Modeling COVID-19 Outbreaks with the SIR Framework

Leena Kang
Department of Mathematics
University of California, San Diego

Abstract

This project explores the effectiveness of the classical Susceptible-Infected-Recovered (SIR) model in capturing and predicting the spread of the COVID-19 virus. While the SIR model is mathematically simpler and widely used in epidemiology, its underlying assumptions (such as and fixed parameters) raise questions about its applicability to real-world, variable disease dynamics. We begin by deriving the SIR model and simplifying it through dimensional analysis, introducing the basic reproduction number R_0 to understand its role in outbreak behavior. The system of differential equations is then solved numerically using Python to simulate disease trajectories under different parameters. To reflect real-world interventions, we extend the model by incorporating a small-scale vaccination term and apply regular perturbation methods to quantify its effects. Finally, using a publicly available COVID-19 cases across the United States from John Hopkins University, we perform parameter estimation by fitting the model to real infection trends, evaluating its predictive capability. By combining theoretical modeling and empirical data fitting, this study assesses the SIR model's strengths and limitations, while exploring how even low-rate interventions can shift epidemic outcomes.

Introduction

Background

The COVID-19 pandemic has been one of the most disruptive global health crises in modern history, affecting millions of lives and fundamentally reshaping education and cultural norms. Its rapid transmission, primarily through close human contact, and the typical recovery time of one to four weeks led to a wide range of policy interventions such as social distancing, mask mandates, quarantines, and large-scale vaccination efforts. In response to this unprecedented challenge, there has been renewed interest in mathematical modeling of infectious diseases to control, predict, and prevent large-scale spreads [1]. One of the most fundamental models in epidemiology is the Susceptible-Infected-Recovered (SIR) model, introduced by Kermack and McKendrick, which uses compartmental dynamics to capture the progression of an epidemic [5]. Despite its simplicity, the SIR model still remains as a valuable framework for modeling disease transmission and serves as a basis for more complex extensions. However, given the inherently stochastic nature of real-world outbreaks, it is worth questioning how well the SIR model reflect reality. Thus, our goal is to explore the dynamics of the classical SIR model, investigate how it changes under intervention strategies such as small-scale vaccination, and evaluate its ability to capture and predict the spread of COVID-19 using real-world data. Specifically, we aim to address the following question: *To what extent can the classical SIR model accurately describe the real-world spread of COVID-19, and how do small-scale interventions affect its predictions?*

Methodology

We will approach the proposed research question at hand in five main key stages that combines theoretical analysis with empirical data fitting. We begin by (1) formulating the classical SIR model as a system of ordinary differential equations, address its assumptions (and hence its limitations to the applicability to the real world), establishing the foundational framework for modeling disease transmission. To simplify the system and better understand its governing dynamics, (2) we perform dimensional analysis, introducing the basic reproduction number $R_0 = \beta/\gamma$ as a key parameter that determines outbreak behavior.To incorporate real-world interventions, we extend the model by (3) introducing a small vaccination term and apply regular perturbation methods to approximate its effects on infection dynamics. (4) The resulting equations are then solved numerically using Python, allowing us to simulate disease spread over time for varying parameter values. Finally, (5) we use publicly available COVID-19 case data curated from

the Johns Hopkins University [13] to fit the model to actual infection curves. This involves estimating transmission and recovery parameters through curve fitting techniques, enabling us to assess the predictive accuracy of the SIR model and evaluate its practical limitations. We will then conclude with some final thoughts, limitations to the fitted model, and possible future directions to continue this work.

Model Formulation

Classic SIR Model Derivation

One of the simplest, yet foundational disease models is known as the standard Susceptible-Infected-Recovered (SIR) model.The SIR model starts with dividing the population into the following three compartments: susceptible (individuals in the population but has not been infected), infected (individuals who have the virus), and recovered (those who got the virus, and no longer can spread it). The change of these three quantities over time is represented by the following set of ordinary differential equations (ODE):

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

Where β is the rate of contact between the susceptible and the infected (transmission rate) and γ is the recovery rate [10].

Assumptions and Limitations

With this simplest form comes with the following assumptions that are worth to consider:

- Population size (N) remains constant, i.e., no individual leaves nor enters a given population . Thus, in our SIR model, we have that $S + I + R = N$, with S always decreasing over time [9].
- The rate of infections if proportional to the rate of contact between infected (I) and susceptible (S) individuals.
- Infected individuals recover at a constant rate (γ is fixed).
- Individuals who are recovered cannot get infected again, i.e., no one can be reverted back to being susceptible again [8].
- Here, we assume that individuals who has recovered and passed away from the virus are in the recovered (R) compartment. Equivalently, we could also assume that all infected individuals fully recover [3].

