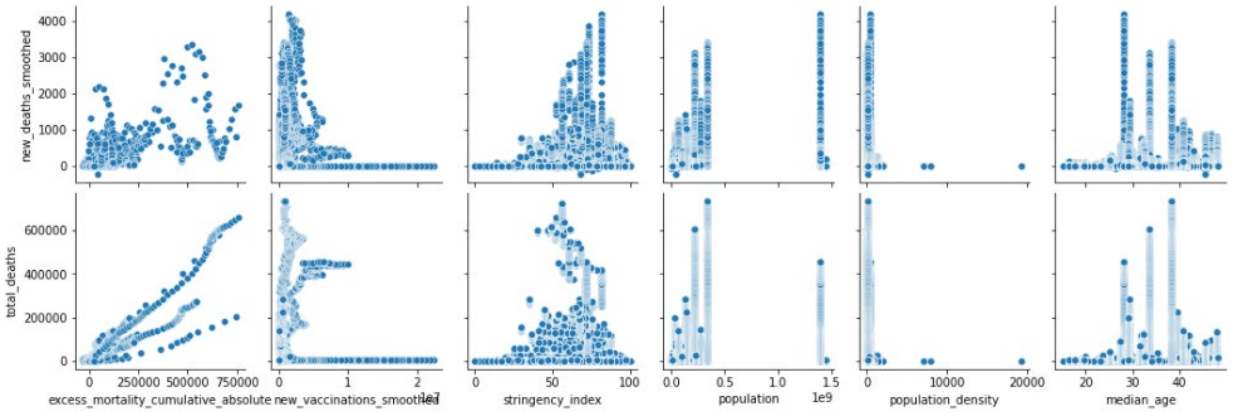
[9] José Manuel Aburto, Jonas Schöley, Ilya Kashnitsky, Luyin Zhang, Charles Rahal, Trifon I Missov, Melinda C Mills, Jennifer B Dowd, Ridhi Kashyap, Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries, International Journal of Epidemiology, 2021;, dyab207, https://doi.org/10.1093/ije/dyab207

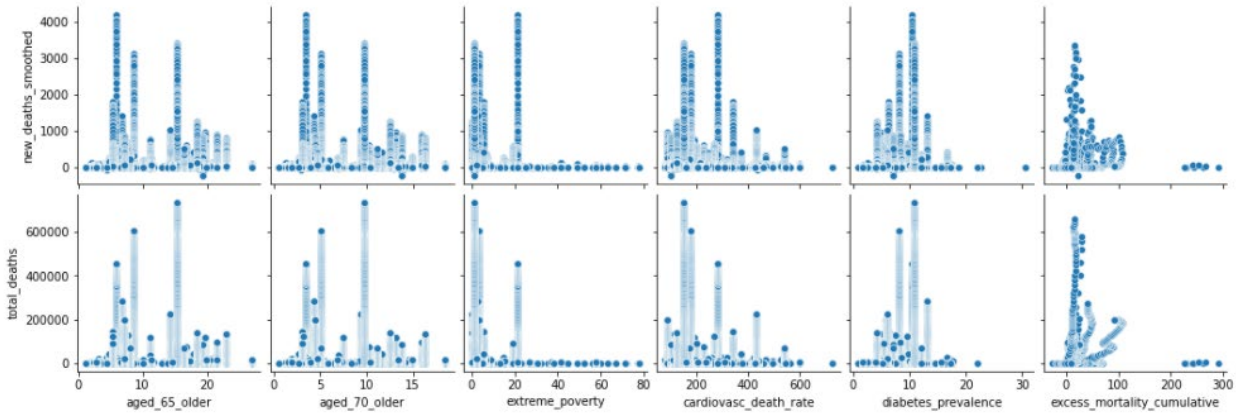
[10] "Herd Immunity And COVID-19 (Coronavirus): What You Need To Know". Mayo Clinic, 2021, https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/herd-immunity-and-coronavirus/art-20486808.

Appendix

Figure 1




new_deaths_smoothed
total_deaths
excess_mortality_cumulative_absolute
new_vaccinations_smoothed
1e7
stringency_index
population
1e9
population_density
median_age


new_deaths_smoothed
total_deaths
aged_65_older
aged_70_older
extreme_poverty
cardiovasc_death_rate
diabetes_prevalence
excess_mortality_cumulative

