Computational and Applied Mathematics (2023) 42:96
https://doi.org/10.1007/s40314-023-02241-w

Regression model for the reported infected during emerging pandemics under the stochastic SEIR

Ivair R. Silva[1] · Yan Zhuang[2] · Debanjan Bhattacharjee[3] · Igor R. de Almeida[1]

Received: 3 September 2022 / Revised: 3 February 2023 / Accepted: 7 February 2023 /
Published online: 24 February 2023

Abstract
The COVID-19 pandemic revealed the necessity of measuring the statistical relationship between the transmission rate of epidemic diseases and the social/behavioral, logistical, and economic variables of the affected region. This paper introduces a regression model to estimate the impact of such covariates on the infectious rate of epidemiological agents. Hidden logistical predictor components, such as weekly seasonality of reported data, can also be accessed with the proposed methodology. For this, we assume that the dynamics of officially reported data of emerging pandemics, related to infected groups, follows a stochastic SEIR model. The main advantage of our method is that it is based on a new three-step algorithm that combines the classical likelihood principle, the minimization of the mean squared error, and a tri-section algorithm to estimate, simultaneously, the coefficients of the covariates and the parameters of the compartmental model. Simulation studies are provided to certify the accuracy of the proposed inference methodology. The model is further applied to analyze the official statistical reports of COVID-19 data in the state of São Paulo, Brazil.

Keywords COVID-19 · Social isolation · Reported infected · SEIR · Social isolation

Mathematics Subject Classification 92B15 · 62J02 · 62M02 · 62M05 · 62M10

1 Introduction

On the study of emerging epidemics, the statistical modeling can be focused on the total number of infected (reported+unreported) individuals, or on the number of reported infected. To model the total number of infections, data information from reported cases, hospital

Communicated by Rafael Villanueva.

✉ Ivair R. Silva
ivairest@gmail.com

1 Federal University of Ouro Preto, Ouro Preto, MG, Brasil

2 Connecticut College, New London, USA

3 Utah Valley University, Orem, USA

admissions, and reported deaths, infection fatality rates, and life tables are often used. Bohk-Ewald et al. (2020) developed a demographic scaling model to estimate the total number of COVID-19 infections, based on data from COVID-19-related deaths, infection fatality rates, and life tables. Reed et al. (2022) used the Bayesian regression framework and other time series analysis to produce global and location-specific estimates of daily and cumulative SARS-CoV-2 infections using data largely from Johns Hopkins University (Baltimore, MD, USA) and national databases for reported cases, hospital admissions, and reported deaths, as well as seroprevalence surveys identified through previous reviews, SeroTracker, and governmental organizations. While both approaches are important for decision-making, the latter, as by definition, is a function of the patients arriving at the hospitals, the demand for tests on drugstores, laboratories, and the number of people seeking health care assistance and medical consultation. That is, the modeling of reported data has a practical appeal. It reflects the number of people using the health care system as a whole since this is the principal way they become officially reported. As such, having an estimate of the overall population's demand for health care, due to an emerging pandemic, is useful to optimize the distribution of resources among the hospitals, clinics and related acute health care facilities.

The acquisition of equipment, hospital materials, and the recruitment of health professionals, all require a realistic prediction on the number of patients using such resources, especially because they become scarce and rationed during the most critical moments of a pandemic. For example, since March 2020, the U.S. Department of Health and Human Services (HHS) is collecting data from hospitals and states to monitor the stress on the health care system in terms of its capacity, and capabilities, as well as the number of patients hospitalized due to COVID-19 (HHS 2022). In addition, reported data have been modeled and predicted by the Institute for Health Metrics and Evaluation (IHME) to study the evolution of the pandemic and the vaccine coverage (IHME 2022).

The present paper introduces an easily amenable and interpretable regression model for the number of reported infected in emerging pandemics, which is a stochastic version of the conventional compartmental disease transmission approach. In face of the terrible impacts of the COVID-19 pandemic, this model is offered as another tool to combat future pandemics. The proposed model, among other features, is useful to understand the statistical relationship between the transmission rate of contagious diseases with social, behavioral, logistical, and economic variables of the affected region.

Compartmental disease transmission models are characterized by the subdivision of the population into compartments. The dynamics behind the transition of individuals from a compartment to another can be modeled as a deterministic pattern (e.g. differential equations) or according to a stochastic law. Kermack and Mckendrick (1927) established the foundations of using compartmental models to epidemiology, arguing that an epidemic may result from a particular relation between the population density, the infectivity, recovery, and death rates.

Since then, compartmental models have been widely used to study the evolution of several recent disease outbreaks, such as the SARS epidemic in 2002–2003 by Chowell et al. (2003), the H1N1 influenza in 2009–2010 by Prosper et al. (2011), and the Ebola outbreak in 2014 by Feng et al. (2016). Most recently, since the COVID-19 pandemic, many researchers studied compartmental models to understand the COVID-19 pandemic. Calafiore et al. (2020) developed a modified Susceptible-Infected-Recovered (SIR) model to understand the COVID-19 contagion in Italy. Wu et al. (2020) used Susceptible-Exposed-Infectious-Removed (SEIR) model to provide an estimate of the size of the epidemic in Wuhan and forecast the extent of the domestic and global public health risks of epidemics, while taking social and non-pharmaceutical prevention interventions into account. Hou et al. (2020) employed SEIR

compartmental model to describe the dynamics of the COVID-19 epidemic based on epidemiological characteristics of individuals, clinical progression of COVID-19, and quarantine intervention measures of the authority.

The number of individuals infected at a given time is an important variable in epidemic models. However, in real epidemic scenarios, only the number of individuals that have been detected "positive" would be counted. As pointed by Calafiore et al. (2020), literature often assumes that the observed cases are the actual ones, which is unrealistic and may lead to wrong epidemiological interpretations and conclusions. In the present paper, we model the number of officially reported infected individuals.

For a discrete calendar time, say day t, based on the four compartments from the population in the stochastic SEIR model, we define: $S(t)$, the number of susceptible to become officially reported infected at day t after the exposure period; $E(t)$, the number of exposed at day t before becoming officially reported infected; $I(t)$, the number of officially reported infected at day t; $R(t)$, the number of removed (deaths + recovered) individuals at day t from the officially reported infected.

Then, we further define: $r(t)$, the number of individuals removed on day t; $i(t)$, the number of officially newly reported infected individuals on day t; d, the average number of days of an individual in $E(t)$ before moving to $I(t)$; q, the average number of days of an individual in $I(t)$ before being removed to $R(t)$; $S(0)$, the number of individuals susceptible to become officially reported infected, before the first infected is reported.

The compartment of infected is calculated as following:

$$I(t) = I(t-1) - r(t) + i(t), \tag{1}$$

for $t \geq 2$. We emphasize that $i(t)$ here denotes the officially reported infected group instead of the actual, unknown, number of new infected individuals at day t.

In a SEIR compartmental model, on the first day, say $t = 1$, when the official report starts to record the data, some individuals will move from $S(0)$ to the exposure compartment $E(1)$. And they will stay inside the exposure compartment for d days. Therefore, only on day $t = d+1$ they will move and be reported as newly infectious individuals, $i(1+d)$. Although we cannot know the amount of individuals leaving $S(0)$ to E at $t = 1$, we assume that they will be reported as newly infectious individuals at time $t = d + 1$. And it can be denoted mathematically as $S(0) - S(1) = i(d + 1)$. Similarly for any given day after $t = d + 1$, we have $S(t - d - 1) - S(t - d) = i(t)$. Therefore, we would denote the relationship as the following:

For $t > d$:

$$S(t-d) = S(t-d-1) - i(t). \tag{2}$$

Assuming it takes d days on average for an individual in $E(t)$ moving to $I(t)$ and it takes q days on average for an individual in $I(t)$ moving to $R(t)$, thus, for $t \geq 1$, the time specific removed individuals is elicited in the follow way:

$$r(t+q) = i(t), \tag{3}$$

and $r(j) = 0$ for $j = 1, \ldots, q$.

