Linear Least Squares Parameter Inference for the SIR Epidemiology Model

Michaël A. VAN WYK [a], André M. MCDONALD [b,1] Mohlalakoma T. NGWAKO [a] Otis T. NYANDORO [a] and Fangfang ZHANG [c]

[a] *School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa*

[b] *Defence and Security Cluster, Council for Scientific and Industrial Research, Pretoria, South Africa*

[c] *Qilu University of Technology (Shandong Academy of Sciences), People's Republic of China*

Abstract. Accurate and early estimation of Susceptible-Infected-Recovered (SIR) epidemiology model parameters in infectious epidemics can enhance planning and resource allocation, thereby mitigating the adverse impacts on affected populations. Focusing on the basic SIR epidemiology model, in this paper we examine the scenario of known incidence rate (e.g., cases per day). Even though the SIR model is *nonlinear*, we obtain an exact least squares solution that is *linear* in simple algebraic functions of the SIR model's parameters, the infection rate, recovery rate and total population. Linear least squares solutions lend themselves to be applied to only a selected time period, to the censoring of unreliable measurements such as obvious outliers as well as to enable iterative update of the parameter estimates as new data (i.e., measurements) become available. We present numerical results for both simulated and real-world COVID-19 data to demonstrate the practical utility and accuracy of the proposed method. The proposed method demonstrates advantages over state-of-the-art approaches while also providing reliable parameter estimates.

Keywords. known incidence rate, novel least squares estimators, SIR model

1. Introduction

Severe diseases, such as COVID-19, chicken pox and Ebola, often require medical, social and public interventions to reduce the risk of mortality among infected individuals. A key challenge is that the healthcare sector operates with a limited allocation of resources. During outbreaks of severe infectious diseases, the demand for these resources increases significantly due to the surge in patient numbers. This elevated demand places considerable strain on the healthcare system.

The failure to manage available resources and to develop effective mitigation strategies can exacerbate the situation. Therefore, mathematical models are needed to un-

[1]Corresponding Author: André M. McDonald, Council for Scientific and Industrial Research, Meiring Naude Road, Pretoria, South Africa; E-mail: amcdonald@csir.co.za.

derstand infectious disease progression, in order to provide government and healthcare authorities with insights to implement policy, and also allocate resources when there are surges in infectious rate [1]. Improved preparedness can be achieved through epidemic studies that involve the accurate estimation of parameters in models such as the Susceptible-Infectious-Recovered (SIR) framework. For instance, [2] uses global epidemic and mobility models to study the impact of travel restrictions on the spread of the COVID-19 epidemic. The aforementioned research framework illustrates how scientific inquiry can inform and influence policy formulation.

The study in [3] presents three estimators to determine the number of susceptible individuals. These estimators are derived using the Peano-Baker series and the Cauchy repeated integral formula. The study in [4], introduces a least-squares estimator to determine the parameters of a discrete-time age-structured SIR model where the SIR model parameters are estimated separately for each age group. A transfer parameter is used to represent cross-compartment contact.

Similarly, [5] presents an SIR model with a varying total population, in which the number of susceptible individuals does not decline monotonically. This design accommodates the emergence of new epicenters at different points in time. The number of deaths is estimated using a trial-and-error nonlinear curve-fitting method, after which the number of recoveries is determined. Other SIR estimator contributions include asynchronous state estimators for nonlinear SIR models, [6], and direct estimation methods based solely on infected population dynamics [7]. The direct estimation approach uses a logarithmic method that relies on estimating the slopes of the increasing and decreasing infected cases, as well as the peak, in order to determine the parameters of the SIR model. This estimator serves as a benchmark for evaluating the performance of the estimator presented in this paper.

The present study applies least-squares parameter estimation within the SIR modelling framework. This approach is employed in a novel manner, as the estimators utilize linear methods on a nonlinear system. The proposed scheme uses the incidence rate as its input, thereby operating on new infections reported by epidemiologists via public platforms and which communicates the severity of an infectious disease to the public. What distinguishes the proposed method from those of the abovementioned authors is that the structure of the nonlinear SIR equations is exploited via Laplace transform analysis to yield a *linear* least-squares solution, which provides the *exact* solution for the theoretically noiseless case. The proposed method is applied to both real-world COVID-19 data from South Korea and artificial data to estimate disease parameters.
The remainder of this paper is organized as follows: Section 2 presents the SIR model. The proposed least-squares estimator is introduced in Section 3, and the results are presented and discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. SIR Model

A simple, deterministic SIR-model comprising three compartments, namely susceptible (S), infected (I), and recovered (R) individuals, is used. The mathematical representation of the model is given by

$$S'(t) = -\frac{\beta}{N} S(t) I(t), \tag{1}$$

$$I'(t) = \frac{\beta}{N} S(t) I(t) - \gamma I(t), \tag{2}$$

$$R'(t) = \gamma I(t), \tag{3}$$

where the five parameters are infection rate (β), recovery rate (γ), total population size (N), initial number of susceptible individuals ($S_0 := S(0)$), and initial number of infected individuals ($I_0 := I(0)$).

