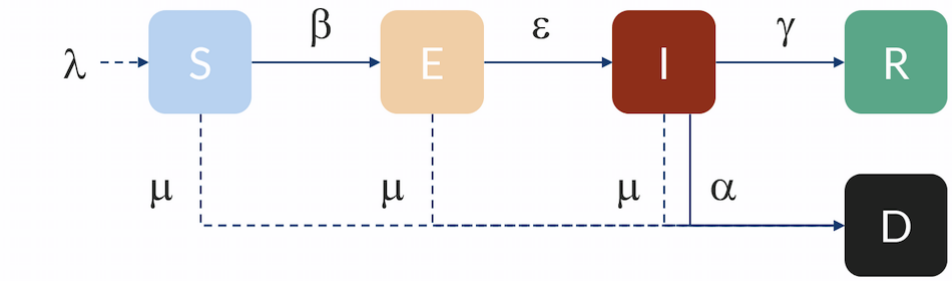
Mathematical Epidemiology: From Data to Model

Stephen Adei

September 21, 2024

Bachelorscriptie Wiskunde

Begeleiding: prof. dr. Ale Jan Homburg



Korteweg-de Vries Instituut voor Wiskunde
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

Abstract

In this thesis we discuss simple epidemic models and how to fit such models to data. By first introducing the Kermack-McKendrick SIR model, we analyse how we can assign individuals to different compartments and how these compartments interact. Afterwards, we explore different ways of expanding this simple model to better accommodate reality, like adding compartments of exposed or quarantined individuals to t model. Throughout the thesis there are several examples of fitting data to an epidemic model, programmed in Python. Other relevant topics like the basic reproduction number $\mathcal{R}_0$, used to determine the number of secondary cases produced by a single individual, and the Akaike information criterion used to compare different models for a fixed data set, are discussed as well.

Titel: Mathematical Epidemiology:
From Data to Model
Auteur: Stephen Adei, stephen.adei@student.uva.nl, 11155000
Begeleiding: prof. dr. Ale Jan Homburg,
Einddatum: September 21, 2024

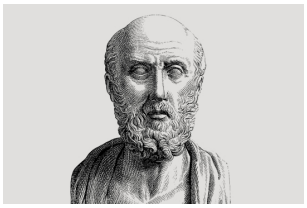
Korteweg-de Vries Instituut voor Wiskunde
Universiteit van Amsterdam
Science Park 904, 1098 XH Amsterdam
`http://www.kdvi.uva.nl`

Contents

1. Introduction

And then there was COVID-19. The year is 2020 and what started out as a new, unknown infection in China became a determining factor in what our day-to-day lives, all over the world, looked like. From voluntarily testing to involuntary lockdowns, the list of measures adopted in attempts to control this elusive nemesis is close to endless. As with most major changes in society, the measures (read: restrictions) imposed by governments were met with great scrutiny. Did the government realise what it means for a restaurant to be closed for the greater part of the year? How come this new disease is approached with so much fear and trembling, while its symptoms look an awful lot like those of the common flu? What is the point of vaccinating an entire population if only a (small) part of it is vulnerable to the disease? What is this "R" I keep hearing about?

This and many other questions make one thing abundantly clear; there is a lot of uncertainty around COVID-19. This is where epidemiology comes into play. Hippocrates (460 - 377 B.C.) is often seen as the father of epidemiology since he was the first to accredit symptoms of a disease to natural rather than supernatural causes. The plague of Athens (430 - 426 B.C.) was the first substantial epidemic that was described by historians, most notably by Thucydides (460 - 400 B.C.).



Hippocrates

Epidemiology has an extensive past. Mathematical epidemiology however is only roughly 350 years old. John Graunt (1620-1674) performed the first statistical study of a disease with his book *Natural and Political Observations Made Upon the Bills of Mortality*, concerning public health statistics.

In today's fight against COVID-19, mathematical epidemiology plays a unique role. Due to technological advances of the past century, governments and scientists have access to an immense amount of data concerning local or global populations. What intrigued me the most was how this data could be of any significance when we know so little about the disease itself. This led me to research what mathematical epidemiology is.

In chapter 2, we introduce the Kermack-McKendrick *SIR* epidemic model. This model was introduced by Kermack and McKendrick in 1927 and raised the bar of what mathematical epidemiology entailed. Afterwards, we put the SIR model to the test by fitting the data of an influenza outbreak at an English boarding school to it.

In chapter 3, we explore several extensions of the Kermack-McKendrick and illustrate what the incorporation of birth and death rates means to such a model. We continue by analysing the prevalence of HIV in the UK by means of fitting data to an *SIIII* model.

Next, in chapter 4, we look at model selection using the Akaike information criterion,

a method from information theory used to determine which model, or models combined, best describe the disease at hand.

In chapter 5, we take a closer look at $\mathcal{R}_0$, or the *basic reproduction number* as it is formally known. We explore the key features of this measure and discuss it's real-world implications.

We conclude this thesis by briefly pondering on the accuracy and effectiveness of such methods and models with regards to the COVID-19 epidemic in the Netherlands.

2. The SIR Model

> The word "epidemiology" is derived from the Greek terms epi, which means "upon," demos, which means "people," and logos, which means "study."

We briefly examine the SIR epidemic model. Hereby we closely follow the findings of Martcheva[1]. The SIR model was proposed by McKendrick and Kermack[2] and is known as one of the first and simplest mathematical epidemic , great for an introduction on the subject.

2.1. Derivation of the SIR Model

First, we assume that the total population N can be separated into three (isolated) compartments:

- Susceptibles (S): These are healthy individuals that can contract the virus.
- Infected (I): These are sick individuals that have contracted the disease and are able to infect others.
- Recovered (R): These are individuals that have recovered from the disease and can no longer contract it.

These compartments together make up the total population. Note that the size of a compartment changes over time, so $S(t)$, $I(t)$, $R(t)$ and their sum $N(t)$ are functions of time t, such that

$$N(t) = S(t) + I(t) + R(t).$$

The number of individuals that are sick at a specific time is called the *prevalence* of the disease and the number of individuals that become sick during a specified period the *incidence* of the disease. In our SIR model, the incidence can be interpreted as

$$S'(t) = -\text{incidence} = -\beta IS,$$

whereas the prevalence is given by

$$I(t) = \text{prevalence},$$

for specific t.

Figure 2.1.: Flowchart of the Kermack–McKendrick SIR epidemic model

The SIR model comes with two assumptions, to simplify reality. The first assumption is that the total population $N(t)$ is constant. The second assumption is that infected individuals are infectious. This means that every individual in compartment $I(t)$ can infect one or more individuals in compartment $S(t)$. The first assumption might seem intuitive but is not always the case. An example is diabetes. This is caused by the body's inability to use blood sugar, not by an infection. Diseases best modelled with the SIR model are infectious diseases that lead to permanent immunity. This occurs most often with childhood diseases like chickenpox and smallpox. We describe the dynamics in each compartment by a system of ordinary differential equations (ODEs). To determine these equations, we analyse how the different compartments change over time and affect each other.

