Enhancing Epidemiological Models with Parameter Estimation and Symbolic Regression: A Case Study on COVID-19

Shila Rezvani* Mostafa Abbaszadeh† Mehdi Dehghan‡

Abstract

This study explores three significant topics in contemporary research: parameter estimation, Symbolic Regression and the COVID-19 pandemic. Parameter estimation is a fundamental aspect of statistical analysis, focusing on deriving unknown parameter values from empirical data. This study employs methodologies such as Maximum Likelihood Estimation (MLE) and advanced modeling techniques to refine the system of equations for a better fit to observed data. In the context of the COVID-19 pandemic, parameter estimation plays a pivotal role in developing epidemiological models that inform public health strategies. The SIDARTHE model, for instance, offers an innovative approach by categorizing individuals based on their infection status and Severity of symptoms, providing crucial insights into the virus's transmission dynamics. In addition to traditional parameter estimation, this work leverages Symbolic Regression (SR) to update and refine the right-hand side of the system of equations based on the data. SR, a machine learning-based regression technique rooted in genetic programming, uncovers new equations. By integrating statistical methods, SR and epidemic modeling, this study highlights the importance of simultaneously updating coefficients and discovering new governing equations to enhance our understanding of disease spread and guide effective intervention measures.

Keywords: Parameter estimation, Covid-19, Symbolic Regression

1 Introduction

1.1 COVID-19:Overview and Modeling Approaches

Coronavirus disease 2019 (COVID-19) is a contagious illness caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Researchers first identified the initial case in Wuhan, China, in December 2019. The clinical manifestations of COVID-19 are diverse, commonly including symptoms such as fever, fatigue, cough, difficulty breathing and loss of taste and smell. These symptoms can appear anywhere from one to fourteen days following exposure to the virus, with at least one-third of infected individuals remaining asymptomatic [8].

Most individuals infected with SARS-CoV-2 experience mild to moderate respiratory illness and recover without needing specialized treatment. However, specific populations, particularly older adults and those with preexisting health conditions such as cardiovascular disease, diabetes, chronic respiratory disorders, or cancer, are at a higher risk of developing severe complications. Importantly, individuals of any age can experience severe illness or death from COVID-19 [2]. The virus primarily spreads through respiratory droplets emitted from the mouth or nose during activities like coughing, sneezing, speaking, or breathing. This highlights the need for public health measures to reduce virus transmission [1].

To effectively manage the COVID-19 epidemic, researchers have developed various models to predict its trajectory and inform control strategies. One such model is the SIDARTHE model, which categorizes individuals based on their infection status distinguishing between diagnosed and undiagnosed cases and assessing the severity of symptoms. This distinction is crucial, as diagnosed individuals are typically isolated, reducing their potential to spread the virus [4].

The SIDARTHE model builds on the classical SIR model (Susceptible, Infected, Recovered) by incorporating a more nuanced understanding of transmission dynamics. It includes eight distinct stages of infection:

S: Susceptible (uninfected)

I: Infected (asymptomatic or mild, undiagnosed)

D: Diagnosed (asymptomatic, detected)

A: Ailing (symptomatic, undiagnosed)

R: Recognized (symptomatic, diagnosed)

T: Threatened(infected with severe symptoms, detected)

H: Healed (recovered)

E: Extinct (deceased)

The model comprises the following ordinary differential equations, each representing the dynamics of these stages

*Department of Computer Science, Amirkabir University of Technology, rezvani8080shila@gmail.com

†Department of Computer Science, Amirkabir University of Technology, m.abbaszadeh@aut.ac.ir

‡Department of Computer Science, Amirkabir University of Technology, mdehghan@aut.ac.ir

over time:

$$
\begin{aligned}
&\dot{S}(t) = -S(t)(\alpha I(t) + \beta D(t) + \gamma A(t) + \delta R(t)),\\
&\dot{I}(t) = S(t)(\alpha I(t) + \beta D(t) + \gamma A(t) + \delta R(t)) - (\varepsilon + \zeta + \lambda)I(t),\\
&\dot{D}(t) = \varepsilon I(t) - (\eta + \rho)D(t),\\
&\dot{A}(t) = \zeta I(t) - (\theta + \mu + \kappa)A(t),\\
&\dot{R}(t) = \eta D(t) + \theta A(t) - (\nu + \xi)R(t),\\
&\dot{T}(t) = \mu A(t) + \nu R(t) - (\sigma + \tau)T(t),\\
&\dot{H}(t) = \lambda I(t) + \rho D(t) + \kappa A(t) + \xi R(t) + \sigma T(t),\\
&\dot{E}(t) = \tau T(t).
\end{aligned}
$$