3. Novel Linear Least-Squares Estimator

In this section we present a least-squares parameter estimator for the case when the incidence rate of the disease is reported. Specifically, we assume that the incidence rate (i.e., the number of newly infected individuals per unit time), say $Q'(t)$, is available. The structure of the estimator selected is obtained from the SIR state space equations with an interchange of certain causes and effects.

From (1) and (2), we obtain

$$Q'(t) = -S'(t) = \frac{\beta}{N} S(t) I(t) \tag{4}$$

and

$$I'(t) = Q'(t) - \gamma I(t). \tag{5}$$

Taking the Laplace transform of the latter and manipulating the result yields

$$\mathscr{I}(s) = \frac{I(0)}{(s+\gamma)} + \frac{\mathscr{Q}'(s)}{(s+\gamma)}, \tag{6}$$

and $\mathscr{Q}'(s)$ is the Laplace transform of $Q'(t)$. Taking the inverse Laplace transform, gives

$$I(t) = I(0)\mathrm{e}^{-\gamma t} + Q'(t) * \mathrm{e}^{-\gamma t}, \tag{7}$$

where $*$ represents convolution. Since, from (1) we have

$$Q'(t) = -\frac{\beta}{N} \left(-S_0 + \int_0^t Q'(\tau) d\tau \right) I(t) \tag{8}$$

it follows that (7) can be written in the form

$$Y(t) := \frac{Q'(t)}{-S_0 + \int_0^t Q'(\tau) d\tau} = -\frac{\beta}{N} \left(I(0)\mathrm{e}^{-\gamma t} + Q'(t) * \mathrm{e}^{-\gamma t} \right). \tag{9}$$

Taking the Laplace transform of this expression, we obtain

$$\mathscr{Y}(s) = -\frac{\beta}{N}\left(\frac{I(0)}{(s+\gamma)} + \frac{\mathscr{Q}'(s)}{(s+\gamma)}\right) \tag{10}$$

which we can express as

$$\big(s\mathscr{Y}(s) - Y(0)\big) + Y(0) = s\mathscr{Y}(s) = -\frac{\beta}{N}\big(I(0) + \mathscr{Q}'(s)\big) - \gamma\mathscr{Y}(s), \tag{11}$$

where $\mathscr{Y}(s)$ the Laplace transform of $Y(t)$. Now, taking the inverse Laplace transform we finally obtain,

$$Y'(t) + Y(0)\delta(t) = -\frac{\beta}{N}\left(I(0)\delta(t) + Q'(t)\right) - \gamma Y(t), \quad t \geq 0\,. \tag{12}$$

For time strictly greater than zero, the Dirac delta functions no longer influence the response and so the last expression reduces to

$$Y'(t) = -\frac{\beta}{N}Q'(t) - \gamma Y(t), \quad t > 0. \tag{13}$$

Notice that this expression is linear in β/N and in γ. Evaluating the latter expression at strictly increasing time instants $\{t_i\}_{i=1}^{k}$ with $t_1 > 0$, we obtain the following simultaneous equations collected in matrix form,

$$\underbrace{\begin{pmatrix} Y'(t_1) \\ Y'(t_2) \\ \vdots \\ Y'(t_k) \end{pmatrix}}_{\mathbf{z}(1:k)} = \underbrace{\begin{pmatrix} -Q'(t_1) & -Y(t_1) \\ -Q'(t_2) & -Y(t_2) \\ \vdots & \vdots \\ -Q'(t_k) & -Y(t_k) \end{pmatrix}}_{\mathbf{Y}(1:k)} \underbrace{\begin{pmatrix} \alpha \\ \gamma \end{pmatrix}}_{\mathbf{b}(k)}, \qquad \alpha := \frac{\beta}{N} \tag{14}$$

and often simply write $\mathbf{z}$, $\mathbf{Y}$ and $\mathbf{b}$ to obtain $\mathbf{z} = \mathbf{Y}\mathbf{b}$. This expression is not a dynamical system since interchanging the individual equations, does not change the solution of the equation.

