Qeios

A Multi-factor Model of COVID-19 Epidemic in California

Ted G. Lewis
tedglewis@icloud.com

Abstract

We describe a multi-factor model of the spread of COVID-19 across the 58 counties of California from March 2020 to June 2023. The model provides estimates of cumulative cases and duration of the epidemic versus 5 independent variables. The independent variables are the following factors: population, population density, family income, Gini coefficient, and land area (size) of each county. The correlation coefficients of these factors are used to reduce the error in our model.

The model produces two linear equations – one for cumulative cases and the other for duration of infection. Cumulative case estimate is highly correlated with population, but the estimate is improved by considering all 5 factors. The duration of infection estimate is improved by considering population and income level. We also find that infection rate varies highly and roughly obeys a normal distribution, suggesting randomness, rather than correlation with one or more of the 5 factors.

What Is a Multi-factor Model?

Many models have been developed for estimating and predicting the spread of contagions via contact. They are essentially diffusion models with compartments devised to capture variable diffusion rates. For example, the famous Kermack-McKendrick differential equation model defined four compartments – susceptible, exposed, infectious, and recovered (SEIR). The SEIR model has been extended in many different ways to capture other properties of the population under study. Typically, diffusion rates for each component are estimated by optimal least squares (OLS) curve fitting to empirical data.

The Kermack-McKendrick model and its many extended derivatives are used to estimate the size and duration of epidemics [1, 2]. The spread of disease via contact is assumed to be principally dependent on the infection rate (probability that contact will transfer the disease), while the death rate depends on a different diffusion rate known as death rate (conditional probability that the patient will die assuming they have the disease). More elaborate models incorporated additional factors such as population size, region, public health policy and political leanings. In most cases, the diffusion within a compartment is used to estimate some parameter of interest using OLS curve fitting to arrive at an estimate or prediction.

In this work, we depart from compartmentalization, diffusion/differential equation modeling, and making assumptions about how COVID-19 spread, and instead apply a pure data science

https://doi.org/10.32388/I4DVAG

approach to modeling the spread of COVID-19 throughout California during the pandemic from March 2020 to June 2023 using data collected by public health for each of the 58 counties of California. [1] That is, our model is based on 58 data points provided by the state of California and correlations calculated by the author. This multi-factor model is purely statistical in the sense that it applies correlation coefficients and OLS linear equations. It does not attempt to relate the spread of COVID-19 to biological or medical processes or public policy.

The 58 counties vary widely in terms of population size and density, income and income inequality (Gini coefficient) and cumulative infections. Specifically, population and population densities range from 1,200 to 9.9 million and 1.6 to 17,688 people per square mile, respectively. Incomes range from $45,000 to $120,000 and Gini coefficients[2] range from 0.389 (below the US average) to 0.507 (above the US average), respectively. Cumulative cases range from 139 to 3.5 million – several orders of magnitude differences. The variations are very large resulting in large variations of cumulative cases, see Figure 1. Figure 1 illustrates the vast differences in cumulative cases across the state for these outliers.

Using time series data from California, recorded over the period of the pandemic (2020 – 2023), we show that estimating the *number of cases and duration* of the epidemic is tractable in terms of several independent variables such as population, population density, family income, Gini coefficient, and size (land area) of the county. We build linear models to estimate the size and duration of California's COVID-19 epidemic, but find that estimating the infection rate is impossible because the infection rate behaves like a random variate.