In these equations, α, β, γ, and δ represent the transmission rates associated with contacts between susceptible individuals and those in various infection states. The parameters ϵ and θ denote the detection rates for asymptomatic and symptomatic cases, respectively. The probabilities that infected individuals, whether aware or not, develop clinically significant symptoms are captured by ζ and η. Furthermore, μ and ν indicate the rates at which undetected and detected individuals develop severe symptoms. The mortality rate for those with life-threatening symptoms is represented by τ, while λ, κ, ξ, ρ, and σ reflect recovery rates across the different infected classifications [4].

1.2 Estimation Theory in Epidemiological Modeling

Estimation theory is a fundamental area of statistics focused on inferring unknown parameter values from empirical data, which often includes random variations. These parameters describe the characteristics of an underlying system, and their values influence the distribution of the measured data. The purpose of estimation techniques is to derive accurate parameter estimates by analyzing this data. Two primary approaches are typically employed in estimation theory [6]:

1. The probabilistic approach, which assumes that the measured data is random and that its probability distribution depends on the parameters of interest [6]. This approach is preferred for its ability to model uncertainty in complex systems rigorously.

2. The set-membership approach, which posits that the measured data vector belongs to a predefined set determined by the parameter vector [6]. This method is beneficial in contexts where data is constrained within known bounds, offering a more deterministic framework for parameter estimation.

Both approaches possess distinct advantages depending on the application context; however, The probabilistic approach is frequently chosen because of its flexibility in managing varying degrees of uncertainty and the broader range of statistical tools available for its implementation. In this discussion, we utilize Maximum Likelihood Estimation (MLE), a widely employed method for parameter estimation within a probabilistic framework. MLE operates by determining the parameter values that maximize the likelihood function, which represents the probability of the observed data given specific parameter values [7]. The point that maximizes the likelihood function serves as the maximum likelihood estimate. MLE is a highly versatile and intuitive method, making it one of the most commonly used techniques in statistical inference.

MLE is particularly well-suited to large datasets and complex models, providing reliable parameter estimates as sample sizes increase. Its properties, such as consistency, which refers to convergence toward the valid parameter values as the sample size increases, and efficiency, which denotes achieving the lowest possible variance among estimators, make it a preferred choice in many scientific and engineering applications. Additionally, the ability of Maximum Likelihood Estimation (MLE) to simultaneously accommodate multiple parameters represents a significant advantage in real-world situations that involve numerous unknown factors.

In practical applications, parameter estimation is essential for identifying the physical coefficients in systems that require real-time monitoring and control. For instance, in sensor-based systems, each sensor has unique physical parameters that describe its behavior. By analyzing the statistical characteristics of the model parameters and the corresponding physical parameters, the system can detect potential faults or malfunctions [5].

This estimation method begins by analyzing the system's operational mechanism to derive the relationship between model parameters and the system's output. Real-time data is then gathered, and the system computes the output, comparing it with the predicted output based on the estimated model parameters. This comparison enables the detection of discrepancies, allowing for early identification of faults or deviations from regular operation. The ability to perform real-time parameter estimation makes this method highly applicable in dynamic systems that require continuous monitoring and adjustment.

In this study, we analyze actual data collected over 46 days to estimate the parameters of the SIDARTHE model. By applying MLE to this dataset, we aim to refine the model's accuracy in predicting COVID-19 dynamics, which can inform public health interventions and improve resource allocation. Integrating parameter estimation techniques into epidemiological modeling is essential for enhancing the robustness of predictions and the efficacy of intervention strategies. By accurately characterizing the transmission dynamics of COVID-19, we can better inform public health responses and ensure a more practical approach to managing the ongoing pandemic.