Let c be the contact rate for an infectious individual. Then cN is the number of contacts an individual makes per time unit. $\frac{S}{N}$ can be interpreted as the probability that an individual encountered, is susceptible. Then, $cN\frac{S}{N} = cS$ is the number of *susceptible* contacts an infectious individual makes. Not every contact will lead to transmission of the disease, since every contact has a probability p of leading to transmission. So, pcS is the number of susceptible individuals that become infected per time unit. If we take βSI as the number of newly infected individuals, then we can interpret β as

$$\beta = pc.$$

If we define

$$\lambda(t) = \beta I,$$

then $\lambda(t)$ is the *force of infection*. In this model the force of infection is rather simple. Later on we will see that for diseases with a more complicated incidence, the force of infection also becomes more intricate.

In short, when an infectious individual encounters a susceptible one, the latter becomes infected with a probability p and moves from the susceptible compartment into the infected compartment. In our model however, we make use of β instead of p because β is a combination of both the contact and the infection rate. Similarly, infected individuals that have recovered move, to the recovered compartment with a probability α.

So, the SIR model is given by the following ODEs:

$$\begin{aligned} S'(t) &= -\beta IS, \\ I'(t) &= \beta IS - \alpha I, \\ R'(t) &= \alpha I, \end{aligned}$$

along with the initial values $S(0)$ and $I(0)$. Naturally, α and β are positive constants. Here α can be interpreted as the average time spent in the infectious compartment. For example, if the infective period is 2.1 days, we find

$$\alpha = \frac{1}{2.1} = 0.476.$$

Now β is the transmission rate constant, the rate at which a susceptible individuals are infected per infected individual. This parameter is often estimated with the help of Mathematica or MATLAB, to match the estimated α and the initial values. Martcheva [1] gives detailed derivations of α and β.

2.2. Mathematical Properties of the SIR Model

From the SIR model we can infer distinctive dynamics. These dynamics can largely be attributed to the assumptions mentioned before. First, the number of susceptible individuals $S(t)$ is declining, monotone and positive. So, we find that the following limit exists,

$$\lim_{t\to\infty} S(t) = S_\infty.$$

Secondly, the number of recovered individuals $R(t)$ is increasing, monotone and bounded above by $N(t)$. So,

$$\lim_{t\to\infty} R(t) = R_\infty$$

also exists.

Lastly, the dynamics of the infected compartment are more ambiguous because this compartment can exhibit either monotone or non-monotone behaviour. Hence, the requirement for an initial increase in the number of infected individuals is

$$\beta S(0) - \alpha > 0,$$

since in this case we have $I'(0) = \beta I(0)S(0) - \alpha I(0) > 0$. We can rewrite this condition to

$$\frac{\beta S(0)}{\alpha} > 1. \tag{2.1}$$

Requirements like these are called *threshold conditions*, they are a prerequisite for the initiation of a disease.

A disease or an epidemic dies out when

$$\lim_{t\to\infty} I(t) = 0.$$

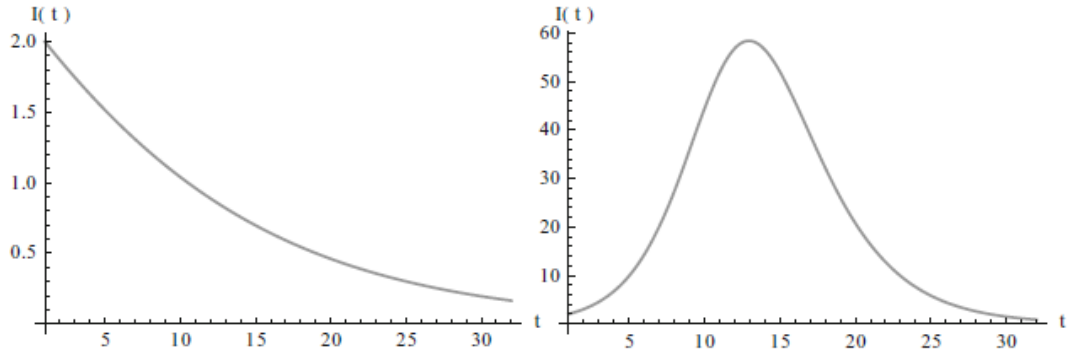


Figure 2.2.: *Left*: shows the prevalence monotonically decreasing. *Right*: shows the prevalence first increasing and then decreasing to zero [1]

Intuitively this makes sense; when there are no infected individuals, the disease can no longer spread.

To solve the system, it is best to omit the $R'(t)$ from the system of equations, since

$$R'(t) = -S'(t) - I'(t).$$

The remaining set of ODEs is then,

$$\begin{aligned} S'(t) &= -\beta IS, \\ I'(t) &= \beta IS - \alpha I. \end{aligned}$$

Solving this gives us, among other results, the following,

$$\frac{\beta}{\alpha} = \frac{\ln \frac{S_0}{S_\infty}}{S_0 + I_0 - S_\infty} \tag{2.2}$$

and

$$I_{\max} = -\frac{\alpha}{\beta} + \frac{\alpha}{\beta} \ln \frac{\alpha}{\beta} + S_0 + I_0 - \frac{\alpha}{\beta} \ln S_0. \tag{2.3}$$

See Martcheva (2013) for details. Note that $S_\infty < S_0 < S_0 + I_0$, because S is a decreasing function. Furthermore, equation 2.3 lets us compute the maximum number of infected individuals reached in the epidemic. This is especially useful information when studying new diseases, since this gives an indication of when the number of infections will start to decline.

2.3. Example: Influenza at West Country English Boarding School

To illustrate the power of modelling a disease, we will fit the SIR model to the data and epidemic scenario as described by Rose[3]. In January 1978, an epidemic occurred at the

Time (in days)	Number cases	Time	Number cases	Time	Number cases
1	2	11	53	21	13
2	5	12	55	22	14
3	10	13	58	23	11
4	12	14	-	24	12
5	14	15	52	25	9
6	15	16	42	26	7
7	-	17	40	27	5
8	31	18	30	28	4
9	42	19	23	29	2
10	45	20	19	30	1
-	-	-	-	32	1

Table 2.1.: Prevalence of influenza in an English boarding school

West Country English Boarding School, which housed 578 boys. The epidemic began on January 15, 1978. For our model, we will simplify the epidemic by ignoring the different vaccines administered to the students. It is possible to model for these variations, as we will see later in this thesis. Nevertheless, in this example we stick to the SIR model as presented by McKendrick and Kermack (1927).

The average length of fever was between four and six days, but the infective period is closer to 2 days. So, we take 2.4 days as our initial guess for the infective period. This means

$$\hat{\alpha} = \frac{1}{2.4} = 0.416667.$$

Alongside the data we get the following initial values:

$$S_1 = 576, I_1 = 2$$

As our initial guesses for α and β we choose

$$\hat{\alpha} = 0.416667, \hat{\beta} = 0.000841081807.$$

We can also estimate the maximum number of infections, namely

$$\hat{I}_{\max} = -\frac{\hat{\alpha}}{\hat{\beta}} + \frac{\hat{\alpha}}{\hat{\beta}} \ln \frac{\hat{\alpha}}{\hat{\beta}} + S_1 + I_1 - \frac{\hat{\alpha}}{\hat{\beta}} \ln S_1 = 250.700052.$$

This number is a lot higher than anything we see in the table.

