A Predictive Model for COVID-19 Infection Rate in Orange County, CA

Vivi Ngo [1]
Advisor: Dr. Sam Behseta

[1]Department of Mathematics
California State University, Fullerton

Abstract

In this work, we build a model for tracking the nuances of COVID-19 infection rates in Orange County, CA since January 2020. To that end, we demonstrate the efficiency of a class of generalized linear models for time series data using which we can gauge the effect of certain mobility parameters, such as traffic flow in the business and residential areas, as well as mass vaccination rates for predicting the infection rate of COVID-19. We show that when it is utilized for forecasting near future patterns, the model does a good job with a high level of confidence.

Background Information and Scientific Goals

On March 11th, 2020, the Coronavirus Disease was declared a pandemic by the World Health Organization, according to the CDC there has been over 15,271,571 cases and 288,762 deaths in the United States within the first ten months the first case was reported. In 2021, variants of the virus that causes COVID-19 are circulating, including in the United States. According to the CDC, the U.S. COVID-19 Vaccination Program began in December 14th, 2020. A total of 47,244,379 COVID-19 cases and 762,994 deaths have been reported as of November 18th,2021. COVID-19 spreads from person to person contact that causes issues within our unprepared immune system and it primarily targets the lower respiratory.

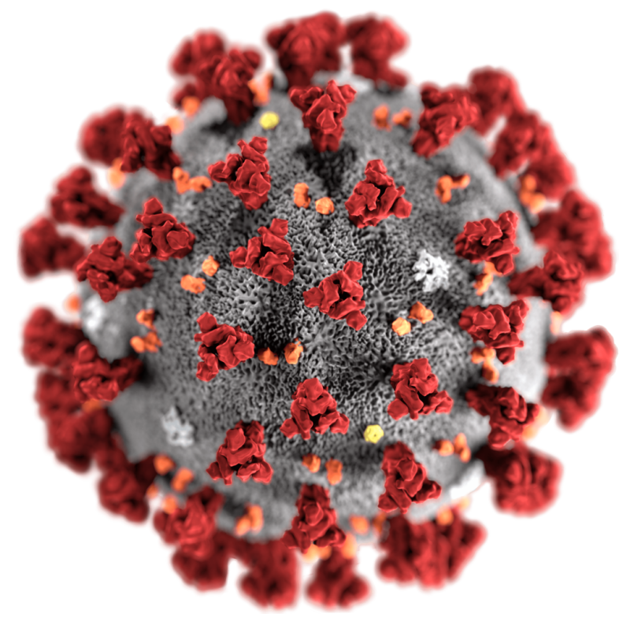


Figure 1. Coronavirus Disease (COVID-19)

In this study we will utilize machine learning, statistical models, including time-series models to address two main goals for this project:

- Generate Forecasting daily COVID cases by county in Southern California.
- Track and Interpret the rate of vaccinations

In order to predict COVID cases for a specific day we have to not only compile the data into one data set, but also consider the following time range from exposure to development of symptoms.

Data Structure, Time-frame, and Covariates

From these sources we created and updated two primary data sets,

1. Mobility Data: Google This data is available by Community Mobility Reports. The reports chart show movement trends over time by geography, across different categories of places. Google collects aggregated data and focuses more on where people spend their time.
2. Mobility Data: Apple Apple's Data has some similarities to Googles. Apple tracks mobility in three categories: driving, walking, and transit. Apple collects the data from requested directions. velit lectus faucibus dolor, quis gravida metus mauris gravida turpis.
3. COVID infection data: USA Facts USA Facts tracks COVID-19 data daily by state and county. It tracks number of cases and deaths.
4. COVID Vaccination Data: OC Health Care Agency The OC Health Care Agency provides updated vaccination data in Orange County, it includes first and second doses administered from Dec. 15,2020.

A Generalized Linear Model for the Time Series of Count Data

Literature Review

- The most frequent theme in all of those is models for prediction of the diagnosis, infection, mortality, and hospitalization rates (Zaobi, 2021). Nevertheless, there is a significantly limited published work, so far, on the statistical or mathematical models for studying the dynamics of COVID-19 fluctuations when viewed through the prism of the economic status of their communities.
- Since economic disparity in the U.S. often correlates with racial disparity, there is an urgent need to tackle the issue, and thereby fill that void in the literature (McLaren, 2020).
- In this work, we draw from the rich and extensive literature on the statistical modeling for the time-series regression, when the response variable is the number of incidents or random occurrences, and the predictors of the model are time dependent as well (for a comprehensive review of the literature, see the references cited in Kedem and Fokianos, 2002).