The least-squares solution to this matrix equation is obtained by application of the Moore-Penrose pseudoinverse,

$$\hat{\mathbf{b}}_{LS} = \mathbf{Y}^{\dagger}\mathbf{z}. \tag{15}$$

We obtain the SIR model parameters, $\hat{\alpha}|_{t_1}^{t_k}$ and $\hat{\gamma}|_{t_1}^{t_k}$ over the duration $[t_1, t_k]$. If there is little chance of confusion, we will simply write $\hat{\alpha}$ and $\hat{\gamma}$.

Clearly, this approach allows us to select the analysis interval to ensure that no structural change occurs during the interval, for example, due to new healthcare restrictions taking effect to curb the spreading of an infectious disease, as these would change the human-disease interaction, thus leading to inaccurate model parameter estimates.

The scheme presented here allows one to also censor the data by removing rows associated with obvious outliers, e.g., the lth row,

$$Y'(t_l) \approx -\frac{\beta}{N} Q'(t_l) - \gamma Y(t_l), \tag{16}$$

from the above matrix equation (14).

With the introduction of $Y(t)$ above, it now becomes apparent that yet another estimator for α can be obtained by simply averaging both $Y(t)$ and $I(t)$, namely,

$$\langle Y \rangle_t = \frac{1}{t} \int_{0^+}^{t} Y(\tau)\, d\tau = -\tilde{\alpha} \frac{1}{t} \int_{0^+}^{t} I(\tau)\, d\tau = -\tilde{\alpha} \langle I \rangle_t, \tag{17}$$

giving

$$\tilde{\alpha} = -\frac{\langle Y \rangle_t}{\langle I \rangle_t}. \tag{18}$$

Assuming an estimate of the population size N available, we obtain,

$$\hat{\beta} = \alpha N, \tag{19}$$

where we can use either $\hat{\alpha}$ or $\tilde{\alpha}$ in the place of α.

4. Results and Discussion

Next, a numerical experiment for characterizing the performance of the novel estimator is performed. This experiment uses the population-normalized incidence rate of daily new COVID-19 cases reported in South Korea between 17 January 2022 and 26 June 2022 [8]. Over this period, the Omicron variant (B.1.1.529) caused record-breaking case numbers in South Korea, including over 600,000 new cases in a single day [9]. The incidence rate for this 160 day period is plotted in Figure 1 (left, blue curve).

Applying the methodology of [10], the SIR model is fitted to the data by numerically searching over the space of candidate parameter values (β, γ, S_0) with the goal of minimizing the ℓ_2 error between the (cumulative) incidence of the data and of the model. This yields the SIR model parameters listed in Table 1. The population-normalized incidence rate (Q'/N) implied by the SIR model fitted to the data is plotted in Figure 1 (left, red curve). The figure shows an accurate fit of the SIR model to the COVID-19 data over both the pre-peak and post-peak time intervals, but with a lower peak value compared to the data. Figure 1 (right) shows an accurate fit between the cumulative incidence implied by the SIR model and that of the data.

The novel parameter estimator derived in section 3 is implemented for *discrete-time* incidence rate. To mitigate degradation of estimator accuracy resulting from the delay inherent in discrete-time numerical integration, as required to compute Y in (9), the daily incidence rate is oversampled by a factor M. Linear interpolation then yields intra-daily incidence rate values $Q[n] := Q'(t)|_{t=n/M}$, where $n = 0, 1, \ldots$ (here, t denotes the time, in days, since outbreak). Evaluation of (9) then yields $Y[n]$, and $Y'[n]$ is subsequently derived from $Y[n]$ using the forward difference approximation to the derivative.

The novel parameter estimator is evaluated by substituting the *daily* values $Y[nM]$, $Y'[nM]$ and $Q'[nM]$ for $n = 0, 1, \ldots$ into (14). This yields *daily* parameter estimates

Table 1. SIR model parameter values obtained from the fit to the South Korea COVID-19 data.

Parameter	Symbol	Value
Infection rate	β	0.4850
Recovery rate	γ	0.3979
Initial fraction of susceptible individuals	S_0/N	0.9998
Initial fraction of infected individuals	I_0/N	0.0002
Initial fraction of recovered individuals	R_0/N	0

