COVID-19 Analysis

CIV 1504 Term Project

Xiaohan Li, 1001357989

Table of Contents

Background

In December 2019, a new type of unknown pneumonia was firstly detected in Wuhan, a city in China. After two months of illness spreading, WHO named this respiratory illness as COVID-19 and confirmed that the outbreak of COVID-19 is a global epidemic. Since the disease has already affected all human-beings in every aspect, such as working, studying, and living, it is necessary to conduct some statistical research by using the daily level data in order to understand and forecast the pattern of the disease spreading[1].

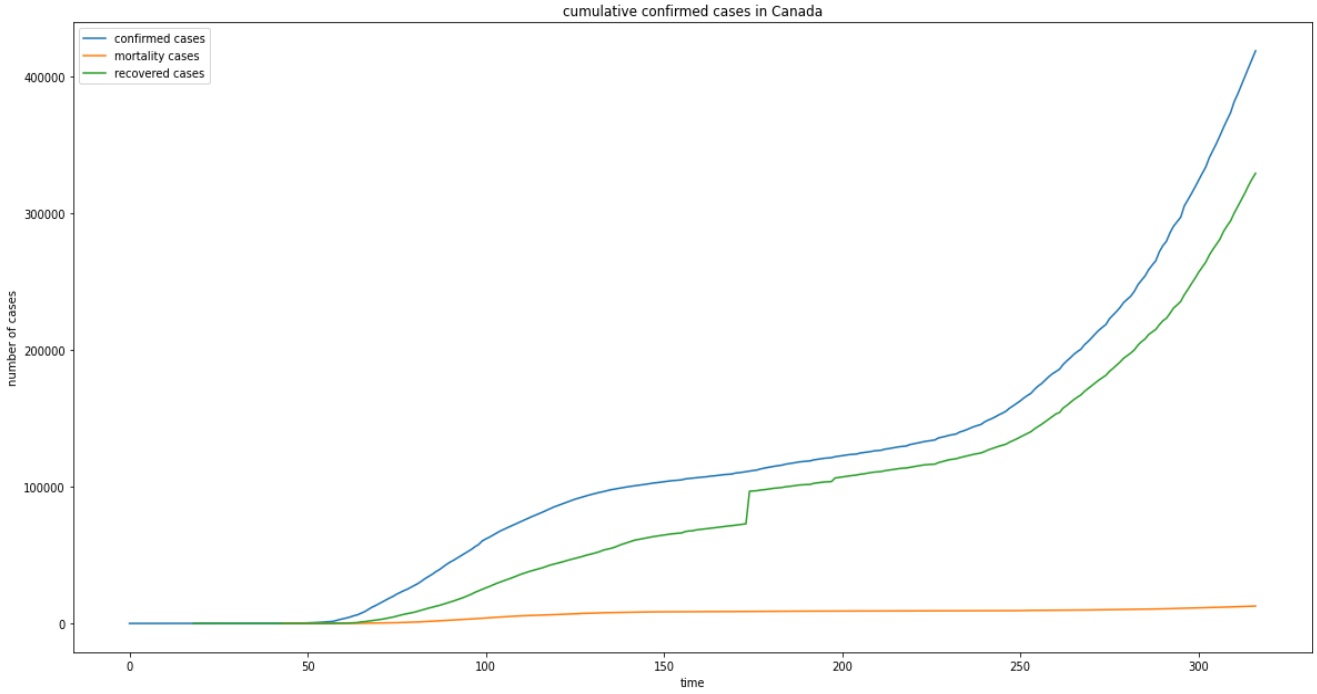
Objective

The primary goal of this project is conducting data analysis on the COVID-19 data by using some statistical techniques covered in CIV1504 courses. The data source is from Kaggle.com[2], and it includes numbers of "Confirmed", "Mortality" and "Recovered" cases in the worldwide area from January 25 to December 06. The other excel files in the data set also include the observation date and province/country, which could support the data analysis. Based on the dataset information, 4 significant questions related to the data of Canada will be addressed:

