Predicting COVID Case Counts in Japan

Ken Mawer

2021-12-05

Being from Japan, I am interested in the COVID situation there. Given that their travel and quarantine restrictions are very strict, I would like to know about whether cases can further go down, which may justify removing barriers for entry. This is because I want to be able to see my family, some of whom is in Japan. Having learned about how hard it is to predict COVID cases during class, I still want to know if Japan is no exception to this difficulty.

Introduction

As older age groups have higher priority for COVID vaccinations in Japan (Source), this study will categories prefectures by the extent to which COVID affected them for the fifth wave, which is after when many people were vaccinated. In addition, it attempts to predict COVID case rates.

Datasets

This project uses three datasets. The Toyo Keizai dataset captures the COVID case rates in each prefecture in Japan. In addition, I am using two Statistics Japan datasets: one captures the population and population densities of each prefecture in Japan, and another captures the distribution of different age groups in Japan.

EDA

I am analyzing the COVID-19 incidence rates in Japan in a two-part analysis. This analysis will compare the state of COVID across each prefecture, or administrative subdivision, across Japan. It will also attempt to find a way to predict future cases based on past cases.

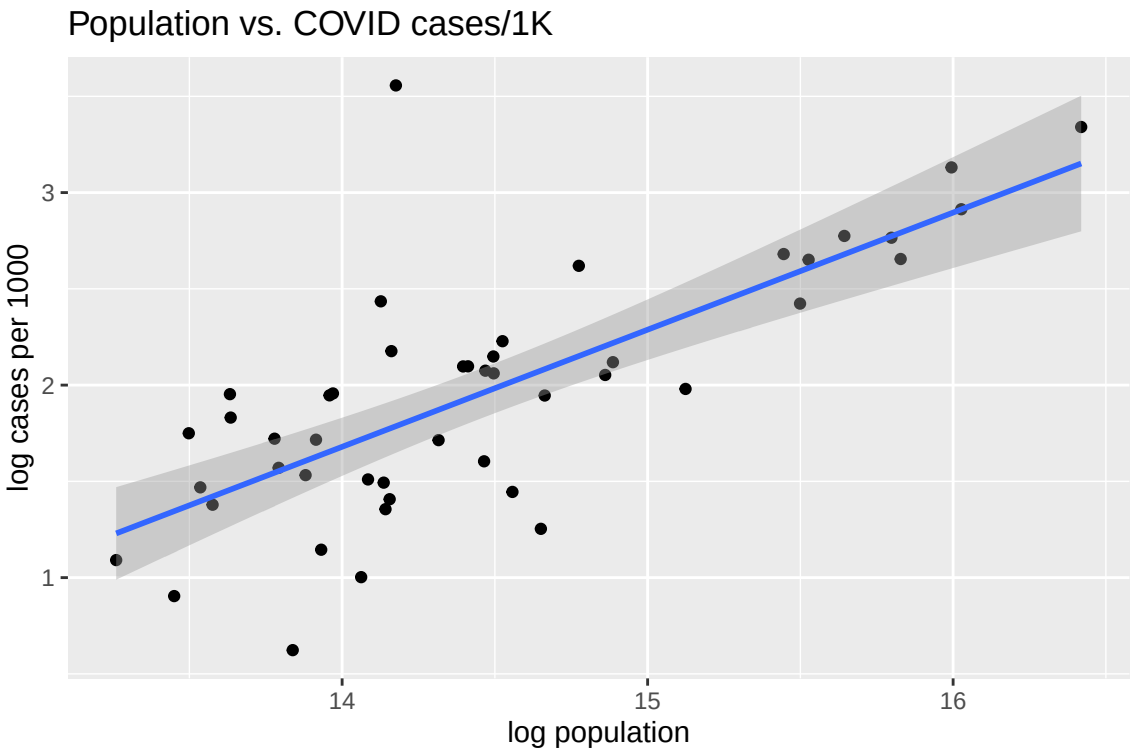
The measurement of impact of COVID-19 in a prefecture will be measured related to COVID cases per capita. In a given prefecture, the response variable is the number COVID cases per 1000 people, which is driectly linearly related to COVID-19 cases per capita by multiplying that number by 1000. The factors include the population of the prefecture and the population density, which may be factored or interacted to improve the model's accuracy. I consider these factors potentially relevant, as I am from Tokyo, the most populated prefecture with a fairly high population density, and this prefecture has had many COVID cases. Additional factors include age demographics of each prefecture, as older populations are prioritized for the vaccine and may be more risk averse.