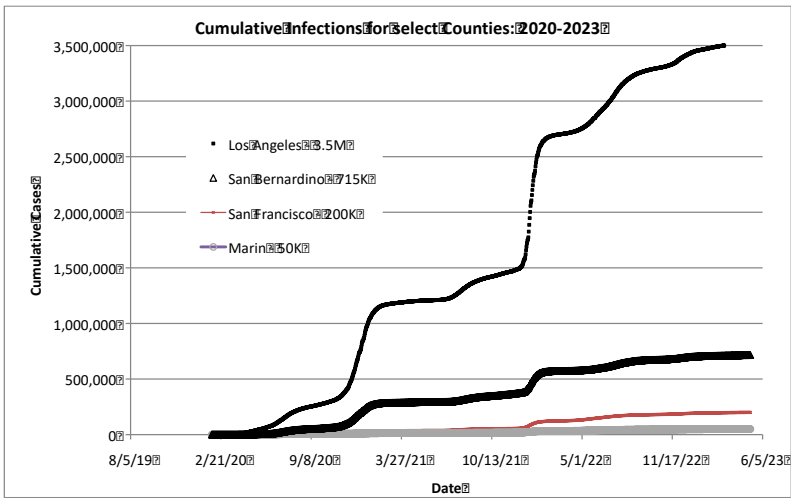

Figure 1. An illustration of the vast differences in cumulative cases: These counties were chosen based on the fact they are the counties with maximum values of factors per Table I. Los Angeles has the largest population; San Bernardino the largest land area (size); San Francisco the highest density and Gini coefficient; Marin the highest family income.

[1] https://data.chhs.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state.
[2] Gini is an international measure of wealth inequality based on Pareto distribution.

Table I. Independent Variables as factors in the formulation of a basis vector, x. Los Angeles has the highest population with nearly 10 million inhabitants. San Francisco has the highest Gini coefficient and population density. Marin is the wealthiest county as measured by family income. San Bernardino County has the largest land area. This variation leads to highly variable cumulative cases and duration.

#i	Factor x_i	County with maximum value
1	Population	Los Angeles
2	Income	Marin
3	Population Density	San Francisco
4	Gini Coefficient	San Francisco
5	Land Area of County (Size)	San Bernardino