Modeling

- We would like to be able to build a regression model whose response variable is the rates of infection, and the predictors in the model are mobility and county values. This means, we model the time series of the cases or the response variable with a so-called non-homogeneous Poisson process (Zeger and Qaqish, 1988).
- This approach allows for building a regression model through a mechanism known as generalized linear models or GLM in short(Nelder and Wedderburn, 1972).
- Let's represent the time series of the response by Y_t, where $t \in \mathbb{N}$, represents the time index. We also let X_t to represent a vector of all covariates at time t. Due to our use of Poisson model, we model the conditional expected value of the time series, given the history of the series or $\mathcal{F}_t$, to have an intensity parameter λ_t, or in general, $E(Y_y|\mathcal{F}_t) = \lambda_t$. We can then write the model as:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k h(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}) + \eta^T \mathbb{X}_t,$$

where h is a transformation function, and g is the so-called *link function* of the generalized linear model. In the case of our model, g is simply the logarithmic function. The parameter vector η reflects the effect of all the covariates, and α and β represent the coefficients for the lagged conditional mean and lagged observations and broadly represent the auto-regressive order of the model and can be determined via the Avutocorrelation and Partial Autocorrelation functions.

Parameter Estimation

All parameters in the model can be updated, iteratively, using a quasi maximum likelihood estimation technique.

Software

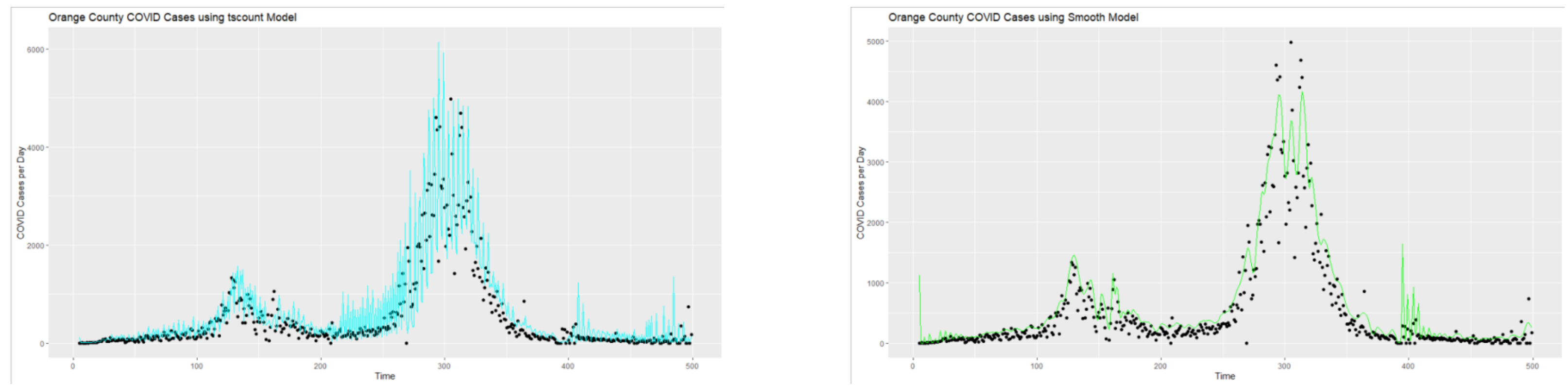
An R package called *tscount* is utilized for modeling and forecasting data. This is coupled with multiple other packages, including the *KernSmooth* package for smoothing splines, and a number of R and Python packages for data wrangling and data preparations.

Predictions and Forecasting

Prediction is implemented in the R package *tscount* using a parametric bootstrap technique for time series.

