603: Statistical Modelling With Data

Modelling the Impact of COVID-19: Preparing for Future Pandemics

2022/12/13

Ann Siddiqui
Kane Smith
Paul Croome
Rodrigo Rosales Alvarez

Introduction

For almost 3 years, COVID-19 has devastated people around the world (personally, societally, and globally).

An increasingly interconnected world means increased virality for viruses and other contagions.

More pandemics like COVID-19 are likely inevitable.

Goals

Better understanding how COVID-19's prevalence and severity in countries around the world related to features of that country's population and pandemic response.

Helping to prepare countries for future pandemics.

Using multiple linear regression to construct models of:

- The overall impact of the COVID-19 pandemic in a country, as measured by their average daily COVID-19 cases and deaths, based on the country's population metrics.
- The daily impact of the COVID-19 pandemic in a country, as measured by daily COVID-19 cases and deaths, based on COVID-19-related variables that were reported daily.

Dataset: COVID: Our World in Data

01

COVID metrics

- Daily cases/deaths, tests, vaccinations, hospitalizations, stringency index, reproduction rate, positive rate

Population metrics

02

- GDP per capita, HDI, median age, population

03

Health metrics

- Life expectancy, diabetes prevalence, smokers

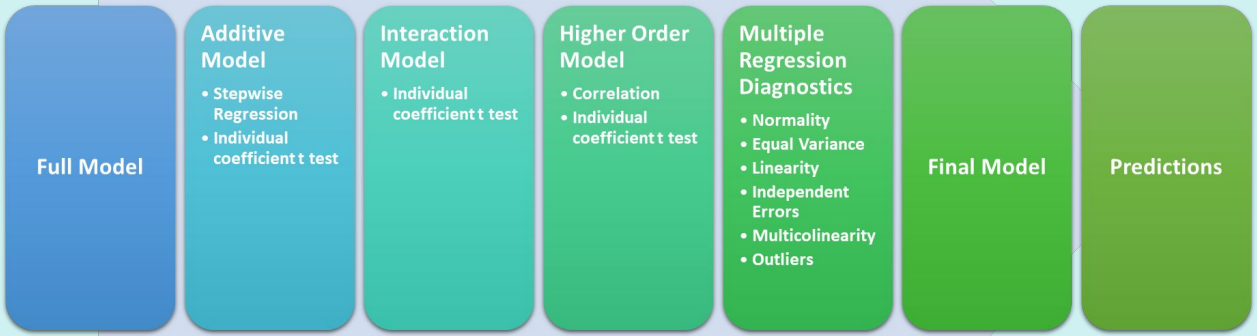
Question 1

What population-related metrics of countries around the world are most strongly related to the prevalence and severity of COVID-19 experienced in a country between February 2020 and October 2022 (as measured by average daily COVID-19 cases and deaths)?

a) What is the best model that can be built from these data for predicting the average daily new COVID-19 cases experienced in a country?

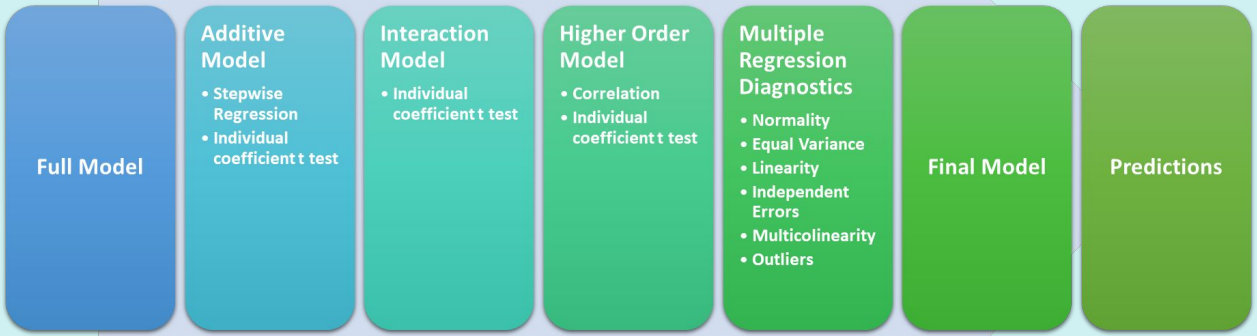
b) What is the best model that can be built from these data for predicting the average daily new COVID-19 deaths experienced in a country?