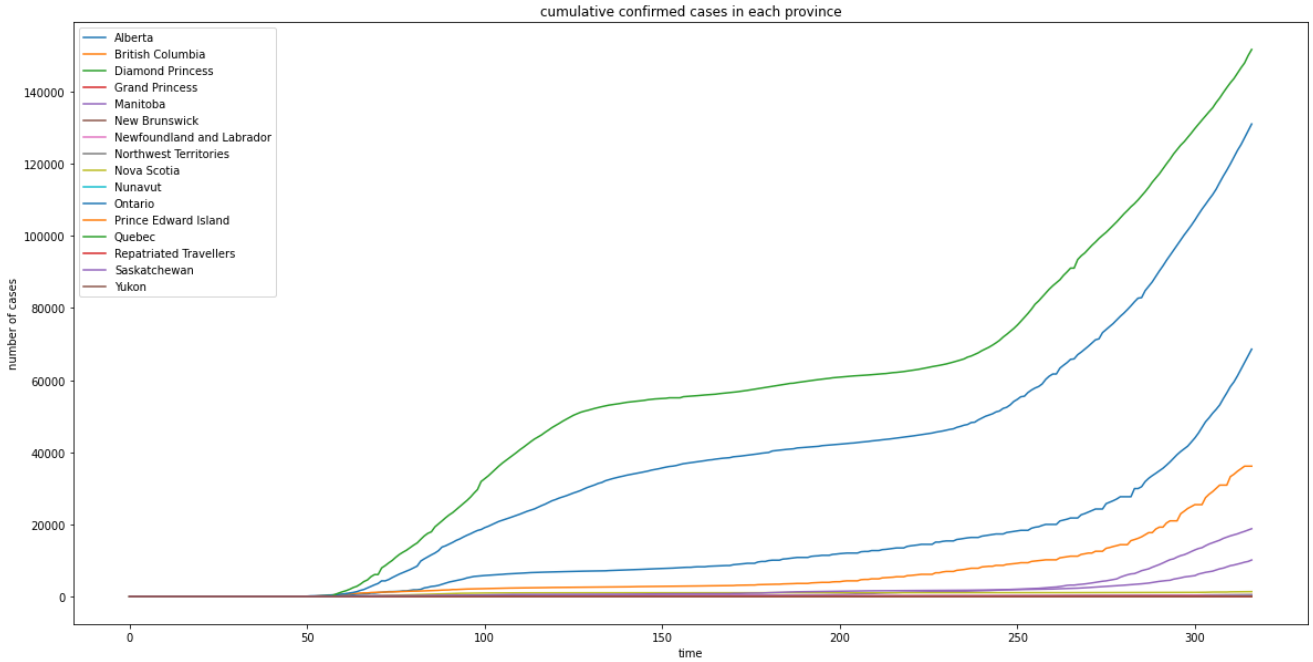
a) Verify that the daily confirmed number of Quebec is larger than Ontario's
b) Is the confirmed cases data of Ontario a typical sample that could represent entire Canada?
c) Does the location affect the daily confirmed cases in Canada?
d) Build a regression model to predict the number of confirmed patients

Analysis

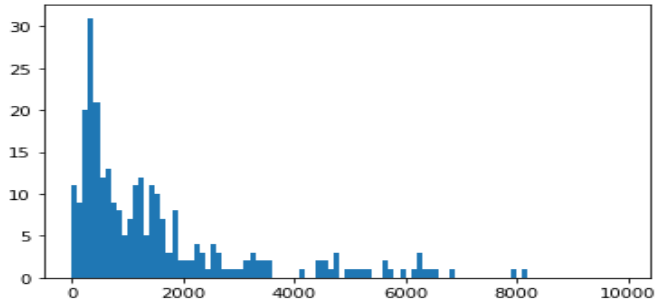
Firstly, it is essential to investigate the general information covered in the data sets. From the figure below, the general trend of the number of patients is increasing over time. In this project, the researching time is from January 25 to December 06. Until the end of the period, the total number of confirmed cases in Canada is 418669, and the recovered and mortality cases are 329138 and 13590, respectively. The increasing speed of the confirmed cases experienced 3 obvious changes that approximately happed at 60 days, 120 days and 240 days. Compared to the rapidly increasing numbers of confirmed cases and recovered cases, the number of mortality cases did not have many undulations.



