Exploration and Criticization of Linear Models in Covid-19 Pandemic Statistics

Echo Zheng
chaoyizheng@ucsb.edu
University of Santa Barbara
Instructor Michael Gu

Abstract

Based on the data collected from the real Covid-19 pandemic statistics, this research paper is going to focused on the construction and assessment of the linear models for the increasing trend of Covid-19 pandemic statistics. Rather than focusing on the data itself, there would be some more explorations about any variables that might affect the result of the confirmed cases. For example, how the location, time, and the infectious period would influence the resulting confirmed proportion. I may collect random samples from the population with the information about whatever variables I am going to analyze. Through constructing the linear regression model between the variables and the response, I could analyze whether there is an obvious relationship based on the coefficients between them after the construction of the models. In this way, I can test whether the variable is useful in predicting the actual response or whether they are correlated utilizing the hypothesis test. Also, there would be multiple indexes like residual standard error used to assess the credibility of the linear model. At last, based on the model and the assessment of it, there would be explanations and conclusions about why the model is fitted or weakly fitted.

Introduction

The outbreak of COVID-19 and its spread across the country and the world have caused huge social impacts. Studying the dynamics of the epidemic's transmission characteristics can help better control and prevent the epidemic. We have developed a discrete variable. The random probability method is used to simulate and predict the development of the epidemic situation in Hubei Province. Based on the number of patients on a daily basis, the changes in the effective infection rate at different stages of the development of the epidemic were inverted, and the future development of the epidemic was predicted accordingly. The peak incidence of disease has passed in early July. Although small ups and downs in the epidemic are not ruled out, as long as strict isolation and control measures are adhered to, the general trend will not change. There are also possibilities that people coming from outside the states or country will induce a large epidemic to rebound. Under such circumstances, some of the countries in the world are at a stage where the epidemic may break out, so they should conduct inspections and quarantine control of entry personnel.

In the article "The Increasing Model construction and prediction of the spread of Covid-19 pandemic", the author Lin Zhang separate the time length of the pandemic into three sections and respectively fit different models for prediction. Inspired by Lin Zhang, I engaged in the analyzation of the trend of the statistics. After the new policies, the growth has slowed down. The general growth model was used to fit the cumulative number of new coronavirus diagnoses across the country from March 2020 to February 2021. The model of the cumulative number of suspects and the number of close contacts nationwide is consistent with the real data released by the official Covid-19 statistics. The fitting results are timely. It reflects the progress of the epidemic prevention and control work, and at the same time provides forecasts and references for the development trend of the epidemic. Scientifically predicting the development trend of the epidemic is essential for epidemic prevention and control. On the basis of the new time-delay dynamic model, a time-delay convolution model and discrete convolution model based on stochastic dynamics are proposed. For this paper, the concept is similar to the paper mentioned above and there would be different approaches in the analysis.

Research Goals

The inspiration of the research goals comes from one of Xiaoli Meng's papers. In the article *Statistical Influence and Paradox in Big Data*, Xiaoli Meng develops his arguments based on methodological and theoretical research including multi-resolution inference, multi-phase inference, and multi-source inference. Although the paper is related to the statistic analyzation of presidential election, I found the methodologies and theories still useful in this analyzation of Covid-19 statistics. The paper puts emphasis on the discussion about quality and quantity of the data. Basically, it analyzes the effectiveness of the big sample on the data evaluation and how the well selection of random sample will contribute to the overall accuracy of the population prediction. Based upon that, Meng applies the principles to the application of the binary outcome of 2016 U.S. presidential election. Specifically, he focuses on assessing the non-response bias and how it affects the effective sample size, thus the margin of error.

Leaning from Meng's research method, the basic goal of this paper is, first of all, to construct a fitted model for the 7-day average confirmed cases of Covid-19 the states in US. Then, based on the linear regression model, there would also be a 95% confidence interval to test the models. After that, a number of test statistics are going to be used to verify the credibility of the models constructed. For example, the residual standard error will be calculated to find out the degree of the residual value. R-squared and adjusted Rsquare will be used to find out the credibility of the linear model. There would also be F statistics to test the hypothesis.

