The effect of lifting mask mandates in the United States

Walter Nelson

April 2022

Preliminary note on reproducibility

This notebook will not run without access to JHU's COVID-19 data. It can be downloaded in ZIP format from their GitHub here: https://github.com/CSSEGISandData/COVID-19 It should then be decompressed so that `here("data/jhu_covid/COVID-19-master")` points to the decompressed directory. The model fitting took less than a half hour on a compute cluster, but doing it in the notebook might be slow. So, alternatively, you can clone the modeling artifacts from GitHub LFS. After installing Git LFS (https://git-lfs.github.com/), they should be pulled down with the repository like so (if cloning from the command line):

```
$ git clone https://github.com/wjn0/sta2201_final_project
$ cd sta2201_final_project && git lfs pull
```

Introduction

Policy responses to the coronavirus disease (COVID-19) pandemic have included non- pharmaceutical interventions in much of the world, such as mask mandates. Unlike pharmaceutical interventions such as vaccination or post-infection treatments, the individual benefit of mask mandates is difficult to quantify. To justify these mandates, policymakers therefore must turn to measures of population-level COVID-19 burden, such as case counts. These case counts can in turn be used to estimate burden in terms of COVID-related death, hospitalizations, or other endpoints of interest.

Due to the nature of infectious disease and ethical constraints, epidemiologists and policymakers cannot employ the standard tooling for treatment evaluation when quantifying the effect of mask mandates: the randomized controlled trial. Instead, they must rely on models which allow them to make robust causal statements from observational data, such as regression discontinuity. While these models are assumption-heavy and vulnerable to confounding in the same ways observational studies are, they have seen extensive use in modeling COVID-19-related and other policy interventions [CITE].

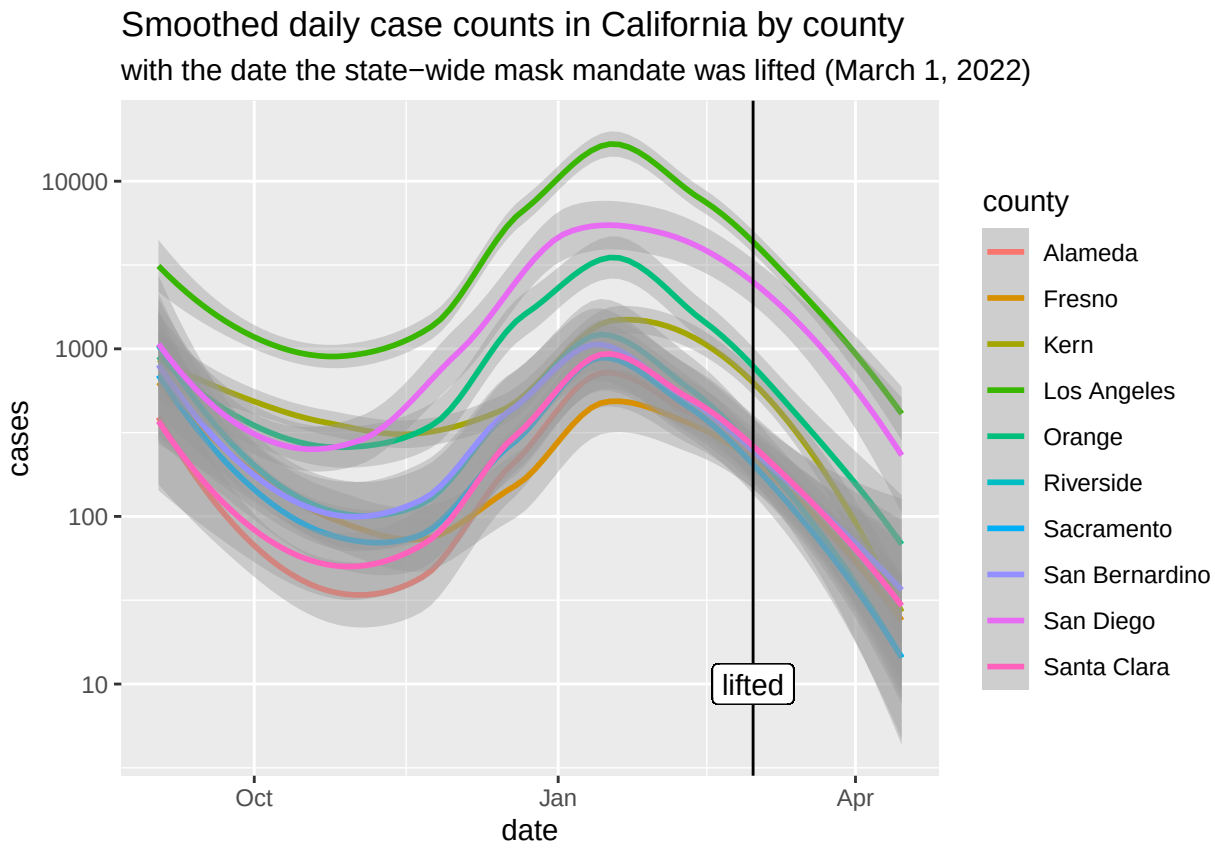
We seek to model the effect of lifting California's mask mandate on 1 March 2022. We begin by briefly summarizing some features of the COVID-19 case data made available by Johns Hopkins University for the counties of California We next simulate data using a simplified the EpiNow model [CITE], a published model of COVID-19 cases that accounts for several of the problems observed in the real COVID-19 case count data. We show that regression discontinuity and regression kink models on the raw case counts fail to recover the simulated effect of a mask mandate. We then show that a Stan implementation of the simplified EpiNow model recovers these effects better, even under mild mis-parameterization and in the presence of hierarchical effects. Finally, we apply this model to the data from the state of California and briefly discuss the implications and limitations of the results.