Secondly, at the provincial level, each provinces' cumulative confirmed cases can be found in the second figure below. It is clear that Ontario and Quebec are the two biggest provinces with the largest number of cumulative confirmed cases at the end of December-06. Therefore, it is essential to analyze these two provinces in terms of daily confirmed cases.

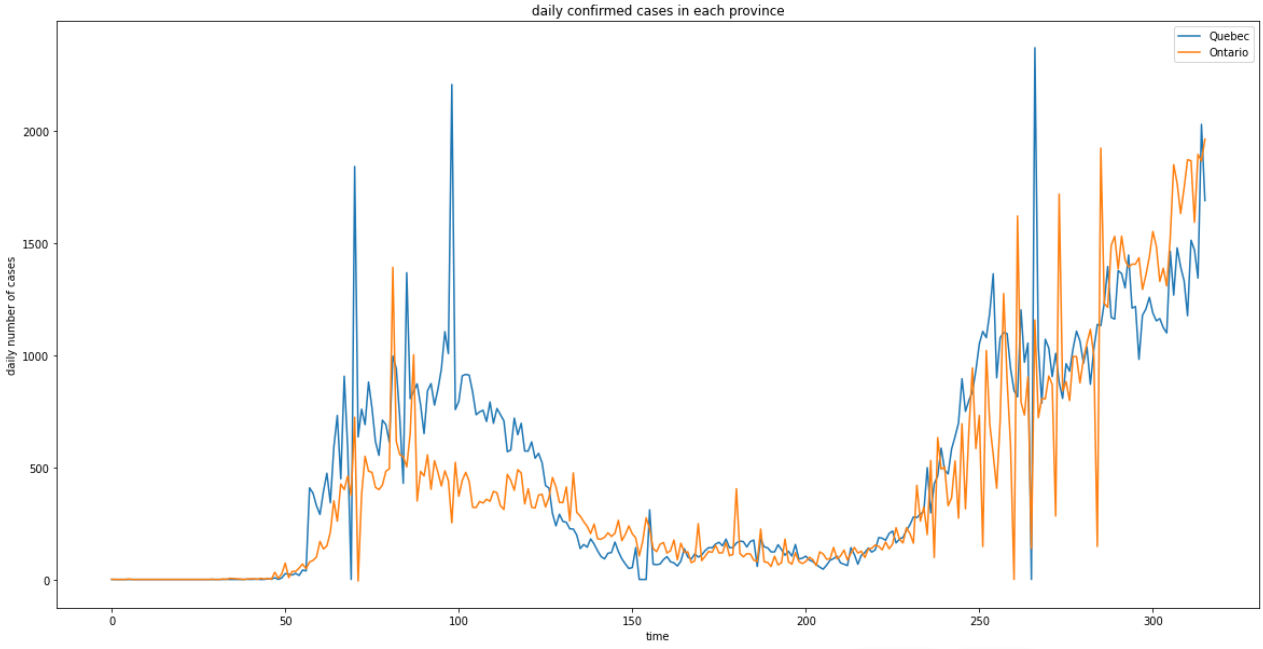


One major concern in this analysis is that the number of daily confirmed cases follows an unknown distribution type. The figure below shows the general distribution shape of daily confirmed cases from January-25 to December-06(which is 317 days in total). However, it also cannot be concluded that the daily number does not follow the normal distribution since the epidemic situation has not ended, and the daily confirmed cases would be extremely increasing if COVID-19 cannot be properly controlled. Therefore, one assumption is that the number of daily confirmed cases of COVID-19 follows a normal distribution to simplify the analysis.