Statistical Models & Analysis

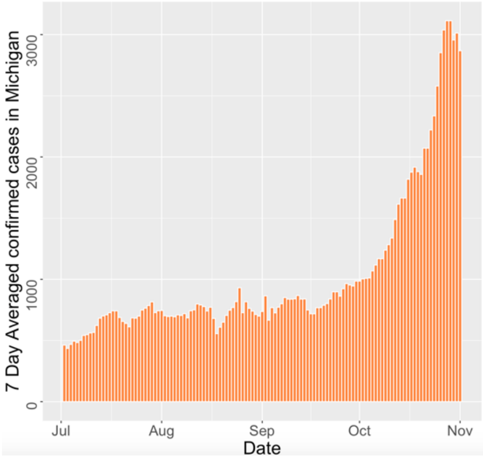
I. Linear Regression Model Construction

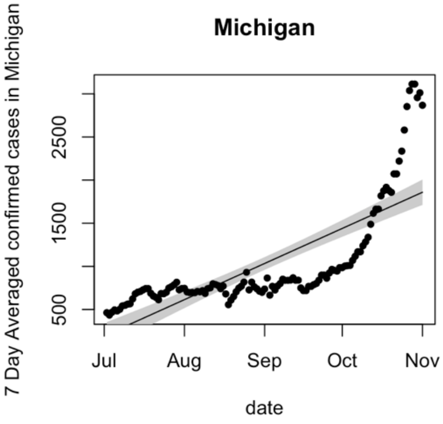
The datasets basically consist of date, states, number of positive cases, number of total test results, etc. The following construction of linear models are going to focused on the response variable confirmed cases in terms of the variable time. Two states in America are taken as an example for analysis, which are Michigan and Florida. After drawing the linear regression model using R, it is obvious that the model follows the equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where Beta0 is the intercept of the linear regression which represents the initial 7-day average confirmed cases in Michigan. Beta1 is the coefficient of the variable X that refers to time here. It implies the correlation between time and the confirmed cases. Every one unit increase in time is associated with Beta1 unit increase in 7-day average confirmed cases.

1. 7-day Average confirmed cases in Michigan



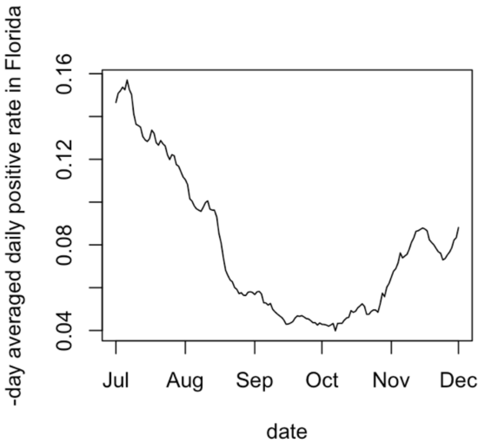


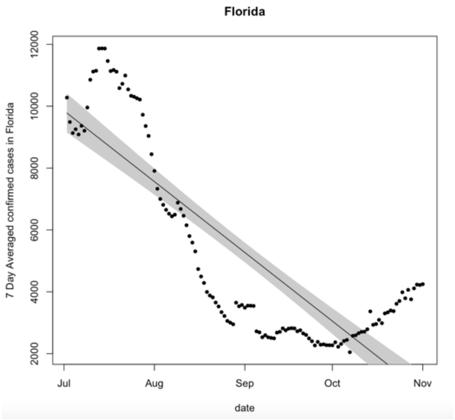
According to the graphs about 7-day average confirmed cases in Michigan above, the slope of the linear model is positive. It means that the 7-day average confirmed cases are increasing as time goes by. The dates in October and November have higher daily confirmed cases, while the dates in April have a higher daily confirmed cases. The reason may because during April and May, the pandemic had not reached its climax, so the confirmed cases is relatively lower compared to the other months.

Near October and November, the pandemic during this time reaches its climax, so there is a large increase in the confirmed case.

What is more, the 95% interval is also constructed, which implies that there would be 95% chance that future data would fall within this interval. Therefore, the linear regression line could serve as a criterion to predict the future trend of the future 7-day average confirmed cases in Michigan.