While the classical SIR model offers a mathematically tractable framework for understanding disease spread, it relies on simplifying assumptions that limit its applicability in real-world environments. First, the assumption of a constant population size excludes births, deaths (*aside from those implicitly moved to the recovered compartment*), and migration/travel, which are significant factors in real-world epidemics where individuals frequently enter or leave a population through travel, relocation, or death. Thus, studies extended this SIR framework in order to improve its applicability to real-world outbreaks like COVID-19, including global dynamics of infection, which has shown to achieve higher predictive accuracies [1]. Second, the model assumes that the rate of new infections is strictly proportional to the product of susceptible and infected individuals, presuming homogeneous mixing, i.e., every individual is equally likely to contact any other, and the chance of an individual getting infected through contact is equally likely. In reality, population structure, behavioral factors, immune systems and health status of the infected, introduce considerable heterogeneity in contact patterns. Third, the model assumes permanent immunity once recovered, which ignores the possibility of reinfection, waning immunity, or new virus variants. Lastly, by aggregating both recovered and deceased individuals into a single compartment, the model loses the ability to distinguish between survival outcomes, which can be valuable information for

public health policy and epidemiological forecasting. Therefore, many studies extend the SIR model by adding interventions (e.g. social distancing, masking protocols, vaccinations, etc.) and new compartments to reflect specific epidemics [8].

Dimensional Analysis

We will begin with the dimensional form of the classical SIR model:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dR}{dt} &= \gamma I,\end{aligned}$$

Recall that:

- S, I, and R are the number of susceptible, infected, and recovered individuals at time t,
- β is the transmission rate [1/person·time],
- γ is the recovery rate [1/time].

Since we know that we are assuming a constant total population $N = S + I + R$, we introduce the following normalized variables:

$$s = \frac{S}{N}, \quad i = \frac{I}{N}, \quad r = \frac{R}{N}$$

Normalizing these variables results in population proportions, thus we now have:

$$s + i + r = 1$$

To non-dimensionalize time, we define:

$$\tau = \gamma t,$$

where τ is the dimensionless time variable. By the chain rule, we have:

$$\frac{d}{dt} = \gamma \frac{d}{d\tau}.$$

We substitute $S = sN$, $I = iN$, and $R = rN$ into the original equations, and convert derivatives with respect to t into derivatives with respect to τ [6]:

$$\begin{aligned}\frac{ds}{d\tau} &= -R_0 si,\\ \frac{di}{d\tau} &= R_0 si - i,\\ \frac{dr}{d\tau} &= i,\end{aligned}$$

where the dimensionless parameter

$$R_0 = \frac{\beta N}{\gamma}$$

is known as the basic reproduction number. In the context of epidemiology, R_0 represents the expected number of new cases from one infected individual [7]. In models where we normalize by population (i.e., s, i, and r are proportions), we typically assume $N = 1$, giving:

$$R_0 = \frac{\beta}{\gamma}.$$

Introducing R_0 and its Significance

Taking the the early-time dynamics of the infected population (i) and using the non-dimensionalized equation:

$$\frac{di}{d\tau} = R_0 si - i.$$

We know that at the start of an outbreak, the fraction of susceptible individuals is approximately 1 (i.e., $s(\tau) \approx 1$). Substituting $s \approx 1$ simplifies the equation to:

$$\frac{di}{d\tau} \approx (R_0 - 1)i.$$

- If $R_0 > 1$, then $\frac{di}{d\tau} > 0$, thus the number of infected individuals will grow over time.
- If $R_0 < 1$, then $\frac{di}{d\tau} < 0$, thus the number of infected individuals will diminish over time.

The basic reproduction number, R_0, thus plays a central role in understanding the dynamics of infectious disease outbreaks. As a dimensionless parameter, its value serves as a generalized threshold indicator for whether an epidemic will grow or die out. By analyzing the early-time behavior of the infected compartment in the nondimensionalized SIR model, we see that if $R_0 > 1$, the number of infected individuals increases exponentially, signaling the onset of an epidemic. Conversely, if $R_0 < 1$, the infection declines and eventually disappears [7]. When $R_0 = 1$, the system lies at a bifurcation point where the infection neither grows nor decays initially. This threshold behavior underscores the importance of public health interventions that aim to reduce R_0 below 1– through strategies such as vaccination, masking, and social distancing– in order to control or prevent further spread of outbreaks. Thus, R_0 can serve as a epidemiological benchmark that guides real-world policy and response.

Analysis and Results

Simulating SIR Model Behavior

Taking the classic SIR Model, we leveraged Python packages (namely scipy) to integrate and solve the system of equations. We first start with the initial conditions $S_0, I_0, R_0 = 1, 0.001, 0$, with $\beta = 0.35$ and $\gamma = 0.1$ (*Note that the* $R_0 = 0$ *represents the proportion of recovered individuals at time* $t = 0$). This represents the start of an outbreak, where 0.1% of the population has the virus. This is to ensure that we simulate an outbreak from start to finish. Simulating over 160 days, we plot the solutions, then observe how the model behavior changes over varying values of β and γ.

Results

Below are the solutions to $S(t), I(t), R(t)$, with the initial conditions defined above.

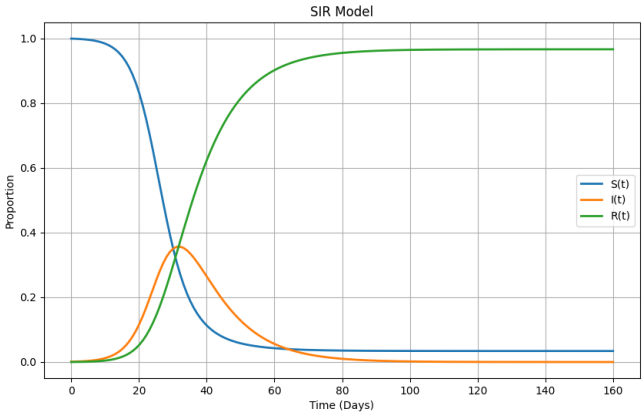

Figure 1: Classical SIR Model

We observe that there is always some noticeable 'peak' in $I(t)$, representing the time of which the largest proportion of individuals were infected with the disease. Simulating with varying values of R_0 (the basic reproductive number), we observe that the absolute maximum of $I(t)$ increases as R_0 increases, as expected.

