Estimating trends in transmission and mortality rates during the Covid-19 pandemic

John Sibert[*]
Joint Institute of Marine and Atmospheric Research
University of Hawai‘i at Mānoa
Honolulu, HI 96822 U.S.A.

August 4, 2020

Abstract

Write me

TO DO:
Rationalize examples
Implement alternate ordinate scale
Make linear plots of I and D?

Introduction

The sudden advent of the Covid-19 pandemic provoked many political jurisdictions to advise people to "shelter in place" and to practice "social distancing". If this advice has been effective, it should be possible to detect the effects of the advice by comparing changes in transmission rates over time and between areas. The SIR models are often applied to the spread of

[*] sibert@hawaii.edu; johnrsibert@gmail.com

epidemics and have certainly been applied to the current Covid-19 pandemic (Chen et al. 2020; Roques et al. 2020). These models divide the affected the effected population into three compartments: susceptible (S), Infected (I) and Recovered (R). SIR models are usually expressed as coupled ordinary differential equations,

$$\begin{aligned}
\frac{dS}{dt} &= -\beta\frac{IS}{N} - \mu S && (1)\\
\frac{dI}{dt} &= \beta\frac{IS}{N} - \mu I - \gamma I && (2)\\
\frac{dR}{dt} &= -\mu R + \gamma I && (3)\\
N &= S + I + R && (4)
\end{aligned}$$

where N is the population size, β is the instantaneous rate ($[t^{-1}]$), μ is the instantaneous mortality rate ($[t^{-1}]$), and γ is the instantaneous recovery rate ($[t^{-1}]$).

Unfortunately, few data sets include data for each of these compartments. The New York Times' "historical" data set[1] is easily accessible source of data and frequently updated. This data set comprise daily totals of "cases" and "deaths" for each county in the United States. I assume that the data included as "cases" are a reasonable approximations of the Infected compartment (I) in a SIR model. There are simply no credible data of comparable scope on either the Susceptible or the Recovered compartments.

[1] `https://github.com/nytimes/covid-19-data/`

Model Structure

I make some simplifying assumptions in the face of incomplete data: (1) The entire population is susceptible so that $S/N = 1$. (2) Over the short term, the size of the Susceptible compartment does not change, $\frac{dS}{dt} = 0 = \frac{dN}{dt}$, eliminating the Susceptible compartment. (3) People who recover from a Covid-19 infection return to the Susceptible compartment, eliminating the Recovered compartment. With these assumptions, and with the addition of a "deaths" compartment, the simplified SIR model is

$$\frac{dI}{dt} = \beta I - \mu I - \gamma I \qquad (5)$$

$$\frac{dD}{dt} = \mu I \qquad (6)$$

and has state variables that might be matched to available observations.

The data available during the initial stages of the Covid-19 pandemic contain measurement errors of various types. Definitions and methods of detecting and reporting the numbers of infected persons and numbers of deaths attributable to Covid-19 have evolved since January of 2020, are continuing to change, and can be expected to change in the future. Reporting protocols also vary between political jurisdictions (or "geographies" in the parlance of the New York Times). Finally, there is additional variability in the biosocial processes that mediate disease transmission.

State-space models separate variability in the biosocial processes in the system (transition model) from errors in observing features of interest in the system (observation model). (See Harvey 1990).

The general form of a state-space process or transition model is

$$\alpha_t = T(\alpha_{t-1}) + \Theta_t \tag{7}$$

where α_t is the state at time t and the function T embodies the dynamics mediating the development of the state at time t from the state at the previous time with random process error, Θ_t.

The transition model for the simplified SIR model is constructed from the explicit finite difference approximations of equations (5) and (6) with associated log-normal random errors.

$$\begin{aligned} I_t &= I_{t-\Delta t}\big(1 + \Delta t(\beta_{t-\Delta t} - \mu_{t-\Delta t} - \gamma_{t-\Delta t})\big)e^{\eta_t} && (8) \\ D_t &= \big(D_{t-\Delta t} + \Delta t \mu_{t-\Delta t} I_{t-\Delta t}\big)e^{\eta_t} && (9) \end{aligned}$$

where η is a normal random deviate, $\eta \sim N(0, \sigma_\eta)$, representing temporal variability in the biosocial factors that mediate the spread of the pandemic. The recovery rate, $\gamma_{t-\Delta t}$, in equation (8) is computed algebraically as

$$\gamma_{t-\Delta t} = \beta_{t-\Delta t} - \mu_{t-\Delta t} + (1 - \frac{I_t}{I_{t-\Delta t}}) \tag{10}$$

I have no particular justification, beyond the parsimony principle, for the assumption that the variance, σ_η, of the processes for I and D, should be the same.

One approach to modeling time-dependent rates of transmission and mortality, β and μ, is to treat them as random effects (Skaug and Fournier 2006). Random effects are appropriate if repeating a time series of observations would not yield the same outcome as the initial observations. Random effects are also appropriate when observing the same process in two different areas. I model the β and μ time series as log-normal random walks. I assume that

$$\begin{aligned}\log\beta_t &= \log\beta_{t-\Delta t}+\varepsilon;\quad \varepsilon\sim N(0,\sigma_\beta) && (11)\\ \log\mu_t &= \log\mu_{t-\Delta t}+\varrho;\quad \varrho\sim N(0,\sigma_\mu) && (12)\end{aligned}$$

A similar approacth has been used in fisheries stock assessment models to estimate time-dependent fishing induced mortalitty (Sibert 2017; Nielsen and Berg 2014).

The general form of the state-space observation model is

$$x_t = O(\alpha_t)+\Omega_t \tag{13}$$

where the function O describes the measurement process with error Ω in observing the state α.

I applied separate observation error models for cases and deaths. The observation model for cases is a simple log-normal error

$$\log\varphi_t = \left(\log\frac{1}{\sqrt{2\pi\sigma_I^2}} - \Big(\frac{\log I_t-\log\widehat{I}_t}{\sigma_I}\Big)^2\right) \tag{14}$$

where I is the observed number of cases and $\widehat{I}$ is the number of cases predicted by equation 8.

Not all those afflicted by Covid-19 have died; there are far fewer deaths than infections. In addition, the observed time series for both I and D begins at the first recorded case, i.e. at time $t=0, I_t\geq 1$. The first recorded death occurs several days or weeks after the first recorded case. Therefore the deaths time-series inevitably contains a substantial number of initial recorded

Table 1: List of model variables for the simple SIR model, `simpleSIR4`. There are two state variables computed from the of estimated parameters and random effects. There are two random effects and five estimated variance parameters. All models variables are represented in the TMB C++ module as their natural logarithms.