To find the best fit for the data, we write a program in Python that determines the optimal values for α and β using the least-squares method. With the least-squares method, we are given n data points

$$\{(t_i, Y_i) \quad | \quad \text{for } 1 \leq i \leq n\},$$

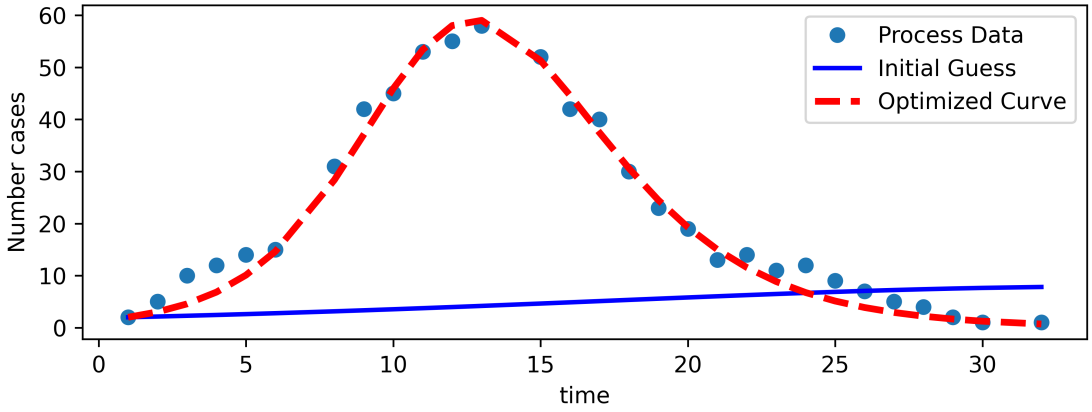


Figure 2.3.: Optimized plot for fitting influenza data to SIR model. All compartments

and the function $I(t)$ that we are fitting. The *sum of squared errors*, also known as the SSE, is the sum of the differences between the observed value from the data and the predicted value that results from our estimate or (initial) guess, at given t, for all t in the data. So,

$$\text{SSE} = \sum_{j=1}^{n} (Y_j - I(t_j))^2.$$

This sum is a function of the parameters of our model, so minimizing it gives us the optimal values for the parameters, as desired.

We roughly proceeds as follows. We compute the $I(t)$ curve, based on the initial guesses $\hat{\alpha}$ and $\hat{\beta}$. Next, we define an objective function which computes the SSE for said $\hat{\alpha}$ and $\hat{\beta}$. Finally, we use the *'L-BFGS-B'* method to minimize the objective function. BGFS, or *Broyden–Fletcher–Goldfarb–Shanno algorithm* is a numerical optimisation algorithm, used to solve nonlinear optimisation problems like ours[4]. The *L-BFGS* algorithm does the same, utilising limited computer memory and the *L-BFGS-B* algorithm is capable of handling simple bounds on the variables[5]. This last property is especially convenient since we want all the variables (the parameters in our model) to be positive. Once the objective function is minimised, we plot the data points (light blue dots), our initial guess (dark blue line) and our optimal estimate (red dashed line) in a single graph. The code can be found in the appendix.

2.3.1. The fitted model

As you can see in figure 2.3, the initial guess is off by quite a bit. Fortunately, our initial guess was of little influence to the optimisation, after which we found the parameters to be

$$\alpha = 0.5994695597295631, \beta = 0.001766602951828876.$$

This result can be interpreted as that the actual infective period is $\frac{1}{0.5994695597295631} = 1.67$ days long instead of 2.4. On the other hand, $\beta = 0.001766602951828876$ can roughly

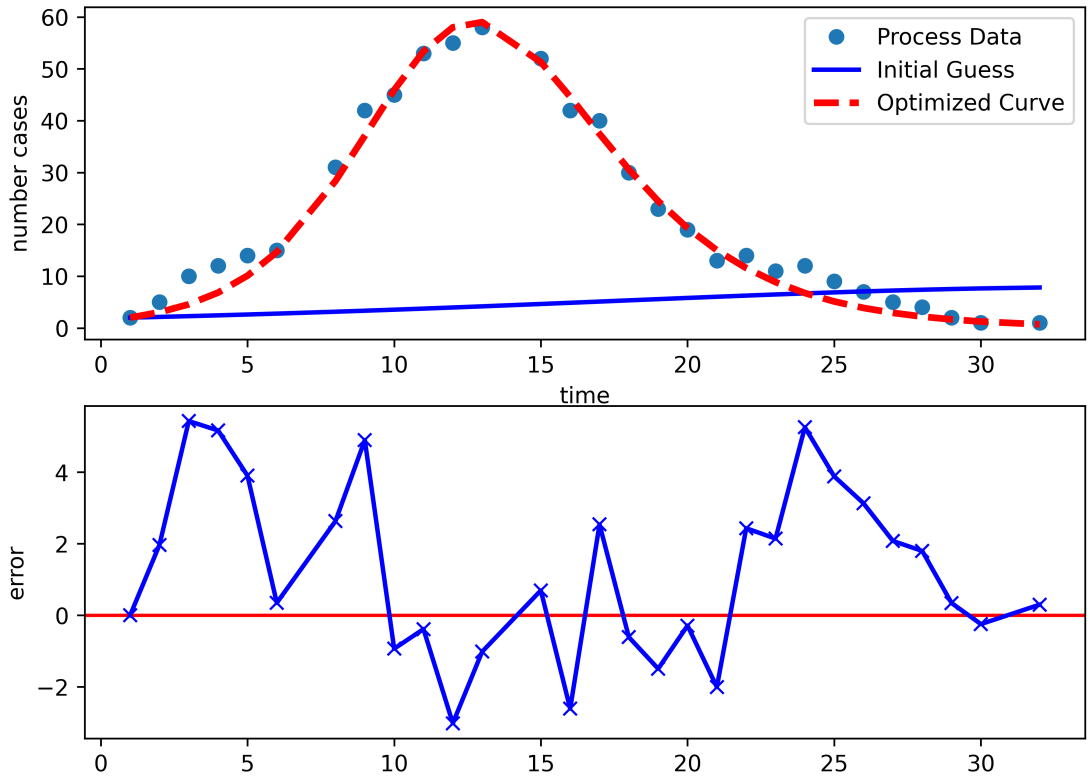


Figure 2.4.: Optimized plot and error for fitting influenza data to SIR model.

be interpreted as the chance that an infected individual infects a susceptible one. Our optimal solution comes very close to most data points, as we can see in the graph, indicating that the found values for α and β are plausible. In support of this, we also find that

$$I_{\max} = -\frac{\alpha}{\beta} + \frac{\alpha}{\beta} \ln \frac{\alpha}{\beta} + S_1 + I_1 - \frac{\alpha}{\beta} \ln S_1 = 59.1163,$$

which matches the data much better than our initial estimate $\hat{I}_{\max}$, indicating that the fitting is successful.

Next, we will look at possible extensions of the SIR model, leading to complex and (for some diseases) more realistic models.

3. More Complex Models

The study of growth and change of human populations is called *demography*.