1.3 Symbolic regression

Symbolic regression (SR) is a machine learning-based regression technique rooted in genetic programming that draws upon methodologies from various scientific disciplines. SR has the unique ability to derive analytical equations directly from data, eliminating the necessity for prior knowledge of the system under investigation. This capability allows SR to uncover deep and complex relationships that are generalizable, interpretable, and applicable across diverse scientific, technological, economic, and social domains [3].
Unlike traditional regression methods (e.g., linear or quadratic regression), which require predefined independent variables and adjust numerical coefficients for optimal fitting, SR simultaneously identifies both the parameters and the governing equations [3].

2 Method

First, the data are organized and prepared to apply the parameter estimation technique. We begin by defining the function that represents the system of equations, along with determining the values of the initial parameters obtained through experiments and previous research. For each equation in the system of ordinary differential equations, we take an approximation for the right-hand side by taking its integral. In the next step, we use the maximum likelihood method and perform optimization considering the negative log-likelihood to minimize the function. This allows us to identify the optimal parameters and update the coefficients in the equations. The results show a clear improvement in accuracy and performance.

In Figure 1, the blue curve represents the real data collected over 46 days. The green curve is based on the system of equations derived from previous research and laboratory samples, using the initial parameter values. Without applying parameter estimation, it is evident that this model does not fit well with the real data. This suggests that the system of equations proposed in earlier studies is not an appropriate model for prediction our data. In contrast, the red curve shows the results after adjusting the parameters. The model now aligns much more closely with the real data, indicating that the modified equation provides a more accurate representation. The same procedure has been applied to the other figures as well. Notably, the system of eight equations was not evaluated individually. Instead, the equations were considered simultaneously, and the parameter estimates were optimized to ensure that all equations fit the data collectively. The initial values of the parameters without applying the parameter estimation technique are given in the captions of the Figures 1-8. After applying the parameter estimation technique, the optimized parameters are in the table below:

Parameter List	
Parameter	Value
α	-1.883329153060913
β	5.4156646728515625
γ	-0.6877768635749817
δ	-0.5272416472434998
ϵ	-0.516588568687439
ζ	3.5108792781829834
λ	1.0128684043884277
η	-3.528564929962158
ρ	0.11585894227027893
θ	0.42635196447372437
μ	-1.0851908922195435
κ	5.254979610443115
ν	-0.1200798898935318
ξ	-7.785756587982178
σ	-5.365184307098389
τ	1.3259201049804688

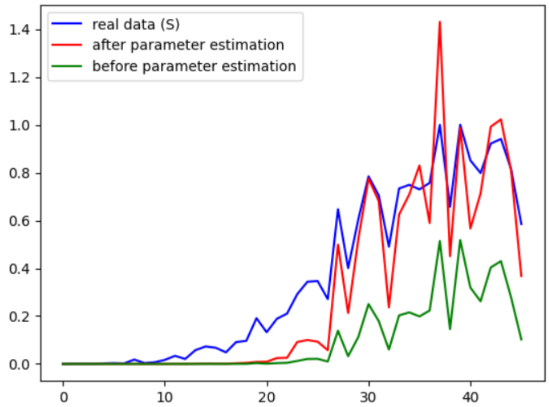

Figure 1: Initial values of the parameters are: α=0.57,β=0.0114,γ=0.456,δ=0.0114.

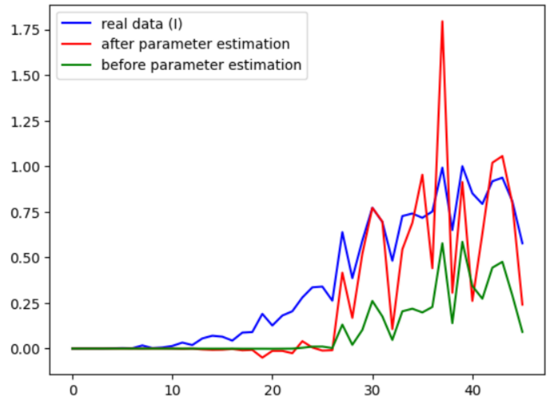

Figure 2: Initial values of the parameters are: α=0.57, β=0.0114, γ=0.456, δ=0.0114, ϵ=0.171, ζ=0.1254, λ=0.0342.