Data exploration

We extract the county-level case counts for California up to 60 days prior to and up to 45 days following the lifting of the mask mandate in California on 1 March 2022. The window size was selected for computational reasons, and because this project was finished on 18 April 2022.

Below, we show the COVID-19 case counts in California by county for the five counties with the largest total number of cases, along with the date of the lifting of the statewide mask mandate. We note that there is no obvious change in case counts associated with the lifting of the mask mandate, at least visually. Rather than re-hash the issues identified in our previously submitted data exploration, we summarize the key potential features of our data below:

1. Due to the lag between infection, case identification, and case reporting, the effect of a mask mandate is difficult to pinpoint in time by case counts using typical regression models.
2. Counties clearly follow similar patterns within California (at least those shown below), therefore there is potential for information sharing between counties in our model.
3. Although not shown in the smoothed case counts, zero days are common and represent reporting issues (such as on weekends).
4. Counties of different sizes may have case counts on a different scale.



Simulation models

EpiNow simulation process

The EpiNow model is a hierarchical model of observed case counts. The model we implement for our simulation process has the form:

$$
\begin{aligned}
k(t, t') &= \sigma \exp\left(-\frac{(t-t')^2}{2\ell^2}\right) \\
\log R(t) &\sim \mathcal{GP}(\mu(t), k) \\
I(t) &= R(t) \sum_{i=1}^{\tau} w_i I(t-i) \\
D(t) &= \sum_{i=1}^{\tau} \xi_i I(t-i) \\
\phi &\sim \text{Exponential}(1) \\
C(t) &\sim \text{NegBin}(D_t \omega_{t \mod 7}, \phi)
\end{aligned}
$$

where k is the kernel of the Gaussian process prior over the time-varying reproduction number $R(t)$, $I(t)$ is the latent (unobserved) infection function, $D(t)$ is the mean of the reported case counts, and the observed case counts $C(t)$ are drawn from a negative binomial distribution where ω is a day-of-the-week factor. Important hyperparameters are ℓ, the bandwidth of the kernel; $\mathbf{w}$, the incubation distribution; ξ, the reporting delay distribution; ω, the day-of-the week effects.

We assume a constant mean function $\mu(t) = 0.3$ for all simulations, corresponding to an average $R(t)$ of 1.35. We assume the standard deviation is $\sigma = 0.4$ for all experiments, and set σ using the median heuristic [CITE]. We set $\mathbf{w}$ and ξ according to the original EpiNow paper, which use discretized log-normal distributions. The window size τ is chosen to minimize information loss caused by the discretization process, while maintaining computational tractability, and is set to 10 for all simulations. We fix two consecutive days per week to have multiplicative factor $\omega_d = 0$ to mimic that case reporting on the weekends often drops to zero; these cases are often then reported on the following Monday, so we set that factor to 3 (covering the three days Saturday, Sunday, and Monday). The other ω factors are simulated from a Uniform$(1/2, 1)$ distribution. Finally, the autoregression in $I(t)$ requires us to simulate the first τ latent infections, we do so from a Poisson(50) distribution for all simulations.

Simulating interventions

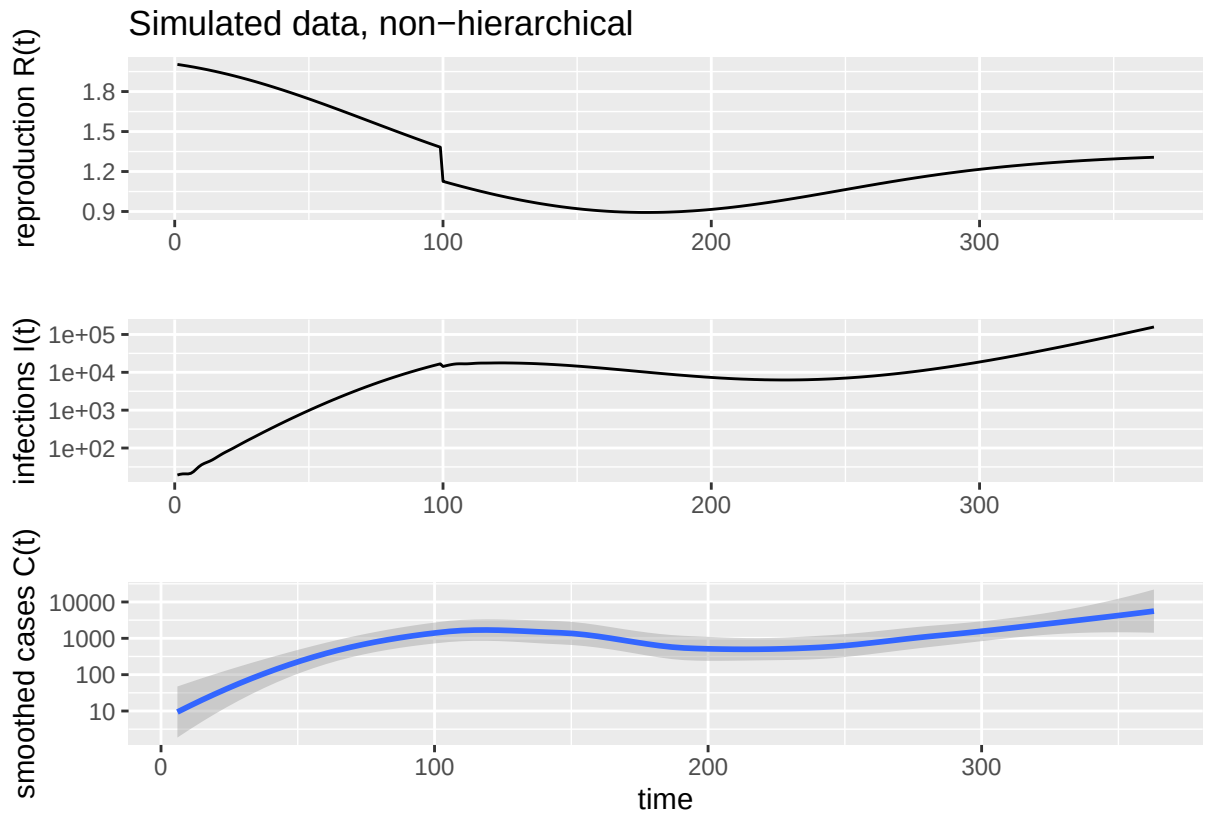
Our hypothesis is that the effect of the interventions of interest is mediated through the reproduction number $R(t)$. Therefore, when modeling interventions, we only need to consider modifications to our parameterization of $R(t)$.

