Coronavirus Disease (COVID-19) Analysis

Kehang Li

Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease and has caused global pandemic. As of 4th May, the total number of COVID-19 confirmed cases in United States is 1.19 million and corresponding death toll is 68797. The agent of COVID-19 is SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2), which are enveloped, non-segmented and positive-sense RNA viruses, belong to broad coronaviruses family that typical causing symptoms related to human respiratory disease. Forecasts of COVID-19 that aim to predict changes in national and state-level cumulative COVID-19 positive cases are very beneficial to help inform public health decision making.

National Level Analysis

1. **Treatment of data**

The original data obtained from the COVID-19 datasets downloaded from the website covidtracking.com on Apr 8, 2020. Missing value and outliers are not detected in the “positive” column and “date” sequence. The descriptive statistics of USA cumulative count of positive cases are examined to check the data range and skewness, the histogram indicates that data distribution is highly right skewed with large spread. Divide dataset into training data that extracted date from March 16th to March 29th and test data that extracted date from March 30th to April 8th.

1. **Variables**

In the national level regression analysis, dependent variable is the USA cumulative count of COVID-19 positive cases, independent variable is the ordinal variable date.

1. **Regression model**
   1. **Model 1: Simple linear regression model**

Building the simple linear regression analysis to examine the relationship between USA cumulative count of COVID-19 positive cases and date based on the training data. Preliminary regression results indicate that USA cumulative count of COVID-19 cases is positively related to the date (coefficient b = 1.018e+04, SE = 8.794e+02, p-value < 7.20e-08). Model 1 is significant (F (1, 12) = 134, p-value = 7.203e-08) and can explain 91.78% the variability of dependent variable.

However, the OLS regression assumptions including linearity, homoscedasticity, normality not met. Model diagnostics show that model residuals have unequal variance, non-linear relationship between variables exists, residuals are not normally distributed with sum to zero.

* 1. **Model 2: Transformed linear regression model**

According to the model 1 summary above, apply box-cox transformation (lam = 0.303) to rescale and normalize highly skewed dependent variable in order to satisfy the OLS regression assumptions. Build model 2 by simple linear regression between transformed USA cumulative positive cases and date. As the table1 showed below, after data transformation, linearity, normality and homoscedasticity assumptions are no longer violated by model 2 and thus the estimates and interpretation of the regression can be regarded as unbiased.

Table 1. Model diagnostics of OLS assumptions results

|  |  |  |
| --- | --- | --- |
| Assumption | Model 1 | Model 2 |
| Linearity |  |  |
| Normality |  |  |
| Homoscedasticity |  |  |

1. **Results**

**4.1 Interpretation of regression**

Model 2 regression results indicate that transformed USA cumulative count of COVID-19 cases is positively related to the date (coefficient b = 6.22, SE = 5.894e-02, p-value < 2e-16). The 95% confidence interval around the regression coefficient is narrow (6.091, 6.348), indicating good precision. Model 2 is significant (F (1, 12) = 1.114e+04, p-value = 2.2e-16) and the association is relatively strong, with 99.89% of the variance in transformed USA cumulative count of COVID-19 positive cases accounted for by date. The significance of coefficient and model both improved and R square is higher compared to model 1, indicating that transformed model fits the data better.

Table 2. Summary of national level regression results

| Independent variable | Model 1: | | | Model 2: | | |
| --- | --- | --- | --- | --- | --- | --- |
| Simple linear regression | | | Transformed linear regression | | |
| B | Std. error | P value | B | Std. error | P value |
| Intercept | -1.87e+08 | 1.61e+07 | 7.2e-08\*\*\* | -1.14e+05 | 1.08e+03 | < 2e-16\*\*\* |
| Date | 1.02e+04 | 8.79e+02 | 7.2e-08\*\*\* | 6.22 | 5.89e-02 | < 2e-16\*\*\* |
| Adjusted R square | 0.911 | | | 0.998 | | |
| F-statistic | F (1, 12) = 134\*\*\* | | | F (1, 12) = 1.11e+04\*\*\* | | |
| \*\*Significant at 0.05, \*significant at 0.1, \*\*\*significant at 0.001 | | | | | | |

**4.2 Prediction**

Build model 3 by linear regression based on the transformed training data that extracted date from March 23rd to March 29th. Using box-cox transformation (lam = 0.263) to rescale and normalize highly skewed dependent variable.