So far, we have looked at models with a fixed population, meaning $N(t)$ is fixed and $N'(t) = 0$. We call these models *epidemic models without explicit demography*. These are best suited for *fast* diseases, like childhood diseases and influenza. Because of their short time span and the fact that children mainly congregate in closed populations like schools, there aren't any significant changes in the population to account for. Conversely, diseases like HIV and tuberculosis are known as *slow* diseases. These develop over a longer period (for an individual), rendering it careless to ignore changes within the population.

3.1. Modelling for changing populations (SIR Model)

To implement the demographic to the previously described SIR model, we need to introduce two new parameters. In addition to α and β, we introduce the birth rate Λ and relative death rate μ. The birth rate is the number of people born per time unit. The relative death rate is the proportion of the population that dies per time unit. The death rate is the same for the entire population. For example, for the susceptible compartment $S(t)$ the death rate becomes μS. This gives us the following ODEs,

$$\begin{aligned} S'(t) &= \Lambda - \beta IS - \mu S, \\ I'(t) &= \beta IS - \alpha I - \mu I, \\ R'(t) &= \alpha I - \mu R. \end{aligned}$$

This model has two assumptions as well. The first is the same as before, infected individuals are infectious. On the other hand, the population is no longer constant, so $N'(t) \neq 0$. More specifically,

$$N'(t) = \Lambda - \mu N,$$

where we still have that $N = S + I + R$. Although the population is not constant, it is asymptotically stable, meaning

$$t \to \infty \implies N(t) \to \frac{\Lambda}{\mu}.$$

For a non-constant population, the incidence is also known as called the *mass action incidence*, because it is can be deduced from the *law of mass action*[6]. So

Definition 3.1.1. Mass Action Incidence = βSI.

Alternatively, it is also common for epidemic models to use the *standard incidence*. This is a normalisation of the mass action incidence. So

Definition 3.1.2. Standard Incidence = $\frac{\beta SI}{N}$.

Making this distinction for a constant population is redundant because the standard incidence and the mass action incidence are the same in that case. The standard incidence is used for diseases with a contact rate that cannot increase indefinitely and is limited, regardless of how large the population becomes. Whereas the mass action incidence is used for diseases for which the contact rate is not limited for an increasing population.

3.2. Various Stages

We've seen that adding demography to a model alters the behaviour of the ODEs. Even so, our population still consists of three compartments. It is also possible to add or remove compartments to better incorporate real life features of the disease. We can split up the possible compartments in three categories: Disease transmission, Control Strategies and Pathogen Heterogeneity. We briefly comment on the interpretation of a few of such compartments:

3.2.1. Disease Transmission

- Exposed Stage: Contrary to assumptions made in the SIR model, there are many diseases where infected individuals don't immediately become infectious. The time between being a healthy susceptible and an infected individual is called the *exposed* stage. This discrepancy between being infected and infectious is not uncommon as it occurs when the pathogen in question needs significant time to establish itself in a new host. Let $E(t)$ be the exposed compartment. Introducing an exposed compartment to the SIR model gives us a model of the form,

$$S'(t) = \Lambda - \beta SI - \mu S, \tag{3.1}$$
$$E'(t) = \beta SI - (\nu + \mu)E, \tag{3.2}$$
$$I'(t) = \nu E - (\alpha + \mu)I, \tag{3.3}$$
$$R'(t) = \alpha I - \mu R. \tag{3.4}$$

  Here $\frac{1}{\nu}$ is the average length of the exposed stage, similar to how we defined α. See figure 3.1 for the corresponding flowchart.

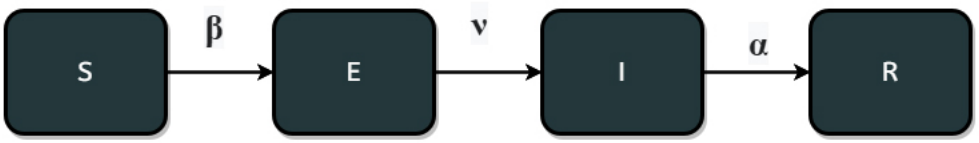


Figure 3.1.: Flowchart of the SEIR epidemic model, demographic rates are not included

- Asymptomatic Stage: *Asymptomatic* infections, also known as subclinical infections, are infections without symptoms. Because of this, asymptomatic infections are harder to detect. However, such infections usually are not exclusively asymptomatic, take for example influenza or COVID-19. An individual with an asymptomatic infection also contributes to the distribution of an infection, just like their classically infected counterparts. Let $A(t)$ be the asymptomatic compartment. Adding an asymptomatic compartment to the SEIR model defined above gives us a new model of the form,

$$\begin{aligned} S'(t) &= \Lambda - \beta S(I + qA) - \mu S, \\ E'(t) &= \beta S(I + qA) - (\nu + \mu)E, \\ I'(t) &= p\nu E - (\alpha + \mu)I, \\ A'(t) &= (1-p)\nu E - (\gamma + \mu)A, \\ R'(t) &= \alpha I + \gamma A - \mu R. \end{aligned}$$

  Here exposed individuals are sent to the infected compartment with probability p, and to the asymptomatic infected compartment with probability $(1-p)$. Furthermore, γ gives the recovery rate for the asymptomatic compartment, so $\frac{1}{\gamma}$ is the average number of days needed to recover from an asymptomatic infection. Usually, the asymptomatic infectious period is shorter than the symptomatic, so $\frac{1}{\gamma} < \frac{1}{\alpha}$. See figure 3.2 for the corresponding flowchart.

3.2.2. Control Strategies

Up to now, all possible extensions of an epidemic model have been related to the natural course of the infection in question. Nonetheless, actions taken to combat the distribution infection, or speed up the recovery rate by medication or hospitalisation, can also drastically alter a model. We call these additions to the model *control strategies*. We will look at the impact one of these measures can have on a model.

- Quarantine: The word *quarantine* comes from the Italian *quarantena*, which means 40-day period. Quarantine is a place, period, or state of isolation. This isolation can be beneficial to both infected and healthy individuals. A healthy person lowers the risk of encountering an infected one by quarantining and vice versa, hence

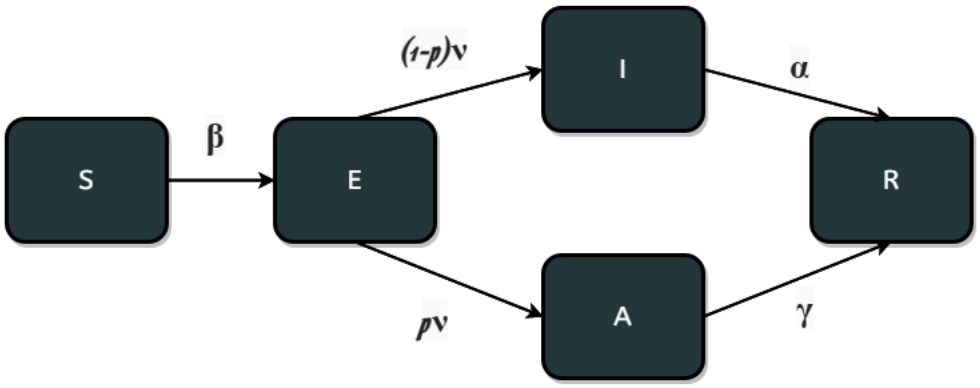


Figure 3.2.: Flowchart of the SEAIR epidemic model with asymptomatic compartment, demographic rates are not included

lowering the distribution rate of the infection. In practice we see that Quarantine is mostly enforced on susceptible individuals that have been in contact with an infected individual. Let $Q(t)$ be the compartment with the quarantined individuals. Before adding a quarantine compartment to the SIR model, it is important to notice that there is an active and an inactive population. Say a part of the population is inactive due to quarantine, then the remaining 'active' population $A(t)$ can be denoted as

$$A(t) = N(t) - Q(t) = S(t) + I(t) + R(t),$$

where $N(t)$ is the total population. Adding a quarantine compartment to the classical SIR model gives us a system of the following form. See figure 3.3 for the corresponding flowchart.

$$\begin{aligned} S'(t) &= \Lambda - \beta SI/A - \mu S, \\ I'(t) &= \beta SI/A - (\alpha + \gamma + \mu)I, \\ Q'(t) &= \gamma I - (\nu + \mu)Q, \\ R'(t) &= \alpha I + \nu Q - \mu R. \end{aligned}$$

Note how here we make use of the standard incidence instead of the mass action incidence.

