evive

COVID-19 Models

The nature, scale, and speed of infection of COVID-19, since it was first recorded in Wuhan, Hubei province, China in December 2019, has caught many governments a bit unprepared in their search for the optimal response. However, as with many other leading academic institutions and healthcare companies, we at Evive firmly believe that publicly available COVID-19 data when coupled with sophisticated mathematical models can lead to immensely useful and timely insights. This can also help guide public policies (duration and nature of lockdown, resources spent on healthcare preparedness, etc.) or individual responses based on personalized needs.

Evive has been one of the pioneers in generating personalized insights at the individual level to help our customers make informed decisions at a time when they most need it. Our efforts have been across many directions—however, in this article, we'll focus on our COVID-19 models and how they are helping end users. The two models as mentioned below take publicly available U.S. data as input and predict the cumulative number of daily cases and deaths in each county for an extended time period (May 2020 end).

An Extended IHME COVID-19 Model

One of the most important prediction models, using a Gaussian CDF-like parametric approach, was used by the Institute for Health Metrics and Evaluation in Seattle, Washington. Currently it is the official model in the U.S. which is being used by both central and state-level governments. The predictions from IHME are available on their site, and can also be downloaded as a CSV file here.

Traditionally, epidemic models were based on the number of people susceptible, infected, and recovered with respect to the epidemic. Such SIR models and their variants have the advantage of being tried and tested along with having a solid mathematical basis. However, since they assume individuals will keep interacting amongst themselves at roughly the same uniformly, random rate (throughout the course of the epidemic), they predict roughly 25% to 75% of the world population will become infected by the time pandemic trails off.

However, it is being observed that both individual and government responses during the course of the epidemic can change the course of it. So instead of modelling the S, I, and R type of parameters and estimating their trajectories (by solving differential equations), the IHME model tries to estimate the growth trajectories for cumulative parameters (e.g., cumulative cases, cumulative deaths, etc.) across locations that were previously hit by COVID-19. While making a strong assumption that the overall shape of the trajectory is invariant across locations, it has the advantage of using government responses as a parameter directly within the model with very minor tweaking. Another assumption is that parameters of the pandemic (death rate, infection rate, etc.) are invariant across locations when conditioned on age and health factors. We believe that given our current knowledge of virus-human interaction, this is quite a reasonable assumption.

Through extensive experiments, it was observed that the cumulative death rate for each location follows a parameterized Gaussian error function.

$$D(t;\, \alpha,\, \beta,\, p) \;=\; (p/2) * \psi(\alpha(t-\beta)) \;=\; (p/2) * (1 + (2/\sqrt{\pi}) \int_0^{(\alpha(t-\square))} exp(-\tau^2) d\tau$$

Where the function Y is the Gaussian error function (written explicitly above), p controls the maximum death rate at each location, t is the time since death rate exceeded 1e-15, β (beta) is a location-specific inflection point (time at which rate of increase of the death rate is maximum), and α (alpha) is a location-specific growth parameter.

Using parameters specific to a given location, it is also possible to estimate cumulative cases, number of ventilators needed, and ICU beds required across time. IHME keeps updating their results after every 2-3 days as more data comes in and they update their model. The granularity of these predictions is currently at the state level. However, being more location-specific, higher

granularity predictions at the county level are more likely to be useful. In our work at Evive, we have attempted to generalize this model to generate both long-term and short-term cumulative death rate predictions by optimizing parameters of a Gaussian CDF type function for each county. It seems these predictions will be more useful because they are more local in nature.

Ref:
[1] Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator days and deaths by US state in the next 4 months, IHME COVID-19 health service utilization forecasting team, University of Washington.

[2] "The Mathematics of Infectious Diseases". SIAM Review. 42 (4): 599–653, Hethcote H (2000)

[3] "Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates". Applied Mathematics and Computation. 236: 184–194. Harko, Tiberiu; Lobo, Francisco S. N.; Mak, M. K. (2014).

[4] "Mathematical models of SIR disease spread with combined non-sexual and sexual transmission routes". Infectious Disease Modelling. 2 (1). section 2.1.3, Miller, J.C. (2017).

Time Series based ARIMA model

Introduction
ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. This model is a popular and widely used statistical method for time series forecasting. It explicitly caters to a suite of standard structures in time series data, and as such, it provides a simple yet powerful method for making skillful time series forecasts.
ARIMA is actually a class of models that explains a given time series based on its own past values—that is, its own lags and the lagged forecast errors—so the equation can be used to forecast future values.

Steps to follow before moving toward ARIMA model
A trend is a long-term increase or decrease in the level of the time series. A time series with a trend is called non-stationary.

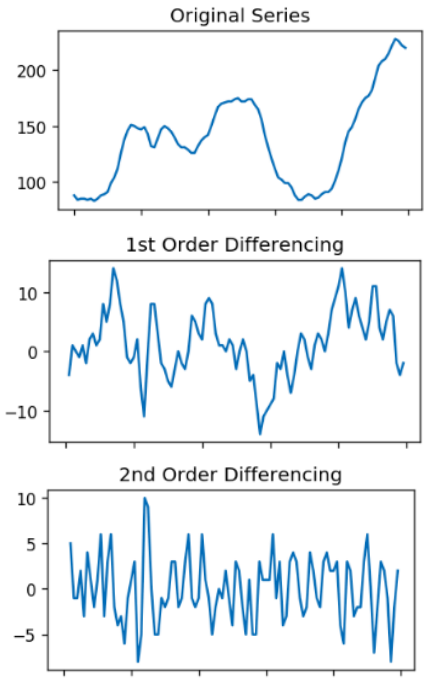
An identified trend can be modeled. Once modeled, it can be removed from the time series dataset. This is called detrending the time series.