| Date | Model 2:  Training on March 16th to March 29th | | | Model 3:  Training on March 23rd to March 29th | | |
| --- | --- | --- | --- | --- | --- | --- |
| Fitted value | Lower limit | Upper limit | Fitted value | Upper limit | Lower limit |
| 2020-03-30 | 165672.3 | 156210.1 | 175526.9 | 166505.8 | 159576.0 | 173655.1 |
| 2020-03-31 | 194245.7 | 183390.7 | 205540.7 | 195882.0 | 187375.1 | 204670.4 |
| 2020-04-01 | 226086.1 | 213676.3 | 238989.6 | 228913.5 | 218498.7 | 239689.8 |
| 2020-04-02 | 261394.7 | 247254.9 | 276088.6 | 265880.8 | 253198.0 | 279025.9 |
| 2020-04-03 | 300373.7 | 284317.0 | 317056.0 | 307073.8 | 291732.7 | 323001.8 |
| 2020-04-04 | 343235.6 | 325055.4 | 362112.7 | 352791.6 | 334369.9 | 371951.1 |
| 2020-04-05 | 390183.8 | 369665.6 | 411482.6 | 403342.3 | 381384.5 | 426218.4 |
| 2020-04-06 | 441431.9 | 418345.3 | 465392.0 | 459043.0 | 433058.8 | 486159.1 |
| 2020-04-07 | 497194.2 | 471294.5 | 524069.8 | 520219.8 | 489682.4 | 552139.0 |
| 2020-04-08 | 557687.3 | 528715.5 | 587747.5 | 587207.8 | 551552.5 | 624534.6 |
| MSPE | 4186229169 | | | 6244159238 | | |
| MSPE = Mean squared prediction error | | | | | | |

Table 3. Fitted values and prediction intervals of USA cumulative positive cases

Predict the cumulative count of COVID-19 positive cases between March 30th and April 8th by model 2 and model 3, compare the fitted values and 95% prediction intervals with true values of positive cases on each date by squared prediction error.

Compare the mean squared prediction errors between model 2 and model 3, based on the prediction values that are transformed to original scale. Model 2 has lower MSPE (model 2 MSPE = 4186229169 < model 3 MSPE = 6244159238) than model 3, indicating that model 2 has smaller deviations to true values and more accurate predictions of USA cumulative positive cases. The reason why model 2 is more accurate should be accounted for its larger training data, the training data for model 2 has 7 more data points than that of model 3, containing more information and thus improve the model fitness.

From the visualization of model 2 and model 3 predictions (Fig 2), it’s obvious that model 3 has higher prediction values of USA cumulative positive cases (red lines and red points) than model 2 (blue lines and blue points), and more biased away from true values than model 2.

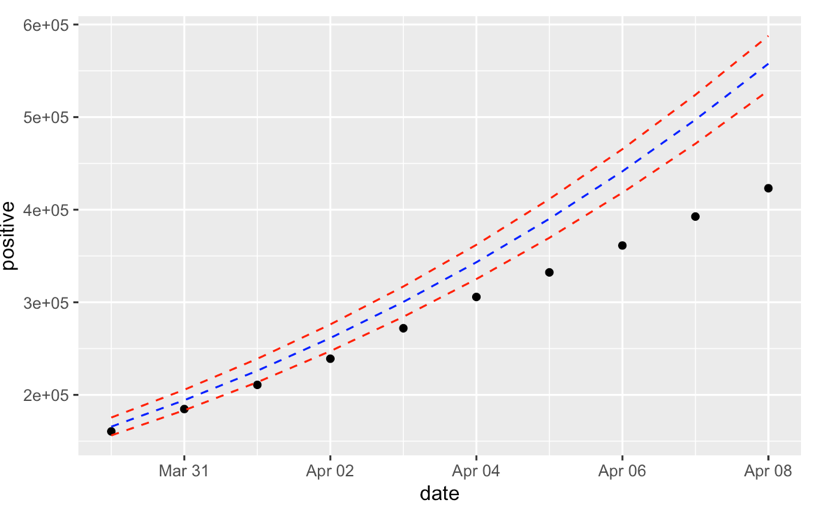


Fig1. Model 2 fitted values and 95% prediction intervals of USA cumulative positive cases

