Data Mining and Analysis of COVID-19 Mortality and Relationship between Country Attributes, Vaccination, and Excess Death

Avery Lawson
Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
avla4927@colorado.edu

Jay Bentley
Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
jabe1234@colorado.edu

Nick O'Connor
Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
nicholas@colorado.edu

Michael Thomas
Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
Michael.M.Thomas@Colorado.EDU

Introduction

In the early months of 2020, a novel pathogen began to spread which would eventually be called COVID-19 and caused a global pandemic at a scale unseen in recent history. Even for countries with developed healthcare infrastructure, COVID-19 brought society across the globe to a standstill. COVID-19 proved to have high mortality rates which only increased the rush to develop vaccinations to help support groups heavily impacted by COVID-19 including the young, the old, and those who were immunocompromised. Once vaccinations were developed and approved, they were rapidly administered and ushered in the beginning of the end of quarantining.

For the next several weeks we will be looking into vaccination rates and mortality rates across the world. We aim to understand what attributes might contribute to higher mortality rates in pandemics and looking at data from multiple countries will help us gain a broader understanding of how COVID-19 affected the world. Understanding what contributed to the impact of COVID-19 will allow us to better our understanding of how global pandemics impact lives and what steps we can take to lessen the impact in the future.

One main challenge with this project also lies in one of its strengths, gathering data from other countries. It should not be difficult to access this data as most of it is open to public access but the challenge comes from within the data itself. Different countries will use different methods for recording the data that we are looking for. In order to ensure the accuracy of our analysis and the conclusions we draw, we will have to clean the data. Once it is cleaned we will be able to use the data for analysis.

Related Work

Since the first identification of COVID-19 in 2019, a significant number of research and related work on the virus has been published globally. They cover a wide range of topics and related issues, reflecting the complexity of the pandemic. Some articles have looked at the data surrounding the rollout and use of COVID-19 vaccinations as a whole[1]. Other papers have analyzed COVID-19 cases and mortality in specific countries such as India[2]. There has been research done on the safety and efficacy of the COVID-19 vaccine[3]. Data mining has been used to research the public option of COVID-19 vaccine on social media to determine whether individuals support the use of the vaccine or not[4]. Closely related to our area of interest, there has been data mining on the mortality, immunity, and vaccine development of the COVID-19 vaccine[5]. We aim to push the literature further by looking at the intersection of COVID-19 morality, vaccination rates, country attributes such as current health expenditure, and excess death in the country due to COVID-19. We are interested in understanding the relationship between these different aspects of health in countries and which factors would be good areas for healthcare intervention.

Proposed Work

Datasets

The primary dataset we will be using is the COVID-19 Cases and Deaths hosted by the World Health Organization (WHO). This gives us insight into the rate of COVID-19 death and mortality in countries around the world. We will also be using a dataset with excess mortality of COVID-19 differentiated by country, age and sex hosted by the WHO[6]. This will allow us to compare death between countries and look at relationships between these attributes that they differentiate on. We plan to use a dataset with COVID-19 vaccination rates hosted by the WHO or the dataset supplied in Homework 2 of this course. This will allow us to analyze the relationship between deaths in countries and their rate of death among other factors. We plan to use datasets with Current Health Expenditures as a percent of GDP by country hosted by the WHO[7]. This will help us because we can look at the relationship between CHE, vaccination rates, and mortality from COVID in different countries. We may also use vaccination rates datasets for other diseases hosted by the WHO to see if COVID-19 vaccination rates were normal for these countries.

Data Cleaning

The first step in gaining the knowledge we're looking for is cleaning the data. In order to do this we will have to address any null values across all datasets, normalize all the data from different countries, and finally we need to use data integration to aggregate the data from all of the different sources/countries. Some countries with more decentralized healthcare infrastructure may need more effort in cleaning than other countries.

Regression Modeling

After cleaning all the data we will run regression models and from that identify what factors correlate with higher mortality rates. We will try linear regression on different attributes such as country, age, and gender to see if we can find any correlations between these and the morality of COVID-19. We can also try polynomial regression if these do not supply a satisfactory linear regression model.

Clustering Analysis

Finally we will run clustering experiments to identify what factors cleanly cluster. We hope to try implementing the k-means clustering algorithm as a simple first clustering algorithm to see if we can cluster based on certain attributes such as country, age, and gender. We also hope to try implementing the hierarchical clustering algorithm. This algorithm makes a tree-like structure of clusters to help see patterns and meaningful proximity of clusters. From this, we will propose reasons as to why these clusters are occurring for certain factors.