When simulating interventions, we draw the base time-varying reproduction number from a GP, $R_b(t) \sim \mathcal{GP}(\mathbf{0}, k)$ as above, but add a linear treatment effect of the form $R(t) = R_b(t) + \beta I[t \geq t_0]$ where t_0 is the intervention time. For all simulations we fix $\beta = 0.2$.

When modeling interventions hierarchically, we draw a base time-varying as above, but model the treatment effects hierarchically. Specifically, we specify the mean effect as $\beta_0 = 0.2$ for all experiments, and simulate the treatment effect for the ith series like $\beta_i \sim \text{Normal}(\beta_0, \sigma^2)$ where $\sigma = 0.05$ for all simulations. The series-specific time-varying reproduction number is then $R_i(t) = R_b(t) + \beta_i I[t \geq t_0]$ where for simplicity we assume the intervention occurs at the same time t_0 for all series.

This simulation model corresponds to the idea that an intervention causes an immediate discontinuous drop in the reproduction number, while all other forces that impact $R(t)$ (such as strain infectiousness, which varies in time, or social contact matrices) are captured in the Gaussian process.

Below, we show the results of simulation at the various levels: the reproduction number $R(t)$, which will later become our primary concern; latent infections $I(t)$, and smoothed observed case counts $C(t)$. We validate our simulation procedure by examining a handful of simulations from different seeds. Things look mostly okay, but occasional outliers do arise so we control for this by specifying seeds where necessary. (For example, this series presents with many zeros in $C(t)$ that are hidden by smoothing in the plots, even more than we would expect in counties with poor reporting – but this is okay). Note the clear discontinuity in $R(t)$ that barely affects the smoothness of the latent infections and observed case counts.