3.3. Repeated or Omitted Stages

Another way to make a model fit the course of a disease better is by repeating or omitting certain stages. In the classical SIR model (without demography), all individuals end up

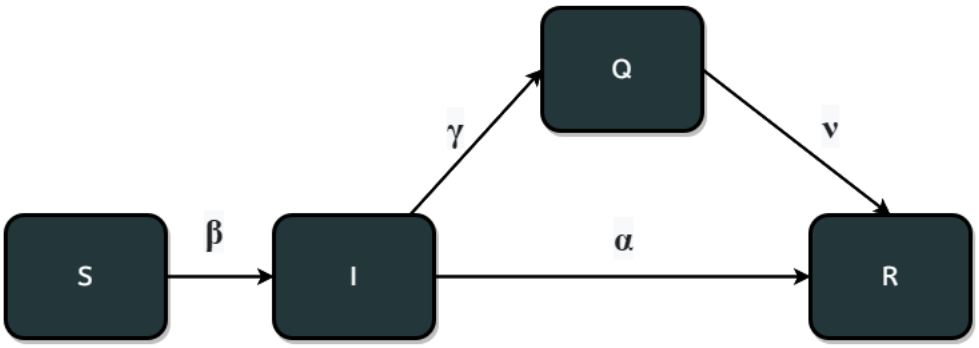


Figure 3.3.: Flowchart of an SIQR epidemic model, Demographic rates are not included

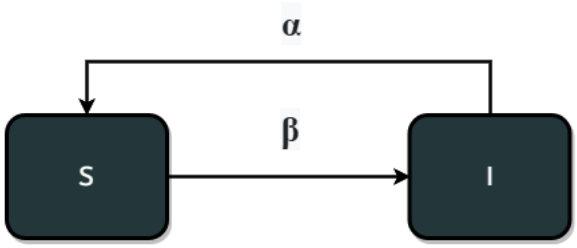


Figure 3.4.: Flowchart of an SIS epidemic model

in the recovered compartment, meaning eventually all individuals are either immune to or never contracted the disease. But not all infections exhibit behaviour that supports the existence of such a recovered compartment, or permanent immunity. An example of this is influenza. We've seen that an influenza outbreak can be fitted to the SIR model. However, a more realistic model is probably the SIS model, which implies that recovered individuals are instantly susceptible again. Such a model has the following form

$$S'(t) = -\beta SI + \alpha I,$$
$$I'(t) = \beta SI - \alpha I.$$

The assumptions for this model are the same as for the SIR model without demography.

Omission of permanent immunity can also mean something else, namely that there is no cure. In this scenario, infected individuals remain infected until they die. A model that exhibits this behaviour is the SI model, see figure 3.5.

Such a model might work for analysing the course of an STD like herpes or HIV, since

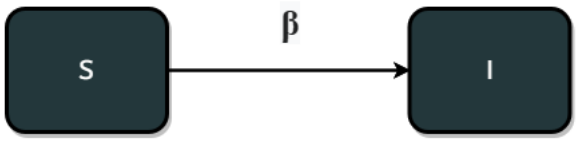


Figure 3.5.: Flowchart of the SI epidemic model.

these have no (known) cure. In the case of HIV however, a simple SI model won't fit data well since it doesn't sufficiently appreciate the complexities of the infected compartment. HIV has a long infectious stage that is subject to different stages in itself. Because all infectious stages affect the incidence of HIV, the force of infection can no longer be represented by a fixed value, as we will see in the following example.

3.4. Example: HIV in UK

Martcheva (2013) states:

> Human immunodeficiency virus (HIV) infection is a disease of the immune system caused by the HIV virus. HIV is transmitted primarily via unprotected sexual intercourse, contaminated blood transfusions, and from mother to child during pregnancy, delivery, or breastfeeding (vertical transmission). After entering the body, the virus causes acute infection, which often manifests itself with flulike symptoms. The acute infection is followed by a long asymptomatic period. As the illness progresses, it weakens the immune system more and more, making the infected individual much more likely to get other infections, called opportunistic infections, that are atypical for healthy individuals. There is no cure or vaccine against HIV; however, antiretroviral treatment can slow the course of the disease and may lead to a near-normal life expectancy.

Here we see a table that gives the number of people living in the UK with HIV.

We want to fit an epidemiological model to the data. To achieve this, we must first determine this model. As discussed before, a simple SI model won't work since this only accommodates one infected compartment. To incorporate the different stages, a common stochastic technique is to use Erlang's *method of stages*. The deterministic variant of this technique allows us to construct the infectious compartment as a series of k compartments, where the duration in each compartment is given by independent identically distributed variables. Isham provides details on this method[7].

We decide to divide our infected compartment into four sub compartments, each with exit rate γ. An individual in any of these sub compartments is infectious. Let $I(t)$ be

Year	Number cases	Year	Number cases	Year	Number cases
1990	21,000	1997	29,000	2004	61,000
1991	22,000	1998	31,000	2005	66,000
1992	23,000	1999	35,000	2006	72,000
1993	23,000	2000	39,000	2007	77,000
1994	24,000	2001	43,000	2008	82,000
1995	25,000	2002	51,000	2009	85,000
1996	26,000	2003	55,000	2010	91,000

Table 3.1.: Number of people living with HIV in the UK

the sum of all infectious sub compartments. Then,

$$\begin{aligned} I(t) &= I_1(t) + I_2(t) + I_3(t) + I_4(t), \\ N(t) &= S(t) + I(t) \end{aligned}$$

Furthermore, we take the force of infection $\lambda(t)$ as non-monotone and given by

$$\lambda(t) = \beta e^{-\alpha I(t)/N(t)} I(t)/N(t).$$

This force of infection can be interpreted as follows. As more individuals contract the virus, making $I(t)$ increase, the healthy susceptible individuals become more aware of the disease and are more cautious in their (sexual) contact, causing the force of infection to decline.

