Research Scientist – Pandemics

Applicant assessment

*This assessment is designed to provide you an opportunity to demonstrate analytical thinking and technical skills relevant to the Research Scientist role on the Pandemics team at IHME. There is not a single “correct” answer; rather, use the challenge below to show how you would approach a problem, and develop an initial attempt at implementing your proposed methodology.*

# Instructions

You have 24 hours to complete this assessment; it is likely to take 3 – 4 hours. Use the final question below to indicate areas for future development or improvements to your approach.

Use the data attached to answer the questions below. For each question please provide the following:

1. A short written answer to the question; this can be typed in a separate document.
2. Where applicable, visuals to show your results, as described in each question. This can be a separate file for each question, but please be conscious of file size and navigability (i.e. do not create a separate file for each location).
3. The code used to create the results and figures –either delivered via a GitHub repo, or attach the files to the submission.

# Prompt

You have been provided actual data from the covid-19 pandemic, as reported by state and national governments. These data have already had spikes or errors in reporting corrected and redistributed. Attached are data from the US (excluding Washington state):

* Cumulative confirmed (reported) cases of Covid 19, by location by date
* Cumulative reported deaths of covid-19, by location by date
* Cumulative hospitalizations for Covid-19, by location by date (NOTE: hospitalization data is not available for all 50 states, only some)
* Population of each location

These are the only 4 inputs that IHME uses for our first-stage deaths model, which produces 14-day forecasts that are then used in an SEIR model.

# Questions

* 1. What is the relationship between cases, hospitalizations, and deaths? Describe how these indicators relate to each other, and visualize the relationships in at least 2 different ways.

Cases and deaths, cases and hospitalizations, and hospitalizations and deaths all appear to have a positive linear relationship with a significant amount of variance. The amount of variance, and slope, differs from state to state. When plotting all observations of daily deaths and daily cases, across all states **(Figure 1)** it appears that daily deaths may have a positive linear relationship with daily cases in some states, and possibly a logarithmic relationship in a different grouping of states. For further investigation it would be helpful to bin states by certain demographic factors or simply population to further tease out this relationship. I made the decision to plot deaths on the x axis here because I planned to predict cases as a function of deaths for question 3. See **cases\_hosp\_plots.pdf** and **deaths\_cases\_plots.pdf** for plots of these indicators broken down state by state.

Chart, scatter chart

Description automatically generated

**Figure 1**

Chart, scatter chart

Description automatically generated

**Figure 1a**

* 1. Fit a curve of daily deaths, utilizing these inputs. Describe the approach you used and visualize the results.

My initial approach was to sum cumulative deaths across all states, then fit a logistic growth model to the cumulative death numbers. I then calculated daily predictions from cumulative predictions. I fit this model using the R optim() function to minimize RMSE. Doing this at the national level resulted in a model fit with a relatively large difference in predicted and observed deaths in the first 30 days of recorded data, and in the most recent 20 days of recorded data. (**Figure 2 and Figure 3).** One issue with my approach to tabulating observed data that immediately jumps out is the negative daily case numbers caused by corrections in cumulative case numbers. Because a logistic growth model’s only input is deaths in the previous time-step, it may perform poorly when deaths are small and its predictions for t+1 are fractional ie: t=5 deaths, t+1=5.2 deaths. The model may be very sensitive to small changes in starting cases such as 3 vs 4, which is a poor reflection of the reality of a pandemic like this one.

Chart, line chart

Description automatically generated

**Figure 2:** Daily observed deaths (all states) with model fit in red.

Chart, line chart

Description automatically generated

**Figure 3:** Cumulative deaths with model fit in red. RMSE: 7247.68

* 1. Create projections for 14-days after the last observed data point. Visualize the result. Describe the benefits and limitations of your approach. Where do you think this approach has performed particularly well? What types of situations cause your model to struggle?

The benefits of a logistic growth model are that the parameters are easily interpretable. In the parameterization below **r** = the expected growth in cumulative deaths per day, **K** = the total number of predicted cumulative deaths. You could see a case where these interpretable values would be useful in making policy decisions – how many deaths could there be in total? How much could cumulative deaths grow each day?

**Parameterization of the logistic growth model:**

N[t+1] = N[t] \* r \* (1 – N[t] / K)

Best fit:

R = 0.056 K = 137,627 N1 = 4,503

Fitting a logistic growth model is clearly performing poorly for forecasting as the observed data has an upwards trend in the most recent days, while the daily predicted deaths are tapering off. **Figure 4** This makes sense as the best fit K is actually less than the observed cumulative deaths. If I were to attempt fitting again, I might constrain this parameter. Another limit of my approach here is that I have not calculated uncertainty around the model parameters which would allow me to forecast upper and lower bounds for cases into the future.

Diagram

Description automatically generatedChart, line chart

Description automatically generated

**Figure 4: Predicted daily deaths in green, diverge from observed deaths (black dots) Modeled past and forecast hospitalizations in purple and cases in orange. Black dashed line separates modeled and observed past from modeled forecasts.**

In order to forecast hospitalizations and cases I fit a linear model with deaths as the independent variable. The strength of this approach is that we have the least missing data for deaths. Another strength is that deaths are more likely to be accurately recorded than either cases or hospitalizations. A weakness of this approach is that all other potential drivers of hospitalization, death and cases are excluded. If I had fit a linear model **cases ~ deaths** / **hospitalizations ~ deaths** for each state – I might be able to observe different coefficients in each state, but would provide no insight into this variation. Masking, behavior, and social-distancing policy might all vary from state to state. It would be useful to know how they vary, and how this impacts the relationship between deaths, cases and hospitalizations.

* 1. Lastly, describe future areas of exploration or improvement for your approach. If you had more time, what would you do next?

If I had more time I would fit a logistic growth model to death data for each state, forecast at the state level, and then aggregate to the national level. Looking at daily deaths, in is possible that a logistic growth model would be better to fit to daily deaths vs cumulative deaths as I have done here.

I would also experiment with different models because, at least at the national level, logistic growth is a clearly a poor predictor. With extra time I would use maximum likelihood estimation in order to fit models, which could allow varying observation error in different states to be considered while fitting the model. In addition, I would use a formalized metric to compare different models – that could be either RMSE, or something like AIC to account for varying numbers of model parameters.

I would also alter my linear model of cases/hospitalizations ~ deaths, and allow the slope and intercept to vary by State. With even more time I could collect additional covariate data for a better understanding of why the functional relationship between deaths, cases and hospitalizations varies state by state.