During class, we learned that it is hard to predict COVID case rates. However, I want to know if Japan is any exception to this, as I have not been able to find anything good in English relating to COVID-19 predictions, as almost everything printed in Japan will be in Japanese. Whether it is related to predicting cases in the future, or finding what factors were involved in the number of COVID cases per capita, it is difficult to find anything done in English.

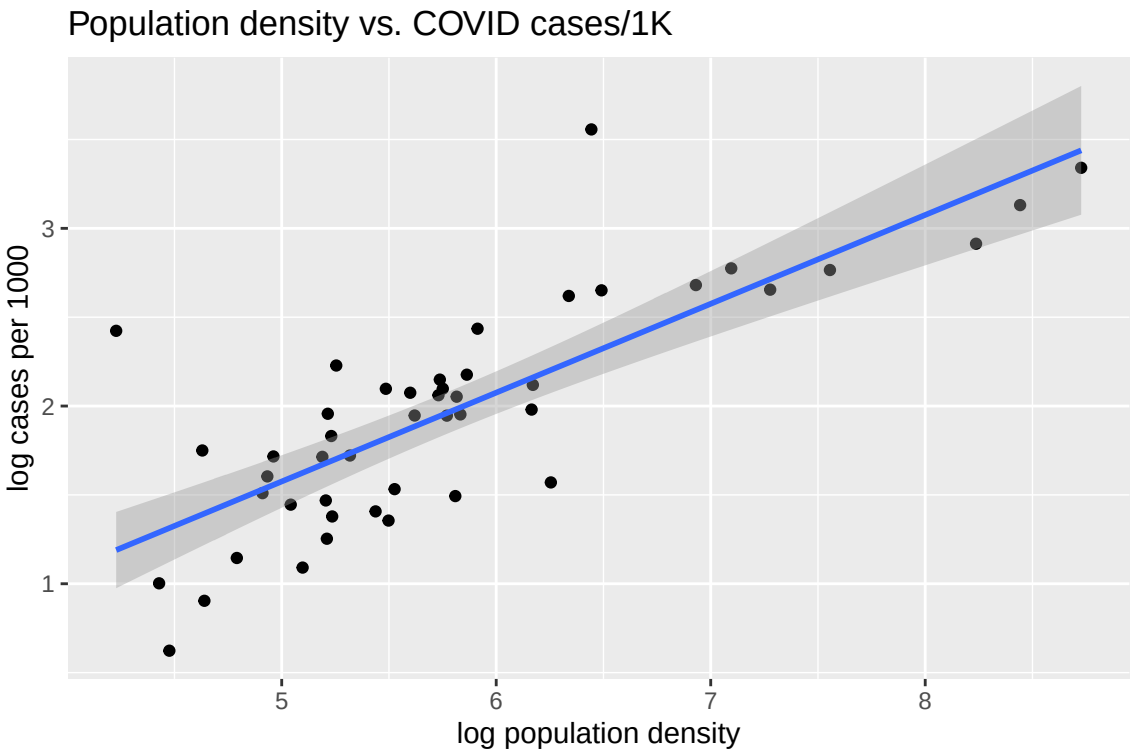
Even if it may be difficult to predict COVID-19 cases, it may be possible to analyze the effects of the pandemic on prefectures (subdivisions of Japan) based on factors such as their population, population density and concentrations of different age groups.

Predicting COVID cases per capita in prefectures of Japan using population and population density

Using a regular linear model, we see that both population and population density have a linear relationship to COVID cases per capita.