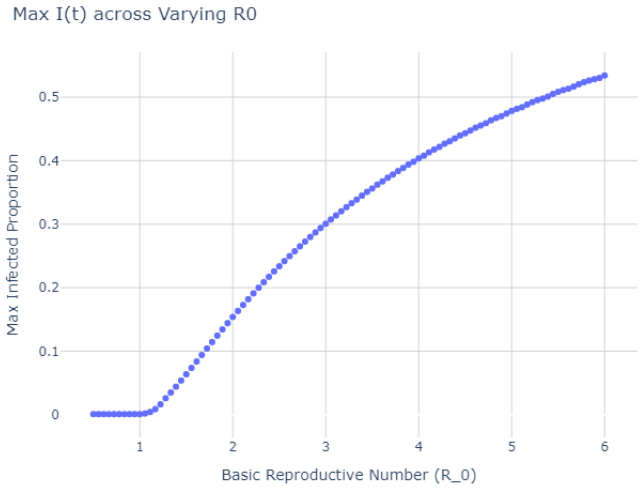

Figure 2: Peak Proportion of Infections Over Varying R_0

To observe how the model changes with varying values of R_0, we simulated 8 more SIR models, keeping γ fixed (Figure 3) then β fixed (Figure 4). Note that we have also kept all other initial conditions the same as the first simulation, ensuring that we are only changing the parameter of interest.

