STEM Fellowship Big Data Challenge 2020

COVID-19 in Canada – Prediction on Canadian Recovery Rate

Yunkun Yang[1] and Jia Xin Li[1]

[1]University of Waterloo

June 1, 2020

Abstract

Coronavirus disease (COVID-19) has been a widespread life-threatening disease in 2020. Under Canadian government's active interventions and the Quarantine Act, the pandemic is increasingly under control, but there is still many people suffering from the symptoms of the disease. We want to know when these people can be fully cured and when the Coronavirus will no longer be a huge public health concern. We used the data from the Government of Canada and constructed a logistic regression model to predict when Canada can recover from the disease.

Keywords
Covid-19; Data Analysis; Logistic Regression

1 Introduction

Coronavirus has become one of the most fatal diseases in 2020. Many people have different symptoms, such as diphtheria and measles, after being infected. The most serious symptom of such a disease is pneumonia, which can be potentially deadly. These virus are spreading through close personal contact and respiratory droplets. To prevent further transmission of the virus, government has issued many alerts and restriction guidelines to request people staying at home. Such alerts have greatly contributed to gradually reducing the epidemic of the Coronavirus.

At the end of May, we see fewer and fewer new confirmed cases each day. However, there are still many people infected and not yet recovered. In this study, we used data published by Government of Canada to analyze what the recovery rate curve looked like, and to predict when patients can recover and the Coronavirus would no longer be a huge threat.

2 Materials & Methods

In this section, we will present how we select the dataset and how we build the logistic regression model, in order to predict the recovery rate for patients in different provinces.

We compared data from multiple sources and found Government of Canada being the most reliable one. The main reason is that Government of Canada provides detailed yet concise provincial and territorial data inside Canada. Most other data sources focus on international data and only have one dimension for Canada (as one single nation). This contradicts to our goal, which is to explore the development phase of COVID-19's in Canada. Another reason is that data released by the government of Canada, as the official data, should be more reliable than other sources from NGOs.

In the raw data, we have done a lot of cleaning and exploratory works. First, We generated many summary statistics for the dataset, trying to understand the significance of different variables. Variables we found crucial to the analysis are province/territory name, date, number of total confirmed cases, number of total cases, number of total tested cases, number of total recovered cases, and number of deaths.

When we do the exploratory works, the greatest issue we found is that there are a lot of missing data in variable "number of recovered cases". Some of them are **N/A** and some of them are just blank in the dataset. These **N/A**s can be divided into two categories:

1. For each province, the first few rows all appear to be NAs. This missing is reasonable, because the very first rows are the very first confirmed cases in Canada. As it's impossible for those patients to be healed right away, there should be no recovered case in the first few days. Hence, we set these **N/A**s to zero.

2. There are also lots of missing points after the first recovered case. We investigated multiple cases by examining data points before and after it, and found these to be loss of records. We cannot treat those as zeros, and did not find other data sources that can fulfill this gap either. After discussion, we decided to use missing data imputation techniques given by forward selection regression to tackle this problem.[1]

Also, data entries in Nunavut, the Yukon, and the Northwest Territories are either zero or really small, so it's hard to tell the trend. In time series data, this can severely harm accuracy of the logistic regression model. Hence, though we still run the model and have the figures saved in figures folder, we will not do individual analysis on those provinces. To clarify, all data from all provinces are included in national-wide study.

The data cleaning works are carried out as described above. Next, we moved forward and conduct a forward selection study, along with correlation analysis, to select key variables that can achieve the highest predicting accuracy. We also tested on multiple combinations of features and different weights. Variables selected are date, number of confirmed case, and number of deaths. Eventually, we are able to build a well-trained logistic regression model to discover the chance of the patients got recovered in each province. [2]

3 Results

In this section, we are going to show our prediction result. We applied the final regression model to provincial and national data and generated Figure 1-6.

The blue horizontal line indicates a 95% recovery rate, which could be a sign that most of the patients have been recovered. We choose a 95% recovery rate as the dividing line because areas that reopen safely, such as New Zealand and Thailand, have a recovery rate greater than 95%. While other countries with a lower recovery rate, such as Brazil, faces a more serious epidemic when they try to reopen [3]. This line comes from a purely mathematical perspective but not from public health experts. Therefore, the line is only there for references but not an official recommendation. The red curve indicates the actual recovery rate until May 31. The green curve establishes our forecast of the upward trend in the recovery rate. Analysis on the charts will be included in Discussion section.