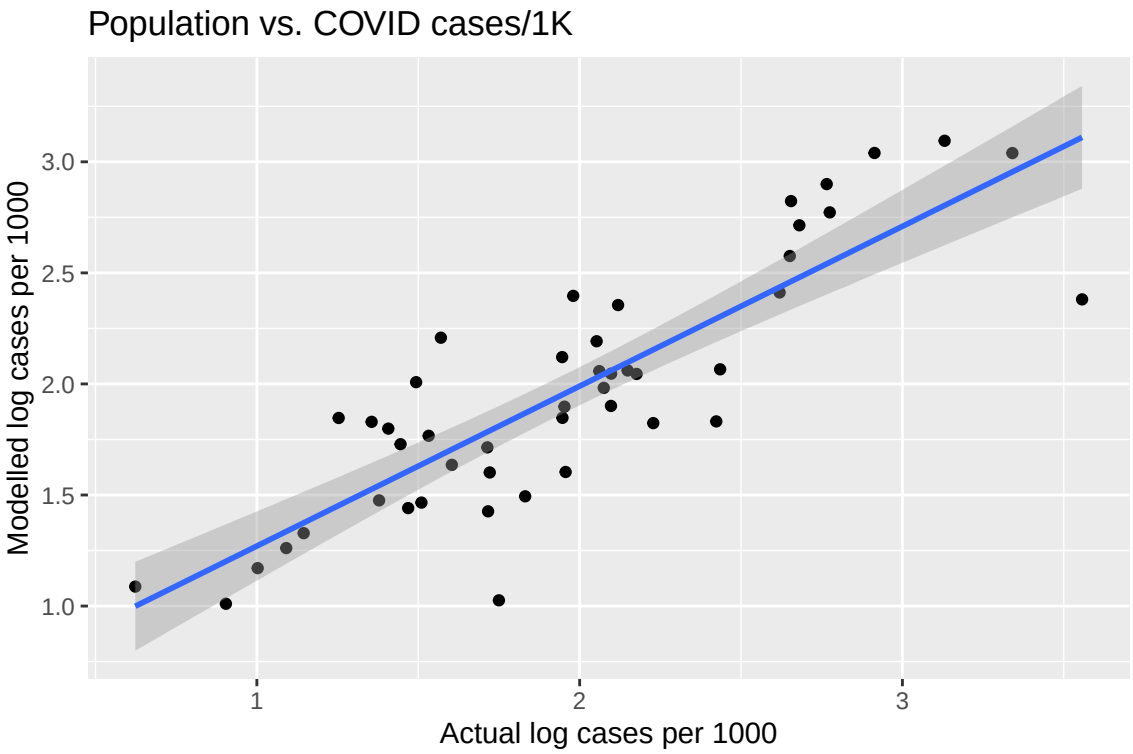


This graph does appear to show a relationship of population and COVID cases per 1,000. However, it is heteroskedastic.



This graph also appears to show a relationship of population density and COVID cases per 1,000. Again, it is heteroskedastic.

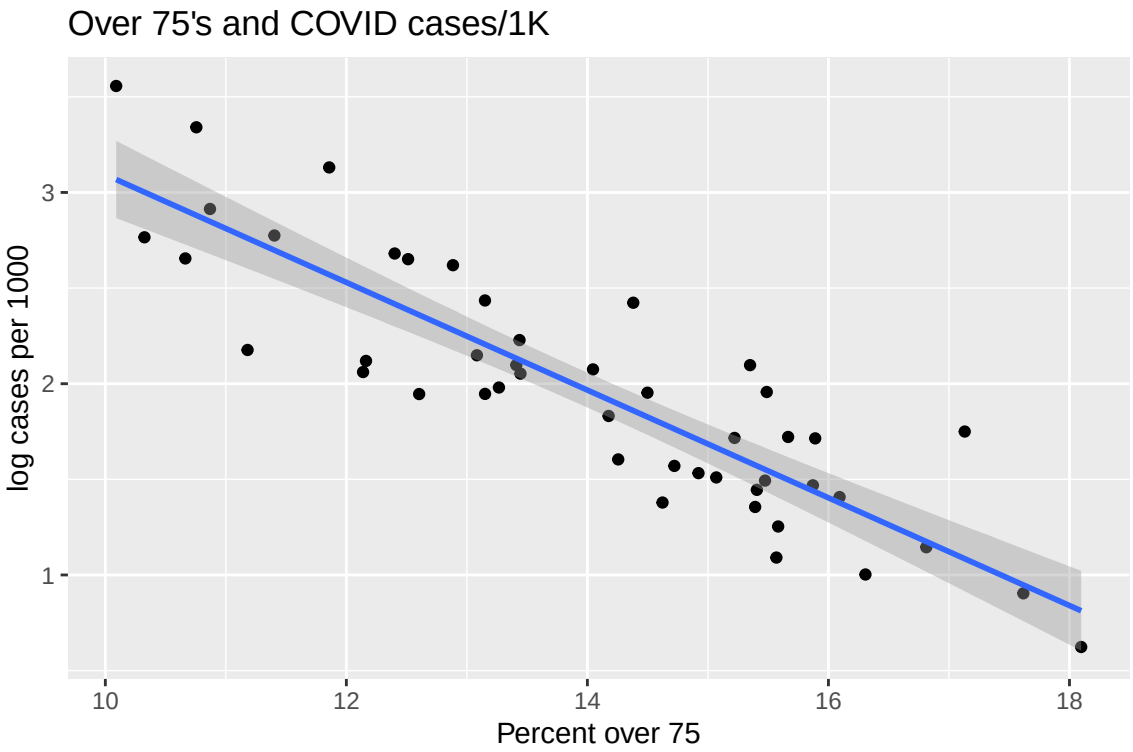
To potentially increase the model's acccuracy and reduce the model's heteroskedasticity, a model with both these variables and interaction terms is used. The adjusted R^2 is 0.699, where both the population and population density are positively correlated with higher COVID cases, but that there is a negative interaction term when those are multiplied:



The graph is an improvement over the other graphs before due to homoskedasticity.

Predicting COVID cases in prefectures by age demographics

Using a model with age demographics involving percentages of each age group (<15,15-64, 65+ and 75+) results in only the percentage of people over 75 being a statistically significant result (adjusted R^2 of 0.7546). Using only the amount of over 75's, we now get an adjusted R^2 . There is evidence to suggest that prefectures with higher percentages of over 75's have lower numbers of COVID cases per capita. This suggests that they are either likely to get themselves infected with COVID due to risk or other reasons, or because they were prioritized for the vaccine.



Cross-validated selection

To improve this model, we use cross-validation rather than select based on p-values.

Can we get the best of both worlds?

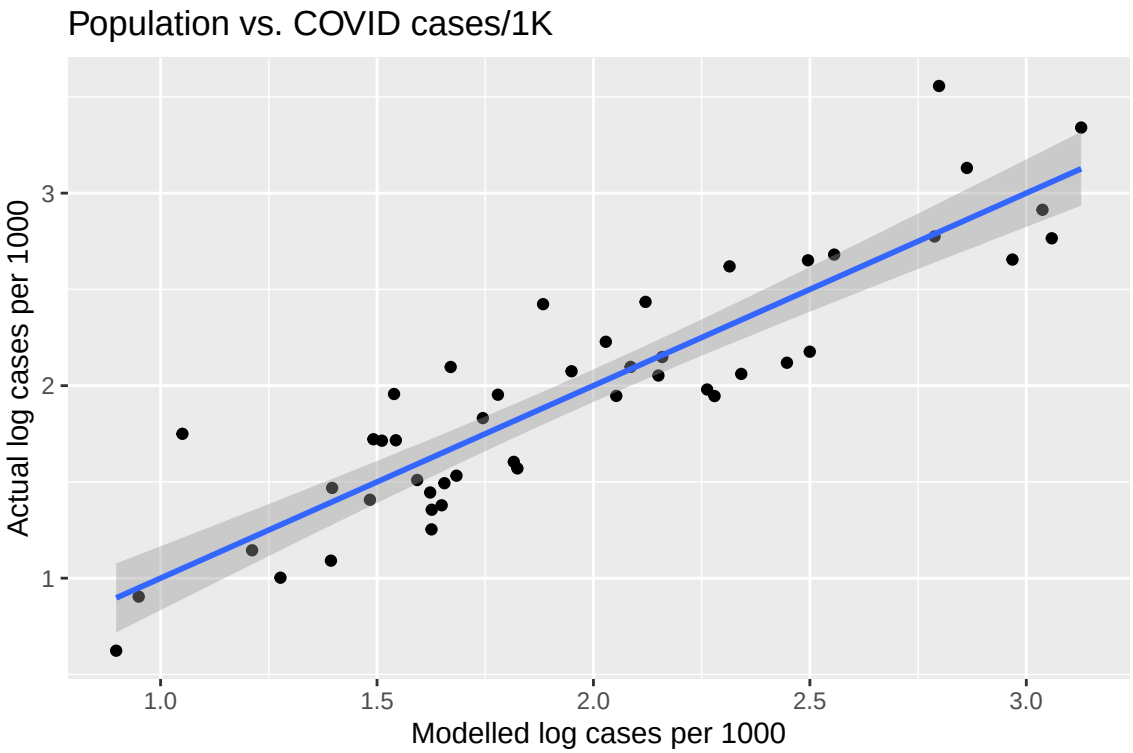
Making a model that includes the logarithm of the population, the logarithm of the population density, the interaction term, and the percentage of people over 75 brings the adjusted R^2 to 0.7859. Indeed, the scatter from the prediction line appears to have somewhat mitigated.