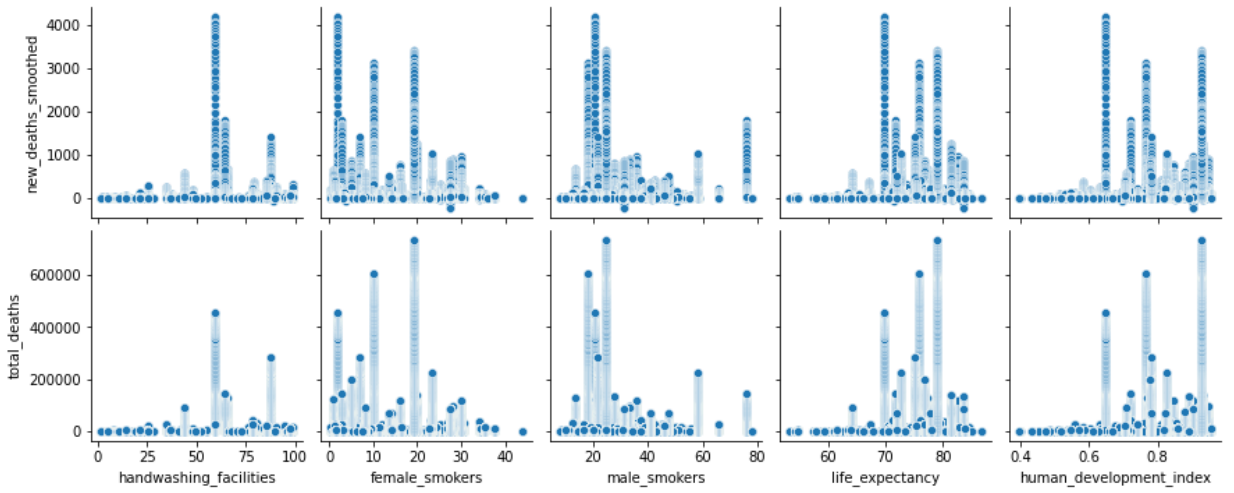

new_deaths_smoothed
total_deaths
handwashing_facilities
female_smokers
male_smokers
life_expectancy
human_development_index