Comparison of the Ontario data and the Quebec data

Comparing the other 14 provinces, Ontario and Quebec are the top 2 provinces with the largest cumulative number of confirmed cases. The daily confirmed cases of each province are plotted in the figure below, and it can be noticed that the general trends of the two provinces are similar. Since the cumulative number of Quebec confirmed cases are larger than Ontario's, it is essential to verify whether the daily confirmed number of Quebec is still larger than Ontario's while assuming the significant level is 10%.



Firstly, test the difference in variances:

$$H_0: \sigma_O^2 = \sigma_Q^2$$

$$H_1: \sigma_O^2 \neq \sigma_Q^2$$

As stated above, the distribution of the daily confirmed cases is following a normal distribution. Therefore, it is reasonable to assume both daily confirmed cases of Ontario and Quebec are normally distributed.

Under H_0, we have $\frac{S_O^2}{S_Q^2} \sim F_{n_O-1,n_Q-1}$

$$F_0 = \frac{S_O^2}{S_Q^2} = \frac{227886.15}{234173.20} = 0.9732$$

While the F-distribution table did not show the exact F-statistics when degree of freedom equals 316, this value should be estimated smaller than $F_{200,200}$ and larger than $F_{500,500}$, which are 1.26 and 1.16, respectively. Therefore, it is reasonable to assume 1.22 as the F-statistics when degree of freedom equals 316.

$$F_{n_O-1,n_Q-1;\frac{\alpha}{2}} = F_{316,316;0.05} = 1.22$$

$$F_{n_O-1,n_Q-1;1-\frac{\alpha}{2}} = F_{316,316;0.95} = \frac{1}{1.22} = 0.82$$

Because 0.9732 is in the range of [0.82, 1.22], it can be concluded that two variances do not have a significant difference under a 10% level of significance.

Secondly, test the mean value of two samples by using a one-tail test.

$$H_0: \mu_O = \mu_Q$$

$$H_1: \mu_O < \mu_Q$$

$$s = \sqrt{\frac{(n_O-1)s_O^2 + (n_Q-1)s_Q^2}{(n_O+n_Q-2)}} = 481.42$$

$$t_0 = \frac{\overline{x_Q} - \overline{x_O}}{s\sqrt{\frac{1}{n_Q}+\frac{1}{n_O}}} = \frac{479.74 - 414.44}{481.42\sqrt{\frac{2}{316}}} = 1.71$$

where the test statistic $t_{632;0.1} \cong 1.28$ which is smaller than 1.71

Therefore, the null hypothesis is rejected at a 10% level of significance. In summary, it is confident to state that the number of Quebec's daily confirmed cases is larger than the number of Ontario's daily confirmed cases throughout the researching period.

Comparison of the Ontario data and the country-level data

According to the cumulative confirmed cases figure above, the Ontario data is one of the major portions of the Canada daily data. In this case, it is important to test whether the number of Ontario confirmed cases and the number of Canadian confirmed cases are from the same population. Another hypothesis test can be applied to verify the statement. Firstly, the test for difference of variances is performed:

$$H_0: \sigma_O^2 = \sigma_C^2$$

$$H_1: \sigma_O^2 \neq \sigma_C^2$$

As stated above, the distribution of the daily confirmed cases is following a normal distribution. Therefore, it is reasonable to assume daily confirmed cases of Canada are normally distributed.