Following the reasoning of Li et al. (2018), given $S(t-d)$, the number of newly infected individuals, $i(t)$ with $t > d$, follows a binomial distribution with parameters $S(t-d-1)$ and $\phi(t)$. That is,

$$i(t) \sim \text{binomial}\left(S(t-d-1), \phi(t)\right), \tag{4}$$

where,

$$\phi(t) = 1 - e^{-\psi(t)\frac{I(t-d)}{N}}. \tag{5}$$

Here, $\phi(t)$ denotes the probability that a susceptible individual gets infected on day t. According to the derivations by Li et al. (2018) the function $\psi(t)$ may be given by,

$$\psi(t) = \beta e^{-\lambda\frac{I(t-d)}{N}}, \tag{6}$$

where N is the population size, and the terms β and λ were interpreted as transmission rate and media impact, respectively. The idea was to capture the reaction of the population after being informed by the media about the epidemic evolution, which can lead the individuals of a region to take protective actions. Thus, Li et al. (2018) used β to represent the baseline transmission rate before the media effect. With media impact, the transmission rate is $\beta e^{-\lambda\frac{I(t-d)}{N}}$. Note that $\lambda(>0)$ can also be interpreted as a coefficient that captures the effect of the integrated auto-regressive covariate $\frac{I(t-d)}{N}$ on the response variable $i(t)$.

In this paper, we extend this notion by considering a multiple regression structure for $\psi(t)$. This new approach allows including observable covariates of interest to the model to precisely capture the transmission rate. The main novelty is the estimation procedure, which is based on a new three-steps algorithm that combines the maximum likelihood method with the minimization of the mean squared error through a tri-section algorithm. This way, the coefficients of the regression model and the parameters d, q and $S(0)$ are estimated simultaneously.

The remainder of this paper is organized as the following: Sect. 2 introduces the SEIR compartmental regression model with a newly constructed $\psi(t)$. The likelihood function and the estimators are derived for all coefficients and unknown parameters. Confidence intervals and hypothesis testings are further provided for statistical inference purposes. We include a thorough simulation study on the models and methodologies in Sect. 3. Prediction estimations are discussed in Sect. 4. The supplementary material addresses model identification and diagnosis. Section 5 analyzes the effect of social mobility on the COVID-19 transmission rate for the state of São Paulo in Brazil. Concluding remarks are briefly made in Sect. 6.

2 Regression model

Let $x_{1,t}, \ldots, x_{k,t}$ denote a sequence of covariates actually observed in the target population for day t. Thus, in the spirit of the generalized linear models, we connect the probability parameter in (5) with these measurements through the following linear functional:

$$\ln \psi(t) = \delta_0 + \delta_1 x_{1,t} + \cdots + \delta_k x_{k,t}. \tag{7}$$

This way, the baseline component for the non-constant transmission rate over time is $\beta = e^{\delta_0}$ when there are no covariates in the model. For the estimation procedure, one needs to consider the auxiliary constant covariate $x_{0,t} = 1$ for each t. For $j > 0$, $x_{j,t}$ can be any potential construct associated with the infection rates namely, (a) a measurable covariate, such as discrete and continuous time-dependent variables, (b) an indicator or deterministic functional, like dummy variables, seasonal or some sort periodic component, and (c) an auto-regressive component for capturing the auto-correlation and moving average structures.

Therefore, the time non-constant transmission rate, among the individuals to appear in the reported statistics, is given by:

$$\psi(t) = e^{\delta_0+\delta_1 x_{1,t}+\cdots+\delta_k x_{k,t}} \equiv e^{\delta_0} e^{\delta_1 x_{1,t}+\cdots+\delta_k x_{k,t}}. \tag{8}$$

Thus, here we model $i(t) \sim$ binomial $(S(t-d-1), p(t))$, where:

$$p(t) = 1 - e^{-\frac{I(t-d)}{N} e^{\delta_0+\delta_1 x_{1,t}+\cdots+\delta_k x_{k,t}}}. \tag{9}$$

One may note that $i(t), t > d$, now follows a binomial distribution with parameters being $S(t-d-1)$ and $p(t)$, where $p(t)$ is a revised probability from $q(t)$ in (5) to include the more general real-valued function $\lambda(X,t) = \delta_1 x_{1,t} + \cdots + \delta_k x_{k,t}$ instead of the univariate term $\lambda \frac{I(t-d)}{N}$, and to consider the reparametrization $\beta = e^{\delta_0}$. It merits to reinforce that we still can consider auto-regressive components when selecting $x_{j,t}$ covariates to include in the model, such as $\frac{I(t-d)}{N}$ for example. With this extended approach, one can insert multiple independent variables representing different auto-regressive lags in time along with, if desirable, other types of predictors, periodic indicators, dummy variables, etc.

An important property of the present regression construction is the interpretability of the coefficients. Here, $e^{\eta\delta_j}$ is the relative change in the transmission rate, denoted by $\Delta\psi(j,\eta)$, when only the jth covariate is changed by η units, while the other covariates are held fixed. To show this, consider to change the state x_l to $x_l + \eta$ for some $1 \le l \le k$. From (8), we have:

$$\begin{aligned}\Delta\psi(l,\eta) &= \frac{e^{\delta_0+\delta_1 x_{1,t}+\cdots+\delta_l(x_{l,t}+\eta)+\delta_k x_{k,t}}}{e^{\delta_0+\delta_1 x_{1,t}+\cdots+\delta_l x_{l,t}+\cdots+\delta_k x_{k,t}}} \\ &= e^{\delta_l \eta}.\end{aligned} \tag{10}$$

Naturally, if the support of the covariate is the real line, then η can be any point of the real set, but it can assume 0 or 1 values only for dummy covariates.

2.1 Likelihood

For the actually reported number of new infected individuals at days $t = d+1, \cdots, T$, denoted by $\tilde{i}(t)$, the likelihood function is given by:

$$\begin{aligned}L\left(\boldsymbol{\theta}|\tilde{i}(d+1), \cdots, \tilde{i}(T)\right) = \prod_{t=d+1}^{T} &\binom{S(t-d-1)}{\tilde{i}(t)} \\ &\times [p(t)]^{\tilde{i}(t)}[1-p(t)]^{S(t-d-1)-\tilde{i}(t)},\end{aligned} \tag{11}$$

where $\boldsymbol{\theta} = (S(0), \delta_0, \delta_1, \ldots, \delta_k, d, q)^T$ is the unknown vector of parameters to be estimated.

Let $l(\boldsymbol{\theta})$ denote the log-likelihood based on the observations $\tilde{i}(t)$. With this, we have:

$$\begin{aligned}l(\boldsymbol{\theta}) = \sum_{t=d+1}^{T} \Bigg(&\ln\binom{S(t-d-1)}{\tilde{i}(t)} + \tilde{i}(t)\ln p(t) \\ &+ \left[S(t-d-1) - \tilde{i}(t)\right]\ln(1-p(t))\Bigg).\end{aligned} \tag{12}$$

Although $\boldsymbol{\theta}$ is not explicit in the right-hand side of (12), it is present in the calculations of $S(t-d-1)$ in (2), and that of $p(t)$ in (9).