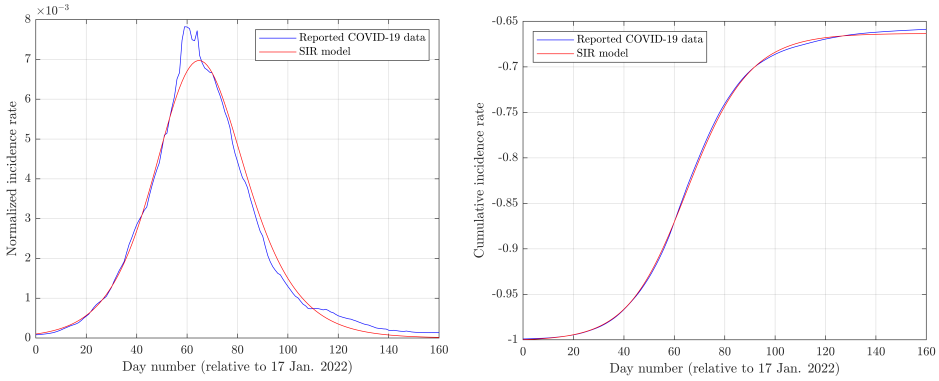

Figure 1. SIR model fitted to COVID-19 data reported from South Korea between 17 January 2022 and 26 June 2022: Normalized incidence rate (left) and cumulative incidence (right),

$\hat{\beta}[n] := \hat{\beta}|_0^n$ and $\hat{\gamma}[n] := \hat{\gamma}|_0^n$. This estimator is evaluated on the real-world incidence rate figures for South Korea with oversampling factor $M = 16$. Figure 2 plots the resulting infection rate and recovery rate parameter estimates, in grey, progressively as a function of time. The figure shows fluctuation in parameter estimates up to the time of peak infection, whereafter the estimator converges to its final value. These abrupt fluctuations are caused by the data matrix being ill-conditioned up to when the peak infection occurs. On day 160, the estimates $\hat{\beta}$ and $\hat{\gamma}$ show only 8.78% and 9.19% error from the true parameter values. The final parameter estimates have been found to be accurate and robust.

The state-of-the-art method, as presented in [7], is applied iteratively to the real-world data to yield daily estimates of the infection rate and recovery rate parameters. These estimates are plotted progressively as a function of time in Figure 2 (orange curves). The figure shows that these estimates do not converge to a final value at day 160. Experimentation shows that this is a result of sensitivity to the estimates of the pre-peak and post-peak slopes of the infection rate curve on the logarithmic scale, as required by the method. Only at a very late stage of the epidemic wave (here, at day 160), the method of [7] yields a relative error of 7.26% and 9.74%, which is comparable to our novel estimator in terms of accuracy, but not in terms of convergence rate.

To further characterize estimator performance, the SIR model is solved numerically for the incidence rate using the parameter values listed in Table 1. The incidence rate is contaminated with additive Gaussian noise to generate an ensemble of artificial incidence rate curves. In order to simulate curves that resemble real-world data more closely, the Gaussian noise process is filtered using an all-pole low-pass infinite impulse response (IIR) filter with a single pole $\rho = 0.75$. The time-dependent variance of the noise is selected to be $\sigma^2[n] = \kappa Q'[n]$ to model the phenomenon that real-world data tend to

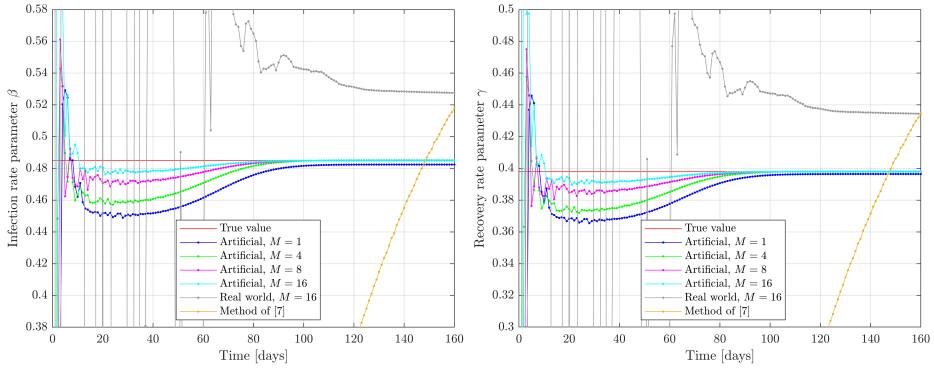

Figure 2. Convergence of the estimates $\hat{\beta}$ (left) and $\hat{\gamma}$ (right) as a function of time, for noise gain factor $\kappa = 0.005$.