Variable	Definition
	State variables:
I	Number of infected individuals
D	Number of deaths
	Random effects:
β_t	Transmission rate; log-normal random walk
μ_t	Mortality rate; log-normal random walk
	Estimated parameters:
σ_I	Infectious compartment estimation standard deviation
σ_D	Deaths compartment estimation standard deviation
σ_η	Standard deviation of transmission and deaths process errors
σ_β	Standard deviation of transmission rate random walk
σ_μ	Standard deviation of mortality rate random walk

zeros. The observation model for deaths accommodates observed zeroes by assuming to be "zero-inflated" log normal likelihood given by

$$\log \varepsilon_t = \begin{cases} D_t > 0: & (1-p_0)\cdot\left(\log\frac{1}{\sqrt{2\pi\sigma_D^2}} - \left(\frac{\log D_t - \log \widehat{D}_t}{\sigma_D}\right)^2\right) \\ D_t = 0: & p_0 \cdot \log\frac{1}{\sqrt{2\pi\sigma_D^2}} \end{cases} \tag{15}$$

where D is the observed number of deaths, $\widehat{D}$ is the number of deaths predicted by equation 9, and p_0 is the proportion of observed deaths equal to zero.

Model parameters are estimated by maximizing the joint likelihood of the process errors, observation errors, and random effects.

$$L(\theta, \alpha, x) = \prod_{t=2}^{m} \left[\phi\big(\alpha_t - T(\alpha_{t-1}), \Sigma_\eta\big)\right] \cdot \prod_{t=1}^{m} \left[\phi\big(x_t - O(\alpha_t), \Sigma_\varepsilon\big)\right] \tag{16}$$

where m is the number of days elapsed since the first recorded case, x_t is the vector of daily observations of cases and deaths, α_t is the vector of the daily calculations of the state variables and random effects, and θ is a vector of model parameters (Table 1). The R package TMB (Kristensen et al. 2016) was used to estimate the parameters of the model. The R and supporting C++ files are available on github.[2]

Results

Six months after the pandemic began spreading in the United States, it was obvious that some areas were more successful than other in controlling the spread of the corona virus. Trends in the per-capita number of cases in the thirtytwo largest counties in the United States are shown in Figure 1. These trajectories fall into two more or less distinct groups: those that are concave downward, e.g. Nassau Co. NY (NaNY), indicating successful control, and those that are concave upward, e.g. Miami-Dade Co. FL (MDFL), indicateing less successful control.

Prevalence histories for two counties representative of concave downward and concave upward trajectories are shown in Figure 2. The 11-day moving averages of the daily increases in cases and deaths indicate daily trends is equivalent to the trends in Figure 1. All histories show extreme day to day

[2] `simpleSIR4` at `https://github.com/johnrsibert/SIR-Models`

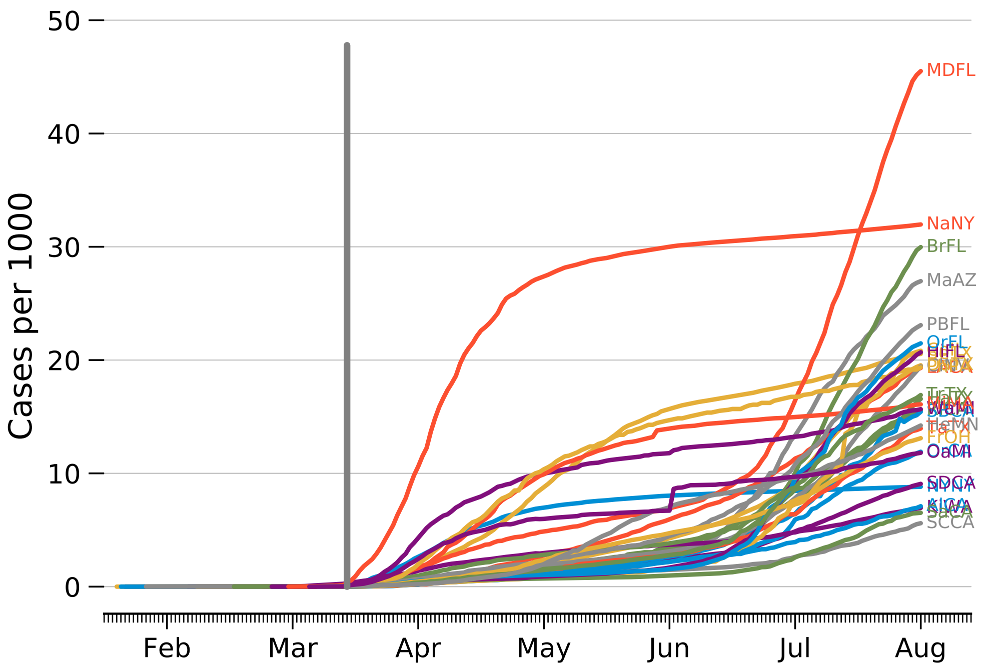
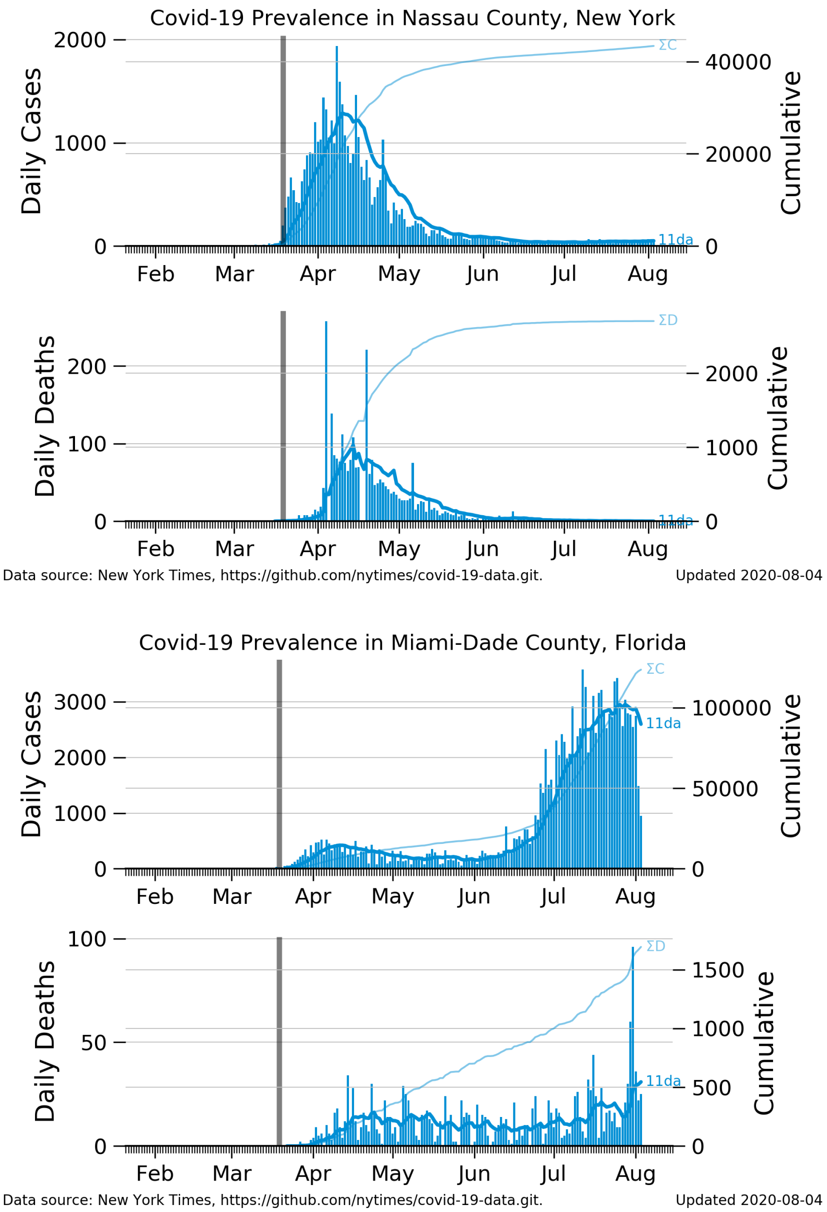