However, the predictors relating to the population and population density now have higher p-values, where only the factor of the percentage of the population being over 75 being statistically significant.

The model is as follows, where x_{75} is the variable relating to the percentage of people over 75, x_p is the population, and x_d is the population density in terms of people per square kilometre. y refers to the COVID cases per 1000 people in a prefecture.

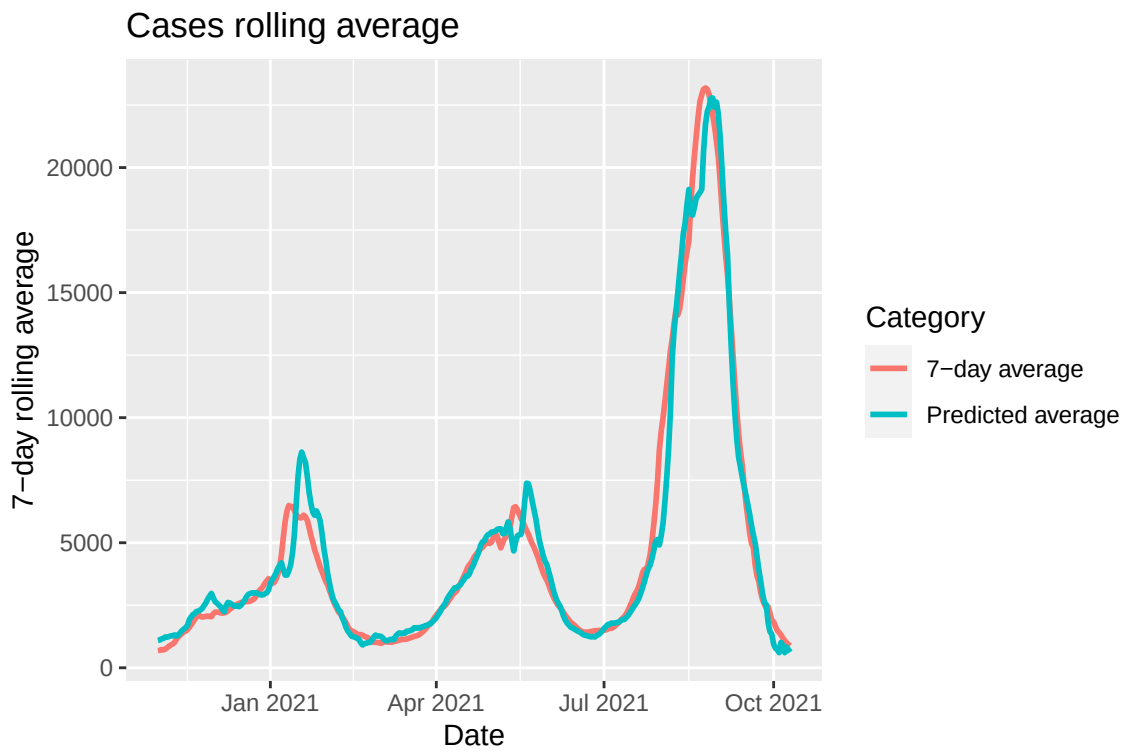
$$\ln y = -0.39298 - 0.19434x_{75} + 0.29737 \ln x_p + 0.53754 \ln x_d - 0.02766 \ln x_p \ln x_d$$

The graph below illustrates the fit of the model with respect to the response variable.