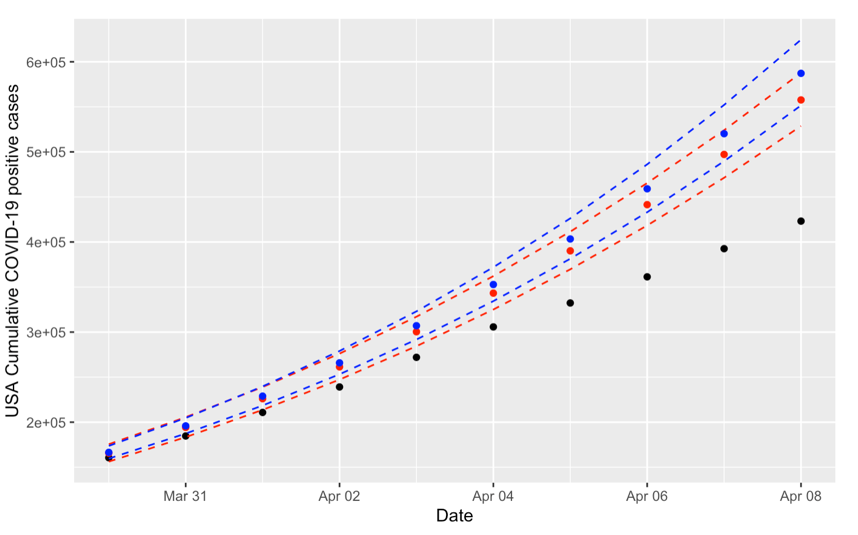


Fig2. Model 2 and model 3 fitted values and prediction intervals of USA cumulative positive cases

State Level Analysis

1. **Treatment of data**

The original data obtained from the COVID-19 datasets downloaded from the website covidtracking.com on Apr 8, 2020, contains the daily statistics for various measures for each state and US territory. Filter the data of states that include New York (NY), New Jersey (NJ), California (CA), Michigan (MI), Massachusetts (MA), which are top 5 states with largest cumulative count of COVID-19 positive cases on March 29th. No missing values and outliers are detected in the data needed for the analysis, cumulative count of COVID-19 positive cases in each state is highly right-skewed according to the descriptive statistics and histograms. Divide dataset into training data that extracted date from March 23rd to March 29th and test data that extracted date from March 30th to April 8th.

1. **Variables**

In the state level regression analysis, the dependent variable is state cumulative count of COVID-19 positive cases, independent variables include ordinal variable date and categorical predictor state, this variable has 5 groups: NY, NJ, CA, MI, MA.

1. **Regression model**

Using box-cox transformation (lam = 0.141) to rescale and normalize highly skewed dependent variable and build model 4 by simple linear regression between transformed state cumulative positive cases and predictors including date and state. Model diagnostics of assumptions imply that transformed linear regression satisfied OLS regression assumptions including linearity, normality and homoscedasticity.

1. **Results**

**4.1 Interpretation of regression**

Model 4 regression results indicate that transformed state cumulative count of COVID-19 cases is positively related to the date (coefficient b = 0.786, SE = 2.778e-02, p-value < 2e-16), one date increase is associated with 0.786 unit increase in transformed state cumulative positive cases, controlling for the state variable. The 95% confidence interval around the regression coefficient is narrow (0.729, 0.843), indicating good precision. Model 4 is significant (F (5, 29) = 1072, p-value < 2.2e-16) and the association is relatively strong, with 99.37% of the variance in transformed state cumulative count of COVID-19 positive cases accounted for by date and state variable.

Define California (CA) as reference state group. The coefficient of state variable MA factor is -0.9415, represents that transformed state cumulative positive cases in Massachusetts is 0.9415 unit lower than that in California on the same date；the coefficient of state variable MI factor is -0.3645, represents that transformed state cumulative positive cases in Massachusetts is 0.3645 unit lower than that in California on the same date; the coefficient of state variable NJ factor is 2.404, represents that transformed state cumulative positive cases in New Jersey is 2.404 unit higher than that in California on the same date; the coefficient of state variable NY factor is 9.210, represents that transformed state cumulative positive cases in New York is 9.210 unit higher than that in California on the same date. All the coefficients are significant at 95% confidence level.