Figure 1: Trends in number of cases per 1000 people in the 30 most populous US counties. The vertical gray bar mark the March 19, 2020 California shelter in place order.



Covid-19 Prevalence in Nassau County, New York

Data source: New York Times, https://github.com/nytimes/covid-19-data.git. Updated 2020-08-04

Covid-19 Prevalence in Miami-Dade County, Florida

Data source: New York Times, https://github.com/nytimes/covid-19-data.git. Updated 2020-08-04

Figure 2: Prevalence trajectories for two US counties. Blue bars indicate daily increases increases in cases and deaths; dark blue lines enclosing the bars indicate 11 day moving averages of daily increases (labeled "11da"); pale blue lines indicate cumulative numbers (labeled ΣC and ΣD); vertical gray bar marks the March 19, 2020 California shelter in place order.

variability. Variability is most notable in the deaths time series, particularly for smaller counties.

The `simpleSIR4` model estimates two random effects and five parameters. In principle, all random effects and parameters are estimated simultaneously. Initial experiments with the model showed that several different numerical algorithms used to find the minimum of the log likelihood function were unable to easily reach a solution. Minima were reached for some counties, but most attempts terminated prematurely. Inspection of the diagnostic plots for the model showed that predicted values of cases and deaths matched observed values almost exactly with unrealistically low estimates of $\sigma_{\ln I}$ and $\sigma_{\ln D}$. The extreme variability in the data is reflected in the extreme variability of the estimated trends in transmission and mortality rates estimated by the unconstrained 5 parameter model. See Appendix A for details.

The `simpleSIR4` model can be configured with selected parameters fixed at constant values. All subsequent analysis focused on models with $\sigma_{\ln I}$ $= 0.223$ and $\sigma_{\ln D} = 0.00953$. These standard deviations are equivalent to measurement errors of approximately 25% in reporting cases and 10% in reporting deaths. The algorithm converges to a solution in all cases, and converges rapidly using gradient methods.

Diagnostics are plotted (figures 3 and 4) on logarithmic scales to both illustrate the lognormal likelihood functions used in the observation model, equations (14) and (15), and to illustrate trends in estimated transmission and mortality rates close to zero. The blue '+' symbols represent the observed cases (I) and deaths (D). The red lines overlaying the symbols are

