University of California, Berkeley

STAT 222 MA Capstone Project

Final Write-Up

Epidemiological Modeling of COVID-19

Kevin Sun

(Mallika Snyder, Ian Shen)

1 Background and Motivation

The impact of COVID-19 on global health, economy, and lifestyle cannot be overstated. Social and policy efforts such as social distancing and mask mandates have been somewhat effective in mitigating the brunt of the pandemic. However, as with any novel virus, the primary answer has always been to implement a vaccine. At the beginning of this year and this project, breakthrough vaccines had just been developed to galvanized the return to pre-COVID normalcy, but shortages in supply raised one of the most important questions for all societies: who should be prioritized with vaccine distribution in order to end the pandemic as quickly as possible, save the most lives, and help the economy?

There are many factors to consider when designing a vaccine roll-out strategy. A direct approach would be to prioritize the elderly and those with underlying medical conditions first in order to reduce overall mortality, since those types of people are most at risk to die from the disease. However, decreasing the likelihood of exposure by targeting those who interact with other people the most would also indirectly help those with the highest risk by decreasing the chance of infection. With this rationale, it would make more sense to first vaccinate healthcare workers and essential workers, which typically are younger people who do not necessarily have the highest mortality from COVID-19.

Throughout the course of this project, we have had the unique experience of watching the results of certain vaccine prioritization schemes unfold before us. As of today, over 29 million Covid-19 vaccine doses have been given out in California, and 47% of Californians have received at least one dose (LATimes, April 30). However, this is not the case in many other countries and contexts and the analysis that we perform for the state of California could be transferred to other countries that haven't yet been able to vaccinate their population like the US has.

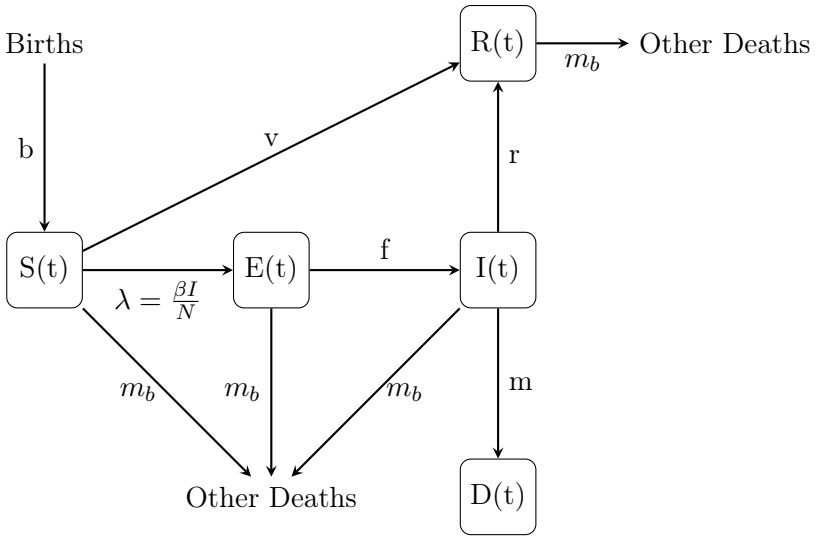
Our project aims to answer the question:

> Which strategy of age-based prioritization for Covid-19 vaccines *would have had* the greatest impact on mortality associated with the disease in California? How does this compare to what happened?

Our approach was to first model the pandemic using an SEIR compartmental epidemiological model, taking special care with the data to have the model reflect reality as closely as possible. Then, we forecasted different vaccine scenarios and compared them with reality.

2 Model

We are primarily utilizing a modified SEIR compartmental epidemiological model. The main compartments are Susceptible, Exposed, Infected, and Recovered. In our case, we split the Recovered category into Recovered and Deaths. Below is a graphical representation of our preliminary compartmental model, where the main compartments are boxed.



The following differential equations describe the rates at which the population progresses through the compartments per age group. Individuals start in the Susceptible category, then become exposed at a rate λ. Exposed individuals then become infected after a period of time f (which represents the incubation period), before either recovering or dying at a rate of r and m. With the introduction of a vaccine, we open a pathway from Susceptible to Recovered with a rate of v. Note that constants with the x subscript are indicated to be age-dependent.