Evaluation

In order to evaluate our success, we are going to use several metrics to measure our models' ability to accurately predict relationships between data attributes. With regression modeling, we have multiple metrics we can use to evaluate our degree of success including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R and R^2.

Mean Absolute Error Formula

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|$$

Mean Squared Error Formula

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)$$

Root Mean Squared Error Formula

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$$

R^2 Formula

$$1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)^2}$$

R Formula

$$\sqrt{R}$$

n = sample size
Y = true value
$\widehat{Y}$ = predicted value

The MAE is very simple to understand as a measure of how well the model predictions match the reality of data we possess. The MAE is less sensitive to outliers so it is less likely to skew extremely but also punishes outliers less than other metrics. The MSE penalizes outliers much more than MAE because of the squaring of the value differences. This makes the MSE less interpretable by eye because it is in squared units. The RMSE tries to find a middle ground between the MAE and MSE as it gives an error metric in the same units as the dataset but also penalizes for larger errors by squaring the difference.

In order to evaluate our success with clustering data, we will be using the Silhouette Score and Davies-Bouldin Index.

Silhouette Score Formula

$$\frac{(b-a)}{max(a,b)}$$

Davies-Bouldin Index Formula

$$\frac{1}{k}\sum_{i=1}^{k} max(\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i,X_j)})$$

a = mean intra-cluster distance
b = mean nearest-cluster distance
k = number of clusters
ΔX_k = intracluster distance within cluster
$\delta(X_i, X_j)$ = intercluster distance between clusters

The Silhouette Score gives a 1 to -1 value making it very clear to understand how well separated the clusters appear. This metric is sensitive to noise and outliers and may not work well for clusters that are not round. Davies-Bouldin Index has similar traits but might offer a method of seeing how similar clusters are to each other.

We plan to use these metrics outlined to assess our success in creating strong regression models and clusterings of our datasets. Strong regression models will exhibit lower error values as their predictions will ideally match true data values. Clustering metrics will demonstrate large distances between cluster groups and strong clustering of data points around those cluster centers.

Milestones

Week 5: Project Proposal
Week 8: Prototype methods of data analyzing
Week 11: Checkpoint due date
Week 14: Draft of final submission
Week 16: Final submission due date

ACKNOWLEDGMENTS

We would like to acknowledge the time and effort of Professor Qin Lv and the TAs Mohsena Ashraf and Julia Romera.

REFERENCES

[1] Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. Nature Human Behaviour, 5(7), 947–953. https://doi.org/10.1038/s41562-021-01122-8

[2] Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. (2021, June 1). TUP Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9343916

[3] Yih, W. K., Daley, M. F., Duffy, J., Fireman, B., McClure, D., Nelson, J., Qian, L., Smith, N., Vazquez-Benitez, G., Weintraub, E., Williams, J. T., Xu, S., & Maro, J. C. (2022). A broad assessment of covid-19 vaccine safety using tree-based data-mining in the vaccine safety datalink. Vaccine, 41(3), 826–835. https://doi.org/10.1016/j.vaccine.2022.12.026

[4] Li, T., Zeng, Z., Sun, J., & Sun, S. (2022). Using data mining technology to analyse the spatiotemporal public opinion of COVID-19 vaccine on social media. The Electronic Library, 40(4), 435–452. https://doi.org/10.1108/el-03-2022-0062

[5] Yih, W. K., Daley, M. F., Duffy, J., Fireman, B., McClure, D., Nelson, J., Qian, L., Smith, N., Vazquez-Benitez, G., Weintraub, E., Williams, J. T., Xu, S., & Maro, J. C. (2022b). A broad assessment of covid-19 vaccine safety using tree-based data-mining in the vaccine safety datalink. Vaccine, 41(3), 826–835. https://doi.org/10.1016/j.vaccine.2022.12.026

[6] Global excess deaths associated with COVID-19 (modeled estimates) (2021) https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modelled-estimates

[7] Current health expenditure (CHE) as percentage of gross domestic product (GDP) (2023) https://www.who.int/data/gho/data/indicators/indicator-details/GHO/current-health-expenditure-(che)-as-percentage-of-gross-domestic-product-(gdp)-(-)