What is the best model that can be built from these data for predicting the average daily new COVID-19 cases experienced in a country?



$$average.daily.cases = -317.9720 - 0.0000004255\, x_{population} + 69.5871\, x_{aged.65.older} - 0.0702\, x_{gdp.per.capita} + 0.000000005053\, x_{population} x_{gdp.per.capita} + 0.0100 x_{aged.65.older} x_{gdp.per.capita}$$

What is the best model that can be built from these data for predicting the average daily new COVID-19 deaths experienced in a country?



$$average.daily.deaths = 0.8449 - 0.000004976\,x_{population} + 10.9375\,x_{human.development.index} + 0.000008337\,x_{population}x_{human.development.index}$$

Results

	Avg. Daily Cases	Avg. Daily Deaths
Adj. R-squared	0.6551	0.5141
RMSE	6,208.4082	87.5619
Assumptions	Normality ❌ Homoscedasticity ❌ Linearity ✅ Independent Errors ✅ Multicolinearity ✅	Normality ❌ Homoscedasticity ✅ Linearity ✅ Independent Errors ✅ Multicolinearity ✅

Question 2

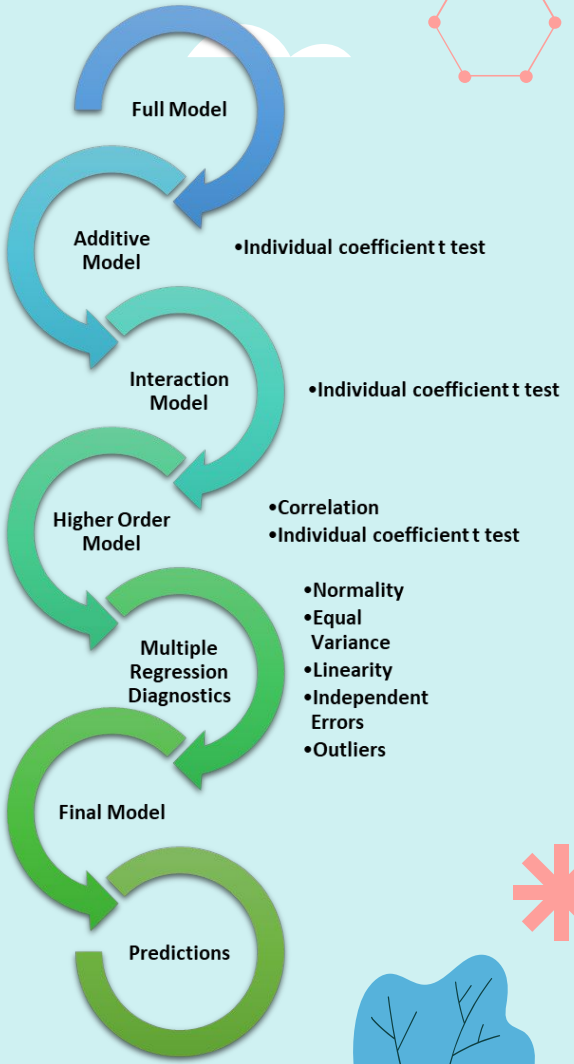
Among countries with reliably reported data relating to cases, positive test rates, vaccinations, and other daily-reported COVID-19 data, which of these variables are most strongly related to the prevalence and severity of COVID-19 experienced in a country between February 2020 and October 2022 (as measured by average daily COVID-19 cases and deaths)?

a) What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 cases in a country?

b) What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 deaths in a country?