$$\frac{dS_x}{dt} = b \times 1(x = 0 - 17) \times N - \frac{\beta_x I}{N} S_x - (m_b + v_x) S_x$$

$$\frac{dE_x}{dt} = \frac{\beta_x I}{N} S_x - (m_b + f) E_x$$

$$\frac{dI_x}{dt} = fE_x - (m_b + (1 - IFR_x) r_x + IFR_x m_x) I_x$$

$$\frac{dR_x}{dt} = (1 - IFR_x) r_x I_x + v_x S_x - m_b Rx$$

$$\frac{dD_x}{dt} = IFR_x m_x I_x$$

For a particular age group, the majority of the population starts in the Susceptible compartment. With the open-SEIR structure, individuals enter the susceptible compartment through birth (for the 0-17 age group). We also start with a small proportion of Infected individuals, which come into contact with those in the Susceptible category and removing them at an infection rate of β_x. The total number of individuals progressing from the Susceptible to Exposed category is dependent on the proportion of individuals in the infected category. Once in the Exposed category, individuals become infectious at a rate f, which is the reciprocal of the latent period, and are then infectious for the infectious period, $1/r_x$. Individuals in the infected period have a chance of IFR or dying from the disease, where IFR is the Infected Fatality Ratio. Only those that survive the disease with proportion $1 - IFR$ are infectious for the whole period, while those that don't survive are removed at a rate $IFR \times m_x \times I_x$. Those that survive throughout the infectious period move into

the Recovered compartment, while those that die are moved into the Dead/Deceased compartment. Outside of the these main interactions, individuals are also removed from each compartment at a constant rate m_b which represents the death rate (independent of COVID-19).

Of particular interest is the vaccine rate v_x, which allows individuals to move directly from the Susceptible to Recovered category, thereby avoiding the disease itself and reducing the number of people who can become infected with the disease.

The compartmental model contains a number of unknown parameters. Most of these parameters are characteristic of the COVID virus itself, and constants that have been measured by the CDC are presented in Table 1. It can be seen that some of the parameters, such as IFR, vary widely across age groups (for the 65+ , the IFR is 4500 times that of 0-17 age group).

Table 1: Parameter constants obtained from the CDC

Variable	Definition	Value	Reference
f	1/Duration of Latent Period	$1/6 = 0.167$ days^{-1}	CDC (2021a)
r	1/Duration of Infectious Period	$1/10 = 0.1$ days^{-1}	CDC (2021b)
m_{0-17}	1/Time from Symptom Onset to Death	$1/10 = 0.1$ days^{-1}	CDC (2021a)
m_{18-49}	1/Time from Symptom Onset to Death	$1/17 = 0.0589$ days^{-1}	CDC (2021a)
m_{18-64}	1/Time from Symptom Onset to Death	$1/19 = 0.0526$ days^{-1}	CDC (2021a)
m_{65+}	1/Time from Symptom Onset to Death	$1/16 = 0.0625$ days^{-1}	CDC (2021a)
IFR_{0-17}	Infection Fatality Ratio	$20/10^6 = 0.00002$	CDC (2021a)
IFR_{18-49}	Infection Fatality Ratio	$500/10^6 = 0.0005$	CDC (2021a)
IFR_{49-64}	Infection Fatality Ratio	$6000/10^6 = 0.006$	CDC (2021a)
IFR_{65+}	Infection Fatality Ratio	$90000/10^6 = 0.09$	CDC (2021a)
b	US Birth rate	11.6/1000 year^{-1}	CDC (2017a)
m_b	US Death rate	8.6/1000 year^{-1}	CDC (2017b)

The parameters that we seek to estimate is β_x for each age group specifically for California. From the β_x we can calculate R_0, which quantifies how contagious COVID-19 is, using the following formula:

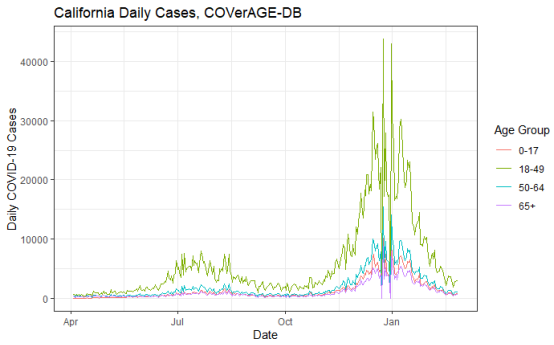
$$\beta_x = \frac{R_0}{1/r_x + 1/f_x}$$

3 Data

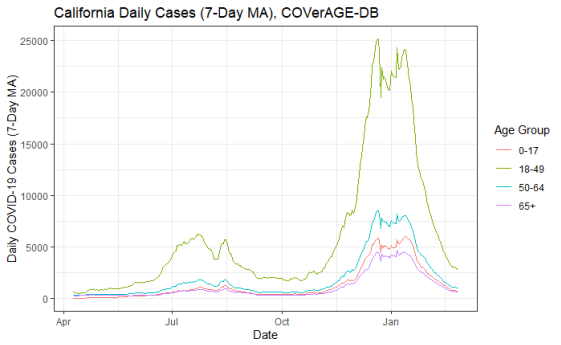
In order to estimate β, we require data on case and death rates for California. We used this type of COVID-19 data from the COVerAGE-DB project, headed by Dr. Time Riffe and Dr. Enrique Acosta at at the Max Planck Institute of Demographic Research. It is a global project that contains data collected from various government sites. In our case, the data for California is collected from the California Department of Public Health and the CDC, and formatted so that each row is a cumulative, aggregated measure of either counts or deaths, dis-aggregated by age group, region, and day. This data set is updated daily, providing an up-to-date view of the data.