Table 4. Summary of model 4 regression results

| Independent variable | Model 4: | | | |
| --- | --- | --- | --- | --- |
| Transformed linear regression | | | |
| B | Std. error | t value | P value |
| Intercept | -1.44e+04 | 5.09e+02 | -28.266 | < 2e-16 \*\*\* |
| Date  State MA  State MI  State NJ  State NY | 0.786 | 2.78e-02  1.76e-01  1.76e-01  1.76e-01  1.76e-01 | 28.295  -5.359  -2.075  13.681  52.421 | < 2e-16 \*\*\*  9.35e-06 \*\*\*  0.047 \*  3.52e-14 \*\*\*  < 2e-16 \*\*\* |
| -0.942 |
| -0.365 |
| 2.404 |
| 9.210 |
| Adjusted R square | 0.9937 | | | |
| F-statistic | F (5, 29) = 1072 \*\*\* | | | |
| \*\*Significant at 0.05, \*significant at 0.1, \*\*\*significant at 0.001 | | | | |

**4.2 Prediction**

Predict the cumulative count of COVID-19 positive cases in each state between March 30th and April 8th by model 4, compare the fitted values and 95% prediction intervals with true values of positive cases on each date by squared prediction error. Calculate the mean squared prediction error of each state and compare the precision of predictions.

Table 5. Fitted values and prediction intervals of cumulative positive cases in each state

| Date | Model 4:  Training March 23rd to March 29th | | | | |
| --- | --- | --- | --- | --- | --- |
| CA | MI | MA | NJ | NY |
| 2020-03-30 | 8003.5 | 7220.1 | 6114.5 | 15240.8 | 72399.2 |
| 2020-03-31 | 9944.6 | 8999.3 | 7660.7 | 18587.3 | 84934.7 |
| 2020-04-01 | 12276.8 | 11142.7 | 9531.4 | 22546.2 | 99290.0 |
| 2020-04-02 | 15063.8 | 13710.3 | 11781.3 | 27208.2 | 115680.6 |
| 2020-04-03 | 18377.5 | 16770.3 | 14472.9 | 32674.9 | 134342.1 |
| 2020-04-04 | 22298.5 | 20398.9 | 17676.2 | 39059.0 | 155531.5 |
| 2020-04-05 | 26917.1 | 24682.1 | 21470.2 | 46485.8 | 179528.7 |
| 2020-04-06 | 32334.2 | 29715.6 | 25943.1 | 55093.8 | 206637.9 |
| 2020-04-07 | 38661.8 | 35606.3 | 31193.7 | 65036.1 | 237189.4 |
| 2020-04-08 | 46024.6 | 42472.6 | 37331.7 | 76481.1 | 271540.8 |
| MSPE | 209466042 | 107138808 | 96893270 | 158225864 | 3732561929 |
| Ratio | 1.2629 | 0.7755 | 0.8985 | 0.3960 | 0.5678 |

Ratio = Root MSPE / Mean cumulative positive cases in each state

According to table 5, compare MSPE of model 4 predictions in each state, NY state has largest prediction error of cumulative positive cases (MSPE = 2195916410), meanwhile MA state has smallest prediction error of cumulative positive cases (MSPE = 57018729), indicating that when predicting smaller cumulative positive cases, the precision and accuracy of model 4 is higher. However, since the prediction error is associated with scale of true value, ratios of root MSPE to mean cumulative positive cases from March 23rd to April 8th in each state are applied to clarify the bias of fitted value more obviously. From this aspect, NJ state has smallest prediction bias while CA state has largest prediction bias in model 4.

As fig 3 showed below, model 4 fitted values biased away from true value as the date grew, prediction errors for each state increased significantly after the date of April 3rd, indicating that model 4 has higher precision when independent variable date closer to training data date. Except for the limit of model training, another reason is that some changes in external factors happened, resulting that model 4 predictions less applicable to the real data and situation.