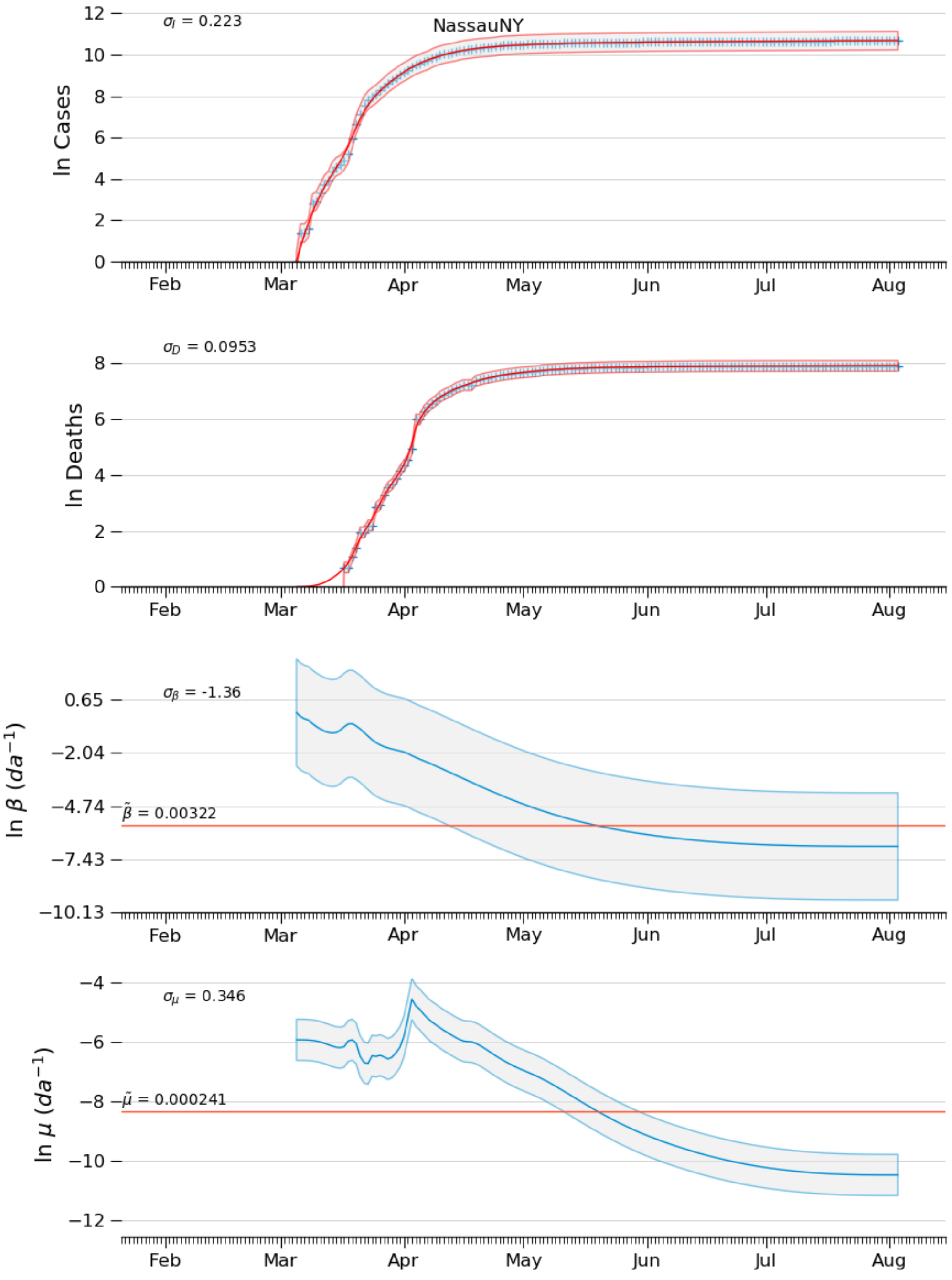


Figure 3: Diagnostic plots of model estimates for Nasssau County, NY, with constraints of the observation model variance, $\sigma_{\ln I} = 0.223$ and $\sigma_{\ln D} = 0.00953$. See page 10 for explanation of figure.

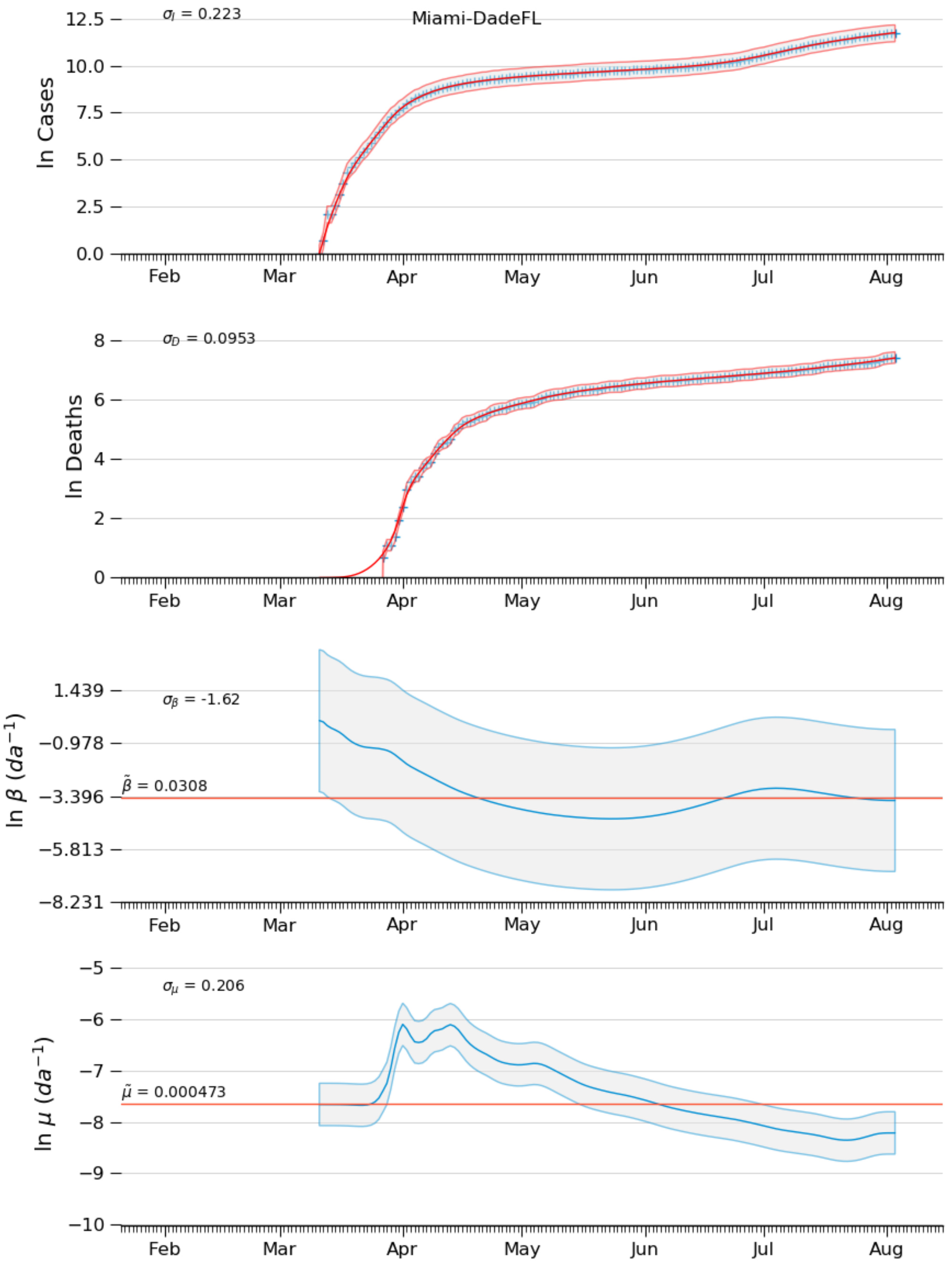


Figure 4: Diagnostic plots of model estimates for Miami-Dade County,FL, with constraints of the observation model variance, $\sigma_{\ln I} = 0.223$ and $\sigma_{\ln D} = 0.00953$. See page 10 for explanation of figure.

model predictions ($\widehat{I}$) and ($\widehat{D}$) of cases and deaths. σ_I and σ_D are the estimated standard deviations of logarithms of cases and deaths in the observation model. The shaded areas bounded by red outlines are ±2 estimated standard deviations around the estimated trends. The solid blue lines in the transmission rate $\ln\beta$ and mortality rate $\ln\mu$ diagnostic plots are the estimated transmission and death rate random effects. The shaded areas bounded by blue outlines are estimated random effects ±2 standard deviations of the generating random walks. The red lines labeled $\tilde{\beta}$ and $\tilde{\mu}$ are the medians of the estimated random effects over the time period.