2. 7-day Average daily positive rate in Florida



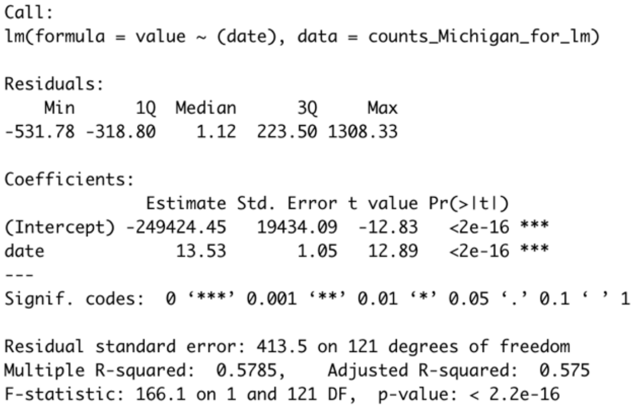


According to the graphs of 7-day average 7-day average daily positive rate in Florida above, the slope of the linear model is negative. It means that the 7-day average daily positive rate are decreasing as time goes by. In the same way, the 95% interval is also constructed and implies that there would be 95% chance that future data of 7-day average daily positive rate in Florida would fall within this interval.

However, the linear regression line can only indicate the information about the general trend of the confirmed cases in Florida, so it cannot tell us any information about some minor changes during this time. Actually, when we look at the first graph, although the positive rate dramatically decreases between July and October, it begins to rise again starting November to December. It is not shown in the linear regression model.

II. Test of Linear Regression Model

After the construction of the linear regression models for Michigan and Florida, it could be observed that there are some major variations that does not fit the model during the life span. For example, in Michigan from September to October, the 7-day average confirmed cases are lower than the linear line of prediction. However, it does not mean that the linear regression models are inaccurate and not reliable. In order to test the credibility of the linear model, we need other statistics to test it.

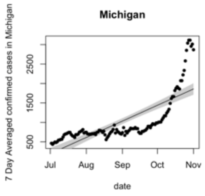


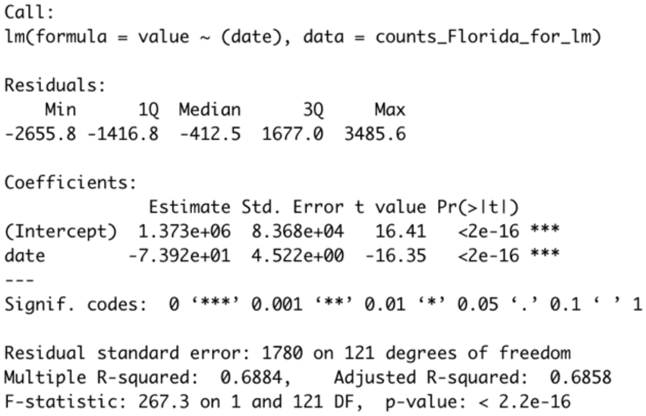
```
Call:
lm(formula = value ~ (date), data = counts_Michigan_for_lm)

Residuals:
    Min      1Q  Median      3Q     Max 
-531.78 -318.80    1.12  223.50 1308.33 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -249424.45   19434.09  -12.83   <2e-16 ***
date             13.53       1.05   12.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 413.5 on 121 degrees of freedom
Multiple R-squared:  0.5785,	Adjusted R-squared:  0.575 
F-statistic: 166.1 on 1 and 121 DF,  p-value: < 2.2e-16
```



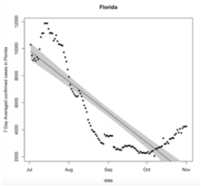


```
Call:
lm(formula = value ~ (date), data = counts_Florida_for_lm)

Residuals:
    Min      1Q  Median      3Q     Max 
-2655.8 -1416.8  -412.5  1677.0  3485.6 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.373e+06  8.368e+04   16.41   <2e-16 ***
date        -7.392e+01  4.522e+00  -16.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1780 on 121 degrees of freedom
Multiple R-squared:  0.6884,	Adjusted R-squared:  0.6858 
F-statistic: 267.3 on 1 and 121 DF,  p-value: < 2.2e-16
```