Altogether, our SIIII model amounts to

$$\begin{aligned} S'(t) &= \Lambda - \lambda(t)S(t) - \mu S(t), \\ I_1'(t) &= \lambda(t)S(t) - (\gamma + \mu)I_1(t), \\ I_2'(t) &= \gamma I_1(t) - (\gamma + \mu)I_2(t), \\ I_3'(t) &= \gamma I_2(t) - (\gamma + \mu)I_3(t), \\ I_4'(t) &= \gamma I_3(t) - (\gamma + \mu)I_4(t). \end{aligned}$$

Before we try to fit the data to this model, we must determine two things, what units to use and which parameters to pre-estimate, our initial guesses. The data is given in units of single people and the population of the UK runs well in the millions. A good middle ground is to use thousands as unit for the number of people. This way we don't need to work with numbers that are too large. Our unit of time is years in the data and that is fine, since we want to analyse the course of HIV over the years. We decide to fix the birth and relative death rate in the model, since these can easily be found, based on empirical data. The population in 1990 was 50.561 million ($N(0) = 50561$) and the relative death rate μ roughly 0.01117. If we take this as the equilibrium population, the asymptotic stability of $N(t)$ implies

$$N = \frac{\Lambda}{\mu},$$

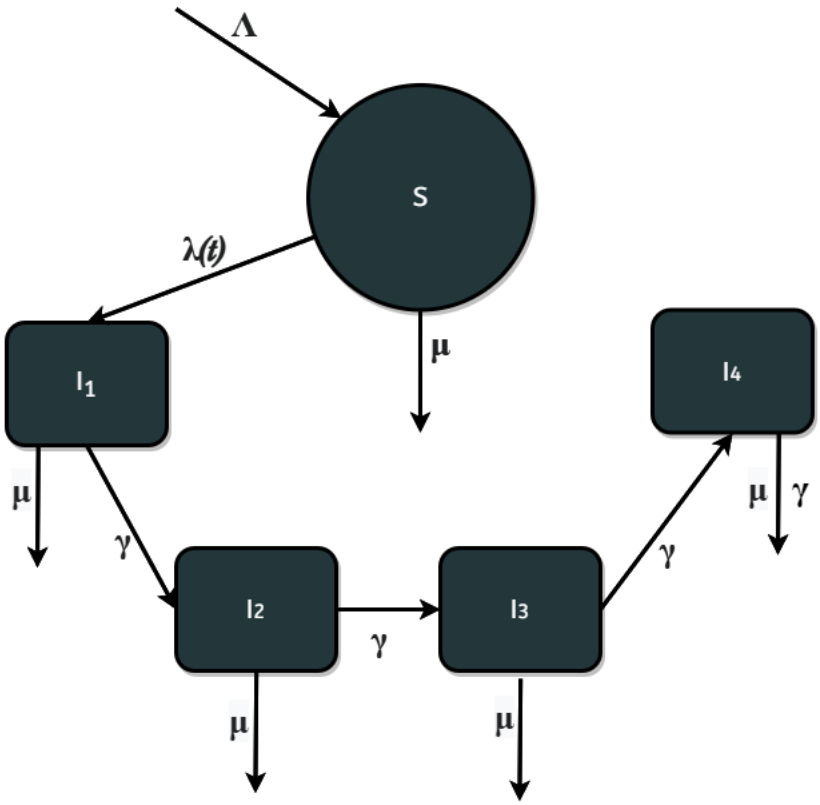


Figure 3.6.: Flowchart of the SIIII epidemic model for HIV

so the birth rate $\Lambda = 564.7663699999999$. We also assume that in 1990, at $t = 0$, all infected individuals were in compartment I_1 and all other infected compartments were empty. So, with the data, our initial conditions become $S_0 = 50540$ and $I_0 = I_1(0) = 21$. Now the only parameters left to fit are α, β and γ. Our initial guesses are

$$\begin{aligned}\hat{\alpha} &= .260492\\ \hat{\beta} &= 0.07\\ \hat{\gamma} &= 0.3399\end{aligned}$$

For the most part, the program used here to perform the fitting is largely identical to that in section 2.3. The main difference here is that we must use the sum of the infected sub compartments instead of simply using a single infected compartment to calculate the objective function.

3.4.1. The fitted model

After the fitting our optimised parameters are

$$\begin{aligned}\alpha &= 0.253992229703288\\ \beta &= 0.0007057834050035877\\ \gamma &= 0.10167018778154496,\end{aligned}$$

Our initial SSE was 25745.569470947547. After the fitting, the SSE became 729.160473442664, quite a bit smaller. See figure 3.7 for the plot of our fitted model along with the graph for residual errors. Parameters α and β are difficult to interpret since they are only a small part in the complicated force of infection function. However, we can interpret $1/\gamma$ as the average time spent in an infected compartment. Note that

$$\frac{1}{\gamma} \approx 9.84\text{years},$$

which means that the average time spent in an infected compartment is almost 10 years. This seems plausible.

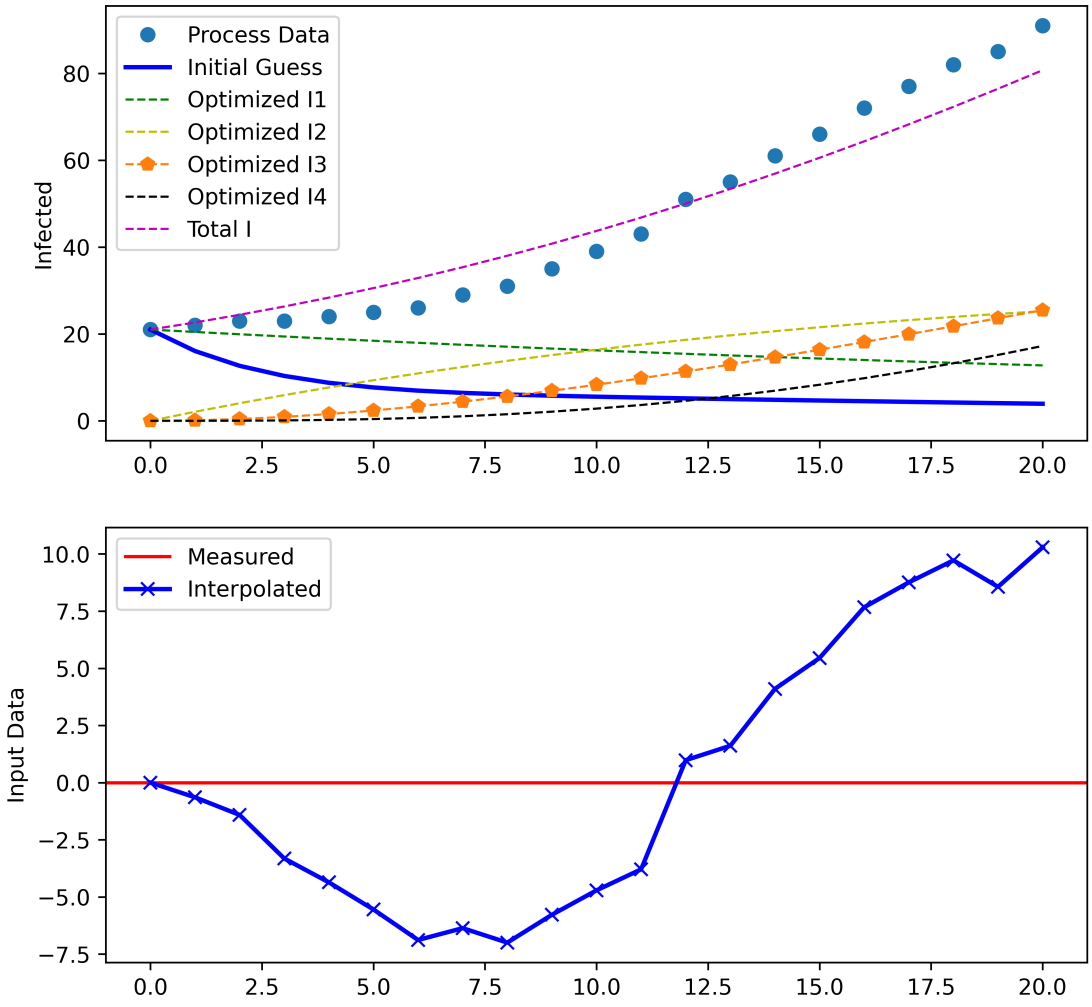


Figure 3.7.: Optimized plot and error for fitting HIV data to SIIII model. All infectious compartments