Despite this, in our initial data analysis, we saw that this set still has some reporting bias, as evidenced by the spikiness in the daily case data (seen with daily case counts in Figure 1a. Due to

this irregularity in reporting conditions, we used a moving average to "smooth" out the time data (Figure 1b), transforming it into a form we can use to fit our SEIR model.



(a) Daily case counts without moving average



(b) Daily case counts with 7-day moving average

Figure 1: Daily case counts for COVerAGE-DB data

4 Fitting

When fitting, we utilize an approach from literature that makes use of the fact that, at the start of an outbreak, there is typically exponential growth from some initial infected group. Taking the logarithmic transform of this results in a linear form, where we can measure parameters with linear regression (Vynnecky and White, 2010).

$$I(t) = I(0)e^{at}; \log I(t) = \log I(0) + at$$

Once an estimate for a is obtained, we calculate R_0 using the following relation:

$$R_0 = (1 + aD)(1 + aF)$$

We can then use the previous formula to calculate β_x. From the COVerAGE-DB data, we identified a period of exponential growth between November 1, 2020 and December 15, 2020. When plotting the log of the daily case, we see a fairly linear trend (Figure 2). Our formal time-frame was from November 1, 2020 to December 15, 2020.

We first performed sensitivity analysis on the moving average "window," or how many days to include in our moving average calculation. For example, a 7-day moving average for day d takes the measurement at d and averages it with the measurements for the 6 days before, up to $d - 6$. We varied the MA window from 2 to 30 days, processed the data with these MA windows. As can be seen in Figure 3, the estimate of β does not vary significantly with the moving average window. Across all age groups, the estimate decreases slightly, but we attribute this to the fact that at higher values of the MA window, we are over-smoothing the curve, effectively linearizing the exponential curve. With this in mind, we decided to stick with using 7 days for our moving average.

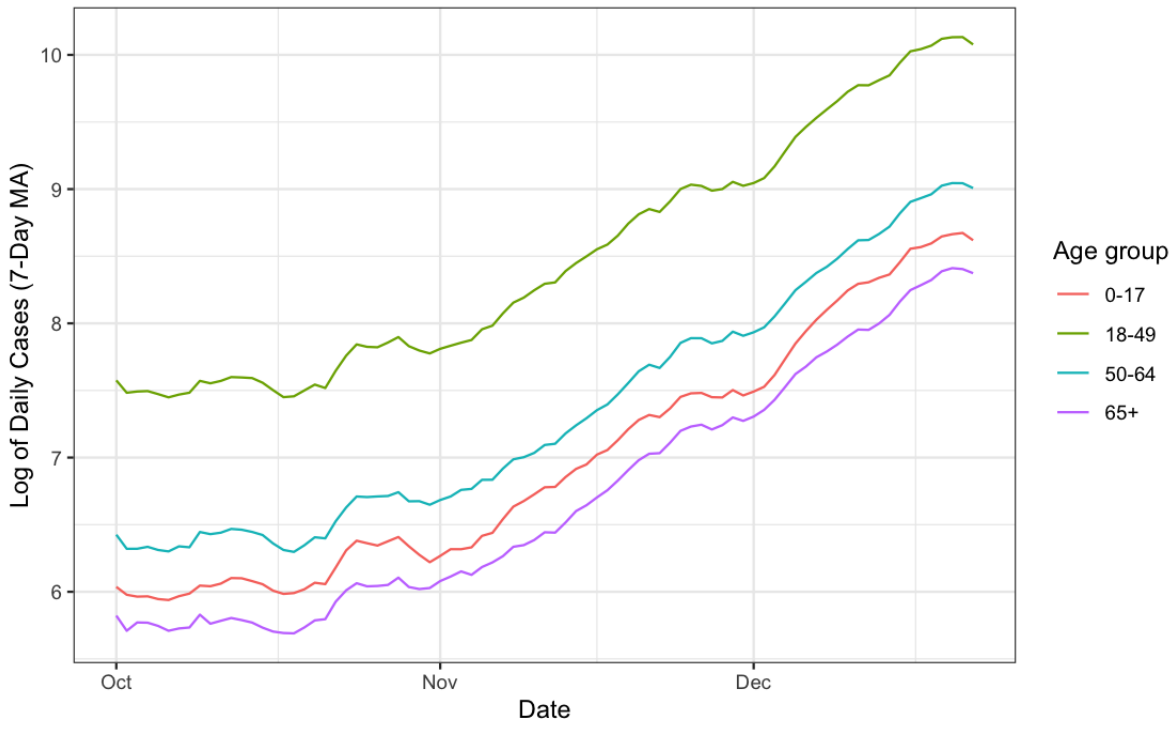


Figure 2: Period of exponential growth. A moving-average (MA) is first applied to the data to smooth out reporting bias. The log of the MA daily case rate results in a linear trend.