Main Results

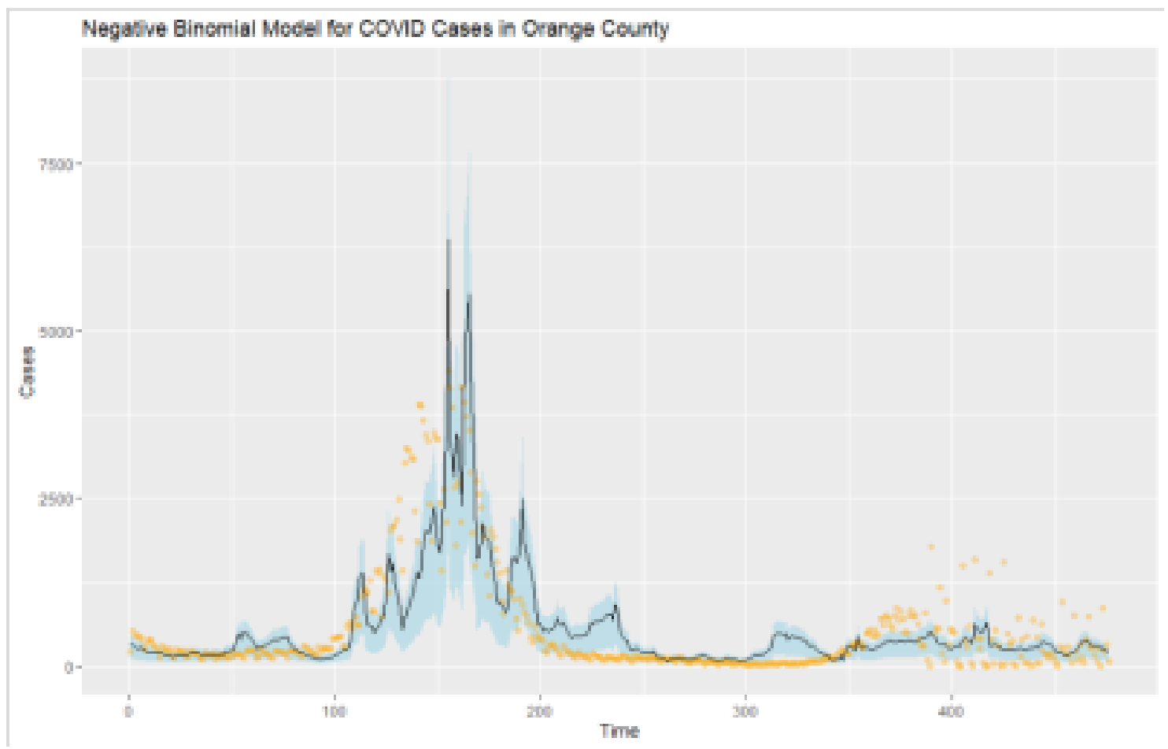
We can summarize the main modeling outcomes of our work via two visuals. In the left panel of figure 2, we demonstrate how well the time series model, depicted in green, mimics the nuances of the infection rates in black. We note that a smoothing process, via kernel smoothing splines, can create a curve that represents the patterns of variation of the infection rate, as shown in the right panel of figure 2.



(a) A subfigure 1

(b) A subfigure 2

Figure 2. Left: time series model models the nuances of COVID-19 infection rates in Orange County, CA. Right: A smoothing splines with a Gaussian kernel smooths out the fitted model, allowing for identifying the overall patterns of variation.



(a) A subfigure 3

Figure 3. Blue: Represents Confidence Intervals for the model Orange: Represent cases in Orange County

Future Work

We will further update the model, particularly with the new vaccination data becoming available to us. We will also look at strategies for making our forecasts more precise by reducing the confidence bands of the future bootstrap predictions.

Acknowledgments

This research has been made possible by the generous grant of Dr. Paul DeLand, Mrs. Sara DeLand. We are utterly grateful for their generosity. The authors wish to thank Dr. Sam Behseta for his vision and mentorship. We would also like to thank the Mathematics Department at CSUF, particularly Mr. Abraham Roldan, and Ms. Gwendolyn Lind for her technical support.

References

[1] B. Kedem and K. Fokianos. *In Regression Models for Time Series Analysis*. Wiley, 2002.

[2] John McLaren. Racial disparity in covid-19 deaths: Seeking economic roots with census data. *The B.E. Journal of Economic Analysis Policy*, 21(3):897–919, 2020.

[3] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A*, 135:370–384, 1972.

[4] Qaqish B. Zeger, S. Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44(4):1019–1031, 1988.

[5] Deri-Rozov S., Shomron N, Zoabi, Y. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *npj Digit. Med.* 4(3)