Diagnostic plots for the constrained model are shown in figures 3 and 4 for convex downward and convex upward trajectories respectively. Estimated cases and deaths agree well with observation throughout the time series.

Figure 5 compares estimated transmission rate among counties. Transmission rates increased rapidly at the beginning of the pandemic exceeding 1da^{-1} ($\ln\beta \approx 0$) in early March, an instantaneous transmission rate equivalent to a doubling time of less than one day. Beginning in April, transmission rates fell substantially, and doubling times increased to longer than 20 days in some counties. Counties with estimated transmission rates less than 0.007da^{-1} (or $\ln\beta \leq 5$) at the end of May correspond roughly to those counties with concave downward prevalence trajectories. Figure 6 is a simplified presentation of estimated transmission rate between counties that compare a county with sustainable suppression of transmission (Cook Co, IL, Nassau Co, NY) with two counties that have suffered a resurgence of cases (Honolulu Co, HI and Miami-Dade Co. FL). The Honolulu example indicates that sim-

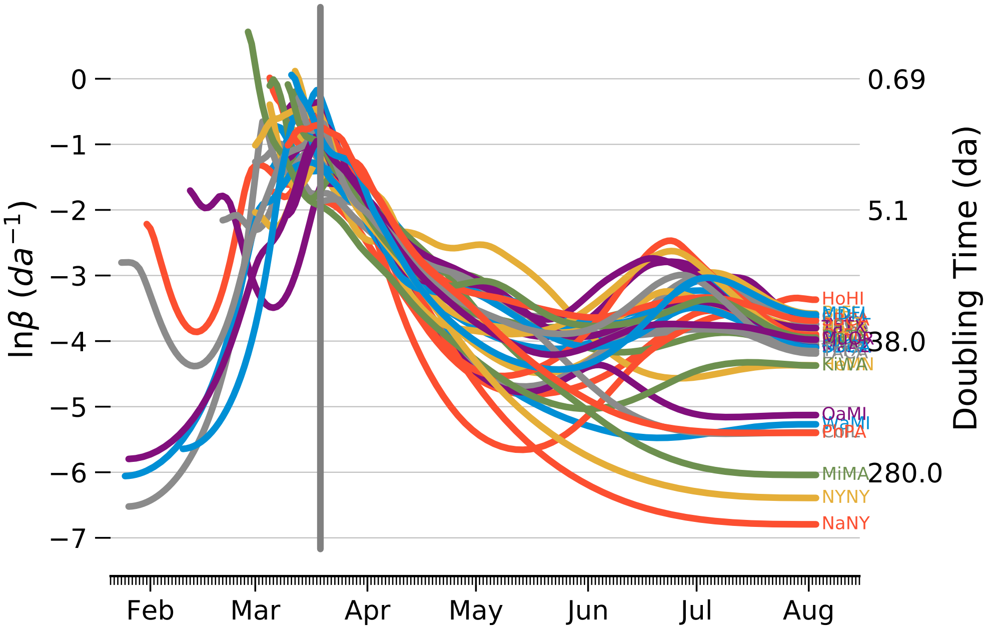


Figure 5: Estimated natural logarithms of the transmission rate for several US counties using the constrained `simpleSIR4` model. Equivalent doubling times ($t_2 = \frac{\ln 2}{\exp(\ln\beta)}$) are shown on the right-hand ordinate.

ply suppressing the transmission rate to a point where the doubling time is greater than 100 days does not ensure a sustainable outcome.

Figure 7 compares estimated mortality rate among counties. Initial mortality rates were around 0.01da^{-1} ($\ln\mu \approx 4.6$) at the beginning of the pandemic but fell steadily to less than 0.001da^{-1} ($\ln\mu \approx -6.9$) in July.

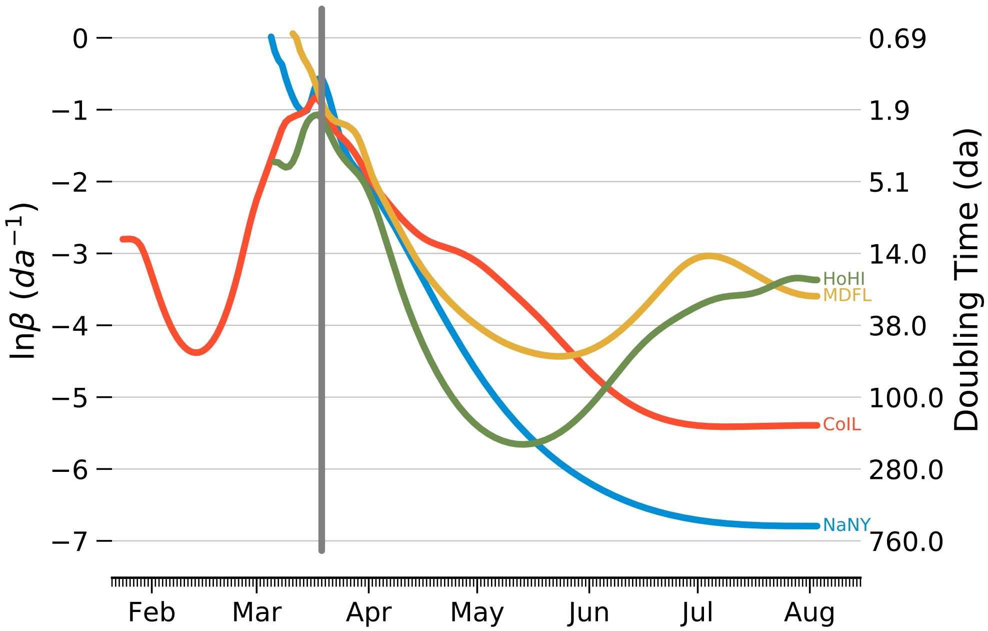


Figure 6: Estimated natural logarithms of the transmission rate for four US counties using the constrained `simpleSIR4` model. Equivalent doubling times ($t_2 = \frac{\ln 2}{\exp(\ln\beta)}$) are shown on the right-hand ordinate.