The point estimation is divided in three algorithms which are discussed in Sects. 2.2, 2.3, and 2.4 respectively. Section 2.2 addresses estimating the parameters δ_j, $j = 0, 1, \ldots, k$, through the Newton–Raphson method for fixed $S(0)$, d and q. Section 2.3 is to estimate d and q for fixed $S(0)$ by maximizing the likelihood with a direct evaluation of a two-dimensional integer grid of meaningful values of d and q. Section 2.4 is based on a tri-section algorithm for estimating $S(0)$ and the implied estimates for the other parameters.

2.2 Estimating δ_j for fixed $S(0)$, d and q

In this section, we use the Newton–Raphson procedure for maximizing the log-likelihood function $l(\boldsymbol{\theta})$ from (12). Note that $l(\boldsymbol{\theta})$ is twice differentiable in δ_j. Hence, a Taylor expansion of $l(\boldsymbol{\theta})$ around the n^{th} iterate $\boldsymbol{\theta}_n$ gives,

$$l(\boldsymbol{\theta}) \approx l(\boldsymbol{\theta}_n) + \nabla_n^T(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)^T \mathbf{H}_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \tag{13}$$

where, the gradient vector ∇_n and the Hessian matrix $\mathbf{H}_n$, evaluated at $\boldsymbol{\theta}_n$ are, respectively, given by,

$$\nabla_n = \left(\frac{\partial}{\partial\delta_0} l(\boldsymbol{\theta}_n), \frac{\partial}{\partial\delta_1} l(\boldsymbol{\theta}_n) \ldots \frac{\partial}{\partial\delta_k} l(\boldsymbol{\theta}_n)\right) \text{ and } \mathbf{H}_n = \left(\left(\frac{\partial^2}{\partial\delta_u \partial\delta_j} l(\boldsymbol{\theta}_n)\right)\right)_{(k+1)\times(k+1)}$$

with,

$$\frac{\partial}{\partial\delta_j} l(\boldsymbol{\theta}) = \sum_{t=d+1}^{T} \left(\tilde{i}(t) \frac{x_{j,t}\psi(t)\frac{I(t-d)}{N}[1-p(t)]}{p(t)} - \left[S(t-d-1) - \tilde{i}(t)\right] x_{j,t}\psi(t)\frac{I(t-d)}{N}\right).$$

and,

$$\frac{\partial^2}{\partial\delta_j\partial\delta_u} l(\boldsymbol{\theta}) = \sum_{t=d+1}^{T} \left(\tilde{i}(t) \frac{x_{j,t}x_{u,t}\psi(t)\frac{I(t-d)}{N}[1-p(t)]\left(p(t) - \psi(t)\frac{I(t-d)}{N}\right)}{(p(t))^2} - \left[S(t-d-1) - \tilde{i}(t)\right] x_{j,t}x_{u,t}\psi(t)\frac{I(t-d)}{N}\right).$$

Here, we define $x_{0,t} = 1$ for each t and j, $u = 0, 1, \ldots, k$. Then, for fixed $S(0)$, d and q, possible candidates for the maximum likelihood estimators of δ_j are obtained by equating the gradient of right hand side of (13) to $\mathbf{0}$ and solving for the next iterate

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{H}_n^{-1}\nabla_n \tag{14}$$

As the extreme points of binomial likelihoods are not located in the boundaries of the parameter space, the roots of (14) are the maximum likelihood estimators of δ_j if the Hessian matrix is a negative definite matrix. This conditions must be checked for each application.

Hereinafter, $\hat{\delta_j}$ denote the maximum likelihood estimator of δ_j for fixed $S(0)$, d and q. This notion is important for the next section.

2.3 Estimating d and q for fixed $S(0)$

In practice, the range of candidates for d and q will be narrow for most of the pandemic dynamics. For example, it is sufficient to evaluate ordered pairs (d_i, q_i), where $d_i = 1, \ldots, 20$ and $q_i = 2, \ldots, 20$. In general, such a two-dimensional grid is supposed to follow the

characteristics of each specific pandemic phenomenon. With that in hand, one can simply apply steps described in Sect. 2.2 to each of the ordered pairs and take the pair of maximum likelihood in the selected grid as the estimate for d and q given $S(0)$.

2.4 Estimating $S(0)$

In order to estimate $S(0)$, we propose to obtain an accurate estimator through minimization of the mean squared error for the fitted transmission rate over the feasible candidates for $S(0)$. For a fixed value $S(0) = s$, the mean squared error, as a function of s, denoted by $MSE(s)$, is given by:

$$MSE(s) = \sum_{t=1}^{T} \left(\hat{\psi}(t) - \tilde{\psi}(t)\right)^2 / T, \tag{15}$$

where $\hat{\psi}(t)$ is the fitted value for $\psi(t)$ based on the maximum likelihood estimates of δ_j, $j = 1, \ldots, k$, d, and q, and $\tilde{\psi}(t)$ is the observed rate. The observed transmission rate is calculated using the observed proportion of new infected in day t, given by $\frac{\tilde{i}(t)}{S(t-d-1)}$, where $\tilde{i}(1) = \tilde{i}(2) = \cdots = \tilde{i}(d) = 0$. Then, from (5), we can derive:

$$\tilde{\psi}(t) = -\frac{N}{I(t-d)} \ln\left(1 - \frac{\tilde{i}(t)}{S(t-d-1)}\right). \tag{16}$$

For finding the estimator of $S(0)$, say $\hat{S}(0)$, one may want to use a naive approach based on running $MSE(s)$ over a small grid of meaningful candidates for $S(0)$, such as a, $a + fN, a + 2fN, \ldots, N$, where $a = \sum_{t=1}^{T} \tilde{i}(t)$, and $f \in (0, 1)$ is and arbitrary fraction of N. The disadvantage of such naive method is that it lacks in precision depending on the choice of f, that is, the difference between the solution and the actual argument of minimum of $MSE(s)$ can reach the magnitude of $f \times N$. On the other hand, this approach is general as it does not require assumptions on the behavior of $MSE(s)$.

Under the assumption that $MSE(s)$ is convex, or monotonous (increasing or decreasing), in the set of feasible candidates for $S(0)$, a more precise estimator can be obtained through a trisection routine. For this, we define the starting tuning parameter $S_4 = \sum_{t=1}^{T} \tilde{i}(t) + 1$, and the starting objective function $y = MSE(S_4)$. This will be updated according to the following trisection routine:

- Step (i) While $MSE(S_4) \leq y$ and $S_4 + I(T) \leq N$, update $S_4 := S_4 + I(T)$ and calculate the new $y = MSE(S_4)$.
- Step (ii) If (i) was finished, because $S_4 + I(T) > N$, then set $S_1 := S_4$ and update $S_4 := N$, otherwise, make $S_1 = S_4 - I(T)$. Then, calculate $S_2 = S_1 + \lceil (S_4 - S_1)/3 \rceil$, and $S_3 = S_2 + \lfloor (S_4 - S_1)/3 \rfloor$.
- Step (iii) For $i = 1,\ 2,\ 3,\ 4$, if $\min_i MSE(S_i) = MSE(S_1)$, then update S_4 as $S_4 := S_2$. If $\in_i MSE(S_i) = MSE(S_2)$, then update $S_4 := S_3$. If $\min_i MSE(S_i) = MSE(S_3)$, then update S_1 as $S_1 := S_2$. But, if $\min_i MSE(S_i) = MSE(S_4)$, then update $S_1 := S_3$.
- Step (iv) Update $S_2 := S_1 + \lceil (S_4 - S_1)/3 \rceil$, and $S_3 := S_2 + \lfloor (S_4 - S_1)/3 \rfloor$.
- Step (v) If $S_4 - S_1 \leq 6$, then stop the iterations and go to Step (vi) below. Otherwise, run Step (iii) again.
- Step (vi) Take $\hat{S}(0) = argmin_{S^* \in \{S_1, S_1+1, \ldots, S_4\}} MSE(S_i)$.