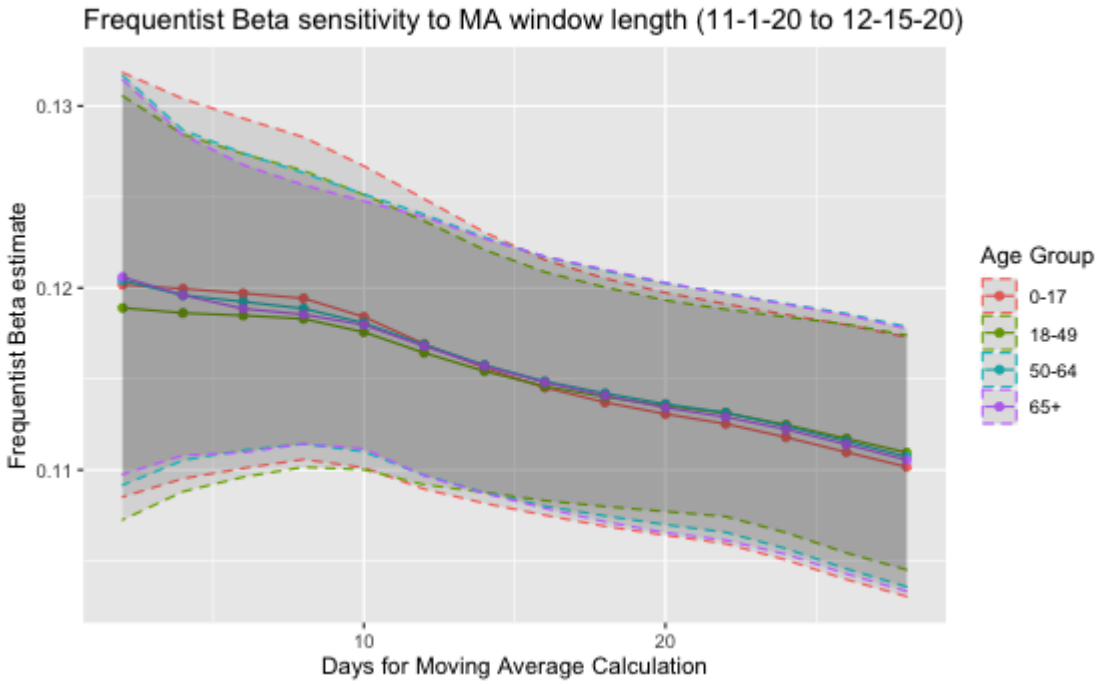


Figure 3: Estimates of β_x by age group, with varying MA window lengths, along with standard errors for each estimate (envelope) obtained from the OLS standard error.

We used a non-parametric bootstrap approach to quantify the uncertainty of our point estimates. For both the start date of Nov 1, 2020 and end date of Dec 15, 2020, we created a 2-week interval from which we sample different start and end dates. For the start date the window ranged from Oct 22, 2020 to Nov 7, 2020, and the end date ranged from Dec 7, 2020 to December 21, 2020. We independently sampled start and end dates uniformly from these 14-day windows for $n = 1000$ times, and repeated the log-transform, OLS fit, and β_x computation for each sample. Figure 4 displays the resulting distributions by age group using this method.