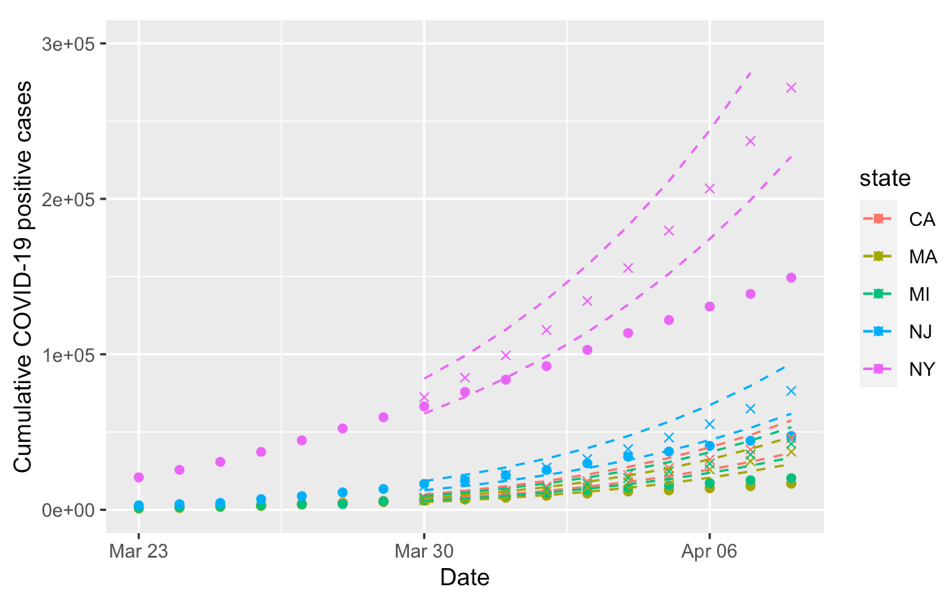


Fig3. Model 4 fitted values and prediction intervals of cumulative positive cases in each state

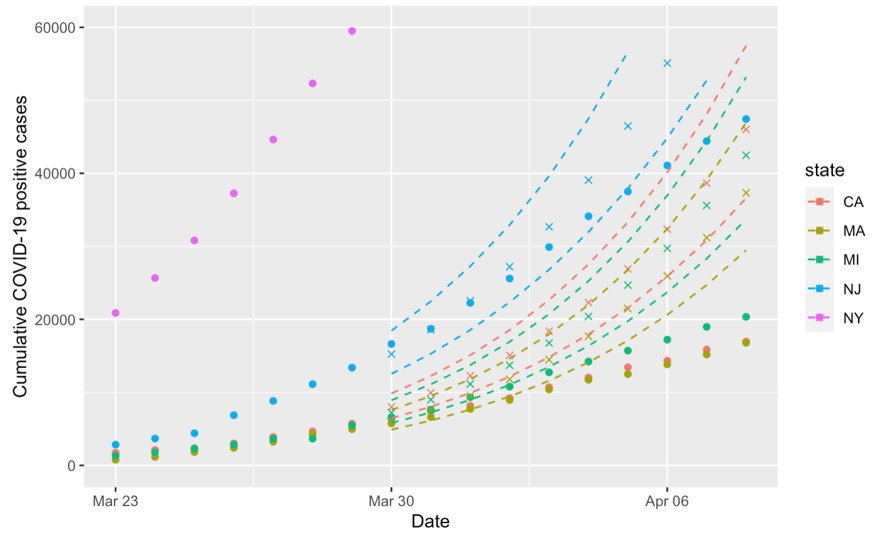


Fig4. Partial view of model 4 predictions in each state cumulative positive cases

Discussion

For national level analysis, the relationship between date and USA cumulative count of COVID-19 positive cases are non-linear, there are 6.22 units increase in box-cox transformed (lam = 0.303) cumulative positive cases every day, compared to the previous date. Reverse the transformation, the functional relationship between date and USA cumulative positive cases is similar to the cubic function, and the derivative of the function gradually becomes larger as the date grows, to be more specific, USA cumulative positive cases rate of growth will become larger and larger over time, according to the characters of optimal linear regression model 2.

For state level analysis, the relationship between date and state cumulative count of COVID-19 positive cases are non-linear. In each, there are 0.768 units increase in box-cox transformed (lam = 0.141) cumulative positive cases every day, compared to the previous date. Reverse the transformation, the relationship between the date and state cumulative positive cases is similar to septic function (7th degree polynomial), it means that the growth rate (deviation) varies from state to state. According to model 4 regression, higher intercept leads to the larger growth rate in state cumulative positive cases and baseline value. Therefore, for the cumulative positive cases in a state, NY state has largest growth rate and baseline value, MA state has smallest growth rate and baseline value.

Both the two types of analysis have similar limitations that models can not reflect the dynamic changes of external factors. At the first stage of COVID-19 pandemic, no efficient social-distancing policy introduced, leading to the exponential increase in cumulative positive cases, and applied as training data for the regression models. As the pandemic deteriorated, US government started to adopt certain isolation related measures to slow down COVID-19 spread, therefore, the growth rate of cumulative positive cases is no longer consistent with that in first stage, also the training data part. As a result, the fitted models are less accurate and have higher prediction error as the date went by. New regression models with larger training data interval should be conducted to adjust the limitations.