This algorithm is a version of the conventional bisection method (Burden and Faires 1985).

The correct interpretation of the estimated $S(0)$ is crucial. Unlike the actual number of people that can be affected by the disease in a population, here $S(0)$ represents the number of individuals susceptible to become reported infected. Note that this number may be much smaller than the overall number of infected persons. For example, the study by Fenga (2021) showed that while there were 12,839 officially reported COVID-19 cases on 12th March, 2020 in Italy, the number of infected could be as high as 105,789, about eight times the reported data. But, the present $S(0)$ parameter has important practical usage because it represents the potential number of people that will actually be need health care assistance, which is a necessary condition to have them among the officially reported cases.

2.5 Confidence intervals and testing

We note that the asymptotic variance-covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is estimated by computing the inverse of the Fisher Information matrix. Theoretically, the Fisher Information matrix can be obtained through the Hessian as,

$$I(\boldsymbol{\theta}) = -E[\mathbf{H_n}]$$

Hence, the inversion of the Wald's test promotes an approximate $100 \times (1-2\alpha)\%$ confidence interval of δ_j, as

$$\left[\hat{\delta}_j - I_{jj}^{-1/2}(\hat{\boldsymbol{\theta}})z_\alpha,\ \hat{\delta}_j + I_{jj}^{-1/2}(\hat{\boldsymbol{\theta}})z_\alpha\right], \tag{17}$$

where $I_{jj}^{-1/2}(\hat{\boldsymbol{\theta}})$ is the inverse of the squared root of the jth, $j = 0, \ldots, k$, diagonal element of the observed Fisher Information matrix, since the observed matrix is a consist estimate for the Fisher information matrix.

Note that, due to the duality between a confidence interval and the related hypothesis testing, and for a fixed value ϕ, an α-level test for

$$H_0 : \delta_j = \phi \text{ versus } H_1 : \delta_j \neq \phi, \tag{18}$$

is obtained by rejecting H_0 if the point ϕ is not covered by the confidence interval in (17).

3 Simulation study

In this section, we provide the results of a simulation study designed to evaluate the performance of the proposed inference methodology under meaningful tuning parameters choices. Each configuration counted with one thousand Monte Carlo databases generated under the proposed regression structure. For N, we selected the small and the big options $N = 5{,}000{,}000$ and $N = 30{,}000{,}000$. For each choice of N, we also considered $S(0) = N$ and $S(0) = 0.8N$. The choice of scenarios for $S(0)$ was a highly subjective step. This parameter may vary a lot depending on the capacity of the health system to serve the population of a country, on the demographic structure of the population that may impede or favor one to have contact with the virus, and on other factors such as culture and political organization. Thus, we used $S(0) = 0.8N$ as a possible value that may occur in some regions of the world, but that is just one example that differs from the extreme case where $S(0) = N$. Although arbitrary, these choices are useful to illustrate the performance of the proposed methodology. For N, we used the values $N = 3{,}000{,}000$ and $N = 30{,}000{,}000$ to show that the method works for very different population sizes. While one can apply the method to analyze the data of a metropolis

of a regular sized population, like the city of Salvador, BA, Brazil, it is also applicable for a huge population, such as that from the São Paulo state, Brazil.

Regarding d and q, we chose $d = 5, 7, 10$ and $q = 5, 7, 10$ based on a thorough literature review. Rădulescu et al. (2020) mentioned that the incubation period (the time from exposure to development of symptoms) of SARS-CoV-2 ranges typically from 2 to 14 days, with a mean of 5.2 days. Feng et al. (2021) chose the midpoint of 7 days because the incubation period of COVID-19 has been reported to be between 2 and 14 days. Udomsamuthirun et al. (2020) discussed average incubation period of 5 days and the average infection period of 10 days. Lai et al. (2021) incorporated in their work with the mean duration from exposed to infectious and from infection to recovery for 5.25 and 7 days, respectively. It is fair to say that our simulation study regarding three choices for d and q, 5, 7, 10 days, is general and representative.

Since we conducted this study in the presence of one covariate, we needed to run configurations for δ_0 and δ_1 only. In order to analyze the performance of the method using near realistic conditions, we used the data of social mobility in the São Paulo state, Brazil, for the covariate, observed from 02-26-2020 to 07-31-2020 ($T = 157$). More details on the social mobility data used here are given in Sect. 5. The usage of an actually observed covariate in the simulation study favored a practical perspective about the size of the effect of a change on δ_1 in terms of effective transmission rate when δ_0 is kept the same, and vice-verse. The effective transmission rate is calculated in the following way:

$$\beta_e(t) = \frac{q \times \psi(t) \times S(0)}{N}. \tag{19}$$

The effective transmission rate represents the expected number of individuals that can be contaminated by a person, during its infectious period, taking in account the fraction of susceptible individuals in the population, in contrast to the basic transmission rate, given by $R_0 = q\psi(t)$. To see the impact of changes in δ_0 and δ_1 on $\beta_e(t)$, we fixed N = 30,000,000, $d = 3$ and $q = 7$, from which we evaluated the evolution of $\beta_e(t)$ with t for three combinations of δ_0 and δ_1. We noted that the $\beta_e(t)$ is highly impacted, e.g. changing in average from the intermediate magnitude of 2 to the high value of 6, or even of 11, when δ_0 or δ_1 jumps from values around 0.01 or 0.1 to 1 as well as to -1. Therefore, the performance of the testing procedure could be considered satisfactory in case of presenting an elevated statistical power for δ_0 and δ_1 between -0.1 and 0.1. If it happens, then the power would be even higher for the relevant values (in terms of $\beta_e(t)$ changes) greater than 0.1 and smaller than -0.1. For this reason, considering the exhaustive number of scenarios to simulate given the many tuning parameters to consider, for parsimony we selected the options $\delta_0 = -0.1, 0.1$ and $\delta_1 = -0.1, 0, 0.1$. Note that the option $\delta_0 = 0$ would not make practical sense as it would imply a zero baseline rate when no covariates are present, that is, there would not exist an epidemic to analyze.

3.1 Accuracy, coverage probability and power

In this section, we discuss the results of our simulation study in terms of the point estimation accuracy, coverage probability of the confidence intervals, and statistical power of the hypotheses tests. The estimates for these three performance measures are summarized in three tables for $\delta_0 = -0.1$ and $S(0) = N$ as the results with the combinations $\delta_0 = 0.1$ and $S(0) = 0.8N$ are fairly similar. The choices of the parameters are also given in the table headings.

