Advanced Statistics Spring 2022 Project Proposal

Title: A Statistical Analysis of Different COVID-19 Aspects in the U.S. and Forecasting the U.S. Mortality-rate Time-series

Rezaur Rashid

Introduction

Coronavirus Disease 2019 (COVID-19) has been one of the biggest catastrophic events since 1921 Influenza Flu and World War II. It was declared as a global pandemic in March 2020, affecting almost every country in the world. As of February 22, 2022, globally over 400 million people have been infected and almost 6 million people have died [1]. According to CDC, over 78 million people have been affected by COVID-19 in the U.S. and almost 1 million people have died [2].

Since this pandemic is still growing, it is important to understand how to combat this disease effectively. A myriad amount of COVID-19 data exists out there and therefore, we can examine and analysis different aspect of the pandemic and prepare for the future crisis.

In this study, we provide a statistical analysis of the COVID-19 pandemic in the U.S. and predict a forecasting time-series on the mortality rate, namely we will use the data available only for U.S.

Dataset

For our project, we will use the COVID-19 data compiled by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [3]. Although, this repository contains COVID-19 data about most of the countries in the world, we will only use the data available for U.S. state level only. The U.S. datasets contains daily cases having different field such as: locations, new cases, deaths, recoveries, testing and hospitalizations etc. for all 50 states in their county level as well as time-series data about confirmed cases and deaths.

Proposed Work

Our proposed work is divided into two parts.

a) Doing a thorough statistical analysis of COVID-19 data in the U.S. that provides a visualized report. Namely, we will try to recreate the analysis presented in the paper [K. Dayaratna and A. Vanderplas, 2021] for the pandemic, but will use updated data up until February 2022 from CSSE repository. We will analyze several distribution models for different aspects of the COVID-19 situations. And this will provide us some insights about the spread of the disease in the U.S., factors that are causing fatality-rate, how hospitals are coping with new cases etc. for the 50 states. Given time, we will try to analyzed the data in county level and will try to find the hotspot of certain region for different pandemic aspects.

b) Secondly, we will try to predict the mortality rate for the U.S. time-series COVID-19 data for the 50 states. We will use a forecasting model in python called ARIMA (Auto-

Regressive Integrated Moving Average) that captures a suite of different standard temporal structures in time series data. We will use different accuracy metrics such as: Mean Absolute Percentage Error (MAPE), Correlation between the Actual and the Forecast (corr), and Min-Max Error (minmax) to compare forecast series for different states as well as test their statistical significance.

Paper

[Kevin Dayaratna and Andrew Vanderplas, 2021
https://www.heritage.org/public-health/report/statistical-analysis-covid-19-and-government-protection-measures-the-us?fbclid=IwAR08FstRcBtO3TtNha50CXMuxe9f1c7RhHVN1DiXub2Y9MSqLkhv3aS8nEQ

References

1. WorldoMeter
https://www.worldometers.info/coronavirus/

2. CDC
https://www.cdc.gov/coronavirus/2019-ncov/index.html

3. CSSE, JHU
https://github.com/CSSEGISandData/COVID-19