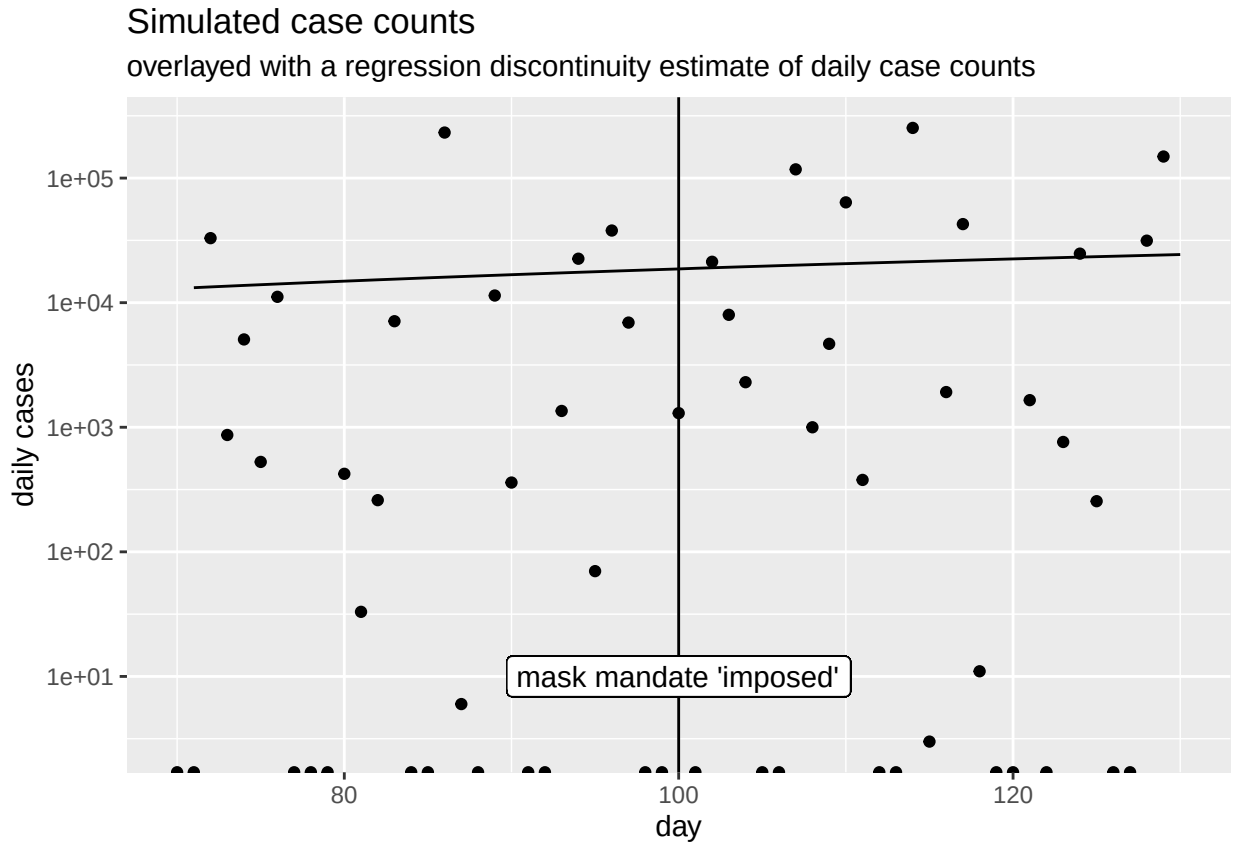


Regression discontinuity model

We begin modeling by attempting to recover our simulated treatment effects in the simple case using a Gaussian observation likelihood, and a linear functional form for the mean of the observed case counts $C(t)$. Mathematically, we assume the following model:

$$
\begin{aligned}
\beta_0, \beta_1 &\sim \text{Normal}(0, 100^2) \\
\alpha &\sim \text{Normal}(0, 1) \\
\epsilon &\sim \text{Normal}^{+}(0, 100) \\
C(t) &\sim \text{Normal}(\beta_0 + \beta_1 t + \alpha I[t \geq t_0], \epsilon^2)
\end{aligned}
$$

where the flat priors are use because $C(t)$ can be quite large, and t_0 is our intervention time. Here, α is the effect of interest; we might interpret $\alpha = 0$ as no effect of the mask mandate. Although this model estimates without significant errors, we see that it fails to capture the correct treatment effect of the *in silico* simulated treatment in its estimate of α: 53.14 (95% CI: -18.33, 212.20). Note that this can be interpreted as the increase in average daily case count before and after our simulated mask mandate was imposed. Reducing the window size for data inclusion on either side of the treatment produces a similar estimate, so there is no evidence that it's a bandwidth issue. Below, we show the learned mean estimate in the model with the original window size. The variance in our estimate of α results in no discnerible effect being detectable visually.



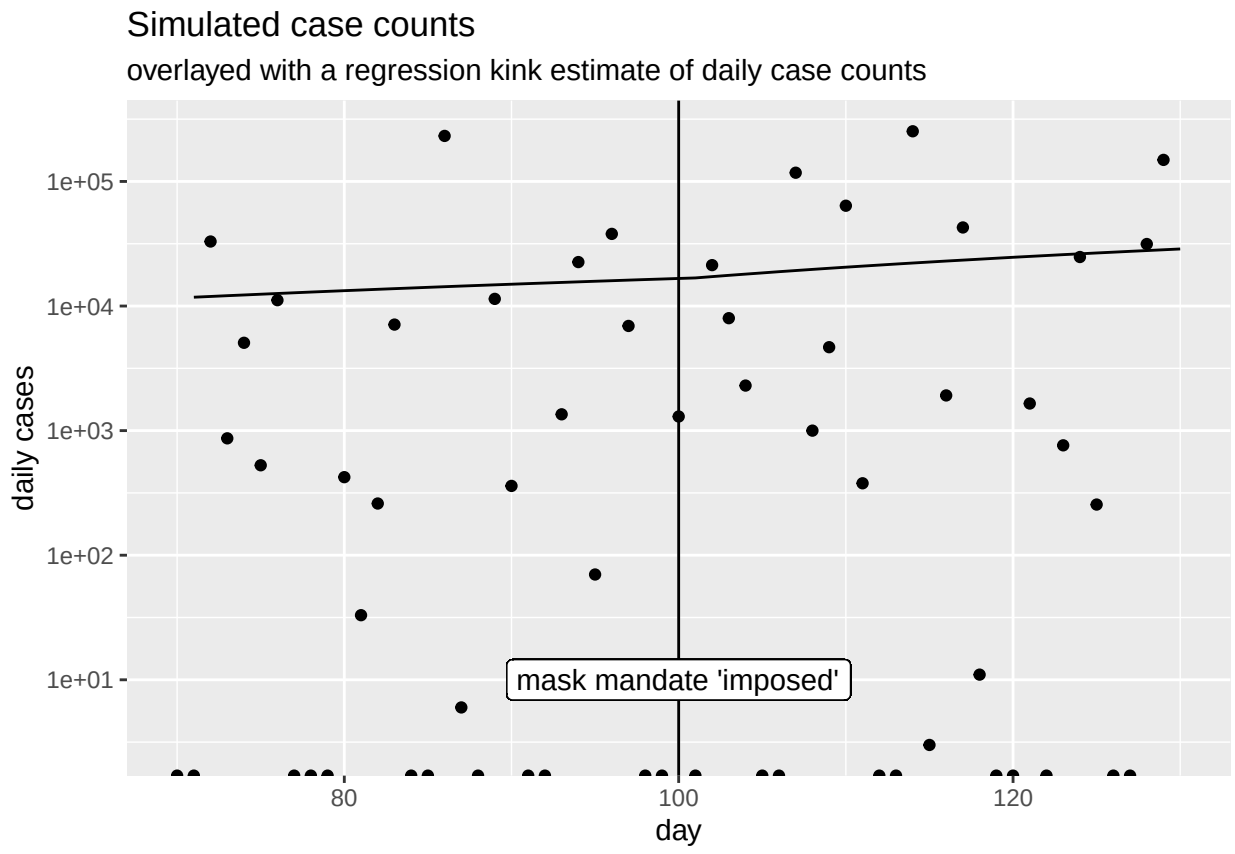
Regression kink model

Noting that the relationship between the treatment and case counts is multiplicative in our simulation model, we might deem it more appropriate to use a regression kink approach. We assume the following model:

$$\begin{aligned}
\beta_0, \beta_1, \alpha &\sim \text{Normal}(0, 100^2) \\
\epsilon &\sim \text{Normal}^+(0, 100) \\
C(t) &\sim \text{Normal}(\beta_0 + I[t < t_0]\beta_1 t + I[t \geq t_0](\beta_1 t_0 + (\beta_1 + \alpha)t), \epsilon^2)
\end{aligned}$$

where we again use flat priors, and t_0 is our intervention time. Here, the $\beta_1 t_0$ term ensures that the function is not discontinuous at t_0, and again the kink is modelled in α. Thus, a frequentist view would say that if the posterior credible interval of α contains 0, we have not identified a significant effect.

This model estimates without errors, but is more complicated to interpret. It is estimated that daily case counts increase in time at a rate of 168.4 per day (95% CI: 149.72 - 187.57) prior to the imposition of the simulated mask mandate. After the simulated mask mandate is imposed, that rate increases by 240.03 cases per day (95% CI: 102.13 - 371.17). This is a sensible estimate of the average daily case count, but it is not correct in the sense of recovering the simulated effect of the mask mandate, which should have reduced case counts. Below, we illustrate the estimated mean relative to the simulated case counts.



Simplified EpiNow model

Gaussian processes are expensive to sample from in Stan. (In a custom MCMC implementation, we might be able to use something elliptical slice sampling to make this more efficient [CITE], but we lose the convenience of Stan). Therefore, we consider the EpiNow model (see the section on simulation) where instead of parameterizing $R(t)$ through a Gaussian process, we use regression splines. Specifically, for a given spline basis matrix B (computed from the integer day each case count was observed TODO what window size?), we have:

$$\begin{aligned} \epsilon &\sim \text{Normal}^{+}(0,1) \\ \alpha_1, \alpha_2 &\sim \text{Normal}(0,1) \\ \alpha_k &\sim \text{Normal}(2\alpha_{k-1} - alpha_{k-2}, \epsilon) \\ \log R(t) &= B\alpha \end{aligned}$$

and the remainder of the model is as in our simulation procedure:

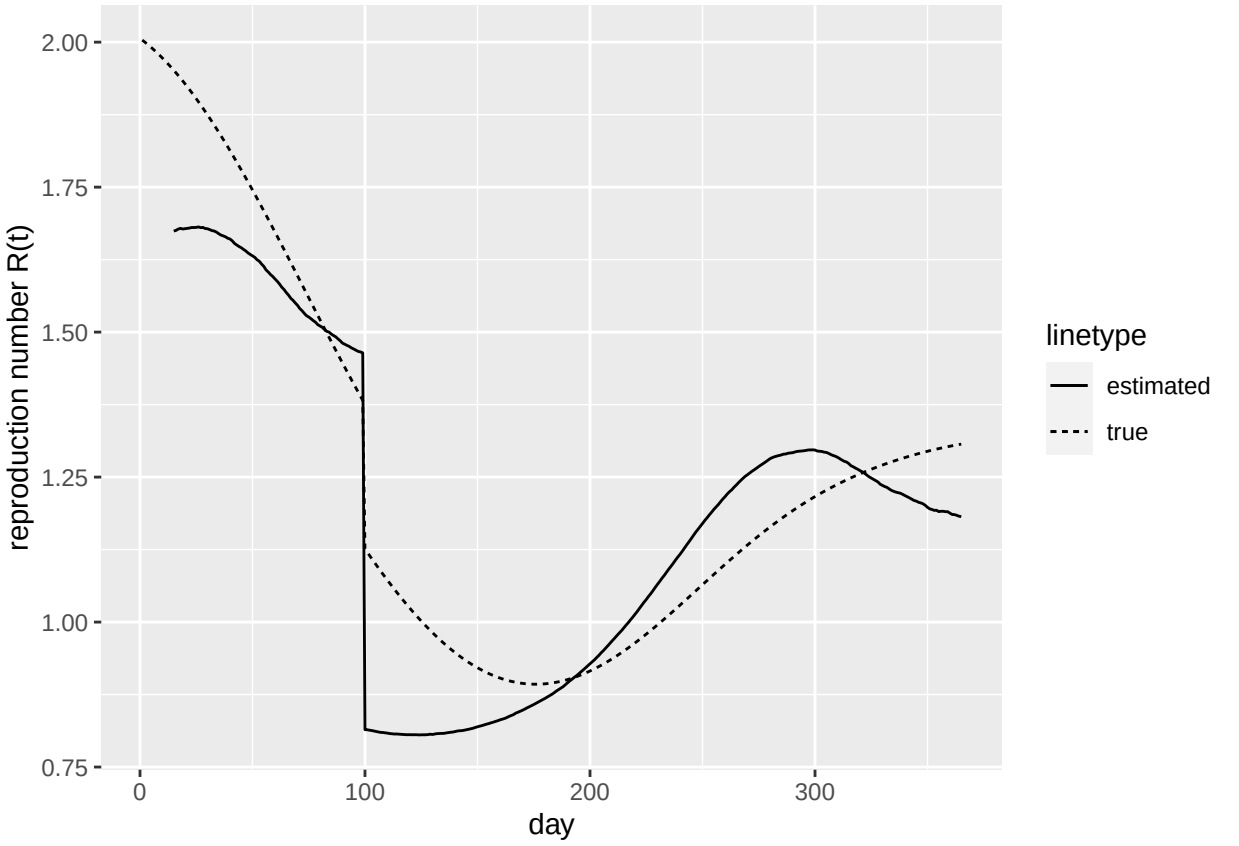
$$
\begin{aligned}
I(t) &= R(t) \sum_{i=1}^{\tau} w_i I(t-i) \\
D(t) &= \sum_{i=1}^{\tau} \xi_i I(t-i) \\
\phi &\sim \text{Exponential}(1) \\
C(t) &\sim \text{NegBin}(D_t \omega_{t \mod 7}, \phi)
\end{aligned}
$$