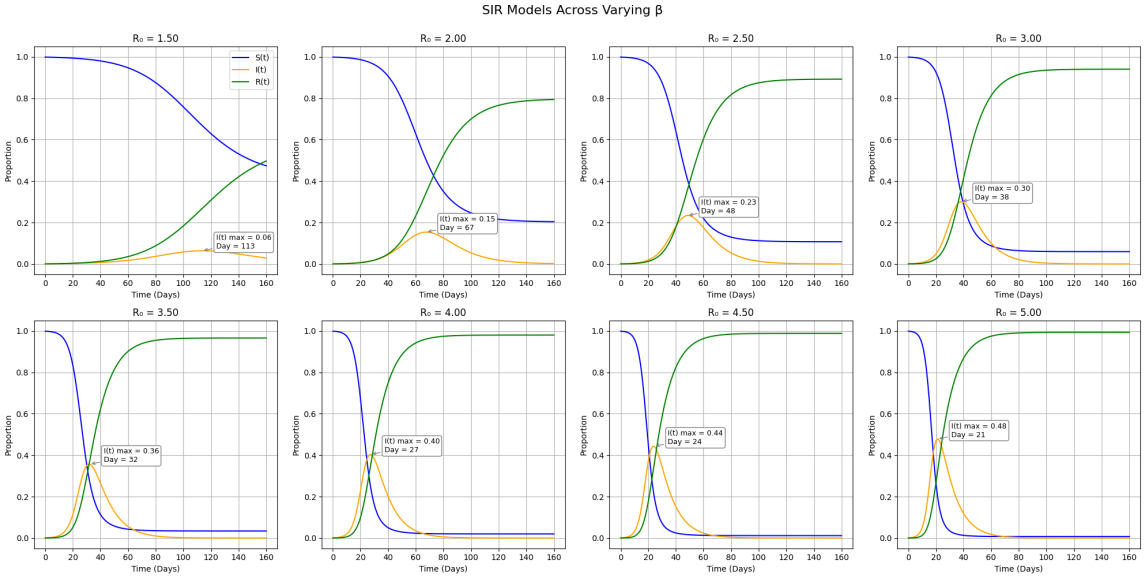

Figure 3: Simulated SIR Models with Varying β

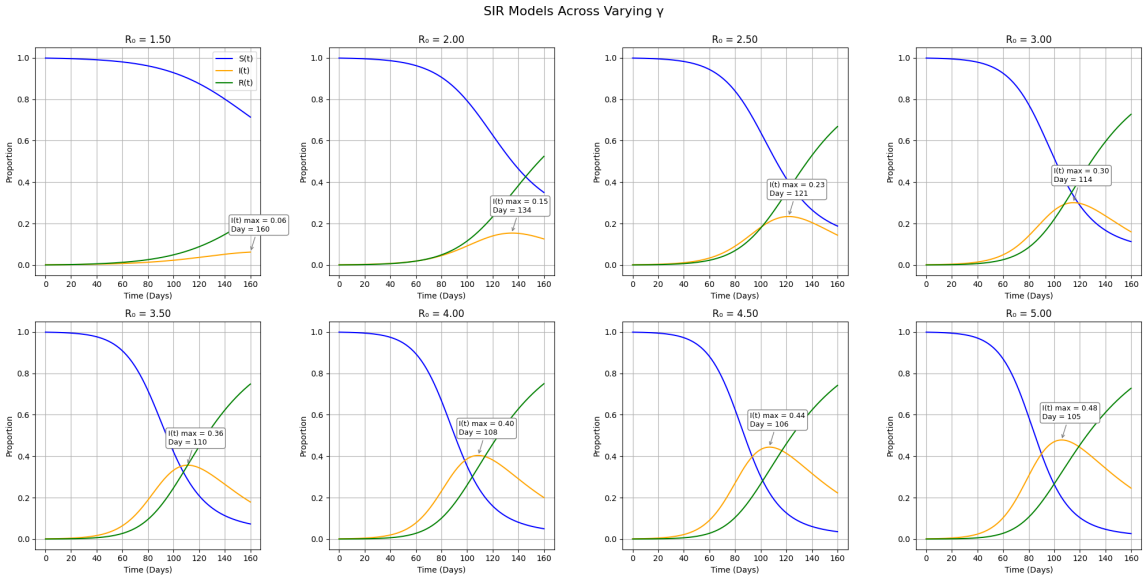

Figure 4: Simulated SIR Models with Varying γ

For both graphs, we osberve that as R_0 increases, the peak of infection $I(t)$ increases, the local maximum of $I(t)$ is reached earlier in time, $S(t)$ decreases at a faster rate, and $R(t)$ increases at a faster rate.

In addition to these shared trends, we also observe noticeable differences in the model's behavior when only increasing β and γ. While we see that R_0 governs how much a disease spreads, the values β and γ individually controls the timescale of that spread, thus R_0 does not inform us how fast an infection occurs. Increasing transmission rate (β) with fixed γ will lead to infections spreading faster, yet with a constant recovery rate, as reflected with shorter and faster outbreaks as β increases in (Figure 3). However, γ with fixed β leads to infections lingering longer for each individual, resulting in a slower, more 'stretched out' outbreak as reflected in (Figure 8).

Vaccine Extension: Perturbation Analysis

Though there are many ways to extend the SIR model to better reflect real-world outbreaks and policy implementations (*e.g. adding a new compartment in addition to SIR*), we investigate the effect of a small-scale vaccination intervention by introducing a constant-rate vaccination term into the classical model. For simplicity, we assume a small fraction of the susceptible population is vaccinated at a rate ε, modifying the system as equations as follows:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI - \varepsilon S,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dR}{dt} &= \gamma I + \varepsilon S.\end{aligned}$$

To ensure that the mathematical derivations and approximations presented in this section were meaningful, we consulted with an AI assistant (ChatGPT) to help validate and guide the application of regular perturbation methods. Given the limited amount of literature applying formal perturbation theory to the SIR model with vaccination, AI assistance was instrumental in verifying the correctness of intermediate steps and approximating solutions with confidence.

Similar to our work before, we define the following terms:

$$s = \frac{S}{N}, \quad i = \frac{I}{N}, \quad r = \frac{R}{N}, \quad \tau = \gamma t, \quad \epsilon = \frac{\varepsilon}{\gamma}, \quad R_0 = \frac{\beta}{\gamma}$$

Substituting these new values leads to the following dimensionless equations:

$$\frac{ds}{d\tau} = -R_0 si - \epsilon s,$$
$$\frac{di}{d\tau} = R_0 si - i,$$
$$\frac{dr}{d\tau} = i + \epsilon s,$$

To study the impact of small ϵ, we apply regular perturbation theory, finding the first-order correction. Thus, we assume that the solution can be written as a power series in ϵ:

$$s(\tau, \epsilon) = s_0(\tau) + \epsilon s_1(\tau) + \epsilon^2 s_2(\tau) + ...$$
$$i(\tau, \epsilon) = i_0(\tau) + \epsilon i_1(\tau) + \epsilon^2 i_2(\tau) + ...$$
$$r(\tau, \epsilon) = r_0(\tau) + \epsilon r_1(\tau) + \epsilon^2 r_2(\tau) + ...$$

Substituting into the system and collecting terms by powers of ϵ, we obtain:

Zeroth-order system (ϵ^0, no vaccination):

$$\frac{ds_0}{d\tau} = -R_0 s_0 i_0,$$
$$\frac{di_0}{d\tau} = R_0 s_0 i_0 - i_0,$$
$$\frac{dr_0}{d\tau} = i_0.$$

First-order correction (ϵ^1):

$$\frac{ds_1}{d\tau} = -R_0(s_1 i_0 + s_0 i_1) - s_0,$$
$$\frac{di_1}{d\tau} = R_0(s_1 i_0 + s_0 i_1) - i_1,$$
$$\frac{dr_1}{d\tau} = i_1 + s_0.$$

We first solve the classical SIR model to obtain (s_0, i_0, r_0), then use these to numerically solve the linear system for (s_1, i_1, r_1). The perturbed solution is approximated to first order as:

$$s(\tau, \epsilon) \approx s_0(\tau) + \epsilon s_1(\tau), \quad i(\tau, \epsilon) \approx i_0(\tau) + \epsilon i_1(\tau), \quad r(\tau, \epsilon) \approx r_0(\tau) + \epsilon r_1(\tau).$$

Similar to our methodology before, we used Python to solve the set of equations for both the zeroth-order and first-order correction, and plotted these solutions.