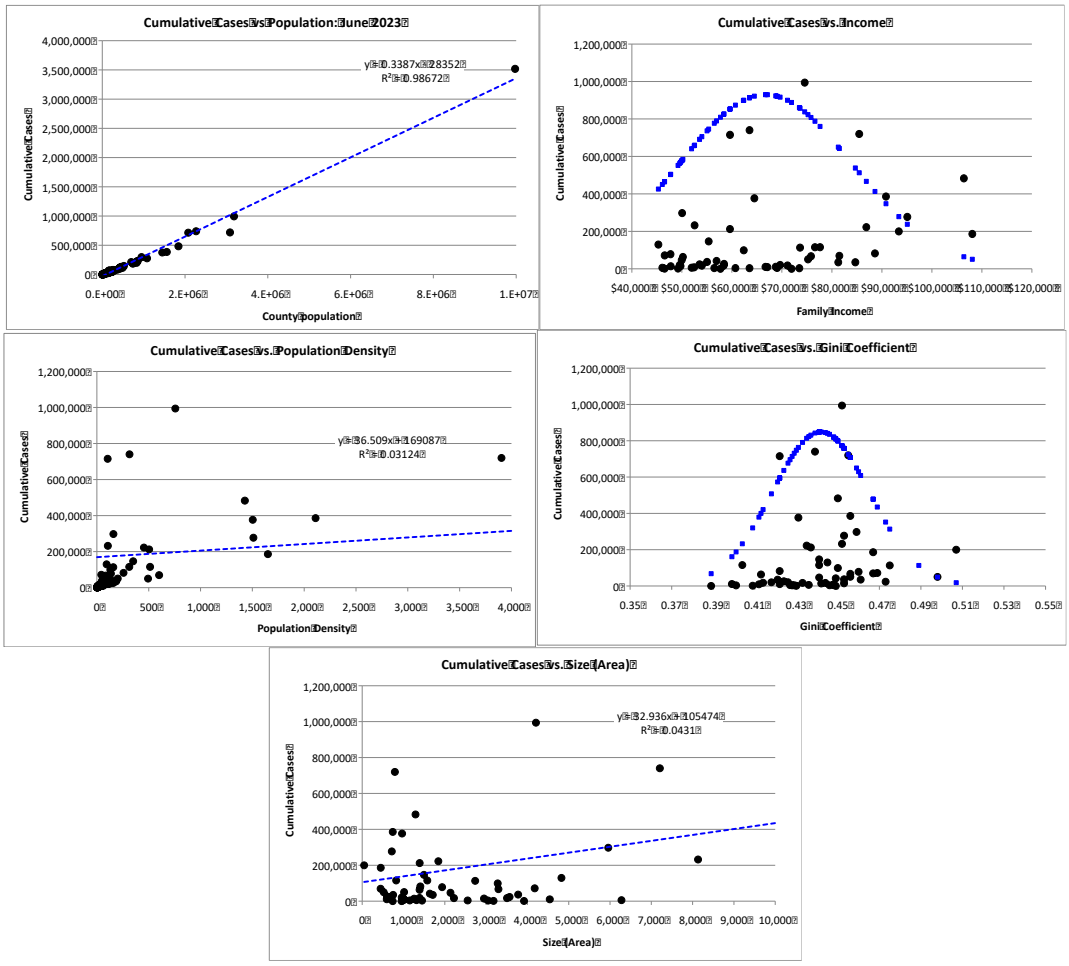

Figure 2. California COVID-19 data shows very little correlation between cumulative cases and independent factors of population density, income, Gini coefficient, and size. Only population is strongly correlated. Income and Gini coefficient even appear to obey a normal distribution and lack any indication of a trend.

Methods and Models

Generally, one purpose of modeling is to reveal cause-and-effect in a phenomenon. In the case of COVID-19, it is suspected that rates of spreading, duration, and ultimate number of infected people are the result of factors such as crowding, economic diversity, and public health policy. Is it possible to construct a model that is based on such factors? In the following, we isolate 5 factors that intuitively impact the number of cases infected and the duration of the epidemic in California. See Table I.

The approach taken here is not unique, but rather it has a precedent. The Fama-French equation uses a similar approach to estimate the value of stocks in a portfolio [3, 4]. However, the Fama-French equation does not use correlations as predictors. As far as the author knows, this is a novel approach.

First, we show that single-factor estimation is insufficient to obtain accurate estimates of duration. Duration of epidemic is improved greatly by using more than one factor and even cumulative case estimation is improved by using multiple factors. Regardless of the number of factors used, the estimation of the infection rate remains beyond the reach of our method, as the infection rate appears to be random across the 58 counties.

Single Factors

Figure 2 shows the results of the OLS curve fitting of cumulative cases versus each of the 5 independent variables. This approach failed. There is little correlation between cumulative cases measured at the end of the epidemic and four of the five independent variables considered one at a time. In cases of Income and Gini coefficient, the dependence appears to be random, roughly obeying a Normal distribution, although the R^2 measures indicate a poor fit.

Table II. Correlation Coefficients for each factor across all 58 counties.

Factor	*Cumulative Cases Correlation*	*Duration Correlation*
Population	0.993	0.323
Income	0.082	0.277
Density	0.177	0.220
Gini	0.177	0.340
Size	0.208	0.140

Combining Factors

An alternative to single-factor modeling is to use many factors in combination to gain more accuracy. The multi-factor model developed here is based on 5 independent variables – x-factors – collected per county. Table I lists the independent variables used to estimate cumulative cases and duration. In each case, we fit a linear equation to a “predictor” function

based on the correlation coefficients obtained by correlating cumulative cases or duration with each of the 5 factors.

The cumulative number of cases and duration at the end of the epidemic are accurately modeled in terms of independent vector x as follows:

$$Cases = 24{,}608x - 21{,}065;\ R^2 = 0.99$$

$$Duration = 2{,}765 + 302.5;\ R^2 = 0.67$$

The infection rate is shown to be uncorrelated and appears as a random variable, roughly obeying a normal distribution across counties.