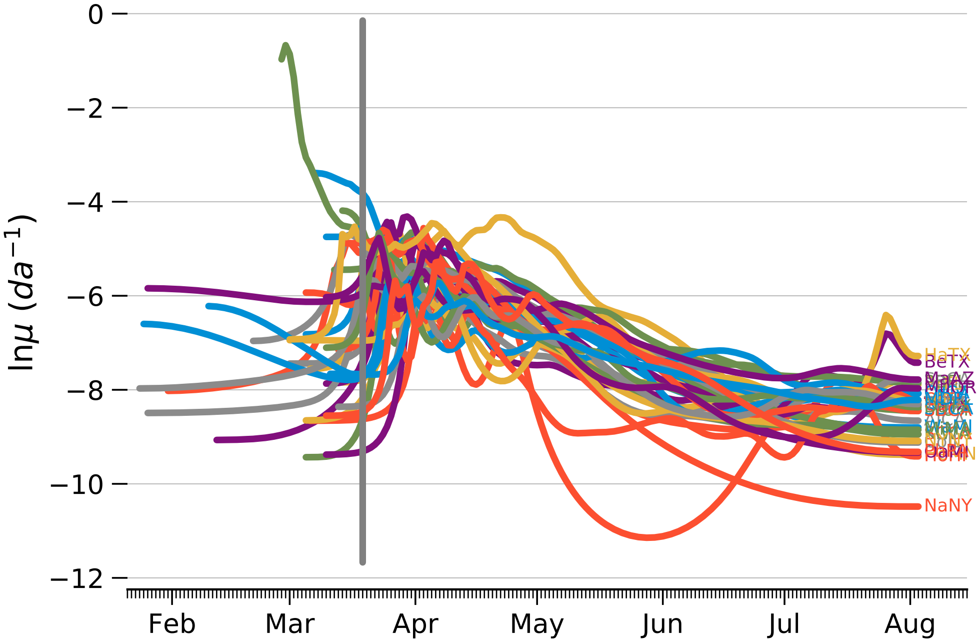


Figure 7: Estimated natural logarithms of the mortality rate for several US counties using the constrained `simpleSIR4` model.

Discussion

Nonlinear statistical models with multiple estimated parameters rely on numerical methods to estimate parameters by searching for minima (hopefully finding only one) in the negative of the likelihood function. The parameter values at the minima are considered to be maximum likelihood estimators. Failure to find well defined minima is usually a cause for concern. The minimization algorithms applied to unconstrained `simpleSIR4` do not reliably converge to a solution. The standard deviations in the observation model are components of the likelihood, and the algorithm therefore pushes these parameters toward zero. Since the estimated parameters in `simpleSIR4` are represented as logarithms, they cannot take values $\leq$ zero. Restricting the values of σ_I and σ_D to small, but non-zero, constants allows the algorithm to estimate the other parameters.

The trends in estimated transmission rate Figure 5 seem reasonable. The extremely high transmission rates in March agree well with doubling times reported in newspaper articles at the time. The steady decline of transmission rates after shelter-in-place advice is also consistent with casual observation. The incubation time of the Covid-19 virus is usually assumed to be about 14 days. The trends in Figure 5 in conjunction with the empirical prevalence trends suggest that sustainable containment of the pandemic does not occur unless the instantaneous transmission rate is forced below $0.018da^{-1}$, that is, unless the doubling time is greater than 35 days, approximately twice the incubation period.

Upward bump in transmission rates consistent with observed increases in cases in July.

Omit: Whether the available data are sufficiently informative to enable estimation of the model parameters is a critical aspect of the evaluation of any statistical model. The speed at which the Covid-19 pandemic spread during the first quarter of 2020 means that the length of the time series doubled during the development of this model. The capability of the model improve conveniently during the model development period, but whether the improvement is attributable to changes in model structure or to the increase in the length of the time series is unclear. This ambiguity influenced the development of the model.

Sibert 2017; Nielsen and Berg 2014

Table 2: Model results. Estimating β and μ trends as random effects with computed γ and constraints on σ_I and σ_D. Data updated 2020-08-04 from https://github.com/nytimes/covid-19-data.git.2020-08-04