The figure below compares the unperturbed SIR model (no vaccination, dashed lines), with first-order corrected model (with small vaccination, solid lines) with $\epsilon = 0.01$. Each subplot shows the change of the three compartments over τ.

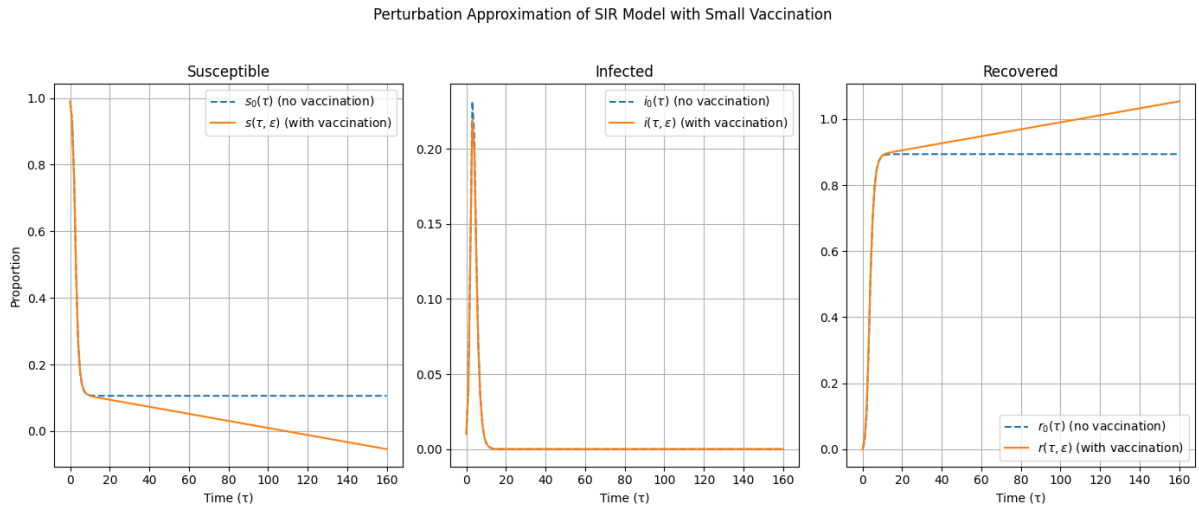

Figure 5: S, I, R Approximation with Small Vaccination

Discussion and Limitations to the Perturbation Approximation

Susceptible Population $s(\tau, \epsilon)$

In the unperturbed model, the susceptible population declines initially but eventually flattens out at a non-zero value. This reflects the fact that once the infection dies out, there are no further transitions between compartments– a portion of the population remains susceptible indefinitely.

In contrast, the first-order corrected curve shows that the susceptible population continues to decline beyond the infection phase. This is due to the vaccination term: even after the disease has mostly vanished, the added vaccination term continues to reduce the number of susceptibles. This leads to the susceptible proportion declining consistently over time, without constraints. We also observe that at some point, the proportion drops below 0, an implausible value.

Infected Population $i(\tau, \epsilon)$

Both curves for the infected population display a typical epidemic peak followed by a sharp decline. However, the perturbed model shows a slightly lower peak. This is expected as it is the intended effect of vaccination: by gradually reducing the number of susceptibles, the model curbs disease transmission, thereby limiting the spread. However, the effect here is substantially small due to the small value of the vaccination rate ϵ.

Recovered Population $r(\tau, \epsilon)$

The recovered population in the unperturbed model reaches a steady value below 1.0, consistent with a fixed population where some individuals remain uninfected. However, in the first-order corrected model, the recovered proportion continues to increase over time, and even exceeds 1.0. This occurs because the model adds vaccinated individuals to the recovered class, and vaccination continues even after the infection is gone.

Limitations

While the perturbation analysis provides valuable insight into the influence of small-scale vaccination, it also introduces notable limitations related to population proportions. Specifically, the first-order approximation results in implausible behavior over long periods of time. In the simulation, the recovered proportion $r(\tau, \epsilon)$ exceeds 1.0, and the susceptible proportion $s(\tau, \epsilon)$ eventually becomes negative, violating the constraint that $s + i + r = 1$ for a closed, normalized population.

These limitations arise because applying regular perturbation methods expands the solution around the classical SIR model without enforcing hard bounds (i.e., ensuring that $s+i+r=1$). The vaccination term continues to remove individuals from the susceptible class and adds them to the recovered class constantly, even after the infection has subsided. Though it was considered to adjust the solutions by enforcing these bounds, this would deviate from the structure of regular perturbation methods. We choose to retain the unmodified approximation to preserve the mathematical consistency of the perturbation expansion and instead acknowledge its limitations when interpreting long-term model behavior.

Model Fit to COVID-19 Data

Data

In order to further analyze the classical SIR model's ability to capture real-world outbreak dynamics, we leveraged a publicly available COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU) [3]. This repository offers comprehensive records of confirmed cases, deaths, and recoveries, compiled from various global and local health authorities. All files are orgaized in CSV formats, with daily reports and time series files available at global, national, and subnational levels. For this study, we focused on the U.S. state-level data, specifically analyzing the daily reports in New York in 2021. The repository includes a CSV file of COVID-19 reports for each US state for each day, from 2020-2022. However, there were sufficient amounts of missing files in 2020, and though it would be more ideal to analyze COVID-19 dynamics in 2020– the year with the first confirmed COVID-19 case in the United States– for better data quality, we have decided to only extract files from 2021.