Under H_0, we have $\frac{S_O^2}{S_C^2} \sim F_{n_O-1,n_C-1}$

$$F_0 = \frac{S_O^2}{S_C^2} = \frac{227886.15}{2566891.62} = 0.089$$

$$F_{n_O-1,n_C-1;\frac{\alpha}{2}} = F_{316,316;0.05} = 1.22$$

$$F_{n_O-1,n_C-1;1-\frac{\alpha}{2}} = F_{316,316;0.95} = 0.82$$

Because 0.089 is not in the range of [0.82, 1,22], the null hypothesis is rejected in favor of the alternative hypothesis. It also indicates that Ontario daily confirmed cases could not represent the trend of the entire country. This is because the country-level data is a composite of information from fourteen provinces, and it cannot be described by a single variance of a single province's data.

Analysis of Variances on location factor

From the figure above, different provinces have different numbers of cumulative confirmed cases. In this case, it is essential to test if the location is a significant factor that could affect the daily confirmed cases. However, the location factor also includes other potential factors such as population density, environment, and climate, which may also affect the result. In addition to that, different provincial governments are having different attitudes and proceeding different policies regarding COVID-19. It is difficult to quantify the effects of each potential factor. Therefore, it is reasonable to assume that all potential factors related to the location can be neglected.

Analysis of Variances will be applied to this test. Firstly, Quebec, Ontario, Alberta, and British Columbia which have the most cumulative confirmed cases are chosen for the test. Each province has 317 data points representing the daily confirmed cases number. The ANOVA table is shown below.

Variation Source	Sum of Sq.	D.O.F.	Mean Square	F-ratio Test
Btw Samples	27366255.55	3	9122085.18	51.75
Within Samples	2.23*10^8	1265	176275.29	

$$H_0: \delta_Q = \delta_O = \delta_A = \delta_B = 0$$

$$H_1: \delta_{locations} \neq 0$$

$$F_0 = 51.75 \; and \; F_{3,1104;0.1} = 3.78$$

Since 51.75 is greater than 3.78, there is a significant variation in location factor at a 10 percent significant level. While the testing result corresponds to the observation, it is important to notice that the potential factors mentioned above also played an important role in resulting this significant variation.

Regression analysis on time series data

Compared to the cumulative confirmed cases, the daily confirmed cases have numerous fluctuations throughout the researching period. These fluctuations add difficulties in building an accurate regression model for predicting future daily confirmed cases. Therefore, it is feasible to conduct a regression analysis by using cumulative time series data.

Firstly, a one-variable linear regression model is constructed, and the general form is shown below:

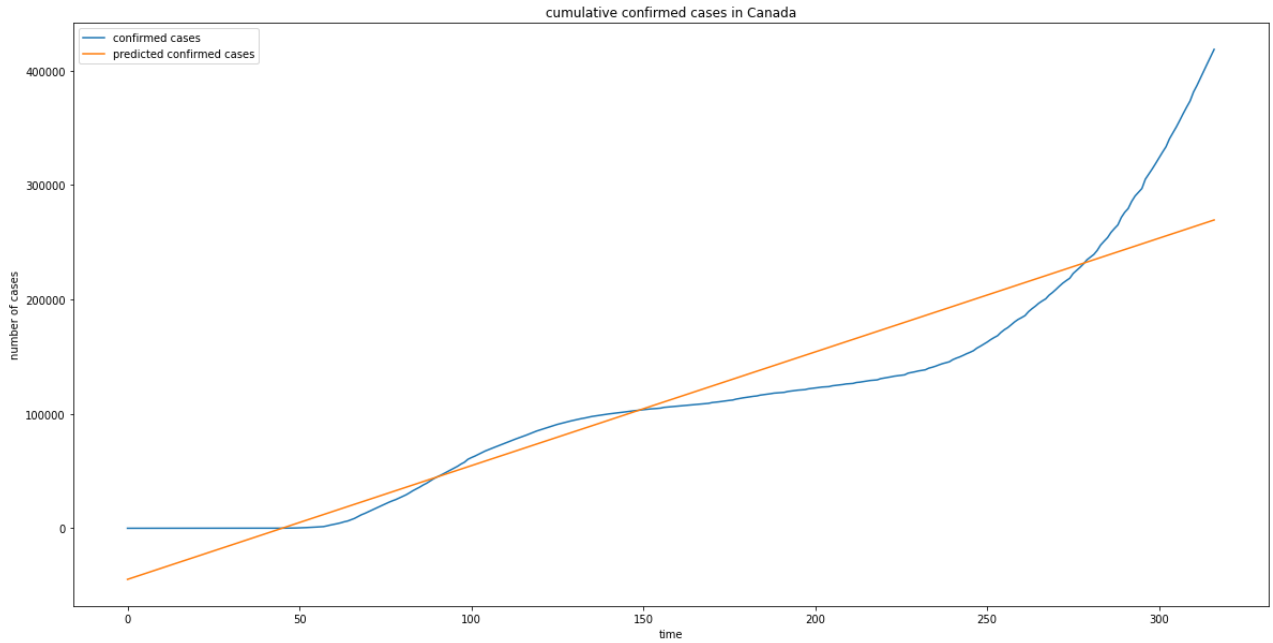
$$y_i = \theta_1^* + \theta_2^* x_i + \varepsilon_i$$