Linear predictor equations use correlated factors to predict key parameters of the epidemic. They are similar to the Fama-French equations used in finance to value a stock. Let P(x) be a linear predictor equation as follow:

$$P(x) = W(c, x)/\sum x$$

where,

$$W(c, x) = \sum_{i=1}^{5} c_i\ x_i$$

$\sum x = \sum_{i=1}^{5} x_i$; x_i value of factor, c_i correlation coefficient

The correlation coefficients for cumulative cases and duration are listed in Table II. The correlation coefficients c_i were tested to determine the best fit for the data. Population and income were the major drivers of cumulative infections and duration, with density and size in W(c, x) marginally improving the prediction of cumulative cases.

For cumulative cases:

$$I(x) = \frac{.993x_1 + .082x_2 + .177x_3 + .177x_4 + x.208x_5}{x_1 + x_2 + x_3 + x_4 + x_5}$$

For duration:

$$D(x) = \frac{.323x_1 + .277x_2 + .220x_3 + .140x_5}{x_1 + x_2}$$

Where, of course, the x's are actual data points taken from the California database. Figures 3, 4, and 5 show the results.

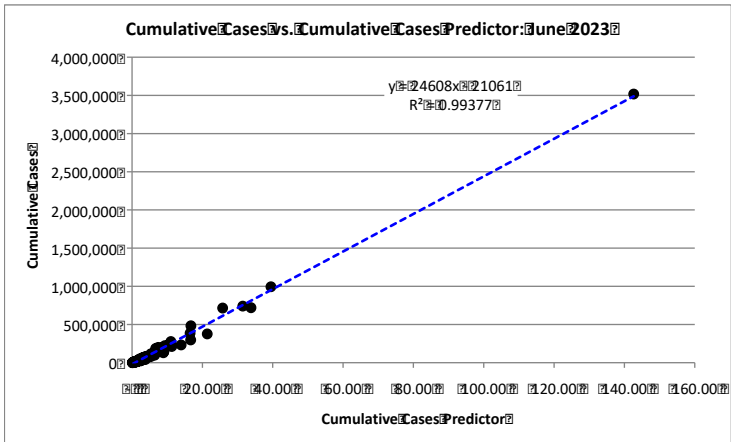

Figure 3. Cumulative cases are accurately estimated by population, income, density, Gini coefficient, and size. Data collected by California is plotted versus cumulative cases predictor I(x).

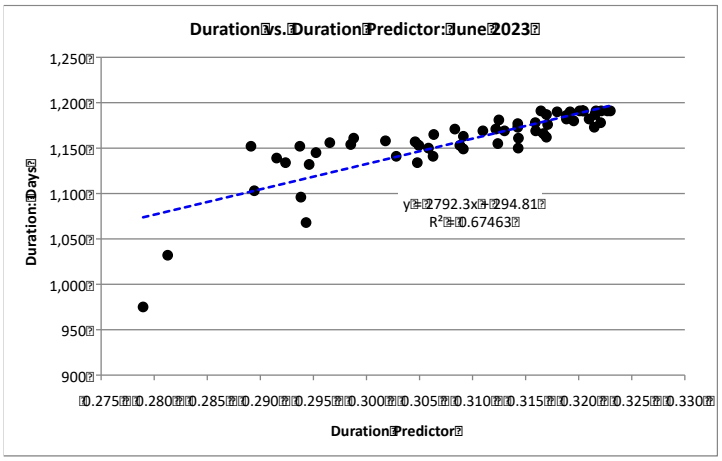

Figure 4: Duration (days) is somewhat accurately estimated by population, income, density, and size. Data collected by California is plotted versus cumulative cases predictor D(x).