Among the 21 columns available for each file, we focus on the following columns below. The descriptions below are adapted directly from the documentation provided in the original CSSE GitHub repository:

- **UID** – Unique identifier for each row entry.
- **Province_State** – The name of the state within the USA.
- **Last_Update** – The most recent date the file was updated.
- **Confirmed** – Aggregated case count for the state.
- **Deaths** – Aggregated death toll for the state.
- **Recovered** – Aggregated recovered case count for the state.
- **Active** – Aggregated confirmed cases that have not been resolved.
- **Incident_Rate** – Number of cases per 100,000 persons.

Below are the first 10 rows reported on 01-01-2021.

Province_State	Last_Update	Confirmed	Recovered	Deaths	Active	Incident_Rate
Alabama	2021-01-02 05:30:44	365747	202137.0	4872	158738.0	7459.3759
Alaska	2021-01-02 05:30:44	47019	7165.0	206	39648.0	6427.3558
American Samoa	2021-01-02 05:30:44	0	nan	0	nan	0.0
Arizona	2021-01-02 05:30:44	530267	76934.0	9015	444318.0	7285.1713
Arkansas	2021-01-02 05:30:44	229442	199247.0	3711	26484.0	7602.9457
California	2021-01-02 05:30:44	2434971	nan	26298	nan	6164.4697
Colorado	2021-01-02 05:30:44	362438	18102.0	5435	314186.0	5854.7744
Connecticut	2021-01-02 05:30:44	185708	9800.0	5995	169913.0	5208.7812
Delaware	2021-01-02 05:30:44	58064	18851.0	1065	38148.0	5962.8411
Diamond Princess	2021-01-02 05:30:44	49	nan	0	nan	nan

Figure 6: COVID-19 Data from 01-01-2021 (Raw)

Approach

In order to fit and parameterize SIR models onto the dataset, we will perform the procedures as follows. First, We will extract all files from 2021, and concatenate them all onto one pandas DataFrame. After performing some elementary data cleaning procedures, in order to get a time series of S, I, R, we will derive the following quantities from the raw data:

- **N**: Total population, estimated as

$$N = \frac{\texttt{Confirmed} \times 100{,}000}{\texttt{Incident\_Rate}}$$

- **S(t)**: Susceptible population, computed as

$$S(t) = N - \texttt{Confirmed}$$

- **I(t)**: Infected population, taken directly from the `Active` case count.
- **R(t)**: Recovered population, computed as

$$R(t) = \texttt{Recovered} + \texttt{Deaths}$$

Following these preprocessing procedures, we get the resulting DataFrame below:

province_state	confirmed	recovered	deaths	active	incident_rate	date	n	s(t)	i(t)	r(t)
alabama	365747	202137.0	4872	158738.0	7459.3759	2021-01-02	4903185.0	4537438.0	158738.0	207009.0
alaska	47019	7165.0	206	39648.0	6427.3558	2021-01-02	731545.0	684526.0	39648.0	7371.0
american samoa	0	nan	0	nan	0.0	2021-01-02	nan	nan	nan	nan
arizona	530267	76934.0	9015	444318.0	7285.1713	2021-01-02	7278717.0	6748450.0	444318.0	85949.0
arkansas	229442	199247.0	3711	26484.0	7602.9457	2021-01-02	3017804.0	2788362.0	26484.0	202958.0
california	2434971	nan	26298	nan	6164.4697	2021-01-02	39500088.9473	37065117.9473	nan	nan
colorado	362438	18102.0	5435	314186.0	5854.7744	2021-01-02	6190469.1182	5828031.1182	314186.0	23537.0
connecticut	185708	9800.0	5995	169913.0	5208.7812	2021-01-02	3565287.0	3379579.0	169913.0	15795.0
delaware	58064	18851.0	1065	38148.0	5962.8411	2021-01-02	973764.0	915700.0	38148.0	19916.0
diamond princess	49	nan	0	nan	nan	2021-01-02	nan	nan	nan	nan

Figure 7: COVID-19 Data from 01-01-2021 (Cleaned)

We observed that nearly all values for infected $i(t)$ and recovered $r(t)$ are missing after March. This made it unfeasible to fit the model using the full system $(s(t), i(t), r(t))$. However, the susceptible proportion $s(t)$ remained consistently available and reliable throughout the year. Thus, moving forward we restricted our model **fitting to the susceptible curve**. However, the methodology for fitting the susceptible can be directly applied to both the infection and recovery compartments.

To fit these curves using the classical SIR model, we first defined the SIR model as a system of ordinary differential equations (ODEs), where the rates of change for the susceptible, infected, and recovered compartments are governed by the parameters β (transmission rate) and γ (recovery rate). Like before, these equations will be numerically solved using `scipy.integrate.odeint`.