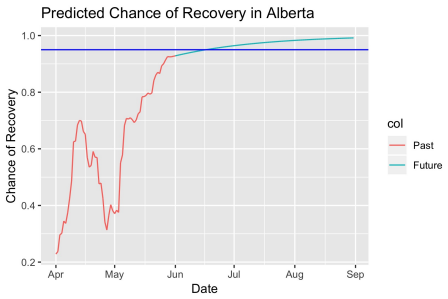

Figure 1: Predicted Recovery Rate in Albert

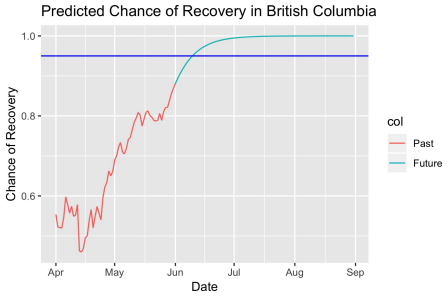

Figure 2: Predicted Recovery Rate in BC

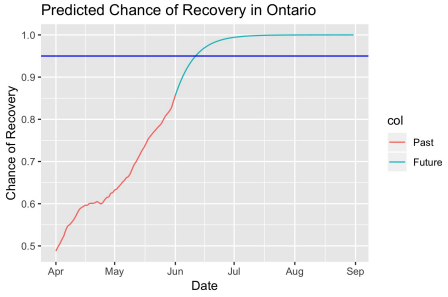

Figure 3: Predicted Recovery Rate in Ontario

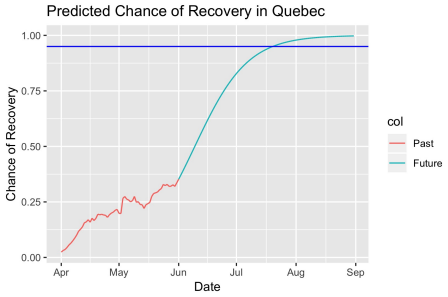

Figure 4: Predicted Recovery Rate in Quebec

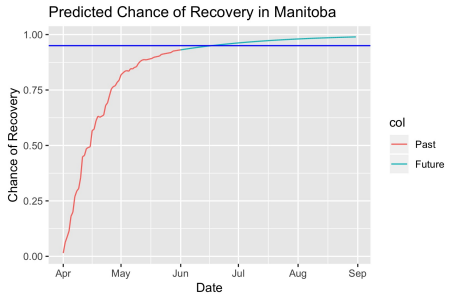

Figure 5: Predicted Recovery Rate in Manitoba

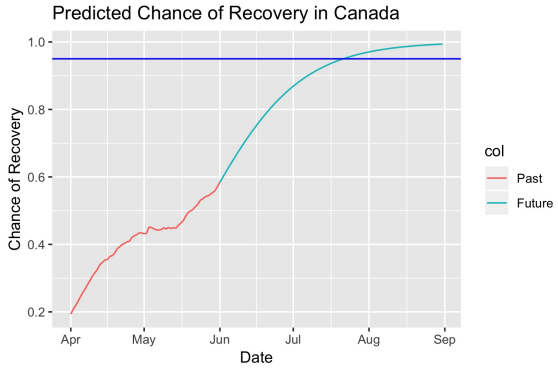

Figure 6: Predicted Recovery Rate Nationalwide

4 Discussion

In the logistic regression model we built, the number of confirmed cases, the number of deaths and the dates are the variants and the recovery rate is the response variable. Due to the limited amount of data and dataset flaws mentioned in Section 2, deep learning models like gradient boosting are not applicable. Logistic regression is chosen because it is the most suitable model for continuous variables that develop over time, and we can overcome the data flaws using regression properties. It is also highly informative and interpretable, so that we can better understand the relationship between number recovered and the variants.