Table 1 Estimated averages $\hat{E}[\hat{\delta}_0]$, $\hat{E}[\hat{\delta}_1]$, $\hat{E}[\hat{d}]$, $\hat{E}[\hat{q}]$, and $\hat{E}[\hat{S}(0)]$ of the estimators for δ_0, δ_1, d, q, and $S(0)$, with $N = S(0) = 5 \times 10^6$ and $\delta_0 = -0.1$

δ_1	d	q	$\hat{E}[\hat{\delta}_0]$	$\hat{E}[\hat{\delta}_1]$	$\hat{E}[\hat{d}]$	$\hat{E}[\hat{q}]$	$\hat{E}[\hat{S}(0)]$
−0.1	5	5	−0.100	−0.100	5.00	5.00	5×10^6
		7	−0.100	−0.100	5.00	7.00	5×10^6
		10	−0.100	−0.100	5.00	10.00	5×10^6
	7	5	−0.100	−0.100	7.00	5.00	5×10^6
		7	−0.100	−0.100	7.00	7.00	5×10^6
		10	−0.100	−0.100	7.00	10.00	5×10^6
	10	5	−0.100	−0.100	10.00	5.01	5×10^6
		7	−0.097	−0.091	10.04	6.86	5×10^6
		10	−0.073	−0.069	10.19	8.67	5×10^6
0	5	5	−0.100	0.000	5.00	5.00	5×10^6
		7	−0.100	0.000	5.00	7.00	5×10^6
		10	−0.100	0.000	5.00	10.00	5×10^6
	7	5	−0.094	0.000	6.57	4.69	5×10^6
		7	−0.100	0.000	7.00	7.00	5×10^6
		10	−0.100	0.000	10.00	9.99	5×10^6
	10	5	−0.100	0.005	10.02	4.99	5×10^6
		7	−0.085	0.023	10.14	6.48	5×10^6
		10	−0.061	0.027	10.24	8.36	5×10^6
0.1	5	5	−0.100	0.100	5.00	5.00	5×10^6
		7	−0.100	0.100	5.00	7.00	5×10^6
		10	−0.100	0.100	5.00	10.00	5×10^6
	7	5	−0.100	0.100	7.00	5.00	5×10^6
		7	−0.098	0.098	7.01	6.99	5×10^6
		10	−0.098	0.098	7.01	6.99	5×10^6
	10	5	−0.096	0.114	10.07	4.90	5×10^6
		7	−0.074	0.128	10.20	6.20	5×10^6
		10	−0.057	0.123	10.24	8.32	5×10^6

Table 1 provides the average values from the simulated estimators of δ_0, δ_1, d, q, and $S(0)$ with N = 5,000,000 and $\delta_0 = -0.1$. With an exception of a few results under $d = 10$, we see that the MLE is practically unbiased.

We summarize the results on coverage probability and power for δ_0 and δ_1 in Tables 2 and 3, respectively. The results are satisfactory in most cases. The coverage probability falls around the target significantly, except that the test size is higher than 0.05 when $d = 10$ and $q = 7$ and 10. The statistical power hangs tightly around 1 for all scenarios, as is shown in Table 3.

Table 2 Coverage probability estimates of the 95% confidence intervals for δ_0 and δ_1

d	q	$\delta_1 = -0.1$		$\delta_1 = 0$		$\delta_1 = 0.1$	
		Coverage prob.		Coverage prob.		Coverage prob.	
		for δ_0	for δ_1	for δ_0	for δ_1	for δ_0	for δ_1
5	5	0.949	0.950	0.957	0.955	0.946	0.947
	7	0.957	0.963	0.956	0.956	0.963	0.962
	10	0.949	0.955	0.957	0.960	0.921	0.928
7	5	0.950	0.954	0.965	0.963	0.947	0.948
	7	0.957	0.956	0.949	0.948	0.941	0.942
	10	0.910	0.902	0.854	0.858	0.804	0.847
10	5	0.945	0.941	0.937	0.939	0.930	0.929
	7	0.880	0.904	0.825	0.848	0.755	0.784
	10	0.724	0.760	0.678	0.719	0.713	0.747

Table 3 Statistical power estimates of the 0.05-level test for δ_0 and δ_1 with $N = 5{,}000{,}000$ and $\delta_0 = -0.1$

d	q	$\delta_1 = -0.1$		$\delta_1 = 0$		$\delta_1 = 0.1$	
		Power of test		Power of test		Power of test	
		for δ_0	for δ_1	for δ_0	for δ_1	for δ_0	for δ_1
5	5	1.000	1.000	1.000	0.045	1.000	1.000
	7	1.000	1.000	1.000	0.044	1.000	1.000
	10	1.000	1.000	1.000	0.040	1.000	1.000
7	5	1.000	1.000	1.000	0.037	1.000	1.000
	7	1.000	1.000	1.000	0.052	1.000	1.000
	10	1.000	1.000	1.000	0.142	1.000	0.945
10	5	1.000	0.998	1.000	0.061	0.999	1.000
	7	0.972	0.998	0.926	0.152	0.860	0.998
	10	0.992	0.957	0.994	0.281	0.997	0.997

4 Predictions

Once the unknown vector of parameters, $\boldsymbol{\theta} = (S(0), \delta_0, \delta_1, \ldots, \delta_k, d, q)^T$, is estimated, one may continue to conduct prediction estimations on the transmission rate, say h days forward. This can be done by means of selected scenarios for the predictors. In situations where auto-regressive predictors are present in the regression model, it will also be necessary to obtain the predictions on the number of officially newly reported infected individuals. For an arbitrary $h \in \{0, 1, \ldots\}$, let $\hat{i}(t+h)$ denote the predicted number of new reported infected at time $(t+h)$. For $h = 1$, we define:

$$\hat{i}(t+1) = \hat{p}(t+1)S(t-d), \tag{20}$$

where:

$$\hat{p}(t+1) = 1 - e^{-\frac{I(t-d+1)}{N}\hat{\psi}(t+1)}, \tag{21}$$

and $\hat{\psi}(t+1)$ is the estimate of $\psi(t+1)$ given the values of the covariates at time $(t+1)$, and calculated with the estimator of θ with the information up to time T.

The prediction $\hat{i}(t+1)$ is then used to calculate the time specific fitted removed $\hat{r}(t+d+q+1)$ following (3), and the fitted susceptible compartment, $\hat{S}(t-d+1)$, for time $t-d+1$:

$$\hat{S}(t-d+1) = S(t-d) - \hat{i}(t+1). \tag{22}$$

The prediction for time $(t+2)$ is calculated as follows:

$$\hat{i}(t+2) = \hat{p}(t+2)\hat{S}(t-d+1), \tag{23}$$

where:

$$\hat{p}(t+2) = 1 - e^{-\frac{I(t-d+2)}{N}\hat{\psi}(t+2)}, \tag{24}$$

This process is repeated for the consecutive horizons until reaching time $t+h$ of interest. It is important to note that, for $h > d$, $I(t-d+h)$ is replaced by $\hat{I}(t-d+h)$, where:

$$\hat{I}(t+1) = I(t) - r(t+1) + \hat{i}(t+1), \tag{25}$$

$$\hat{I}(t+l) = \hat{I}(t+l-1) - r^*(t+l) + \hat{i}(t+l), \text{ for} l \geq 2 \text{ integer}, \tag{26}$$

$$r^*(t+l) = r(t+l) \text{ for } l \leq d+q, \text{ and} r^*(t+l) = \hat{r}(t+l) \text{ otherwise.} \tag{27}$$

This way, the fitted ($h = 0$) and the predicted ($h > 0$) values are calculated in general with:

$$\hat{i}(t+h) = \hat{p}(t+h)\hat{S}(t-d+h-1). \tag{28}$$

Finally, the predicted transmission rate h days forward, $\hat{\psi}(t+h)$, can be calculated using the parameters estimates, reasonable scenarios on the values of predictors and, when present, the predicted values for the autogressive terms based on (28) iteratively.