I. Residual

The above graphs are the result from running the summaries of linear models of Michigan and Florida in R and we can get the information we need from that. First of all, it consists of different quantiles of the residual value. For example, in Michigan, the smallest residual is -531.78 and the largest residual is 1308.33, which means in Michigan during July to November, the actual 7 day-average confirmed cases that is closest to the prediction is 531.78 cases less than the prediction. In the same way, the actual 7 day-average daily positive rate that is farthest from the prediction is 1308.33 cases more than the prediction. It provides us with an insight about the value of deviation. The smaller the deviation, the more accurate the linear model. Compared to the result in Florida, it is found that Florida has more extreme residual values. Therefore, from the aspect of residuals, the linear model for Michigan is more fitted compared to that of Florida.

II. Coefficient

The coefficient of the time variable is an essential indication of the relationship between time and the 7-day average confirmed cases or the 7-day average confirmed cases. The coefficient in Michigan is 13.53 and it is -73.92 in Florida. It indicates that every extra day is associated with 13.53 increase in 7-day average confirmed cases in Michigan. Also, every extra day is associated with 73.92 decrease in 7-day average daily positive rate in Florida.

III. Residual Standard Error

Residual standard Error is simply the standard Error is simple the standard error of the average of the sum of squared residuals which has the formula below:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{SSR}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$$

It provides us with the information about the general deviation of the absolute value of residuals. In the case of Michigan, the RSE is 413.5 with 121 degrees of freedom; in the case of Florida, the RSE is 1780 with 121 degrees of freedom. Actually, these numbers are pretty large. However, it does not directly imply that the model is not appropriate since the daily confirmed cases are relatively large numbers. Hence, we still need other statistics to test the linear model.

IV. R-Squared

The statistics of R-Squared could give us information about the proportion of the variability in Y that can be explained using X. It has the formula:

R squared (R^2)

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0,1]$$

The closer the R-squared to 1, the larger correlation between the independent variable and response variable, then the more convincing the linear model. As we can see, Michigan has the R-squared of 0.58 and Florida is about 0.69. These numbers are relatively far from 1, so we might say that there are weak correlations between time and the confirmed cases. The reason might be that there are too much extreme values that influence the value of R-squared.

V. F-Statistics

The F-statistics work in a similar way compared to Z-test or T-test. Under the significance level of 0.05, the null hypothesis that there is no correlation between the independent and response variable is rejected if the p-value is smaller than 0.05. It is show that the p-value in Michigan and Florida are all smaller than 0.05. Therefore, we could say that time explains a significant proportion of the variations in 7-day average confirmed cases.

Conclusion & Discussion

In conclusion, the current analysis constructs linear regression models and 95% confidence intervals for the 7-day average confirmed cases respectively in Michigan and Florida. After that, there are a series of statistics used to test the correlation between the variable time and the confirmed cases including the standard residual error, R-Squared, and F-statistics. It could be concluded that the correlation of them is validated based on these statistics. Therefore, the general trend of confirmed cases is increasing as time goes by in Michigan. Also, the trend of positive rate is decreasing in Florida.

All these test methods or indexes of indications have their advantages and disadvantages, so sometimes we might need to look at the model from different perspectives and draw the conclusion comprehensively. Besides, the model constructed is useful to make predictions for future statistics. In this way, government policies as well as medical assistances can adjust according to the prediction. For example, according to the linear model of Florida, if I am the governor of Florida, I would release policies about social distancing and self-quarantine as much as possible. For example, each person can only go out of their home three times a week to buy food. Both the counties on the lower right part and upper left part of Florida should have more restrictive measures and other parts can have less restrictions.

Reference

Lin, Zheng.“The Increasing Model construction and prediction of the spread of Covid-19 pandemic”，2020. https://xueshu.baidu.com/usercenter/paper/show?paperid=1u0g0640uq2100s0px450gx036038770. Accessed on Mar. 20th, 2021.

Xiaoli, Meng. “Statistical Influence and Paradox in Big Data”，2018. https://statistics.fas.harvard.edu/files/statistics2/files/statistical_paradises_and_paradoxes.pdf. Accessed on Mar. 20th, 2021.