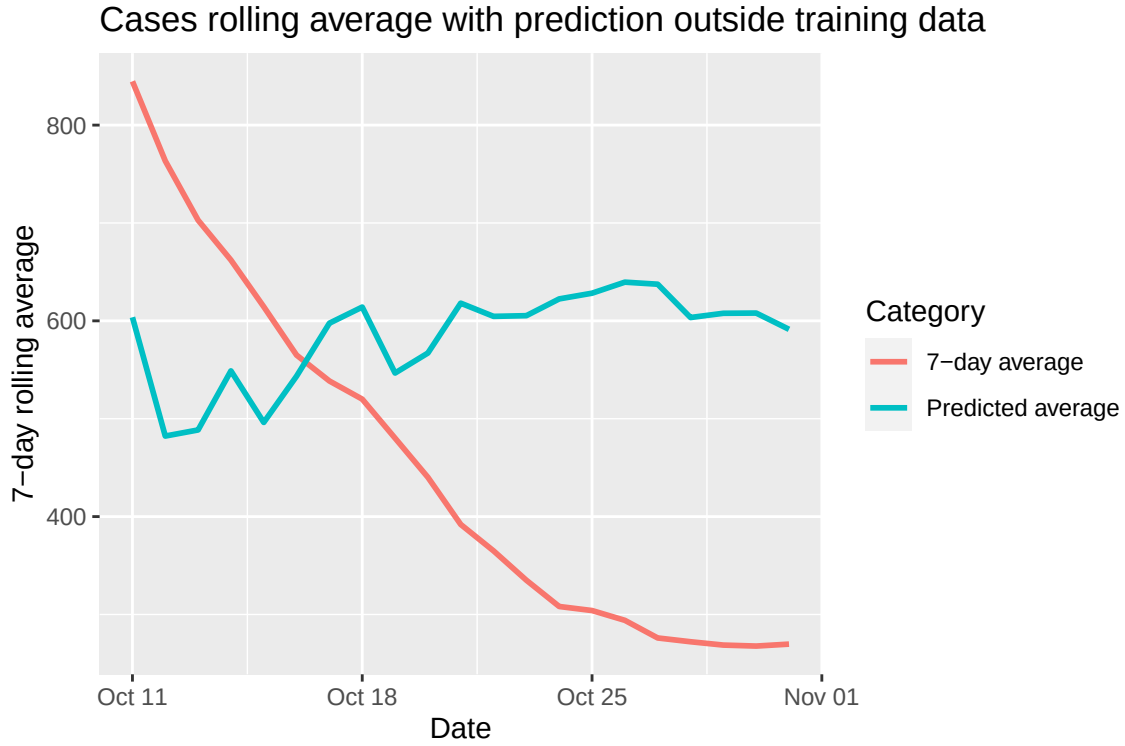


Predicting COVID trends in Japan over time.

Predicting each prefecture is arguably more inconsistent when there are 47 prefectures in Japan. Moreover, many policies are dependent upon the country's situation itself. Thus, I am predicting the COVID trends in Japan overall, rather than individual prefectures.



This model is accurate to an adjusted R^2 with 0.9626. Taking out the last three weeks for this to be used for testing data, we have that the accuracy of the training model is 0.9612. The accuracy may not be the same when we test the data outsite that has been extrapolated from the time series.



However, further effort needs to be done to select a model that does not overfit. By merely using linear

regression and extrapolation without cross-validation, the model accuracy seems very poor, as shown above. Even when the cases decrease, the model used does not seem to account for this, but instead produces a jagged line. This suggests that the model is overfitting the data. It may be possible to improve the model through cross-validation, as to reduce overfitting. However, it is important to remember that the model would still utilize extrapolation, which is inherently risky. We want to create a model that reduces problems relating to overfitting.

Results

Further research

More insight may be possible by analyzing COVID cases before the commencement of vaccination for seniors, as this may provide implications relating to the effectiveness of the COVID vaccine itself. However, this would omit the fifth wave of COVID, which happened after vaccination and resulted in a significant increase in COVID cases. The omission of this fifth wave may even be useful in describing the effectiveness of the COVID-19 vaccine through comparing demographics, which would suggest different vaccination rates. This would likely necessitate also gathering Japan's COVID-19 vaccination rate, as to ensure an accurate comparison.

Citations

https://toyokeizai.net/sp/visual/tko/covid19/en.html https://www.stat.go.jp/data/nenkan/66nenkan/zuhyou/y660203000.xls

https://www.storybench.org/how-to-calculate-a-rolling-average-in-r/