4. Model Selection

> *Model selection* is the task of selecting a mathematical model from a set of candidate models, given data.

Now we have seen how an epidemic can be modelled and solved mathematically. Once the model has been formed, simulations and analysis can be performed on it. It is however evident that multiple models can be created for the same disease, depending on how you view the biological scenario. An example of this is HIV. One can model HIV with a simple SI model, an SI model with vertical transmission or as we have seen above an SI model with multiple (in our case 4) infective stages. The question remains, which model best fits the disease at hand.

With model selection, we take the data as given and want to find the model that best describes the data. The researcher must determine a set of reasonable candidate models, reasonable for the scenario being studied. Once this selection is made the researcher can then perform an analysis to determine the best model. Previously, when estimating the best parameters for our model, we used the SSE to determine the optimal values. We could attempt arranging each model by their minimal SSE, in which case the model with the smallest SSE best fits the data. Now it is generally known that the more parameters your model has, the more accurate your model will be. However, these additional parameters need to mean something. The less parameters your model has, the easier it is to interpret each parameter. So, our selection criterion must be based on two things

- Goodness of fit
- Simplicity

Goodness of fit can be interpreted as the model with the smallest minimal SSE and simplicity means a minimal number of parameters. In other words, a good selection technique must choose the simplest model that best fits the data.

There are several statistical criteria that can be used to decide on the best model. We will look at the Akaike information criterion.

4.1. Akaike Information Criterion

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from, considering both the SSE and the number of parameters being fitted.

It is important to note that the AIC does not tell us whether the model is reasonable or how well it fits the data in an absolute sense. The AIC only determines which of the candidate models fits the data best. There will always be a best model, or a best combination of models, even if none of the models properly matches reality. This can be a result of analysing the course of the disease improperly or a mere lack of information due to the novelty of the disease in question.

Definition 4.1.1. The *Akaike information criterion* (AIC) is a measure of the relative goodness of fit of a mathematical model.

$$\mathrm{AIC} = n\left[\ln\left(\frac{\mathrm{SSE}}{n}\right)\right] + 2k,$$

where n is the number of data points in the data set, k is the number of parameters fitted plus one, and SSE is the least-squares error.

Say we have a few candidate models. The model with the best fit, is that with the smallest AIC. From the definition we can also again conclude that a having a lower SSE (least-squares error) or k (number of fitted parameters) makes a model more likely to be appropriate for the data, confirming that having more (fitted) parameters doesn't result in a more accurate model.

The rudiments of the AIC can be found in information theory. Hence, the AIC is only a measure for information loss and as such, is only valid when the number of fitted parameters is sufficiently lower than the number of data points. A rule of thumb commonly used to deduce the validity of the AIC is the following:

$$\frac{n}{K} > 40, \tag{4.1}$$

where n is the number of data points and K the number of parameters in the most complex model among the candidates. If this condition isn't met, then it is recommended to us a modified AIC instead, the AICc.

Definition 4.1.2. The *modified Akaike information criterion* (AICc) is a measure of the relative goodness of fit of a mathematical model.

$$\mathrm{AICc} = n\left[\ln\left(\frac{\mathrm{SSE}}{n}\right)\right] + 2k + \frac{2k(k+1)}{n-k+1},$$

where n is the number of data points in the data set, k is the number of parameters fitted plus one, and SSE is the least-squares error.

The AIC and AICc on their own aren't enough to determine the best model because they are an arbitrary scale[1]. We need two more definitions.

Definition 4.1.3. The distances of the AIC or AICc for any model to the model with minimal AIC or AICc is defined as

$$\Delta_j = \mathrm{AIC}_j - \mathrm{AIC}_{\min},$$

where AIC_j is the AIC or AICc of the jth model M_j, and $\mathrm{AIC}_{\min}$ is the AIC or AICc of the model with minimal AIC.

The Δ_J play an important role when it comes to deciding which of the candidate models have relative support in the data:

- $\Delta_j \leq 2$: Model j has substantial support in the data.
- $4 \leq \Delta_j \leq 7$: Model j has considerably less support in the data.
- $\Delta_j > 10$: Model j has no support in the data.

Definition 4.1.4. The Akaike weight of the ith model M_i is defined as follows:

$$w_i = \frac{e^{-\Delta_i/2}}{\sum_{j=1}^{J} e^{-\Delta_j/2}}.$$

The sum of the Akaike weights is equal to one:

$$\sum_{j=1}^{J} w_i = 1.$$

We want our best fitted model j to have an Akaike weight of $w_j > 0.9$, since then concrete conclusions can be made based on this single best fitted model. If none of the models meets this requirement, then we will need multiple models to make robust conclusions. Burnham and Anderson detail how single and multimodel inferences can be made[8].

4.2. Example: Comparing Models for Influenza

In our first example we fitted an influenza outbreak to the SIR model (M_1). To showcase the power of model selection, let's perform the same experiment on different models. We are going to fit the influenza outbreak from section 2.3 to the SEIR (M_2) and the SIQR (M_3) models. These are the systems of ODEs we are going to compare:

$$M_1 := \begin{cases} S'(t) &= -\beta IS \\ I'(t) &= \beta IS - \alpha I \\ R'(t) &= \alpha I, \end{cases} \tag{4.2}$$

$$M_2 := \begin{cases} S'(t) &= -\beta SI, \\ E'(t) &= \beta SI - \nu E, \\ I'(t) &= \nu E - \alpha I, \\ R'(t) &= \alpha I. \end{cases} \tag{4.3}$$

$$M_3 := \begin{cases} S'(t) &= -\beta SI/A, \\ I'(t) &= \beta SI/A - (\alpha + \gamma)I, \\ Q'(t) &= \gamma I - \nu Q, \\ R'(t) &= \alpha I + \nu Q \end{cases} \tag{4.4}$$

Model	Fitted parameters	SSE	AICc	Δ AICc	w_i
M_1	α, β	208.74	64.20	0	0.94
M_2	α, β, ν	332.91	80.44	16.24	$2.81 * 10^{-4}$
M_3	$\alpha, \beta, \nu, \gamma$	208.74	69.85	5.65	0.056

Table 4.1.: Model selection table

Our data set consists of 21 data points so requirement 4.1 is not met. We will use the AICc instead of the AIC to determine the Akaike weights. The specifics of each of the models can be found in the previous chapters. We fitted each of these models to the data with the same method as section 2.3. For each model our initial values were zero except for S_0 and I_0. The results of our comparison can be found in table 4.1. The plots for each model can be found in the appendix.