If a dataset does not have a trend or we successfully remove the trend, the dataset is said to be trend stationary.

Model formulation
An ARIMA model is characterized by 3 terms: p, d, q; where 'p' is the order of the AR term, 'q' is the order of the MA term, and 'd' is the number of differencing required to make the time series stationary.

The value of 'd,' therefore, is the minimum number of differencing needed to make the series stationary. And if the time series is already stationary, then d = 0.

evive



The first step to building an ARIMA model is to make the time series stationary, because the term 'Auto Regressive' in ARIMA means it is a linear regression model that uses its own lags as predictors. Linear regression models work best when the predictors are not correlated and are independent of each other.

The order of the 'Auto Regressive' (AR) term, 'p', refers to the number of lags of Y to be used as predictors. And the order of the 'Moving Average' (MA) term, 'q', refers to the number of lagged forecast errors that should go into the ARIMA model.

$$\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

is the general equation of the ARIMA model, after making the time series stationary by differencing 'd' (0) times.
Here is the ARIMA model in words:

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

So, (p+q+1) parameters are to be estimated from the above model, using the time series data; p0, q0.

Results

The general structure of this model is not intended for making predictions on epidemic/pandemic situations like COVID-19, so we are not trying to predict for the long term; but this model is quite good for making short-term predictions.

As this model uses values corresponding to recent past time points for predicting points in the future, this model is quite powerful for making short-term predictions.

Using this model, we are making predictions of daily "Number of Confirmed Cases" and "Number of Deaths" in the U.S. county-wise, for the upcoming 5 days.
MAPE is Mean Absolute Percentage Error:

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{A_i - F_i}{A_i} \right|$$

It's the most commonly used metric for comparing performance of a predictive time series model.

MAPE for all the counties lies within 10%, and among them for the most counties, MAPE lies within or around 5%; even less than 1% for many counties.

Around 5% MAPE implies the model is around 95% accurate in predicting the next 5 observations.

Ref:
[1] Dehesh, Tania & Mardani-Fard, H.A. & Dehesh, Paria. (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. 10.1101/2020.03.13.20035345.

[2] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, Massimo Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data in Brief, Volume 29, 2020, 105340, ISSN 2352-3409

evive

Logistic Growth Model

Introduction
Mr. Verhulst enhanced the exponential growth theory of population, as saying that the population's growth is not always growing, but there is always a certain Limit or a Carrying Capacity to the exponential growth. Combining the exponential growth with a limit, it's then called the Logistic Growth.

Why Logistic Growth Model?
Logistic Growth is a mathematical function that can be used in several situations. It is characterized by increasing growth in the beginning period, but a decreasing growth at a later stage, as it gets closer to a maximum.

Eventually, the growth rate will plateau or level off, making an S-shaped curve.

The reason to use Logistic Growth for modeling the Coronavirus outbreak is that epidemiologists have studied those types of outbreaks and it is well known that the first period of an epidemic follows Exponential Growth and that the total period can be modeled with a Logistic Growth.

Model formulation
The equation of Logistic Growth is given by:

$$\frac{dP}{dt} = rP \cdot \left(1 - \frac{P}{K}\right)$$

where P is the "Population Size" (i.e., number of infected persons), t is "Time", r is the "Growth Rate", K is the "Carrying Capacity" (or limiting value for number of infected persons).

Interpretation
At any given point in time during a population's growth, the expression (K - P) tells us how many more individuals can be added to the population (i.e., number of infected) before it hits carrying capacity.

(K - P)/K, then, is the fraction of the carrying capacity that has not yet been "used up." The more carrying capacity that has been used up, the more the (K - P)/K term will reduce the growth rate.

When the population (i.e., number of infected) is tiny, N is very small compared to K. The (K - P)/K term becomes approximately (K/K), or 1, giving us back the exponential equation.

In other words, the (1 - P/K) determines how close the population size is to the Limit K, which means as the population gets closer and closer to the limit, the growth gets slower and slower.

Results
Using this model, we are making predictions of daily "Number of Confirmed Cases" and "Number of Deaths" in the U.S. county-wise, for the upcoming 15 days.
MAPE is Mean Absolute Percentage Error:

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{A_i - F_i}{A_i} \right|$$

MAPE for almost all the counties lies within 10%, and among them for the most counties, MAPE lies within or around 5%; even less than 1% for many counties.

Around 5% MAPE implies, on average, that the model is around 95% accurate in predicting the Number of Cases/ Deaths for next 15 days.

Ref:
[1] Batista, Milan. (2020). Estimation of the final size of coronavirus epidemic by the logistic model.

[2] Tsoularis, AN & Wallace, James. (2002). Analysis of Logistic Growth Models. Mathematical
biosciences. 179. 21-55. 10.1016/S0025-5564(02)00096-2.