Looking at the result figures, we can see that most of the patients are going to be completely cured (recovery rate approaches 100%) in mid-July or early-August. Meanwhile, the nationwide recovery rate forecast indicates that COVID-19 outbreak is very likely to be over by September.

4.1 Analysis by Regions

In the graph of recovery rates for Albert, British Columbia and Quebec (Figures 1, 2 and 4), we observed obvious fluctuations, which is caused by data flaws - missing values.

Another reason for Alberta's obvious fluctuation (Fig 1) is the small number of total cases. Consequently, some missing data may induce a large spike in the recovery rate curve. But overall, it is moving towards a completely cured community. We anticipate an acceptable reopening time in mid-June.

British Columbia (Fig 2) has also found lots of infected cases, but it has more medical resources than most other regions. Hence, its recovery rate is increasing at a faster pace than Alberta's. We predict that BC can slowly reopen starting from early June. According to news, starting May 16, gatherings with two to six guests are allowed for dinner parties and backyard barbecues in British Columbia. Hotels, resorts and overnight camping in some parks will open in June[4]. Though a little bit to fast to allow barbeques in May, overall we find this to be a safe reopen pace.

Ontario (Fig 3) has a smoother curve that other regions, since the number of samples in Ontario is much larger than other regions. The curves turned flat and even dropped for a bit at end of April, and then went back up. The reason is that Ontario shutdown starts at mid-April. The consequence started showing up after COVID-19's incubation period, which is two week. If the shutdown remains in place, we predicted that the recovery rate is able to go above 95% in mid-June. This is actually consistent with Ontario's self-isolation policy – governor has claimed that Ontario can start reopening in June in different phases[5]. Though details are not disclosed yet, districts with smaller population and higher recovery rate will open first and more densely populated districts are likely to remain in shutdown until next phase.

Comparing to the other provinces, Quebec (Fig 4) appears to be the most affected region in Canada. Though some people declare the reason to be a higher testing coverage, we did not observe that in data. Quebec has about 430k tested cases while Ontario has approximately 680k. The recovery rate is under 30 percent by May 31st, which is significantly lower than any other provinces. According to Premier Legault, "high schools, junior colleges and universities in the province will not reopen until September"[4]. This is a wise decision and we recommend non-necessity business in Quebec remains closed until early August.

Overall in Canada (Fig 6), if all the restrictions remains the same, the disease outbreak will likely to be over by late July and almost certain to be gone by September.

4.2 Limitation

There are also some limitations in this study. All limitations come from the fact that there are more factors that can influence the spread of COVID-19, such as government policy, public consciousness, and weather conditions.

These factors cannot be measured or predicted. For example, we all know shutdown policy induced positive impacts on the number of total infected cases. It is much more complicated than choosing between"shutdown" and "open", because different regions are having very different policies, and each policy comes into place in discrete scales and varied time. In terms of inability to predict, the input variable "number of confirmed cases" is facing similar difficulty. We tried to predict it based on past trends of confirmed cases, but found significant inconsistency in the data. The line before and after mid-April looks very different. The turning point is likely to be each region's shutdown start time plus COVID-19 incubation period. This makes it implausible to predict number of confirmed cases using regression. We are able to overcome this difficulty by making it a flat variable, but we cannot do the same to other immeasurable factors like government interventions.

Therefore, the inability to count in these immeasurable and uncontrollable factors can hinder the model's predicting power. Meanwhile, these are unavoidable issues for most models in real life. So as long as the model matches the reality in most circumstances, the model and the prediction is still a reliable reference for us to learn about COVID-19.

Conclusions

Using data released by the Government of Canada, we built a logistic regression model to predict future recovery rate. The epidemic is likely to calm down in late July and almost certain to be gone by September. Different regions have very different COVID-19 development curve, so all disease-related policies should be customized to meet that region's specific needs. But overall in Canada, the government measures have worked out and put the outbreak under control. We look forward to seeing regions starting to reopen slowing in their own pace.

References

[1] Therese D Pigott. A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383, 2001.

[2] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[3] Iman Ghosh. The road to recovery: Which economies are reopening?, May 2020.

[4] Brooklyn Neustaeter. The 'new normal' begins: Where each province and territory stands with reopening, May 2020.

[5] *Province considering a regional approach to reopening: Ford*. May 2020.