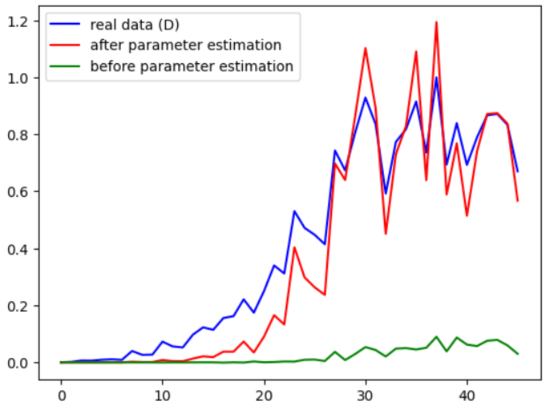

Figure 3: Initial values of the parameters are: ϵ=0.171, η=0.1254, ρ=0.0342.

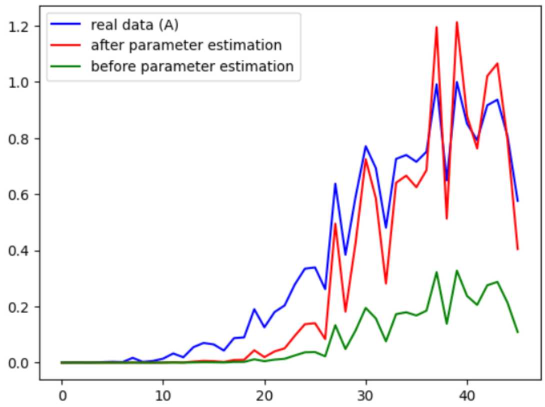

Figure 4: Initial values of the parameters are: ζ=0.1254, θ=0.3705, μ=0.0171, κ=0.0171.

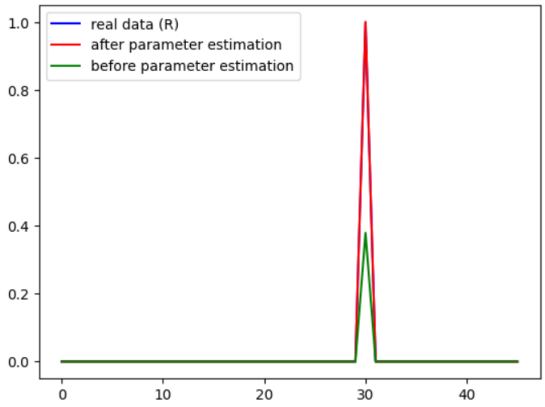

Figure 5: Initial values of the parameters are: η=0.1254, θ=0.3705, ν=0.0274, ξ=0.0171.

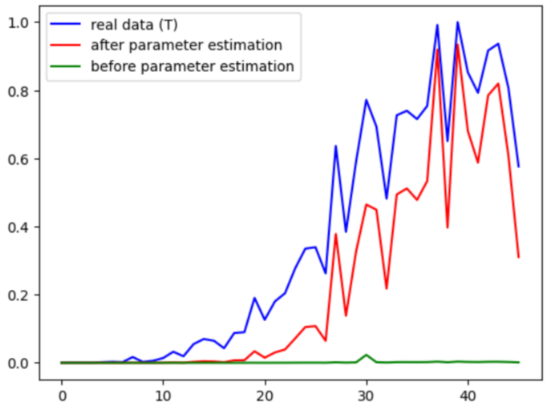

Figure 6: Initial values of the parameters are: μ=0.0171, ν=0.0274, σ=0.0171, τ=0.01.

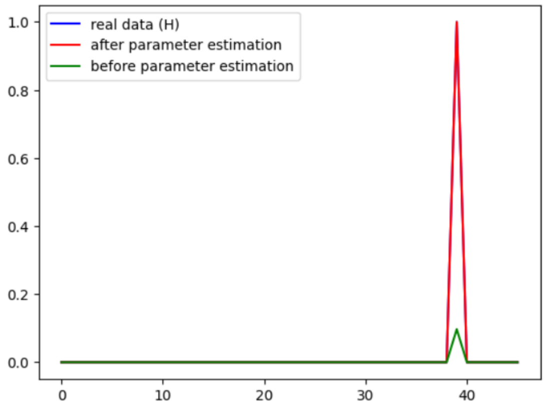

Figure 7: Initial values of the parameters are: λ=0.0342, ρ=0.0342, κ=0.0171, ξ=0.0171, σ=0.0171.