Using a Gaussian approximation for the distribution of the maximum likelihood estimators, considering that $\hat{\psi}(t+h)$ is a linear combination of the parameters estimators, and for fixed values of the predictors at time $t+h$, an approximate $100 \times (1-2\alpha)\%$ predictive interval can be calculated as follows:

$$[\hat{\psi}(t+h) - \hat{\sigma} z_\alpha,\ \hat{\psi}(t+h) + \hat{\sigma} z_\alpha], \tag{29}$$

where $\hat{\sigma} = \sum_{t=1}^{T}(\hat{\psi}(t) - \bar{\psi})^2/T$, and $\bar{\psi} = \sum_{t=1}^{T}\hat{\psi}(t)/T$.

5 Analyzing the COVID-19 infection rate in the State of São Paulo, Brazil

We analyze the daily number of reported COVID-19 infections in the State of São Paulo observed from 02-26-2020, when the Government started to make it publicly available, to 07-31-2020 ($T = 157$). This period was selected in order to avoid confounding information on the infectious rate as the proposed regression model is not designed for reinfections, which occurs when the surviving removed individuals return to the susceptible compartment. According to Santos et al. (2021), the first reinfection in Brazil probably occurred in July, 2020. This data is freely available at https://www.seade.gov.br/coronavirus. Currently, the population of São Paulo is about 46.6 millions.

An important aspect to investigate is the impact of the social mobility on the infectious rate. The reduction of the social mobility has been treated by governments as an effective public policy to mitigate the risk of transmission in the current COVID-19 pandemic. Aiming to capture such an effect in the State of São Paulo, we use the isolation index provided by mobile telecommunication devices in the city of São Paulo as a proxy for the State. For this, we use

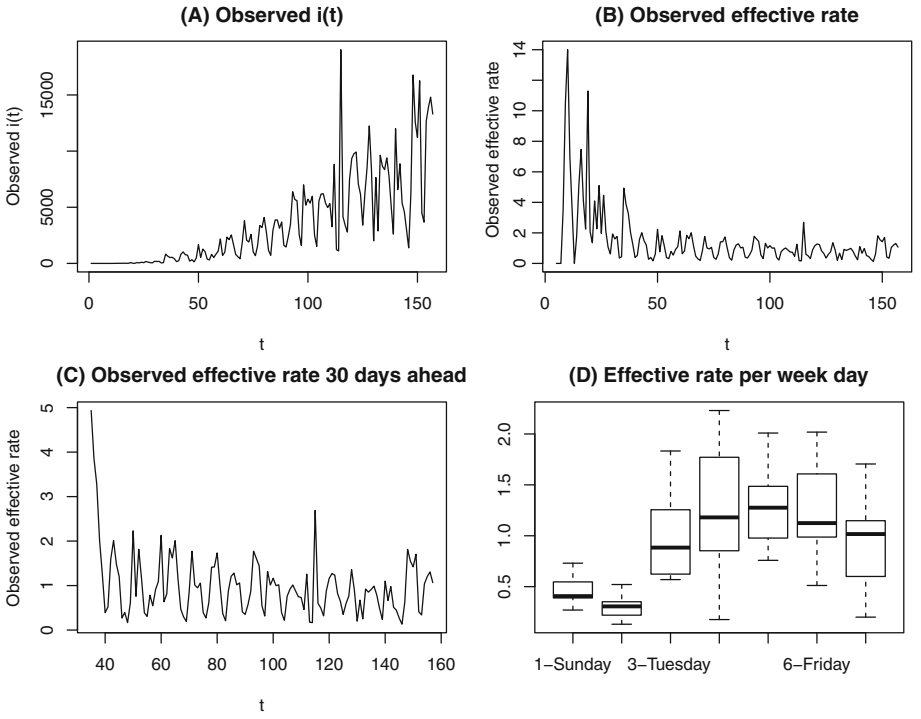

Fig. 1 Reported number of daily-specific COVID-19 cases in the State of São Paulo, Brazil, from 02-26-2020 to 07-31-2020 ($T = 157$) and effective transmission rate using $d = 3$ and $q = 7$ and $S(0) = N = 46{,}600{,}000$

the SIMI (Information System and Intelligent Monitoring), which is formed by aggregated anonymous information about mobility, health and other data of agencies and entities of the State Public Administration (Palhares et al. 2020). The index produces a number in the [0, 1] interval, which in practice represents the fraction of the population that is not moved from the home Cell Site. Regarding the evolution of the isolation index in the city of São Paulo, Brazil, from 02-26-2020 to 07-31-2020 ($T = 157$), we noted that it jumped from a baseline of about 0.3 to an oscillation about 0.5 after the first 30 days of the pandemic in São Paulo.

5.1 Analyzes results

Figure 1A shows the daily evolution of $\tilde{i}(t)$. The observed effective transmission rate is shown in Fig. 1B, which was calculated using $d = 3$ and $q = 7$, revealing a weekly seasonality. This becomes more evident when we remove the first 30 days of data, as shown in Fig. 1C. It merits to mention that the State of São Paulo started an official quarantine in May 24, 2020 (AL 2022).

Moreover, the permanent work for updating and informing the population on the pandemic status, provided by the international and the Brazilian media, apparently favored a change in the population's behavior weeks earlier to the first day of the official quarantine. This may explain the variation stability in the Fig. 1C in comparison to the first thirty days of reported data shown in Fig. 1B. For this reason, we used the period 03-27-2020 to 07-31-2020 ($T = 127$) for the regression analysis.

Now, we perform a descriptive exploration of the data to evaluate the existence of autocorrelation structures in $\tilde{\psi}(t)$. This demands a previous step for removing possible periodic structures that may act as confounding components. Actually, the existence of a deterministic weekly seasonal component in the number of reported infections is a well-known fact in Brazil, where the reported data for Saturday, Sunday and Monday are under-reported. This may occur, because the public agencies are closed in the weekends (www.seade.gov.br/covid-19). Only for this descriptive part, the week seasonality was removed using the ordinary mean squares regression, which was handle through dummy variables for Tuesday to Sunday. The autocorrelation function (ACF) of the seasonally natural log adjusted effective rate revealed a relevant moving average component of lag 1, and the partial autocorrelation function (PACF) indicated autocorrelation components of lags 1 and 7.

The impact of Governmental actions, such as the implementation of active social isolation, may only take effect a few days after the moment of its implementation. We have explored the delay that the social isolation index takes to impact the transmission rate for the lags $t-2,\ t-3,\ t-4,\ t-5,\ t-6$. It seemed that the effect of a variation in the logarithm of the social isolation in time t is perceived in the reported transmission rate after four or five days. This effect becomes more evident on the Sundays and Saturdays, which is an indication of interaction between the weekend factor with the isolation index.

Naturally, these descriptive insights using $d = 3$ and $q = 7$ will not coincide exactly with the predictors structure entering in the regression model estimate, but they are useful initial steps for the model identification. Based on these exploratory evidences, we tried a series of models with the following basic structure:

$$\ln \psi(t) = \delta_0 + \sum_{j=1}^{6} \delta_j D_j(t) + \delta_8\, lSI(t-l) + \delta_9\, lSI(t-4)\ D_6(t) + \\ + \delta_{10}\, lSI(t-4)\ D_1(t) + \delta_{11} \ln \tilde{\psi}(t-1) + \delta_{12} \ln \tilde{\psi}(t-7), \qquad (30)$$

where, $D_j(t) = 1$ if t corresponds to the week day represented by j, $j = 1, \ldots, 6$; $D_j(t) = -1$ if t corresponds to Monday; and $D_j(t) = 0$ otherwise. The $lSI(t)$ is the natural logarithm of the social isolation index, and $lSI(t-l)$ denotes the lag of $(t-l)$ with $l = 3, 4, 5, 6, 7$ denoting different models respectively. More details on this model and its formulation is discussed in the following sequel. Here, we use $j = 1$ for Tuesday, $j = 2$ for Wednesday, ..., $j = 6$ for Sunday. The coefficient of $D_j(t)$ can be interpreted as the average change promoted by the jth week day. This interpretation is possible by construction because the format of $D_j(t)$ ensures $\sum_{j=1}^{7} \delta_j = 0$, thus the average change of Monday is given by $\delta_7 = -\sum_{j=1}^{6} \delta_j$.