County	n	p_0	f	C	σ_η	σ_β	σ_μ	σ_I	σ_D	$\tilde{\gamma}$	$\tilde{\beta}$	$\tilde{\mu}$
Nassau, NY	151	0.0789	-348	0	0.14	0.256	0.346	0.223	0.0953	-1.22e-08	0.00322	0.000241
New York City, NY	155	0.0833	-310	0	0.161	0.217	0.36	0.223	0.0953	-2.36e-08	0.00533	0.000405
Wayne, MI	146	0.0544	-336	0	0.14	0.229	0.16	0.223	0.0953	-1.8e-08	0.00619	0.000875
Oakland, MI	146	0.068	-332	0	0.137	0.226	0.426	0.223	0.0953	-1.62e-08	0.0101	0.000594
Philadelphia, PA	146	0.102	-340	0	0.126	0.175	0.425	0.223	0.0953	-2.42e-08	0.0106	0.000532
Middlesex, MA	151	0.105	-357	0	0.124	0.237	0.371	0.223	0.0953	-1.25e-08	0.0107	0.000451
King, WA	157	0.00633	-431	0	0.126	0.234	0.339	0.223	0.0953	-8.63e-09	0.013	0.000481
Franklin, OH	142	0.0629	-419	0	0.102	0.157	0.3	0.223	0.0953	-1.76e-08	0.0208	0.00101
Honolulu, HI	150	0.166	-484	0	0.0721	0.227	0.49	0.223	0.0953	-5.15e-08	0.0215	0.000174
Cook, IL	192	0.275	-473	0	0.101	0.234	0.217	0.223	0.0953	-2.2e-07	0.0226	0.000466
Alameda, CA	155	0.141	-465	0	0.0808	0.133	0.248	0.223	0.0953	-3.45e-08	0.0231	0.000527
Multnomah, OR	146	0.0272	-497	0	0.0798	0.181	0.318	0.223	0.0953	-5.07e-08	0.0233	0.000362
Santa Clara, CA	185	0.204	-586	0	0.071	0.237	0.274	0.223	0.0953	-1.55e-07	0.0246	0.000352
Los Angeles, CA	190	0.236	-454	0	0.102	0.305	0.244	0.223	0.0953	-3.45e-07	0.0249	0.000423
San Diego, CA	175	0.244	-430	0	0.0956	0.277	0.317	0.223	0.0953	-2.63e-07	0.0267	0.000719
Riverside, CA	149	0.06	-482	0	0.09	0.138	0.183	0.223	0.0953	-2.72e-08	0.0294	0.000895
Palm Beach, FL	144	0.069	-378	0	0.111	0.166	0.157	0.223	0.0953	-1.54e-08	0.0302	0.000942
Harris, TX	151	0.0921	-373	0	0.102	0.197	0.322	0.223	0.0953	-2.44e-08	0.0302	0.000301
Miami-Dade, FL	145	0.11	-342	0	0.133	0.197	0.206	0.223	0.0953	-1.08e-08	0.0308	0.000473
Orange, FL	143	0.0208	-418	0	0.101	0.21	0.435	0.223	0.0953	-1.98e-08	0.0314	0.00024
Clark, NV	151	0.0724	-436	0	0.1	0.158	0.218	0.223	0.0953	-3.3e-08	0.0319	0.000743
Travis, TX	143	0.0972	-359	0	0.1	0.19	0.271	0.223	0.0953	-1.73e-08	0.0327	0.000319
Tarrant, TX	146	0.068	-414	0	0.1	0.138	0.42	0.223	0.0953	-3.12e-08	0.0328	0.000335
Broward, FL	150	0.0728	-391	0	0.107	0.167	0.447	0.223	0.0953	-2.15e-08	0.0333	0.000411
Orange, CA	191	0.313	-519	0	0.0769	0.234	0.266	0.223	0.0953	-3.73e-07	0.0334	0.000732
Dallas, TX	146	0.0612	-388	0	0.11	0.17	0.309	0.223	0.0953	-1.43e-08	0.0336	0.000504
San Bernardino, CA	141	0.0634	-406	0	0.0999	0.14	0.198	0.223	0.0953	-2.06e-08	0.0343	0.000794
Sacramento, CA	164	0.109	-574	0	0.0666	0.174	0.252	0.223	0.0953	-8.02e-08	0.0343	0.000428
Hennepin, MN	144	0.103	-359	0	0.114	0.207	0.402	0.223	0.0953	-1.24e-08	0.036	0.000917
Hillsborough, FL	155	0.16	-461	0	0.0813	0.189	0.227	0.223	0.0953	-7.15e-08	0.0373	0.00065
Maricopa, AZ	190	0.283	-482	0	0.0897	0.235	0.158	0.223	0.0953	-4e-07	0.0422	0.00188
Bexar, TX	173	0.224	-478	0	0.0745	0.213	0.376	0.223	0.0953	-7.61e-08	0.0461	0.000393
Median	150.5	0.09465	-418.5	0	0.101	0.202	0.3045	0.223	0.0953	-2.43e-08	0.0298	0.000477

References

Baudin, Michale (2010). "Nelder-Mead User's Manual". In: April, p. 119.

Chen, Yi-Cheng, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu (2020). "A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons". In: pp. 1–18. arXiv: `2003.00122`. URL: `http://arxiv.org/abs/2003.00122`.

Harvey, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press. ISBN: 978-0521321969.

Kristensen, K., A. Nielsen, C.W. Berg, H.J. Skaug, and B.M. Bell (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70, pp. 1–21. DOI: `doi:10.18637/jss.v070.i05`.

Nielsen, Anders and Casper W. Berg (2014). "Estimation of time-varying selectivity in stock assessments using state-space models". In: *Fish. Res.* 158, pp. 96–101. ISSN: 01657836. DOI: `10.1016/j.fishres.2014.01.014`. URL: `http://dx.doi.org/10.1016/j.fishres.2014.01.014`.

Roques, Lionel, Etienne Klein, Julien Papa, and Samuel Soubeyrand (2020). "Modele SIR mecanistico-statistique pour l'estimation du nombre d'infectes et du taux de mortalite par COVID-19". In: pp. 1–11. arXiv: `arXiv:2003.10720v2`.

Sibert, John (2017). "Assessing of a portion of the Pacific Thunnus albacares stock : Ahi in the Main Hawaiian Islands". In: *arxiv.org* arXiv:1702. arXiv: `arXiv:1702.01217v1`.

Skaug, Hans J and David A Fournier (2006). "Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models". In: *Comput. Stat. Data Anal.* 51.2, pp. 699–709. ISSN: 01679473. DOI: `10.1016/j.csda.2006.03.005`.