Since the cumulative confirmed cases are starting from 0, there should be no interception on the y axis. However, the constant item is kept for a better comparison with the other higher-order models. Because time is the only independent variable in this model, the structure of X is composite of 1 column and 317 rows. After adding a column of 1 into the independent matrix, the $\hat{\theta}$ is calculated as follow:

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

$$\hat{\theta} = \begin{bmatrix} -44682.47 \\ 994.37 \end{bmatrix}$$

The detailed matrix calculation is conducted by python notebook which can be found in the appendix file. Based on the result, this one independent variable linear model has 994.37 as the slope and -44682.47 as interception. Compared to the actual value of the cumulative confirmed cases data, the model is shown below:



For this simple linear model, the error variance is calculated as:

$$\hat{\sigma}^2 = \frac{\sum_i r^2}{n - p} = 1275085221.96$$

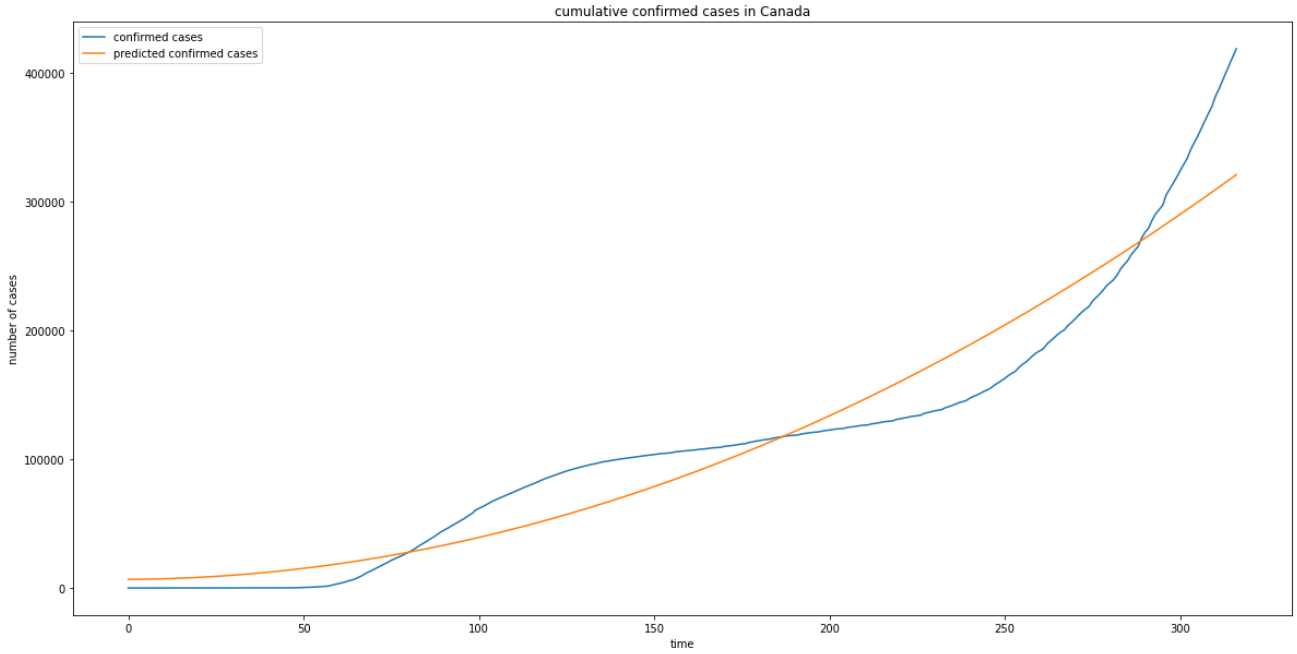
As time increases, the bias between the actual confirmed cases and the predicted cases would become larger