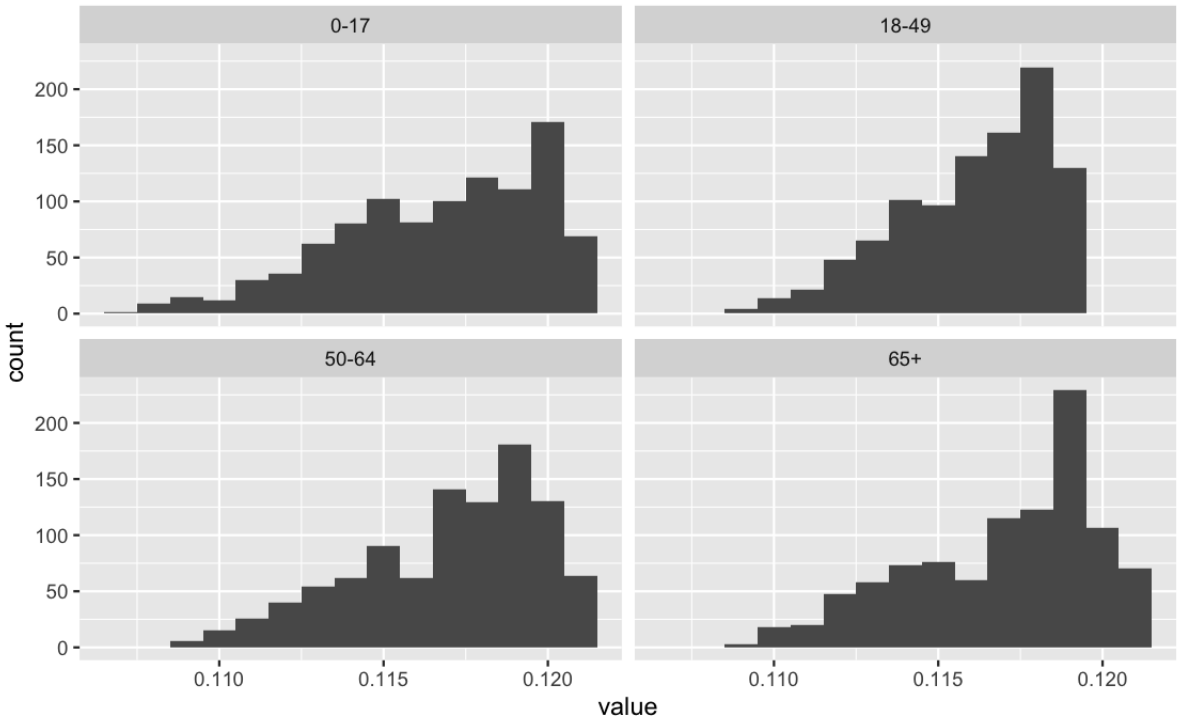


Figure 4: Bootstrap distributions of β by age group.

The skewed nature of the histograms is most likely due to the fact that, as we push the end date further past December 15, the exponential growth begins to taper off. This results in lower values of a when regressing on the log of the daily cases, resulting in a wider range of smaller estimates. Additionally, due to the skewed nature, we use percentile-based confidence intervals for our estimate of β as opposed to normal-based confidence intervals.

We computed 95% confidence intervals for our estimates by computing the 2.5% and 97.5% quantiles in the above distributions, arriving at the final estimates for β using this method (Figure 5).

As seen in the figure, we found that there was no significant difference in the β estimates across the different age groups. This is likely due to the fact that individuals interact across age groups in many complex ways, whereas the SEIR model assumes that individuals act independently within their respective age groups. Additionally, with more time, it has been observed that the major differences for COVID-19 across age groups has been in the mortality rate and not necessarily the rate at which it spreads.