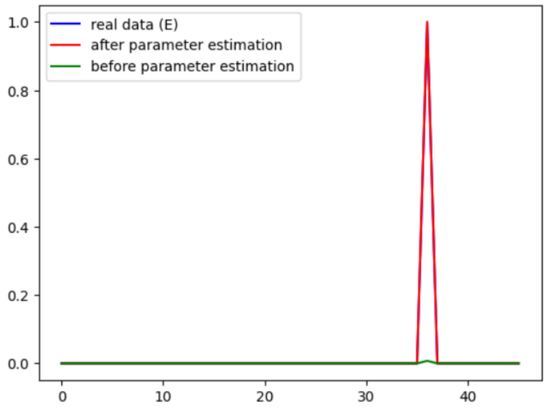

Figure 8: Initial values of the parameters are: τ=0.01.

After updating the coefficients with parameter estimates for the new system of equations, we apply a machine learning technique called symbolic regression to adjust the right-hand sides of the equations, align them with real data, and identify new relationships between variables. This process is done using real data and functions available in Python libraries (gplearn). Then the right side of the equations is replaced by the newly derived equations. In Figure 9, which corresponds to the first equation of the system related to the S, we have plotted the values on the right-hand side of the equation based on the real data collected over 46 days. The blue points represent these values so that the horizontal axis indicating time (in days) and the vertical axis showing the function values on the right-hand side of the equation. The orange points, meanwhile, represent the values obtained after applying symbolic regression to the proposed function, which approximates the right-hand side of the equation. The same procedure has been used for other Figures 10-16 as well.

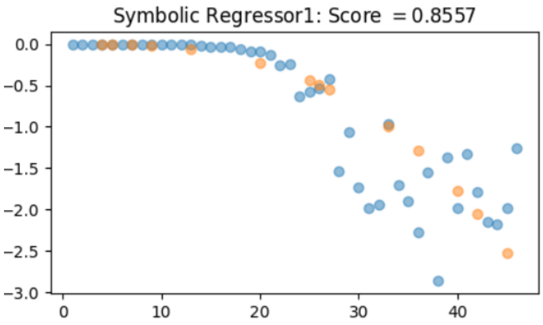

Figure 9: The blue points are the real data for S and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

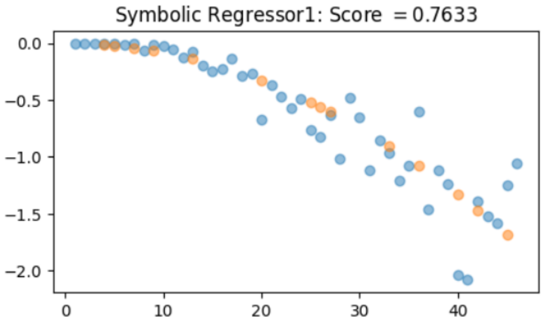

Figure 10: The blue points are the real data for I and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

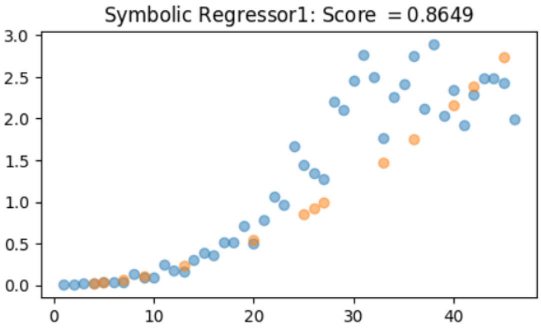

Figure 11: The blue points are the real data for D and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

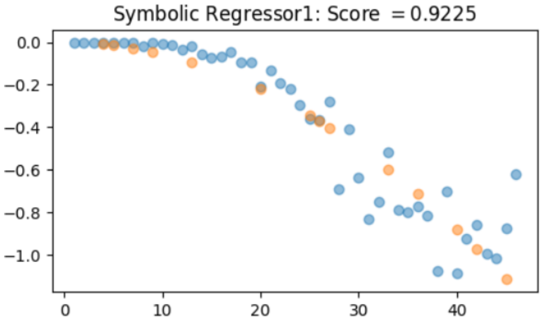

Figure 12: The blue points are the real data for A and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