‘

Appendix

A Unconstrained Estimates

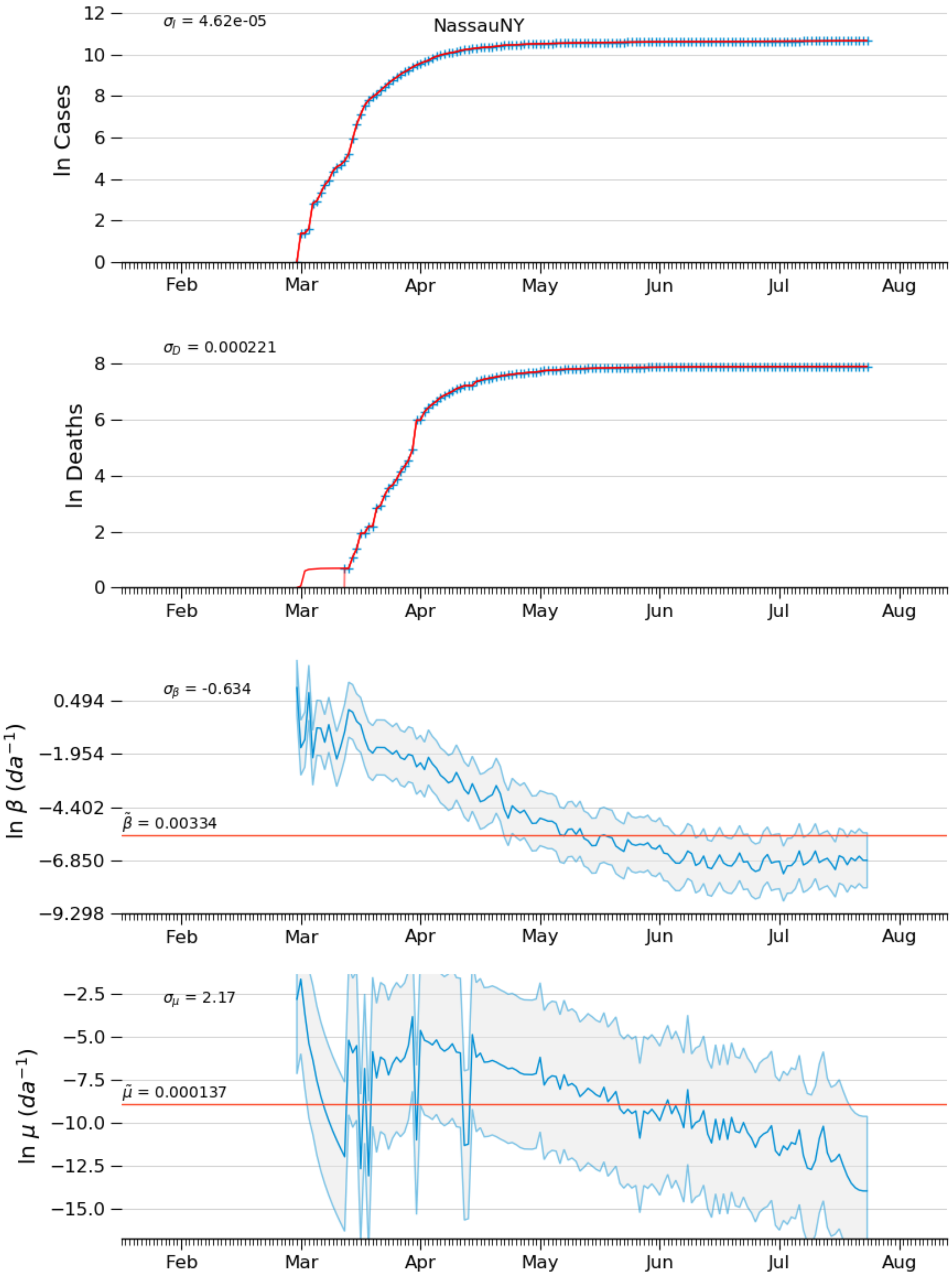


Figure A.1: Diagnostic plots of model estimates for Nasssau County, NY, without constraints of the observation model variance. See page **??** for explanation of figure.

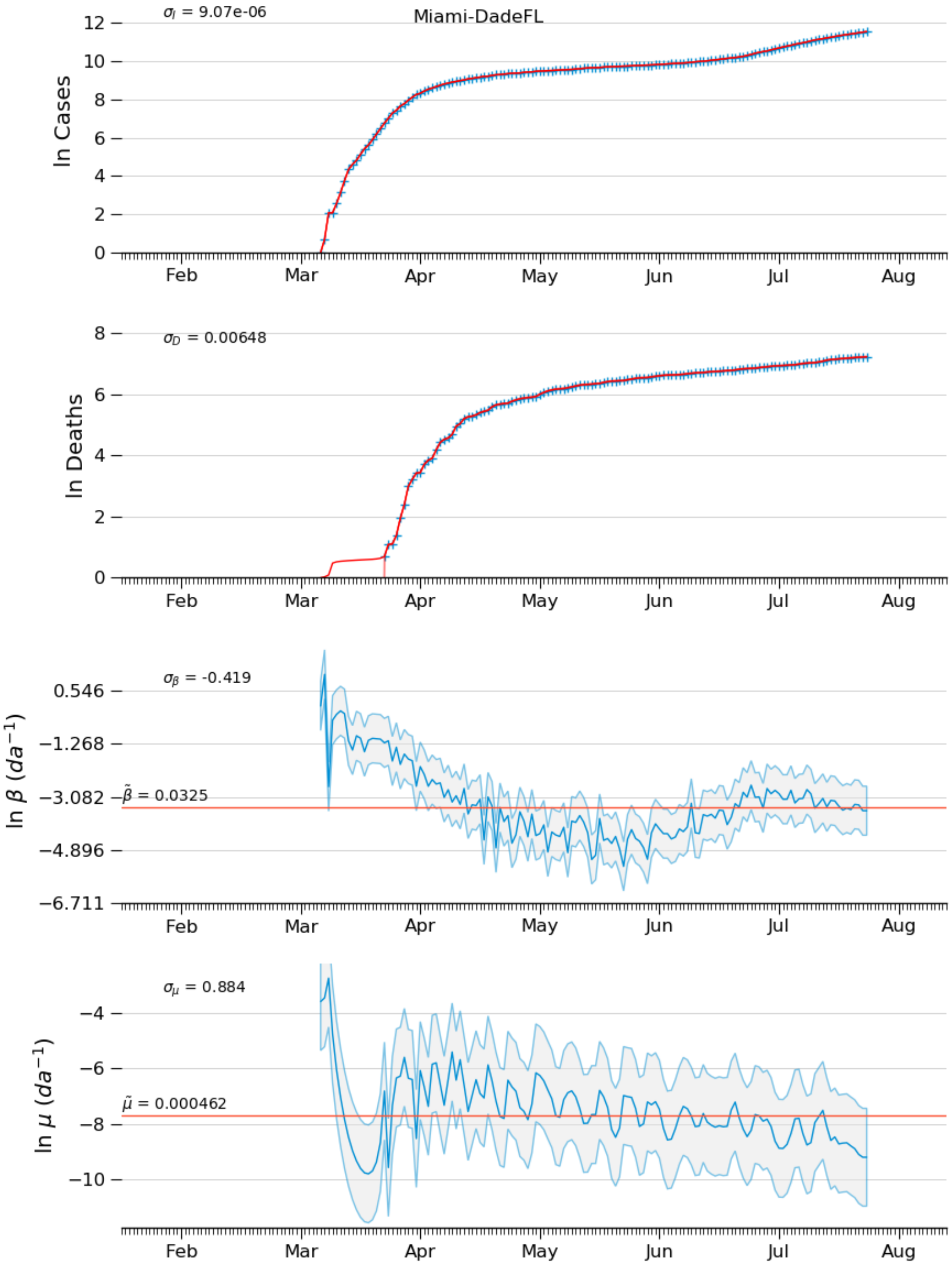


Figure A.2: Diagnostic plots of model estimates for Miami-Dade County,FL, without constraints of the observation model variance. See page 10 for explanation of figure.