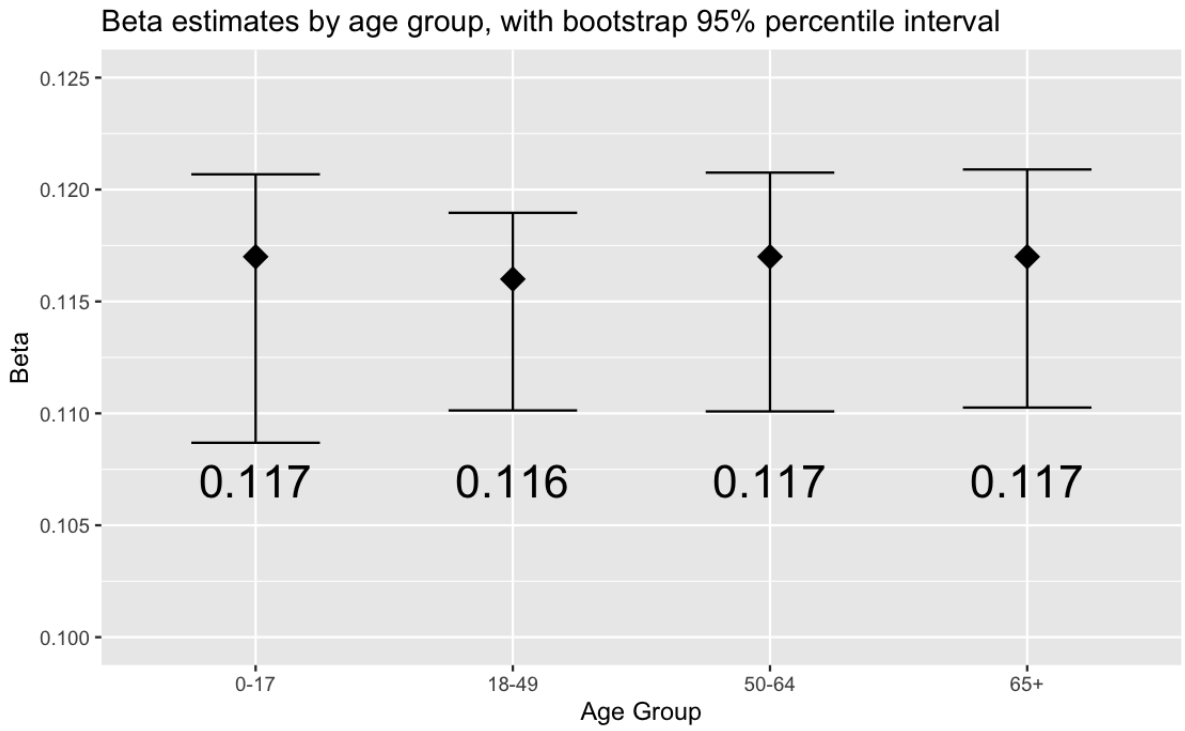


Figure 5: β estimates by age group, with 95% percentile confidence intervals.

While β does not seem to vary much across age-groups, it still provides an estimate for California specifically. To put the final estimate of $\beta \approx 0.117$ in context, Rădulescu et al. (2020) estimated a β of 0.1, lowered to home exposure rate of 0.08 for New York and also saw fairly limited variation between ages (0.08-0.1), dependant on destination. The IHME COVID-19 forecasting team (2021) also utilized a similar frequentist regression strategy that estimated β to be around 0.4 in California in August 2020, which is significantly higher than The CDC Pandemic Planning Scenarios (2021) currently estimate an R_0 of 2.5 (based on multiple international studies), which divided by the latent and infectious period (10+6 = 16) is 0.16. In this context, our estimate of around 0.117 with limited age variation is within the bounds of these other studies and therefore not unreasonable.

5 Prediction and Simulation

When we plot trajectories of vaccination by age from mid-December 2020 to end-April 2021, we see that for California, individuals in the 18-49 and 50-64 range were first prioritized to some extent (Figure 6). These individuals were most likely healthcare workers and essential workers. After this initial spike, the elderly (65+) were prioritized while those in the other two adult age groups were still receiving vaccines. Once a significant portion of the elderly population was vaccinated, we see very recently an increase in vaccinations for adults in general.