We emphasize that it is not indicated to use the social isolation index in its original $(0,\ 1)$ scale. This is so because the estimation of its coefficient would not take in account that the range of variation is bounded by 0 and 1, which can lead to misleading interpretations. This problem is solved by using the covariate $lSI(t)$, which is the natural logarithm of the social isolation index. The support of $lSI(t)$ is the real line, a convenient scale choice for the interpretation of its coefficient in the model.

The interaction terms $lSI(t-4)\ D_6(t)$ and $lSI(t-4)\ D_1(t)$ are meant to capture the exceptions for the association between $\ln \psi(t)$ and $lSI(t-4)$ as identified during our descriptive analysis. The auto-correlation terms $\ln \tilde{\psi}(t-1)$ and $\ln \tilde{\psi}(t-7)$ follow from the stylized ARMA process indicated by the ACF and PACF.

After running the model for $l = 3,\ 4,\ 5,\ 6,\ 7$ in the lagged term $lSI(t-l)$, the observed log-likelihoods were -43383.1 $(l = 3)$, -42857.68 $(l = 4)$, -43193.12 $(l = 5)$, -43331.34

Table 4 Point estimates, 95% confidence intervals, and standard errors for the coefficients of the model in (30)

Coefficients	Point estimate	Confidence interval		Standard error
		Lower bound	Upper bound	
Intercept	−0.4944	−0.5367	−0.4521	0.0216
Dummy Tuesday	0.2788	0.2705	0.2872	0.0043
Dummy Wednesday	−0.5396	−0.6241	−0.4552	0.0431
Dummy Thursday	0.6696	0.5825	0.7567	0.0444
Dummy Friday	0.9280	0.8434	1.0126	0.0432
Dummy Saturday	0.0032	−0.0042	0.0106	0.0038
Dummy Sunday	−0.4342	−0.4427	−0.4256	0.0043
$lSI(t-4)$	−1.1542	−1.2108	−1.0976	0.0289
Dummy Saturday × $lSI(t-4)$	−1.1235	−1.2330	−1.0140	0.0559
Dummy Sunday × $lSI(t-4)$	0.3622	0.2493	0.4751	0.0576
$\ln\tilde{\psi}(t-1)$	0.0425	0.0367	0.0483	0.0030
$\ln\tilde{\psi}(t-7)$	−0.1133	−0.1186	−0.1080	0.0027

($l = 6$), and −43744.85 ($l = 7$). Following the principle of parsimony, we adopted the model with the single $lSI(t-4)$ as it was the one with the highest likelihood. The residuals of the fitted model presented a regular behavior around a zero mean. Dispersion plots between $\ln\tilde{\psi}(t)$ and $\hat{e}(t-j)$, for $j = 1, \ldots, 7$, dispensed the inclusion of more auto-regressive terms in the model. Also, there is no clear association between the residuals with $lSI(t-4)$, hence, it is unnecessary to try other types of transformations in the social isolation covariate besides the logarithm function. In addition, the periodogram of the residuals has not presented a prominent value for a specific frequency, hence the seasonality is apparently well fitted with the model.

Table 4 contains the point estimates and the 95% confidence intervals for the coefficients of the model. The last column of this Table also contains the standard errors for the point estimators. Note that the only confidence interval covering the zero point is the coefficient for the dummy of Saturday, but we prefer to keep that in the model as its interaction with the $lSI(t-4)$ is significant under a 0.05-level, just like the other coefficients of the model.

It is important to emphasize the estimated relative change in the transmission rate followed by a change, say η, in the logarithm of the social isolation. From (10), and based on the coefficients estimates shown in Table 4, such a relative change is given by $\Delta\psi(8,\eta) = e^{(-1.1542-1.1235)\eta}$ for Saturdays, $\Delta\psi(8,\eta) = e^{(-1.1542+0.3622)\eta}$ for Sundays, and $\Delta\psi(8,\eta) = e^{-1.1542\eta}$ for Monday to Friday. For example, for Mondays to Fridays, and holding everything else constant, the state $lSI(t-4) = \ln(0.3)$ provokes a transmission rate that is 80.33% more than that with $lSI(t-4) = \ln(0.5)$ due to the fact that $e^{-1.1542(\ln(0.3)-\ln(0.5))} = 1.8033$. Conversely, the transmission rate with the isolation of 0.5 is about 1/1.8033=0.555 that one with an isolation of 0.3.

The estimates for d and q are $\hat{d} = 3$ and $\hat{q} = 5$, respectively. These estimates are consistent with the results of many studies on the behavior of the COVID-19 pandemic as discussed in the first paragraph of Sect. 3.

The estimated susceptible population is $\hat{S}(0)$ = 4,600,000. Again, we stress that $S(0)$ here represents the number of susceptible to become reported data. In Brazil, as well as in other countries, the number of reported infected individuals is much smaller than the total number

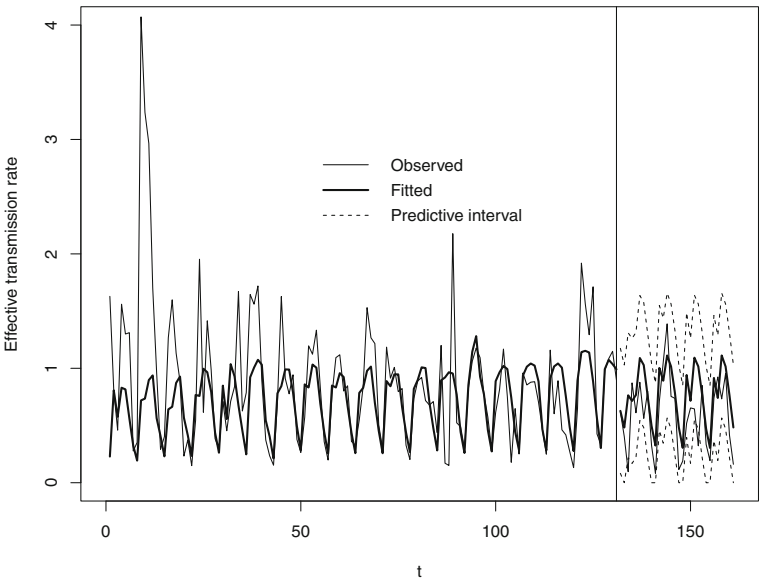

Fig. 2 Observed, fitted and predictive effective transmission rate

(reported + non-reported) of infected. Li et al. (2020) estimated the ratio of confirmed cases to actual cases to be only 14% during the early outbreak in China. Anand et al. (2020) estimated that only 9.2% of actual infections were laboratory-confirmed in the U.S. until July 2020. Jungsik and Gaudenz (2021) concluded that actual cumulative cases were estimated to be 5 to 20 times greater than the confirmed cases in 25 out of the 50 countries they studied. Hence, the number of susceptible to become reported will be much smaller than the total number of individuals susceptible to the disease as well. Although several confounding factors are present in the reported data during the years of 2020 and 2021, such as reinfection, new variants of the virus, and vaccination, it merits to mention that, by the moment at which these authors are writing the present paper, January of 2022, the cumulative number of reported infected individuals in the State of São Paulo is about 4,534,000 (https://github.com/CSSEGISandData/COVID-19), which is very close to the estimate $\hat{S}(0) = 4{,}600{,}000$.