where we fix $\mathbf{w}$, ξ, and ω to their known values. Stan is capable of estimating this model, and despite the slight mis-parameterization of $R(t)$, recovers sensible estimates (not shown).

We also consider the case of a treatment effect under the following model:

$$
\log R(t) = B\alpha + \beta I[t \geq t_0]
$$

where the priors on α are as before. Recalling that our simulated effect is -0.2, this model recovers a directionally consistent effect of -0.584 (95% CI: -1.29, 0.389). This effect is clearly too large in magnitude, and visually it appears that this is because the model is unable to correctly apportion the complexity in the $R(t)$ function between the treatment effect and the spline. Additionally, this particular simulation was "difficult" for the model, because the model has increasingly less robust estimates of $R(t)$ closer to the start of the time series. These are all things to bear in mind as we transition to running the model on real data.



Application of this model to real data revealed strong posterior correlation between the αs and β, making it difficult to assess the treatment effect. Therefore, in the hierarchical case, we make a number of changes which result in our final model.

Hierarchical case

For the final hierarchical model, we consider a non-parametric regression discontinuity in $R(t)$. Specifically, we consider two sets of spline coefficients α_μ and β_μ, both with RW(2) priors as before. The coefficients α model the $R(t)$ to the left of the treatment, while the β model to the right. Our model is then:

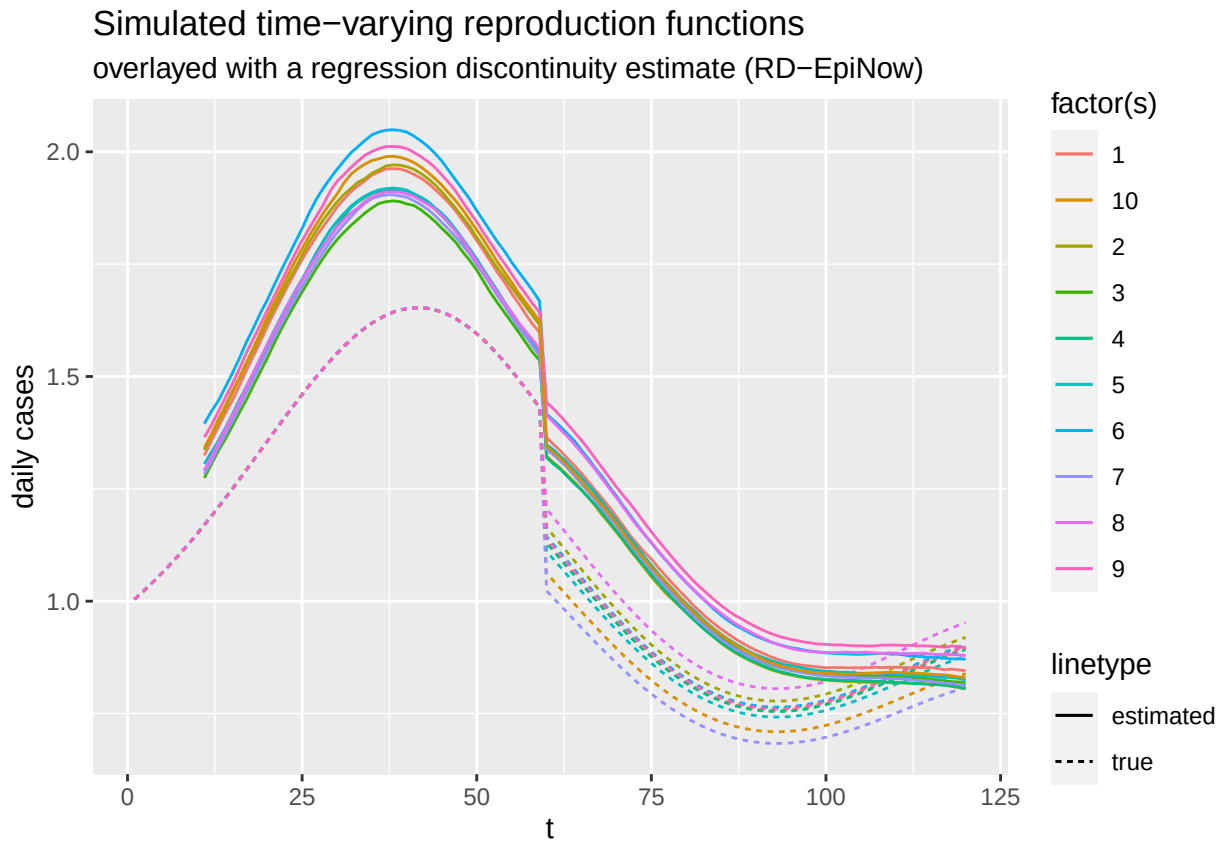
$$\begin{aligned}
\epsilon &\sim \text{Normal}^+(0, 1) \\
\alpha_i &\sim \text{Normal}(0, \epsilon) \\
\beta_i &\sim \text{Normal}(0, \epsilon) \\
\log R(t) &= B_t\dot{(}\alpha_m u + \alpha_i \mathbf{1})I[t < t_0] + B_t\dot{(}\beta_\mu + \beta_i \mathbf{1})I[t \geq t_0]
\end{aligned}$$

and the treatment effect for series i can be computed from a Monte Carlo sample as:

$$\rho_i = B_{t_0}\dot{(}\beta_\mu - \beta_i \mathbf{1} - \alpha_m u + \alpha_i \mathbf{1})$$