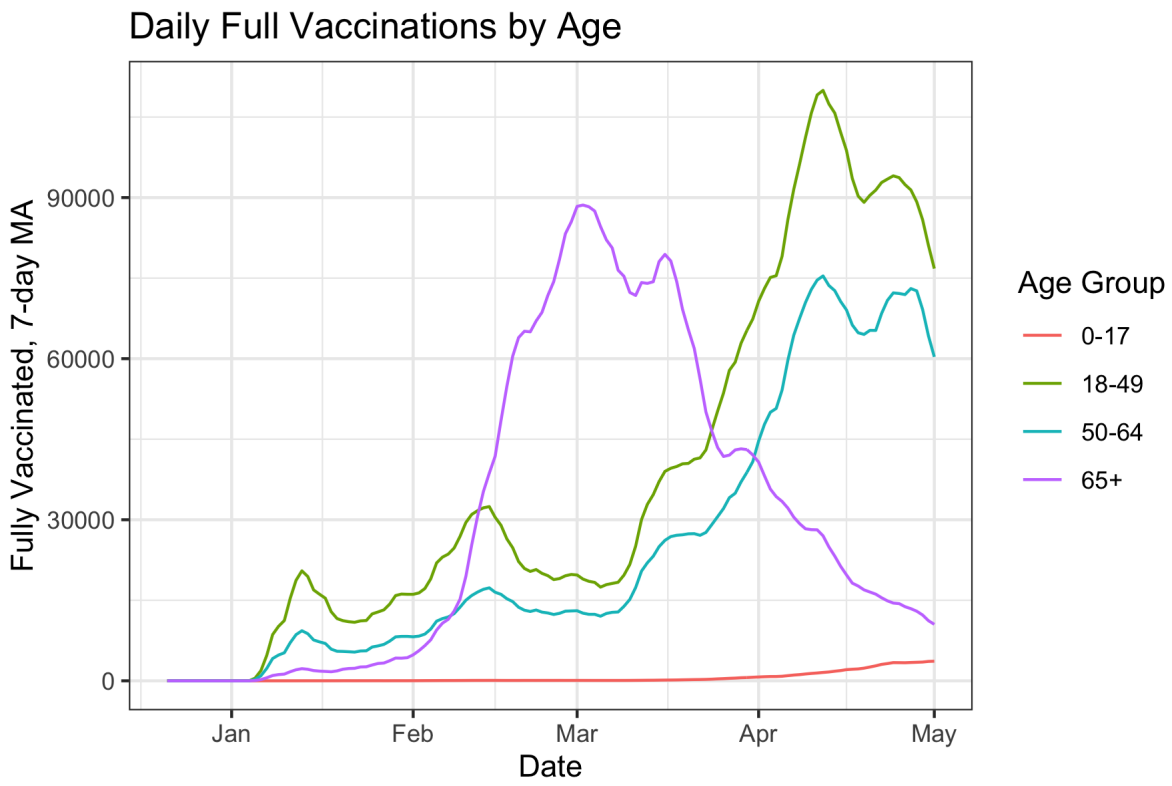
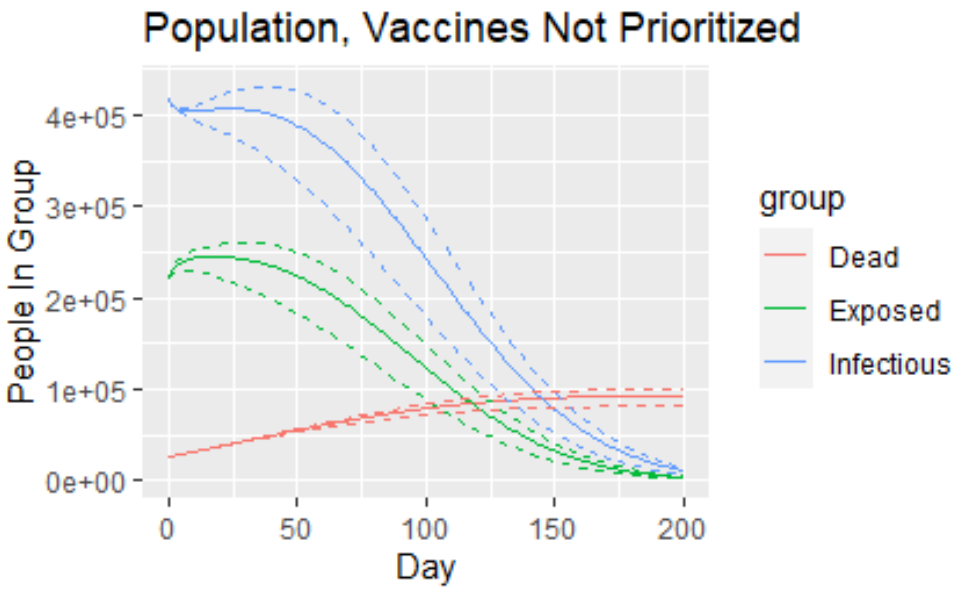


Figure 6: Vaccination rates by age group, California.

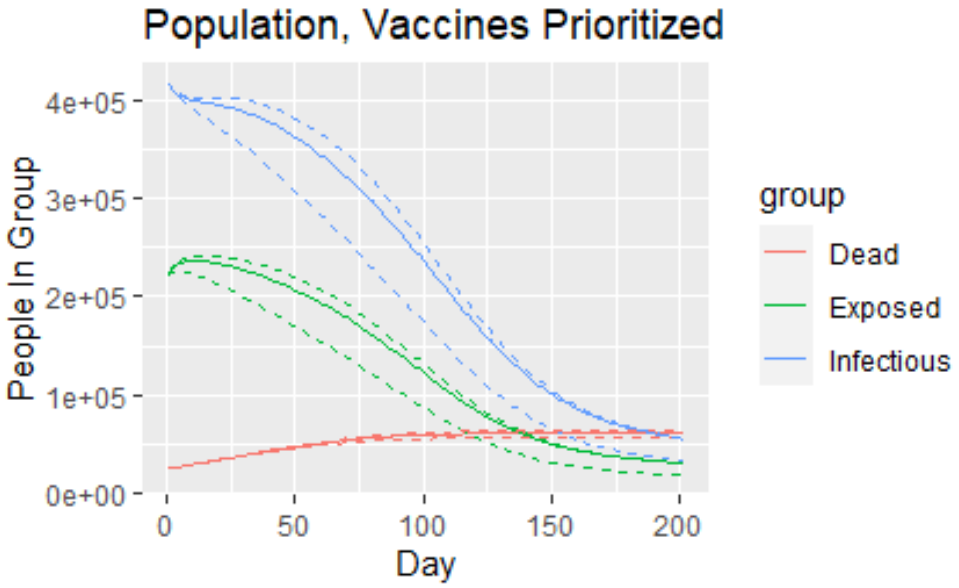
For our simulations, we simulated two extreme vaccinations scenarios that encompass the reality: one in which the most vulnerable age groups (the elderly) are completely prioritized, and one where all adults receive the vaccine with equal prioritization. In the first scenario, individuals in the 65+ age category are prioritized until all 65+ individuals have been vaccinated, at which point we begin to to vaccinate those in the 50-64 age group.