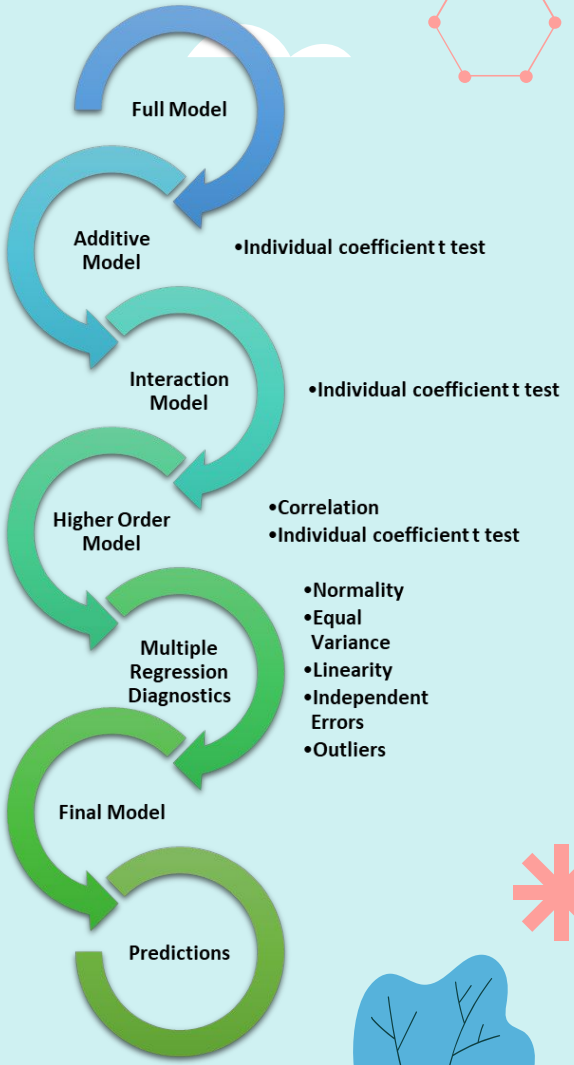
What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 cases in a country?

$$
\begin{aligned}
new.cases = {} & -956.9219 - 177.4731\, x_{reproduction.rate} + 0.0103 x_{new.tests.smoothed} \\
& - 782.4864 x_{positive.rate} + 0.0443 x_{new.vaccinations.smoothed} \\
& - 0.1002 x_{new.people.vaccinated.smoothed} - 35.6258 x_{stringency.index} \\
& + 1.207 x_{hosp.patients} - 0.0028 x_{reproduction.rate} x_{new.tests.smoothed} \\
& - 0.0114 x_{reproduction.rate} x_{new.vaccinations.smoothed} \\
& + 0.0505 x_{reproduction.rate} x_{new.people.vaccinated.smoothed} \\
& + 58.5161 x_{reproduction.rate} x_{stringency.index} - 1.4194 x_{reproduction.rate} x_{hosp.patients} \\
& + 0.9823 x_{new.tests.smoothed} x_{positive.rate} \\
& - 0.00000001809 x_{new.tests.smoothed} x_{new.vaccinations.smoothed} \\
& + 0.000000008441 x_{new.tests.smoothed} x_{new.people.vaccinated.smoothed} \\
& - 0.0001008 x_{new.tests.smoothed} x_{stringency.index} \\
& + 0.0000001965 x_{new.tests.smoothed} x_{hosp.patients} \\
& - 0.05656 x_{positive.rate} x_{new.vaccinations.smoothed} \\
& - 0.03826 x_{positive.rate} x_{new.people.vaccinated.smoothed} + 2.0004 x_{positive.rate} x_{hosp.patients} \\
& + 0.000000005571 x_{new.vaccinations.smoothed} x_{new.people.vaccinated.smoothed} \\
& - 0.0004916 x_{new.vaccinations.smoothed} x_{stringency.index} \\
& + 0.0000005381 x_{new.vaccinations.smoothed} x_{hosp.patients} \\
& + 0.0009311 x_{new.people.vaccinated.smoothed} x_{stringency_{index}} \\
& - 0.0000007765 x_{new.people.vaccinated.smoothed} x_{hosp.patients} \\
& - 0.002075 x_{stringency.index} x_{hosp.patients}
\end{aligned}
$$