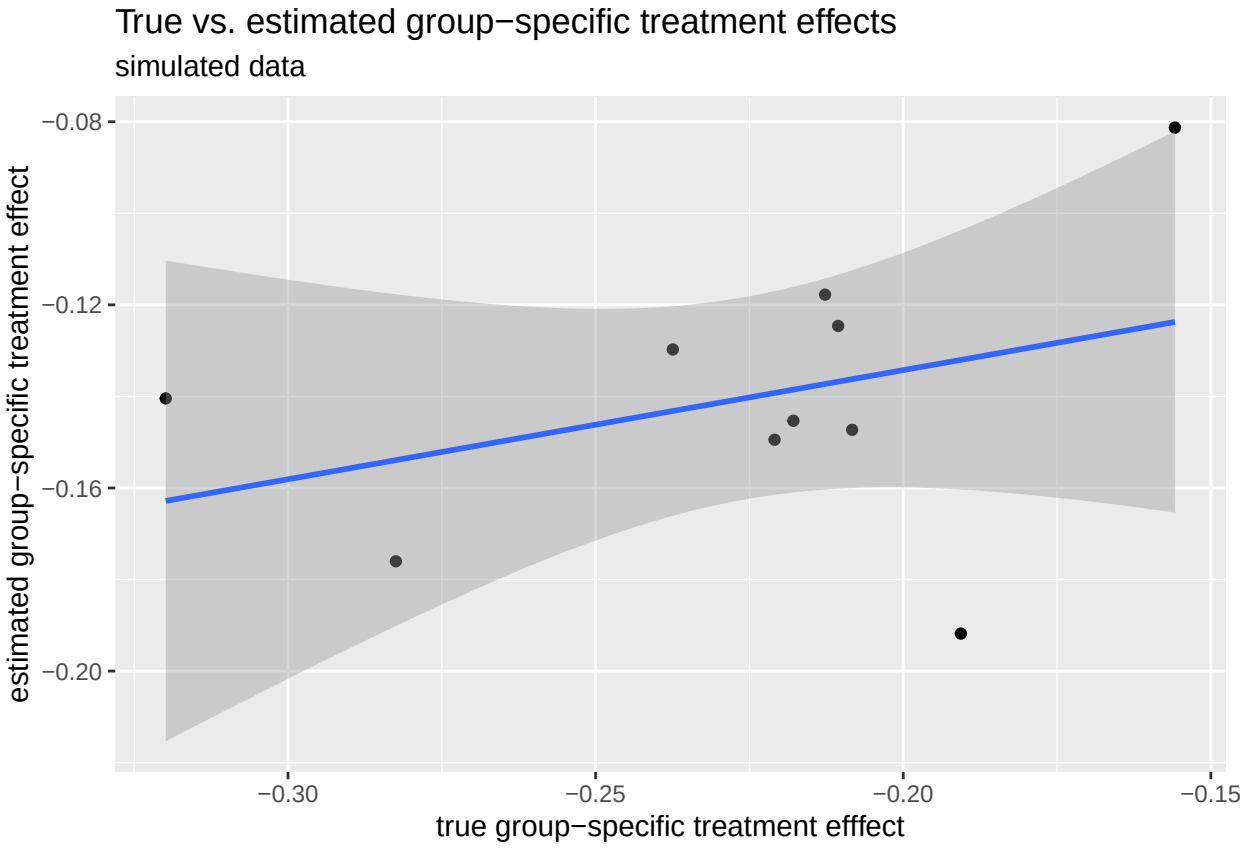
which is intuitively the difference between the spline estimated from the post-intervention data evaluated at t_0, and the spline estimated from the pre-intervention data evaluated at t_0.

Below, we plot the time-varying reproduction number for each of the 10 simulated groups against its true value. Although we capture the correct overall shape, and the estimated treatment effects look reasonable, we systematically overestimate $R(t)$.