While the SEIR model treats vaccination as immunity, in reality we have seen that the COVID-19 vaccine generally requires two doses and, as a reuslt, a delayed vaccine effect. we delay our vaccination function by 14 days and divide it by 2, that is for an intended vaccine distribution function $f_x(t)$, we have $v(t) = \frac{1}{2} f_x(t - 14)$, $v(t) \geq 0$.

We use our estimate of β obtained from above, and other parameters such as latent/infectious periods and IFR for each age group are from the CDC estimates. We begin simulations on January 1, 2021, using the COVerAGE-DB data to get starting points of infected and deceased individuals. The number of deceased individuals was obtained from the original COVerAGE-DB data (which was presented in cumulative cases/death counts), and the number of infected individuals was calculated using a 10-day rolling sum since the infectious period is 10 days. We ran simulations using our estimated β values, as well as the upper and lower limits of our confidence interval, in order to obtain confidence intervals with our simulations.



(a) Unprioritized vaccine scenario for all age groups (all ages prioritized equally).



(b) Prioritized vaccine scenario for all age groups (65+ prioritized).

Figure 7: Infected, exposed, and deceased for combined age groups with two different vaccination prioritization strategies.

As seen in Figure 7, we see a drastic reduction in the number of deaths for the elderly population when that population is prioritized, with almost a 50% reduction. With the prioritized method, the number of deaths at the end of the simulation lies in the interval of (57195, 64297), while with the unprioritized method, the number of deaths lies in the interval (82016, 101054).

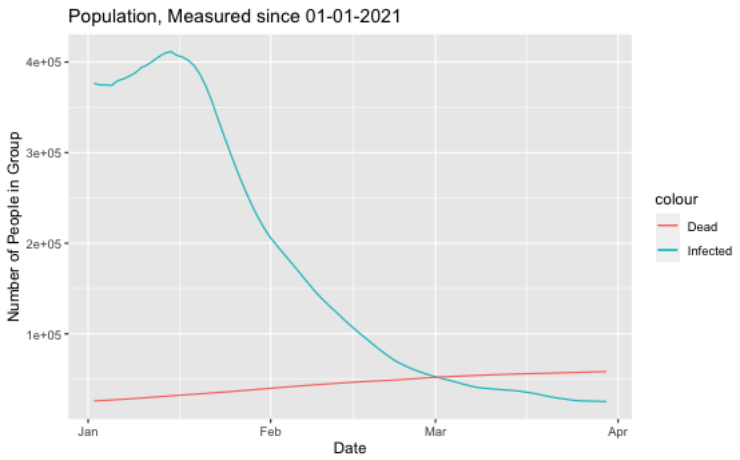


Figure 8: Measured number of infected and deceased for California population, all age groups.

We compare these simulations with reality by using the same rolling sum approach for estimating the number of infected individuals at any time, as seen in Figure 8. As can be seen in the graph, the death rate approaches California's current number of 62,215 COVID deaths.

6 Conclusion

In this project, we applied frequentist statistical methods to a very relevant epidemiological problem, COVID-19. Using the COVerAGE-DB data set of daily case and death counts for California, we estimated the rate of infection across different age groups for the state and found that β did not vary significantly by age. Using this estimate and other estimates provided by the CDC, we simulated two different outcomes using two extremes approaches to vaccine prioritization, and found that these two approaches roughly line up with reality, with the worst case scenario significantly overshooting reality. While California is on track to be mostly vaccinated very soon, and move past the COVID-19 pandemic, other countries are not so fortunate and do not have as much access to the vaccine yet, and these methods could be used in those countries in order to help determine how vaccines should be prioritized to most effectively reduce mortality.

7 Works Cited

CDC (2017a). Births and Natality. National Center for Health Statistics.

CDC (2017b). Deaths and Mortality. National Center for Health Statistics.

CDC (2021a). COVID-19 Pandemic Planning Scenarios. *COVID-19*.

CDC (2021b). Interim Guidance on Duration of Isolation and Precautions for Adults with COVID-19. *COVID-19*.

Rădulescu, Anca & Williams, Cassandra & Cavanagh, Kieran. (2020). Management strategies in a SEIR-type model of COVID 19 community spread. *Scientific Reports*. 10. 10.

Vynnycky, E. and White, R.G. (2010) An Introduction to Infectious Disease Modelling. New York: *Oxford University Press*.