Figure 2

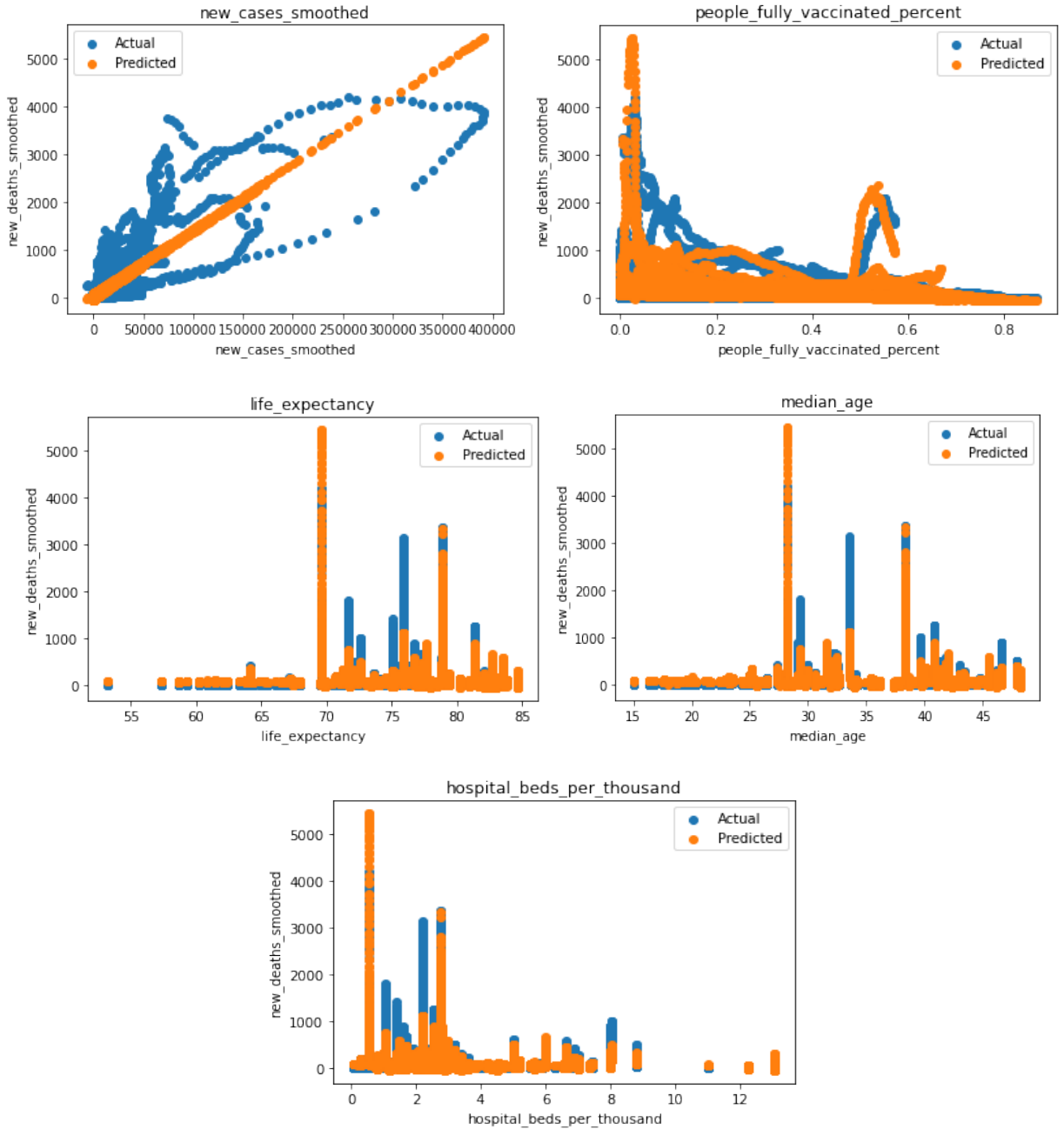

new_cases_smoothed
Actual
Predicted
new_deaths_smoothed
new_cases_smoothed
people_fully_vaccinated_percent
people_fully_vaccinated_percent
life_expectancy
life_expectancy
median_age
median_age
hospital_beds_per_thousand
hospital_beds_per_thousand