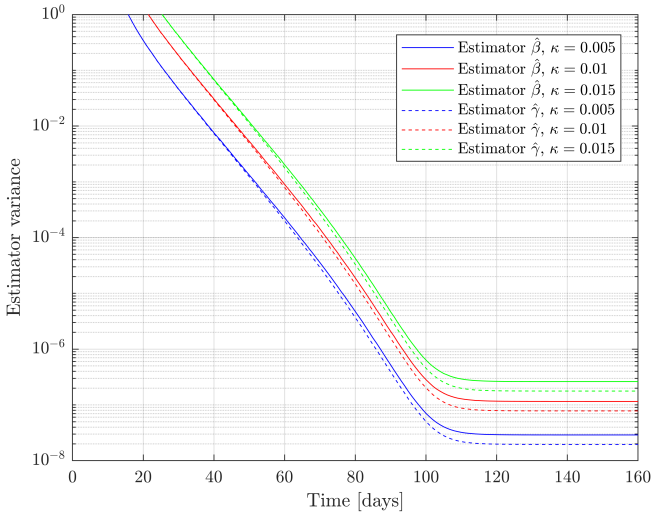

Figure 3. Variance of the estimates $\hat{\beta}$ and $\hat{\gamma}$ as a function of time, for distinct noise gain factors κ.

exhibit larger noise as the infection rate and infection numbers grow. Here, we refer to the scaling parameter $\kappa > 0$ as the *noise gain factor*.

The estimator is evaluated on the artificial data for noise gain factor $\kappa = 0.05$ and for distinct oversampling rates M. The mean value of each estimate, computed over the ensemble, is plotted progressively as a function of time in Figure 2. The figure shows that the mean value of both estimates converges rapidly for the artificial data. Oversampling both reduces the residual bias and improves the convergence rate.

The variance of the estimates, as evaluated over the artificial data with oversampling factor $M = 16$ and for distinct noise gain factor values κ, are plotted progressively as a function of time in Figure 3. Also evident is the exponential decay in variance for all noise factors considered as the estimates converge.

5. Conclusion

This paper presented a novel linear least-squares estimator for the infection and recovery rate parameters of the SIR epidemiology model, derived from reported incidence rate data. Numerical evaluation using COVID-19 case data from South Korea, for the Omicron variant during the first half of 2022, demonstrated robust estimator performance, achieving less than 10% error in parameter estimation. Although the accuracy of the novel estimator is comparable to that of a state-of-the-art method that requires infection rate measurements, which are unavailable in practice, the proposed approach offers several distinct advantages. In contrast to the state-of-the-art method, numerical experiments show that our novel estimator converges to a final value after the time of peak incidence and yields reliable parameter estimates. Furthermore, the novel estimator is directly applicable to reported incidence figures, has the flexibility to generate daily estimates as the outbreak evolves, and permits censoring. Collectively, these features position the proposed scheme as a practical and effective tool for real-time epidemic monitoring.

For future work we plan to address the data conditioning matter using well-established methods used in adaptive filter theory. Additionally, we will work with experts in healthcare towards enhanced predictive modelling to inform public health planning and support timely resource allocation during future infectious disease outbreaks.

Acknowledgment

We acknowledge the support of the Carl and Emily Fuchs Foundation.

References

[1] Acemoglu D, Chernozhukov V, Werning I, Whinston MD. Optimal Targeted Lockdowns in a Multigroup SIR Model. American Economic Review: Insights. 2021 December;3(4):487–502.

[2] Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395-400.

[3] Van Wyk MA, McDonald AM, Rubin DM, Zhang F. Novel Estimators for the Number of Susceptible Individuals in SIR Models of Infectious Epidemics. In: 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2024. p. 1-6.

[4] Cantó B, Coll C, Sánchez E. Estimation of parameters in a structured SIR model. Advances in Difference Equations. 2017;2017(1):33.

[5] Cooper I, Mondal A, Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons & Fractals. 2020;139:110057.

[6] Song X, Peng Z, Song S, Stojanovic V. Asynchronous state estimation for switched nonlinear reaction–diffusion SIR epidemic models with impulsive effects. Biomedical Signal Processing and Control. 2025;105:107600.

[7] Schmitt FG. An algorithm for the direct estimation of the parameters of the SIR epidemic model from the $I(t)$ dynamics. Eur Physical J Plus. 2022, Art no 57;137:1-16.

[8] Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, et al.. Coronavirus Pandemic (COVID-19). Our World in Data; 2020. Accessed: 2025-09-09. Available from: `https://ourworldindata.org/coronavirus/country/south-korea`.

[9] Seung-Yeon K. South Korea reports over 600,000 daily COVID-19 cases amid Omicron surge. Yonhap News Agency; 2022. Accessed: 2025-09-09. Available from: `https://en.yna.co.kr/view/AEN20220317002652320`.

[10] Batista M. Estimation of the final size of the coronavirus epidemic by the SIR model. ResearchGate; 2020. Accessed: 2025-09-09. Available from: `https://www.researchgate.net/publication/339311383_Estimation_of_the_final_size_of_the_coronavirus_epidemic_by_the_SIR_model`.