Figure 2 shows the observed and the fitted effective transmission rates. The vertical line marks time T (July 31, 2020) in order to highlight the predictions from time T for the effective transmission rate on time $t+h$, with $h = 1, \ldots, 30$. The predictive 95% confidence intervals are illustrated with the dashed line. Note that the predicted values are tight to the observed effective transmission rates.

6 Concluding remarks

The stochastic SEIR regression proposed in this paper was designed to be very practical as it only requires the actual reported number of infected individuals, the population size of the target region, and the assumption that the dynamics of the transmissions follows a SEIR model. The estimation of the key parameters of the SEIR structure, namely $S(0)$, q, and d, are embedded in the estimation algorithm for the regression equation. The point and

interval inference approaches has been proved to be fairly accurate. Another strength of the proposed modeling is the practical interpretation of the regression coefficients, which are easily convertible to the relative change in the transmission rate when only one of the covariates is dislocated by arbitrary unities.

The proposed regression model is useful to identify effective actions to mitigate the transmissions when the disease is emerging and the SEIR model is a realistic assumption. As a simple rule of thumb, we recommend the proposed methodology for applications where the removed individuals can only be re-conducted to the susceptible compartment after a period of time greater than $T + h$, where h is the desired horizon of time for forecasting. Caution is also needed when a vaccination is administrated on the target population as this can greatly influence the transmission rate. In such scenarios, the information on the periods at which the vaccine was administrated, the rate of vaccine shots per day, and the percentage of the population's adherence to the vaccine campaign, could be considered as possible covariates to the model.

We have used the proposed model to analyze the reported COVID-19 infections in the state of São Paulo, Brazil, observed during the first four months of the pandemic in the state. The results indicate that a social isolation of 0.5 could potentially reduce the transmission rate on about a half of that under a isolation of 0.3. The identification of an autoregressive term of lag 7 is a novelty encountered with this analysis, which is useful to improve the performance of short term forecasts. Still, future investigations could focus on using covariates missed by the present study. Since the start, the pandemic situation keeps evolving. Besides emergence of new variants, the scenario keeps changing particularly with vaccination (partial or complete) and re-infection. Hence, caution is needed when interpreting these results based on the data of São Paulo when so the inferences are not confounded by the above mentioned factors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40314-023-02241-w.

Funding The first author has received support from Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, grant #301391/2019-0, and from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil, grants #PQ-00787-21 and #RED-00133-21.

Data Availability The data that support the findings of this study are available in Fundação Sistema Estadual de Análise de Dados (SEADE) at https://www.seade.gov.br/coronavirus/.

Declarations

Conflict of interest The authors declared no potential conflicts of interest and no competing interests with respect to the research, authorship, and/or publication of this article.

References

AL (2022) Repositório da assembleia legislativa do estado de São Paulo. https://al.sp.gov.br/repositorio/legislacao/decreto/2020/decreto-64881-22.03.2020. Accessed 16 January 2022

Anand S, Montez-Rath M, Han J, Bozeman J, Kerschmann R, Beyer P et al (2020) Prevalence of sars-cov-2 antibodies in a large nationwide sample of patients on dialysis in the USA: a cross-sectional study. Lancet 396:24–30

Bohk-Ewald C, Dudel C, Myrskylä M (2020) A demographic scaling model for estimating the total number of covid-19 infections. Int J Epidemiol 49:1963–1971

Burden RL, Faires JD (1985) Numerical analysis. PWS Publishers, Boston

Calafiore G, Novara C, Possieri C (2020) A modified sir model for the covid-19 contagion in Italy. 2020 59th IEEE Conference on Decision and Control (CDC), pp. 3889–3894
Chowell G, Fenimore P, Castillo-Garsow M, Castillo-Chavez C (2003) Sars outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. J Theoret Biol 224:1–8
Feng S, Feng Z, Ling C, Chang C, Feng Z (2021) Prediction of the COVID-19 epidemic trends based on SEIR and AI models. PLoS One. https://doi.org/10.1371/journal.pone.0245101
Feng Z, Zheng Y, Hernández-Cerón N, Zhao H, Glasser J, Hill A (2016) Mathematical models of Ebola-consequences of underlying assumptions. Math Biosci 277:89–107
Fenga L (2021) Covid-19: an automatic, semiparametric estimation method for the population infected in Italy. Bioinform Genom. https://doi.org/10.7717/peerj.10819
HHS (2022) Covid-19 guidance for hospital reporting and faqs for hospitals, hospital laboratory, and acute care facility data. https://www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf. Accessed 12 May 2022
Hou C, Chen J, Zhou Y, Hua L, Yuan J, He S, Guo Y, Zhang S, Jia Q, Zhao C, Zhang J, Xu G, Jia E (2020) The effectiveness of quarantine of Wuhan city against the corona virus disease 2019 (COVID-19): A well-mixed SEIR model analysis. J Med Virol 92:841–848
IHME (2022) The institute for health metrics and evaluation. https://covid19.healthdata.org/united-states-of-america?view=daily-deaths&tab=trend. Accessed 12 May 2022
Jungsik N, Gaudenz D (2021) Estimation of the fraction of covid-19 infected people in U.S. states and countries worldwide. PLoS One. https://doi.org/10.1371/journal.pone.0246772
Kermack WO, Mckendrick À (1927) A contribution to the mathematical theory of epidemics. Proc R Soc A Math Phys Eng Sci 115:700–721
Lai CC, Hsu CY, Jen HH, Yen AM, Chan CC, Chen HH (2021) The Bayesian susceptible-exposed-infected-recovered model for the outbreak of covid-19 on the diamond princess cruise ship. Stoch Env Res Risk Assess. https://doi.org/10.1007/s00477-020-01968-w
Li C, Pei Y, Zhu M, Deng Y (2018) Parameter estimation on a stochastic sir model with media coverage. Discrete Dyn Nat Soc. https://doi.org/10.1155/2018/3187807
Li R, Pei S, Chen B, Song Y, Zhang T, Yang W et al (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-COV-2). Science 368:489–493
Palhares G, Santos A, Ariente E, Gomes J (2020) A privacidade em tempos de pandemia e a escada de monitoramento e rastreio. Estud Av 34(99):175–190. https://doi.org/10.1590/s0103-4014.2020.3499.011
Prosper O, Saucedo O, Thompson D, Torres-Garcia G, Wang X, Castillo-Chavez C (2011) Modeling control strategies for concurrent epidemics of seasonal and pandemic h1n1 influenza. Math Biosci Eng 8(1):141–170
Rădulescu A, Williams C, Cavanagh K (2020) Management strategies in a SEIR-type model of covid 19 community spread. Sci Rep 10:21256
Reed RM et al (2022) Estimating global, regional, and national daily and cumulative infections with sars-cov-2 through nov 14, 2021: a statistical analysis. Lancet (London, England) 399:2351–2380
Santos L, Filhoa P, Silva A, Santos J, Santos D, Aquinoa M, Jesus R, Almeida M, Silva J, Altmann D, Boytond R, Santos C et al (2021) Recurrent covid-19 including evidence of reinfection and enhanced severity in thirty Brazilian healthcare workers. J Infect 82:399–406
Udomsamuthirun P, Chanilkul G, Tongkhonburi P, Meesubthong C (2020) The reproductive index from SEIR model of covid-19 epidemic in Asean. medRxiv. https://doi.org/10.1101/2020.04.24.20078287
Wu JT, Leung K, Leung G (2020) Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study. Lancet (London, England) 395:689–697