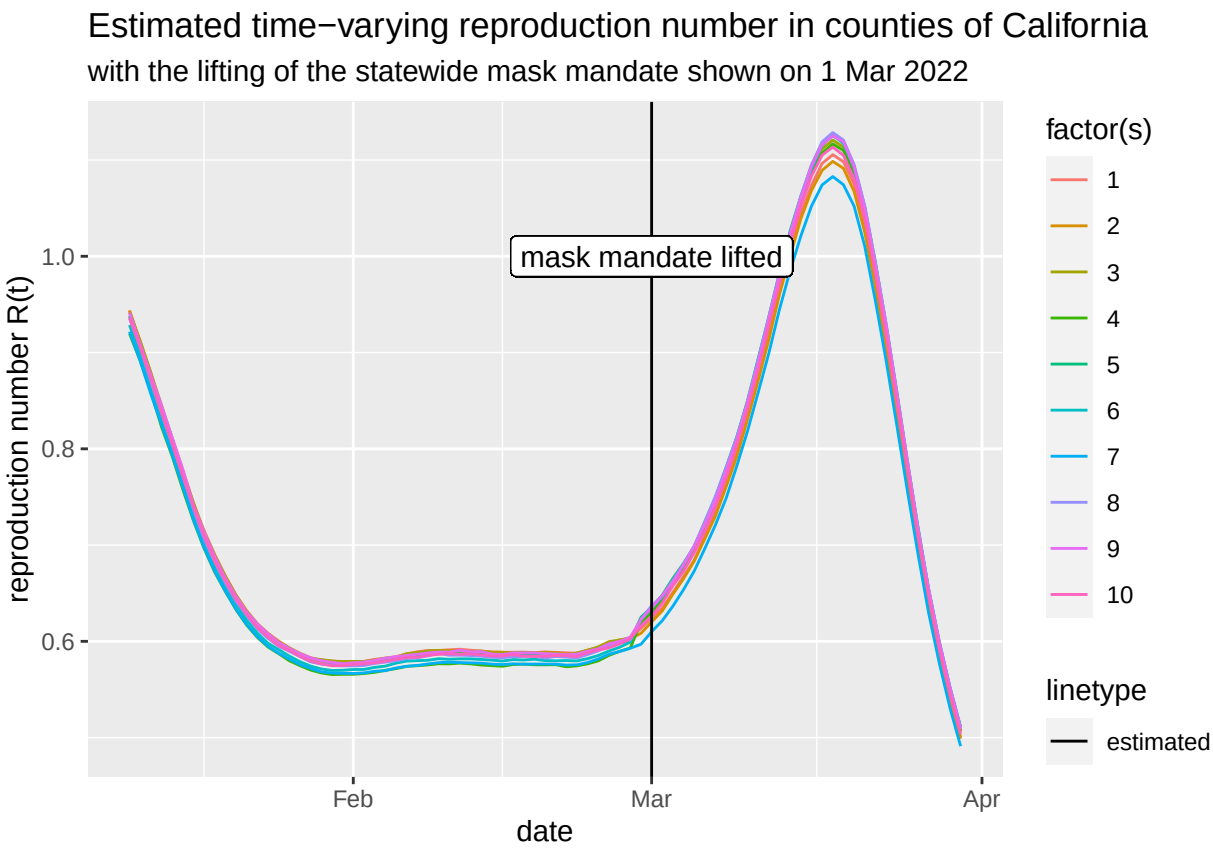
Below, we plot the true vs. estimated discontinuity effects. We see that there is no meaningful correlation between our estimates of the individual, group-specfic effects and the true effects. The mean estimated

reduction in $R(t)$ on the log-scale at the discontinuity is -0.14 (95% CI: -0.75, 0.616), and our point estimate is close to the true simulated mean treatment effect of -0.20.



Mask mandate in California

Despite these issues, we apply our model to the data around the time California lifted its mask mandate in the 10 counties we explored initially (we restricted the model to these counties due to computational budget). Below, we show the estimated $R(t)$, which has an invisible discontinuity the date the mask mandate was imposed.



Our estimate of the effect of lifting of the mask mandate is 0.026 (95% CI: -1.39, 1.50) which corresponds to an increase in $R(t)$ of 2.6% (95% CI: -75.1%, 348.1%).

Discussion

We do not claim that our model has recovered a causal effect of lifting the mask mandate, or that the effect is meaningful.

The main reason for this is that we have not sufficiently validated the approach to $R(t)$ estimation. The use of splines to model $R(t)$, with arbitrary hyperparameter selection, is a significant deviation from the original, validated EpiNow model. Further, for computational reasons, we have reduced the volume of data, and smoothed the input data in a particular way that may be invalid.

Further, the uncertainty in our estimate makes it functionally useless. It is not immediately clear how this could be ameliorated, but a reasonable place to start would be to examine whether the hierarchical model of the spline coefficients even makes sense. Some of this uncertainty is likely due to the fact that the mandate was lifted a little over 6 weeks ago, and this may not provide sufficient data to estimate the post-treatment $R(t)$ function.

Future work

Future work would likely address the following points.

Validation of spline-based modeling of $R(t)$. Use tooling such as leave-one-out cross-validation to select parameters such as window size, the type of spline, and the best way to code hierarchical splines in this

context. We opted for a linear difference between a "mean" set of spline coefficients, where this difference was allowed to vary between groups. This is probably bad.

Better quantification of treatment effect. We have quantified the treatment effect by examining the discontinuity in $R(t)$. However, our model is also a regression kink model, and we could take advantage of the differentiability of splines to quantify that kink. Our estimate in the California data shows a clear difference in slope on either side of the treatment effect, but this is not a realistic biological phenomenon and should be interrogated further.

Examine the endpoints. Strange things happen at the beginnings and ends of the examined time series, because we lack observations to inform the latent infection function (at the start of the series) and the reproduction number (at the end of the series). This seems to impact model fitting.

Divergent transitions. The hierarchical model does exhibit a few divergent transitions, and occasionally estimates a high Rhat (1.05) on the log probability. This is strange, because the Rhat on the parameters look OK.

Factors which inform treatment effect size. If a large hierarchical model could be fit, say, to multiple states, it would be interesting to assess whether demographic factors affect the magnitude of the treatment effect (e.g. whether a higher proportion of low-income individuals in a given county results in a larger jump in $R(t)$, perhaps due to closer working conditions).

Conclusions

The EpiNow model of time-varying reproduction number is amenable to parameterization with splines. This parameterization allows us to recover treatment effects in the simulated case, albeit with a high degree of uncertainty. Application of this model to the ten counties in California with the largest number of reported COVID-19 cases provides no evidence that lifting the statewide mask mandate in California had a significant effect on the time-varying reproduction number.