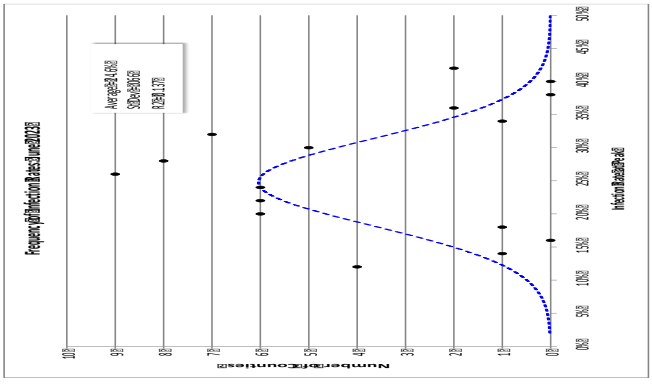

Figure 5. The infection rate varies like a random variable that most likely obeys a Normal distribution with an average value of 24%.

Factors in High Cumulative Cases

We divided the counties into a bottom half and a top half by cumulative cases to determine if counties with a high number of infections were different from counties with low numbers. The bottom half incurred 139 to 42,001 infections while to top half incurred 46,777 to 3.5 million infections. The correlation results are shown in Table III.

Correlation with population remains nearly the same as the state-wide result, but the other factors vary greatly. For the bottom half (fewer cases of infected people), population density (size) and income (Gini) are more strongly correlated. Specifically, lower population density and lower inequality (lower Gini) resulted in a lower number of infections. Cumulative cases are negatively correlated with size.

The top half (greater number of cases) correlates strongly with the population as before, but much less than with the other factors. In fact, Income is negatively correlated with a high number of cases. Generally, the Gini coefficient shows a difference when dividing the counties along cumulative infection lines. Compared with the state-wide correlations, this suggests that density plays a bigger role in infections than income or the Gini coefficient.

Estimating infection rate, an essential element in most diffusion modes, is made more difficult because there is no known correlation with basic factors like population, population density, income, Gini, or size. Infection rates for COVID-19 in California ranged from 10% to 42% – a very large spread.

Improved models for low- and high-infection counties are easily obtained using the correlations in Table III.

Table III. Correlation Coefficients: Cumulative cases bottom half versus top half.

Counties in…	*Population*	*Income*	*Density*	*Gini*	*Size*
Bottom Half	0.970	0.227	0.799	0.374	(0.183)
Top Half	0.994	(0.061)	0.093	0.241	0.214

Conclusions

It is possible to improve the accuracy of epidemic models by using multiple factors rather than a single factor or classical SEIR equations. Using population, population density, income, income equality measure Gini, and size, we obtained very high accuracy for cumulative cases and modest accuracy for estimated duration. Population and density are strongly correlated with

cases reported in low-infection counties, while other measures such as the Gini coefficient and size play a secondary role.

We provide a novel prediction model based on correlation coefficients, but not that the infection rate behaves as a random variable, meaning we are not able to estimate it. The prediction model is a weighted sum of factors where correlation coefficients are the weights. This approach is scalable – it can be applied on a country, county, or local level whenever data are available.

References

[1] Carcione, J. M., Santos, J. E., Bagaini, C., & Ba, J. (2020). A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model. *Frontiers in Public Health, 8.* https://doi.org/10.3389/fpubh.2020.00230

[2] Pei, Y., Li, J., Xu, S., & Xu, Y. (2022). Adaptive Multi-Factor Quantitative Analysis and Prediction Models: Vaccination, Virus Mutation and Social Isolation on COVID-19. *Frontiers in Medicine, 9.* https://doi.org/10.3389/fmed.2022.828691

[3] Fama, E. F., & French, K. R. (1998). Value versus Growth: The International Evidence. *The Journal of Finance, 53*(6), 1975-1999.

[4] Fama, E. F., & French, K. R. (1996). Multifactor Explanations of Asset Pricing Anomalies. *The Journal of Finance, 51*(1), 55-84.