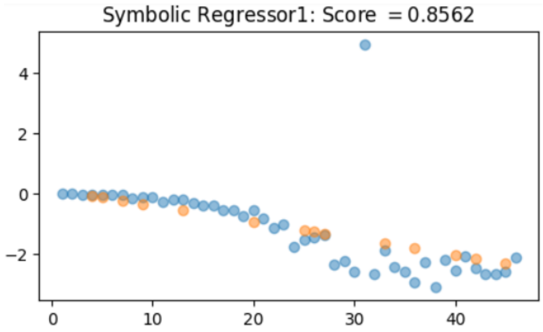

Figure 13: The blue points are the real data for R and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

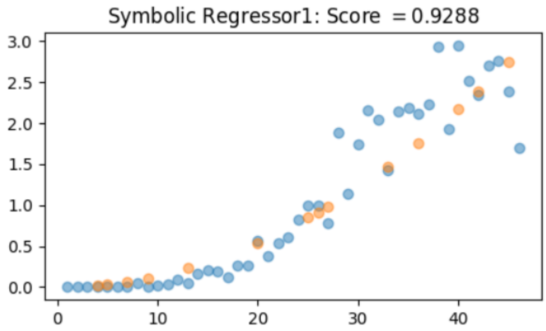

Figure 14: The blue points are the real data for T and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

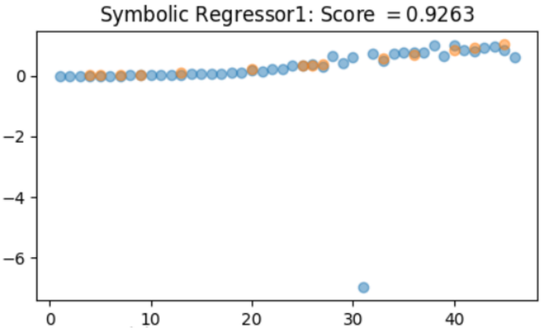

Figure 15: The blue points are the real data for H and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

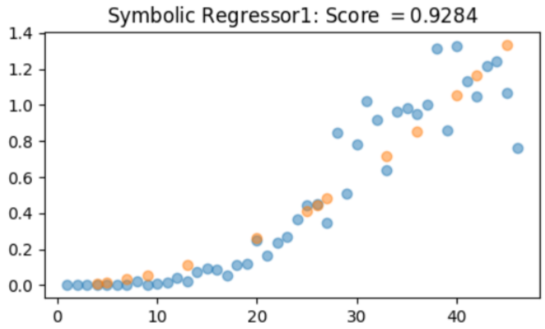

Figure 16: The blue points are the real data for E and the orange points are the estimates of the symbolic regression function calculated with the score mentioned above.

References

[1] Centers for disease control and prevention. how covid-19 spreads., `https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html` (2021).

[2] Centers for disease control and prevention. people at increased risk for severe illness., `https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-increased-risk.html` (2021).

[3] D. Angelis, F. Sofos, T. E. Karakasidis, Artificial intelligence in physical sciences: Symbolic regression trends and perspectives, Archives of Computational Methods in Engineering 30 (6) (2023) 3845–3865.

[4] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di Filippo, A. Di Matteo, M. Colaneri, Modelling the covid-19 epidemic and implementation of population-wide interventions in italy, Nature medicine 26 (6) (2020) 855–860.

[5] D. Li, Y. Wang, J. Wang, C. Wang, Y. Duan, Recent advances in sensor fault diagnosis: A review, Sensors and Actuators A: Physical 309 (2020) 111990.

[6] Wikipedia contributors, Estimation theory — Wikipedia, the free encyclopedia, `https://en.wikipedia.org/w/index.php?title=Estimation_theory&oldid=1224571595`, [Online; accessed 3-October-2024] (2024).

[7] Wikipedia contributors, Maximum likelihood estimation — Wikipedia, the free encyclopedia, `https://en.wikipedia.org/w/index.php?title=Maximum_likelihood_estimation&oldid=1243279128`, [Online; accessed 3-October-2024] (2024).

[8] World Health Organization, [who]. covid-19: Symptoms, `https://www.who.int/health-topics/coronavirus` (2020).