and larger. Therefore, the model would not make a reliable future prediction based on the large mean squared error. However, this model correctly describes the trend of the time series data. Therefore, another quadratic polynomial model is constructed to fit the data. The general form of this model is shown below.

$$y_i = \theta_1^* + \theta_2^* x_i + \theta_3^* x_i^2 + \varepsilon_i$$

Similar to the first model, the least square estimator of θ is calculated, and the model figure is shown below.

$$\hat{\theta} = \begin{bmatrix} 6672.50 \\ 16.18 \\ 3.10 \end{bmatrix}$$



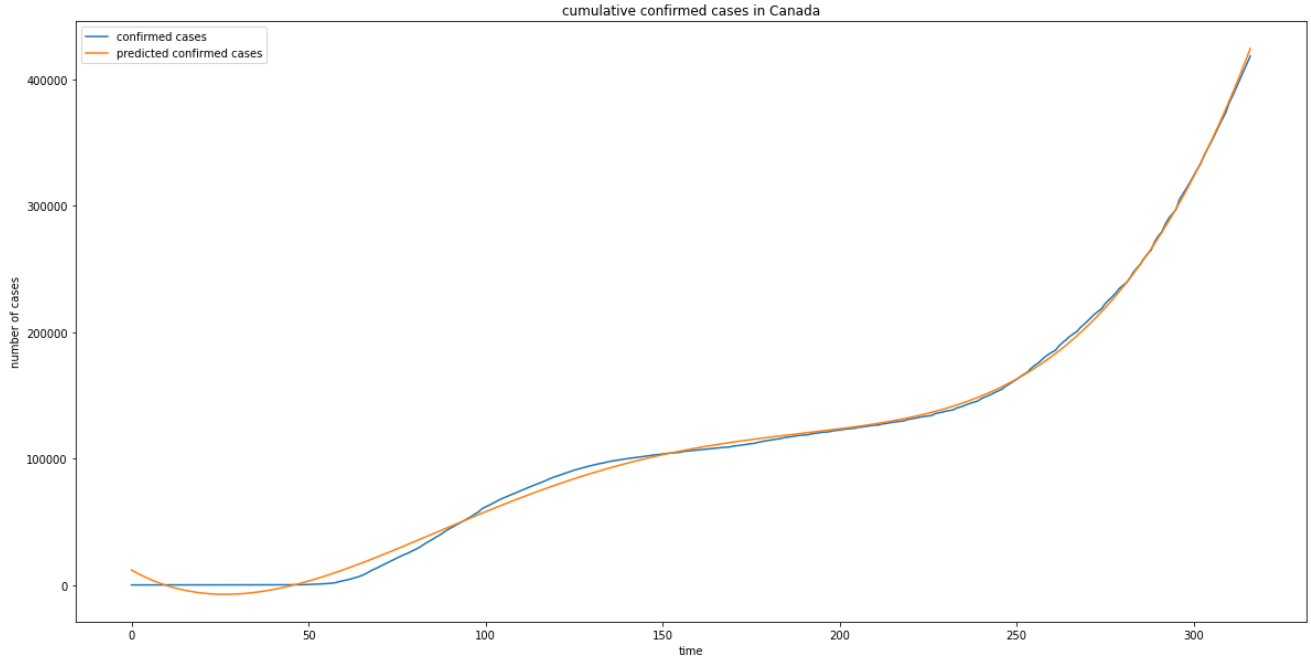
While this model's mean squared error is 734126770.10 which is smaller than the MSE of the first model, the quadratic model is still simple and cannot fit the fluctuations of the data. In this case, a higher degree polynomial model is required. The minimum possible degree of the polynomial can be calculated as the turning point of the data plus one [3].