We fit the model to the observed $s(t)$ curve by minimizing the mean squared error (MSE) between the simulated and empirical values (*note we computed the proportion of susceptible individuals, by dividing each `s(t)` with `N`, which is assumed to be a constant value*). This optimization was performed using `scipy.optimize.minimize`, with bounded constraints and initial guesses for β and γ.

Using these best-fit parameters, we simulated the SIR model and compared the resulting $s(t)$ trajectory to the observed data.

Results and Model Evaluation

Performing these procedures on New York reports in 2021, we get the following curves:

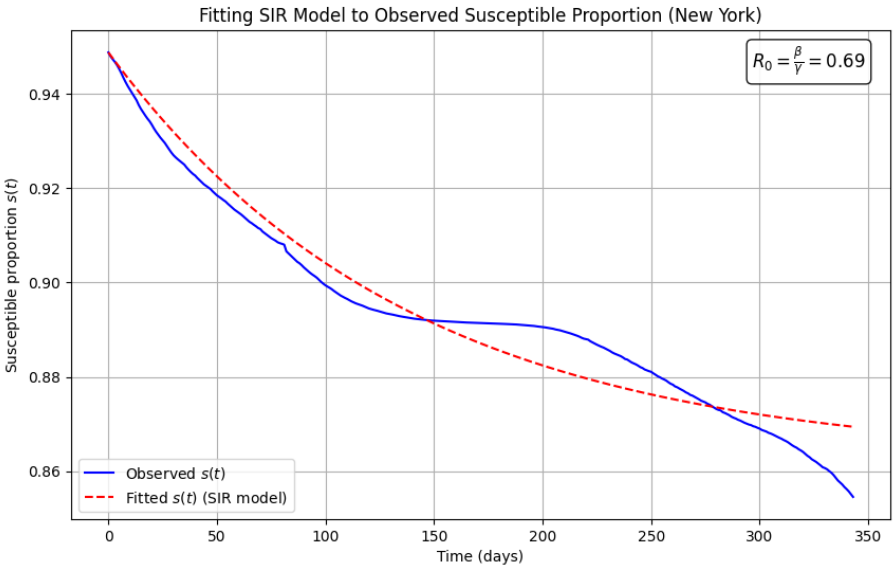

Figure 8: Fitted SIR Model for New York 2021

Figure 8 shows the observed susceptible proportion $s(t)$ for New York over the course of 2021, alongside the fitted curve generated by the classical SIR model using the fitted parameters β and γ. While the general trend of declining susceptibility is captured by the model, there are clear discrepancies in the curvature and timing of the decline. The fitted model produces a smoother, exponential-like decay, whereas the empirical data reveals inflection points and periods of slower decline that the SIR model fails to replicate.This divergence suggests that although the model provides a useful first approximation, it does not fully capture the dynamics of the real-world outbreak. These inconsistencies become more pronounced in the latter half of the year, indicating that additional factors may be influencing the trajectory of $s(t)$ in ways not accounted for by the basic model structure.

Limitations

Thus, several limitations should be acknowledged in interpreting these results:

- **Noisy and Incomplete Data:** Real-world COVID-19 reporting is prone to inconsistencies, delays, and missing values. In particular, data for the infected $i(t)$ and recovered $r(t)$ compartments were largely null after March, making it infeasible to fit the full system and restricting our analysis to the susceptible curve alone.
- **Derived Quantities Based on Assumptions:** Our definitions of $S(t)$, $I(t)$, and $R(t)$ rely on assumptions about how reported columns map to SIR compartments (*e.g., assuming active cases represent infections and deaths are included in the recovered class*). These mappings introduce structural uncertainty into the model.
- **External Real-World Factors:** The classical SIR model does not account for public health interventions such as vaccinations, mask mandates, social distancing, or behavioral changes such as quarantine mandates. These external factors significantly influence disease spread and likely contributed to deviations between the model and the observed data.
- **Mid-Epidemic Data Window:** The first confirmed COVID-19 case in the United States was reported to be in the year of 2020 [10]. However, our dataset focuses on case trends during 2021, meaning we are fitting the model to the middle of an epidemic rather than modeling the full trajectory from initial outbreak to resolution. This affects the interpretability of parameters (such as R_0) and limits the model's capacity to simulate early outbreak dynamics. Especially since the classical SIR model represents early to late outbreak dynamics, it could be unfeasible to attempt to fit the observed data with the SIR framework, unless we either (1) find new data that includes

the early stage of an outbreak, or (2) modify our SIR model so that we align it according to the time/stage we are given.

Overall, these limitations highlight the importance of interpreting simple compartmental models cautiously when applied to real-world data, particularly under complex and evolving public health conditions.