What is the best model that can be built from these daily-reported data for predicting the daily new COVID-19 deaths in a country?

$$
\begin{aligned}
new.deaths = {} & 58.4964 - 35.3352\, x_{reproduction.rate} + 0.0004054 x_{new.tests.smoothed} \\
& - 100.6623 x_{positive.rate} - 0.000004938 x_{new.vaccinations.smoothed} \\
& - 0.001109 x_{new.people.vaccinated.smoothed} - 0.8754 x_{stringency.index} \\
& + 0.1136 x_{icu.patients} - 0.0002561 x_{reproduction.rate} x_{new.tests.smoothed} \\
& + 0.0005209 x_{reproduction.rate} x_{new.vaccinations.smoothed} \\
& - 0.0006226 x_{reproduction.rate} x_{new.people.vaccinated.smoothed} \\
& + 0.4797 x_{reproduction.rate} x_{stringency.index} - 0.09270 x_{reproduction.rate} x_{icu.patients} \\
& + 0.0007177 x_{new.tests.smoothed} x_{positive.rate} \\
& - 0.00000000006414 x_{new.tests.smoothed} x_{new.vaccinations.smoothed} \\
& + 0.0000000008364 x_{new.tests.smoothed} x_{new.people.vaccinated.smoothed} \\
& - 0.000001561 x_{new.tests.smoothed} x_{stringency.index} \\
& + 0.000000008622 x_{new.tests.smoothed} x_{icu.patients} \\
& + 0.003121 x_{positive.rate} x_{new.people.vaccinated.smoothed} \\
& + 3.0156 x_{positive.rate} x_{stringency.index} - 0.1275 x_{positive.rate} x_{stringency.index} \\
& - 0.0000000001269 x_{new.vaccinations.smoothed} x_{new.people.vaccinated.smoothed} \\
& - 0.00001073 x_{new.vaccinations.smoothed} x_{stringency.index} \\
& + 0.00000001979 x_{new.vaccinations.smoothed} x_{icu.patients} \\
& + 0.00002719 x_{new.people.vaccinated.smoothed} x_{stringency_{index}} \\
& - 0.00000005890 x_{new.people.vaccinated.smoothed} x_{icu.patients} \\
& + 0.0008637 x_{stringency.index} x_{icu.patients}
\end{aligned}
$$



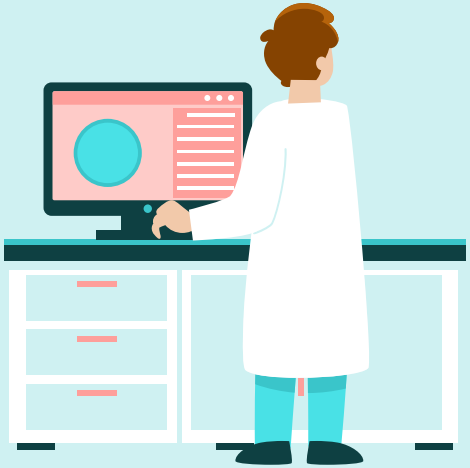
Results

	Avg. Daily Cases	Avg. Daily Deaths
Adj. R-squared	0.9918	0.9197
RMSE	3,582	85.25
Assumptions	Normality ❌ Homoscedasticity ❌ Linearity ✅ Independent Errors ✅ Multicolinearity ✅	Normality ❌ Homoscedasticity ❌ Linearity ✅ Independent Errors ✅ Multicolinearity ✅

Conclusions

Found that population, elderly population, GDP per capita and HDI are most important population metrics in average daily cases/deaths.

Constructed four models which each pass at least 3/4 regression assumptions and explain at least 50% of data.



Found that vaccinations, government regulations, and COVID-19 testing were all important predictors of daily COVID-19 deaths and cases.

Thank You

Any Questions?