Figure 3

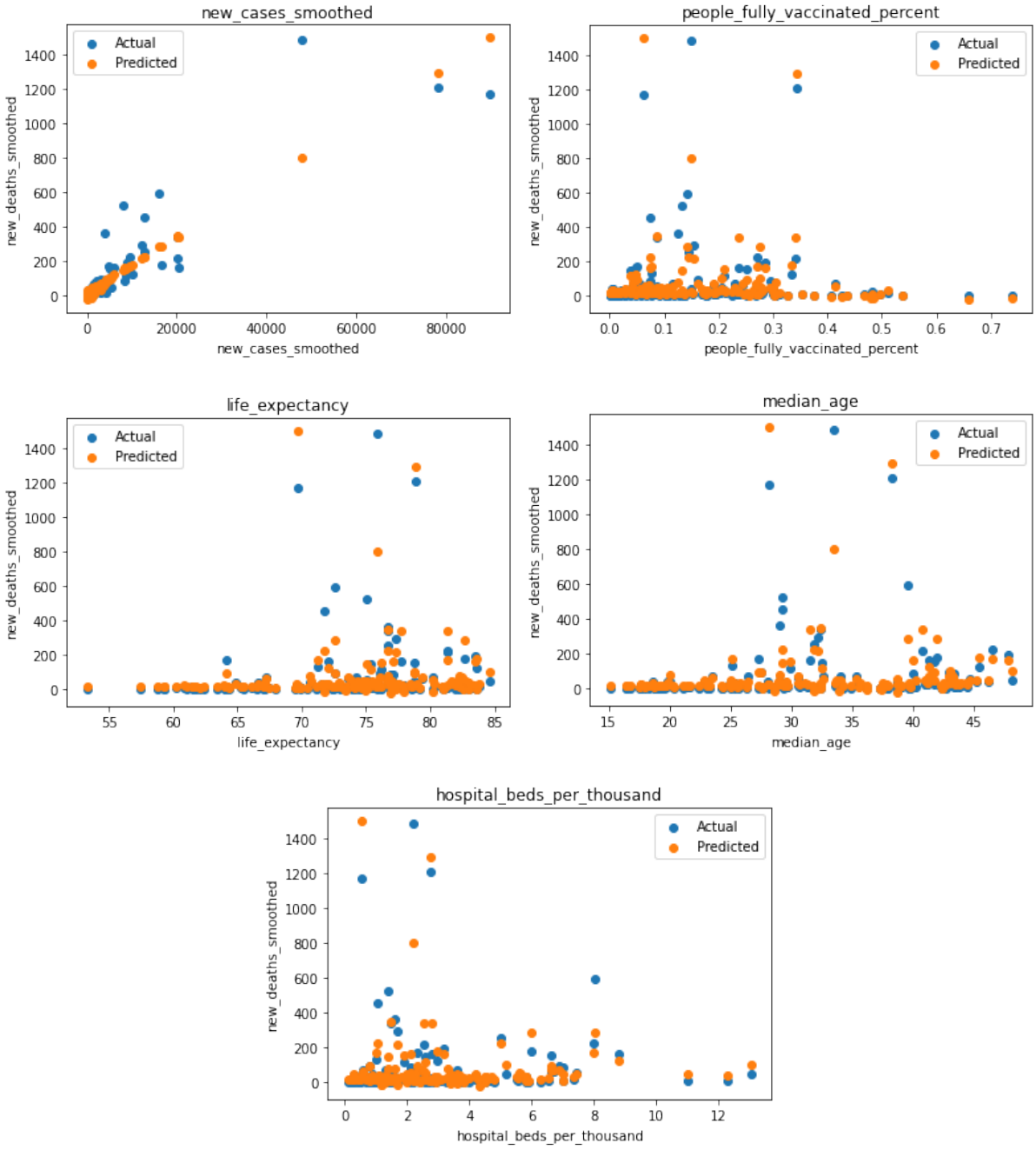
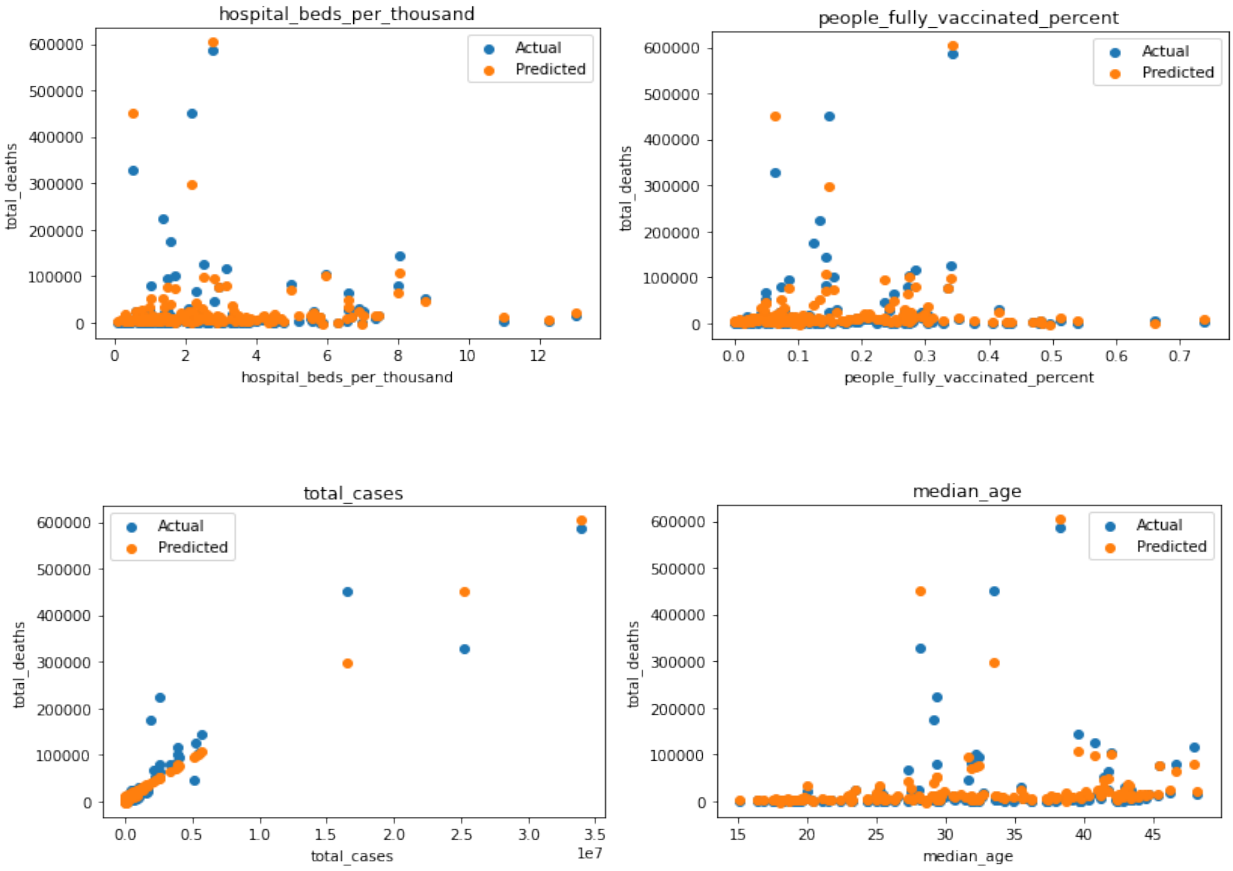

new_cases_smoothed
Actual
Predicted
new_deaths_smoothed
new_cases_smoothed
people_fully_vaccinated_percent
Actual
Predicted
new_deaths_smoothed
people_fully_vaccinated_percent
life_expectancy
Actual
Predicted
new_deaths_smoothed
life_expectancy
median_age
Actual
Predicted
new_deaths_smoothed
median_age
hospital_beds_per_thousand
Actual
Predicted
new_deaths_smoothed
hospital_beds_per_thousand

Figure 4


hospital_beds_per_thousand
Actual
Predicted
total_deaths
hospital_beds_per_thousand
people_fully_vaccinated_percent
Actual
Predicted
total_deaths
people_fully_vaccinated_percent
total_cases
Actual
Predicted
total_deaths
total_cases
1e7
median_age
Actual
Predicted
total_deaths
median_age