Conclusions and Future Work

We investigated the extent to which the classical Susceptible-Infected-Recovered (SIR) model can accurately describe and predict the spread of COVID-19. To start, we derived the model and addressed its underlying assumptions, applied dimensional analysis to simplify the system, and introduced the basic reproduction number R_0 and demonstrated that it serves as a key parameter for understanding outbreak dynamics. Through numerical simulation and perturbation analysis, we examined how the model behaves under various parameter regimes and under small-scale vaccination interventions. We then fitted the model to COVID-19 reports from New York in 2021, estimating γ and β by minimizing error (MSE) between the observed and simulated susceptible proportions.
Our findings demonstrate that while the classical SIR model provides a foundational and interpretable framework for modeling epidemics, its simplicity limits its ability to fully capture the nuanced behavior of real-world outbreaks. In particular, the model struggled to replicate the curvature and inflection points observed in the empirical data, especially in the presence of external interventions and reporting inconsistencies. Additionally, the perturbation analysis revealed limitations in long-term dynamics, where population proportions could violate feasibility bounds due to continued vaccination in a closed system.
In addition, there are several promising avenues for future work that could address the limitations identified in this study:

- **Incorporating Stochasticity:** The deterministic nature of the classical SIR model does not account for random fluctuations inherent in real-world epidemics. Future studies could explore stochastic SIR models, which introduce noise and probabilistic transitions to better capture uncertainty and variability in transmission and recovery.
- **Extending the Model Structure:** Many studies have enhanced the SIR framework by adding compartments to represent additional epidemiological stages and public health interventions. In fact, many sources literature extends the SIR model by adding new compartments that best reflect real-world outbreaks. These extended models can include Susceptible-Exposed-Infected-Recovered (SEIR) models (adding an exposed (E) compartment for individuals who are infected but not yet infectious) [10], SIQR Models (adding a quarantined compartment to reflect populations that were quarantined while they were infected) [12], and extending the model to account for travel between regions [10].
- **Improved Data Collection and Broader Application:** A key limitation of our work was the lack of complete data across all compartments and the use of data from the middle of an epidemic. Future work could prioritize acquiring datasets that capture the full outbreak trajectory, i.e., from initial spread to eventual decline, enabling a more robust validation and generalizability of the modeling framework.

By advancing in these directions, future studies can build upon the classical SIR model to develop richer, more realistic, and policy-relevant tools for understanding and managing infectious disease outbreaks.

References

[1] Kermack, W. O., & McKendrick, A. G. "A contribution to the mathematical theory of epidemics." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721, 1927. https://doi.org/10.1098/rspa.1927.0118

[2] Cauchemez, S., Ferguson, N. M., Wachtel, C., Tegnell, A., Saour, G., Duncan, B., & Nicoll, A. "Closure of schools during an influenza pandemic." *The Lancet Infectious Diseases*, *9*(8), 473–481, 2009. https://doi.org/10.1016/S1473-3099(09)70176-8

[3] Taghvaei, A., Georgiou, T. T., Norton, L., & Tannenbaum, A. *Fractional SIR epidemiological models* (Version 2) [Preprint], 2020, April 30. medRxiv. https://doi.org/10.1101/2020.04.28.20083865

[4] Panovska-Griffiths, J. "Can mathematical modelling solve the current COVID-19 crisis?" *BMC Public Health*, *20*, 551, 2020. https://doi.org/10.1186/s12889-020-08671-z

[5] Dong, E., Du, H., & Gardner, L. "An interactive web-based dashboard to track COVID-19 in real time." *The Lancet Infectious Diseases*, *20*(5), 533–534, 2020. https://doi.org/10.1016/S1473-3099(20)30120-1

[6] Carvalho, A. M., & Gonçalves, S. "An analytical solution for the Kermack–McKendrick model." *Physica A: Statistical Mechanics and its Applications*, *566*, 125659, 2021. https://doi.org/10.1016/j.physa.2020.125659

[7] Lord, A. G. *Modeling COVID-19 data using an SIR model. Citations: Journal of Undergraduate Research*, *18*, 132–136, 2021. https://www.lagrange.edu/academics/undergraduate/undergraduate-research/$_i mages/19 - Citations_2021_A ly_L ord.pdf$

[8] Melikechi, O., Young, A. L., Tang, T., Bowman, T., Dunson, D., & Johndrow, J. "Limits of epidemic prediction using SIR models." *Journal of Mathematical Biology*, *85*(4), 36, 2022. https://doi.org/10.1007/s00285-022-01804-5

[9] Prodanov, D. "Analytical solutions and parameter estimation of the SIR epidemic model." In *Mathematical Analysis of Infectious Diseases*, 163–189, 2022. https://doi.org/10.1016/B978-0-32-390504-6.00015-2

[10] AlQadi, H., & Bani-Yaghoub, M. "Incorporating global dynamics to improve the accuracy of disease models: Example of a COVID-19 SIR model." *PLOS ONE*, *17*(4), e0265815, 2022. https://doi.org/10.1371/journal.pone.0265815

[11] Paul, J. N., Mbalawata, I. S., Mirau, S. S., & Masandawa, L. "Mathematical modeling of vaccination as a control measure of stress to fight COVID-19 infections." *Chaos, Solitons & Fractals*, *166*, 112920, 2023. https://doi.org/10.1016/j.chaos.2022.112920

[12] Kalachev, L., Landguth, E. L., & Graham, J. "Revisiting classical SIR modelling in light of the COVID-19 pandemic." *Infectious Disease Modelling*, *8*(1), 72–83, 2023. https://doi.org/10.1016/j.idm.2022.12.002

[13] Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. *COVID-19 data repository* (Version as of May 1, 2025) [Data Repository], 2024. https://github.com/CSSEGISandData/COVID-19

Appendices

The full source code, data processing scripts, and Jupyter notebooks used in this report are available at the following GitHub repository: `https://github.com/leena-kang/SIR_Modeling_Analysis`