Since there are 3 turning points on the actual data, the minimum possible degree of the polynomial should be 4. Therefore, the general form of the model is:

$$y_i = \theta_1^* + \theta_2^* x_i + \theta_3^* x_i^2 + \theta_4^* x_i^3 + \theta_5^* x_i^4 + \varepsilon_i$$

The result of optimal parameters is shown as follow:

$$\hat{\theta} = \begin{bmatrix} 11820.36 \\ -1592.37 \\ 38.10 \\ -0.21 \\ 0.00039 \end{bmatrix}$$



From the figure above, it can be seen that the 4th degree polynomial model can approximately cover the major features of the data. Compared to the two simple models constructed above, the mean square error of this model is 18335921.41 which is around 40 times smaller than the quadratic model. In order to test whether the model is valid, the coefficient of determination, R^2, is calculated as below.

$$SS_p = \left(\hat{Y}\right)^T Y - \frac{(\sum y_i)^2}{n} = 3020602158440.70$$

$$SS_E = Y^T Y - \left(\hat{Y}\right)^T Y = 5775815251.84$$

$$R^2 = \frac{SS_p}{SS_p + SS_E} = 0.998$$

Based on the value of R^2, it indicates that 99.8% of variability of the Canadian cumulative confirmed cases are explained. Therefore, it is reliable to make future predictions. For example, the cumulative confirmed cases of Canada on December/15/2020 is 478992. By using the model, the time is equal 326 days from the researching starting point, and the predicted confirmed cases are:

$$y_{predict} = (1 \quad 326 \quad 326^2 \quad 326^3 \quad 326^4) \begin{bmatrix} 11820.36 \\ -1592.37 \\ 38.10 \\ -0.21 \\ 0.00039 \end{bmatrix} \approx 505256$$

Conclusion

At the beginning of this project, four statistical researching topics were addressed in order to obtain a meaningful insight into COVID-19 data of Canada. Based on the analysis, it can be concluded that the confirmed number of Quebec is still larger than Ontario's in terms of both daily and cumulative level. Comparing Ontario data to the Country-level data, the daily confirmed cases of Ontario are not typical enough to represent the whole country. After conducting an ANOVA analysis, the location factor is found to have a

significant variation at 10% significant level. Lastly, a 4^{th} degree polynomial model is constructed, and the R^2 coefficient is calculated as 0.998. Another important thing needs to be discussed is that the regression model's degree could be higher than 4, but it might cause overfitting problems. How to find a balance between overfitting and underfitting is critical research.

Reference

[1] “How COVID-19 spread”, Available: https://cdc.gov

[2] “Novel Corona Virus 2019 Dataset”, Available: https://kaggle.com

[3] “Degree, Turnings and ‘Bumps’”, Available: https://purplemath.com

Appendix A

All the data used for analysis is in these four excel files from Kaggle.com. They can be found in the submission package.

1) cases_timeseries_canada.csv

2) time_series_covid_19_confirmed.csv

3) mortality_timeseries_canada.csv

4) recovered_timeseries_canada.csv

Appendix B

All the data manipulation is accomplished in CIV 1504 TermProject.ipynb. This can be found in the submission package.