4.2.1. The models compared

If we look at the Akaike weights W_i in table 4.1 we clearly see that M_1 is the winner and fits the data best. However, there is something interesting going on with M_3 as well. Looking at the SSE for M_1 and M_3, we see that these values are identical. This seems strange since M_3 has twice the number of fitted parameters that M_1 has. Let's take a look at the values of these fitted parameters. For M_1 we find

$$\begin{aligned} \alpha_1 &= 0.5994695597295631, \\ \beta_1 &= 0.001766602951828876, \end{aligned}$$

identical to what we found in chapter 2. For M_3 we find

$$\begin{aligned} \alpha_3 &= 0.5994766372843041, \\ \beta_3 &= 1.02110677162648, \\ \gamma_3 &= 0, \\ \nu_3 &= 1.3159467866633427. \end{aligned}$$

The most notable takeaways from this fitting are how α_1 and α_3 lie very close to each other, suggesting that their mean time spent in the infected compartment is almost identical as well. Moreover, we have a zero-parameter, namely γ_3. There are several explanations for this. One is that individuals that are put in quarantine recover instantly. Another theory is that none of the infected individuals are sent to the quarantined compartment. In both cases the class is effectively negligible, insignificant. We might as well omit the class from the SIQR model. But then the resulting model looks identical to M_1. Hence, their SSEs are identical as well.

5. The basic reproduction number $\mathcal{R}_0$

> The *basic reproduction number* $\mathcal{R}_0$, sometimes referred to as just the *reproduction number*, is a measure for the transmission potential of a disease. It is the number of secondary cases one infectious individual produces in a population with exclusively susceptible individuals.

In the previous chapter, we've seen how best to approach model selection when faced with multiple models. Here the focus lies on finding the best model for a fixed data set. There are of course other comparisons possible when it comes to epidemiological modelling. Although the AIC is a measure for information loss, it gives rather little information on the specifics of the disease in question. A key property which the AIC overlooks, is how infectious a disease is. This information of course is crucial to fighting any (infectious) disease. This is where the *basic reproduction number* comes into play:

Notice how the definition of $\mathcal{R}_0$ is given in words and not in relation to properties of the model at hand. That is because $\mathcal{R}_0$ does not have a single mathematical definition that is applicable to all models. However, there are a few conventions the basic reproduction number usually satisfies. First, the reproduction number should be zero if there is no transmission. Secondly the reproduction number need to be directly correlated with the number of secondary infections. And finally, the reproduction number should only be non-negative if the parameters in the system are. The mathematical interpretation on the other hand, is less ambiguous. The reproduction number is a threshold value by which we can determine the dynamics of the system at $t = 0$. Say $I^* = \lim_{t\to\infty} I(t)$. If $\mathcal{R}_0 < 1$, a single infectious individual infects less than one person in their lifetime, a population of just susceptibles. Hence, the disease never turns into an epidemic because the number of infected individuals does not grow over time. On the other hand, when $\mathcal{R}_0 > 1$, a single infectious individual infects *more* than one person in a population of solely susceptible. Subsequently, the number of infected individuals grows over time and we can speak of an epidemic. Thus, $\mathcal{R}_0$ determines whether a disease is an epidemic or not.

For the SIS model in figure 3.4, we find that when $\mathcal{R}_0 > 1$, the disease does not die out and the number of infected individuals stabilises around a value K, such that

$$I^* = K.$$

In this case, we say the disease is *endemic* to the population and $I^* = K$ is the endemic equilibrium. Conversely, if $\mathcal{R}_0 < 1$, the number of infected individuals progressively declines, causing the disease to fade out of the population. In this case we have a *disease-free* equilibrium, because $I^* = 0$.

There are several ways to compute $\mathcal{R}_0$. We illustrate a few methods by computing $\mathcal{R}_0$ for the SIR and SEIR models.

5.1. $\mathcal{R}_0$ for SIR model

First, let's look at the SIR model 4.2. As mentioned before, the incidence, the number of freshly infected people per time unit, is βSI. We are focusing on how infectious the individual is, so $I = 1$. So, the number of secondary infected individuals becomes βS. Now if the entire population consists of susceptible individuals, this number becomes βN. Recall that $\frac{1}{\alpha}$ is the number of time units, usually days, years, or something in between, an individual is contagious. So in the lifespan of an individual, the individual infects

$$\mathcal{R}_0 = \frac{\beta N}{\alpha}.$$

Note how this looks very similar to inequality 2.2.

For our example at section 2.3, we found that our fitted coefficients are

$$\alpha = 0.5994695597295631, \beta = 0.001766602951828876,$$

and our population is $N = 578$. So the reproduction number for this model becomes

$$\mathcal{R}_0 = \frac{\beta N}{\alpha} = 1.70333.$$

Note that we take the entire population as consisting of susceptibles. We can conclude that there was indeed an epidemic at the school. Figure 2.3 supports this as well.

So, we see that for model 4.2 it is not too complicated to compute $\mathcal{R}_0$, because our model is simple. Models of higher dimensions require more complex computations of $\mathcal{R}_0$. One of these methods makes use of the Jacobian of a system as we will see in the following example.

5.2. $\mathcal{R}_0$ for SEIR model

Definition 5.2.1. The *Jacobian* or the *Jacobian matrix* of a system if ordinary differential equations is the matrix that consists of the partial derivatives of the differential equations.

For this example we consider model 3.1 and we want to compute $\mathcal{R}_0$ at the disease-free equilibrium $\mathcal{E}_0 = (S^*, E^*, I^*, R^*) = (\frac{\Lambda}{\mu}, 0, 0, 0)$. The Jacobian of this model becomes:

$$J = \begin{pmatrix} -\mu & 0 & -\beta S^* & 0 \\ 0 & -(\nu + \mu) & \beta S^* & 0 \\ 0 & \nu & -(\alpha + \mu) & 0 \\ 0 & 0 & \alpha & -\mu. \end{pmatrix}$$

To ensure that the found equilibrium is stable, we pose the condition that the eigenvalues of the Jacobian matrix have a negative real part. In a 2×2 matrix A this means:

- $\mathrm{Det}A > 0$.
- $\mathrm{Tr}A < 0$.

By first expanding $|J - \lambda|$ in terms of the first column, and then in terms of the last column, we find the following eigenvalues:

$$\lambda_1 = \lambda_2 = -\mu.$$

The final two eigenvalues can then be found by finding the eigenvalues for the following matrix,

$$J^- = \begin{pmatrix} -(\nu + \mu) & \beta S^* \\ \nu & -(\alpha + \mu) \end{pmatrix}.$$

Applying the conditions mentioned before, we want Tr $J^- < 0$ and Det $J^- > 0$. In particular, the second condition gets us $(\nu + \mu)(\alpha + \mu) - \nu\beta S^* > 0$. There are several ways to rewrite this inequality as $\mathcal{R}_0$. We choose to define $\mathcal{R}_0$ as follows

$$\mathcal{R}_0 = \frac{\nu\beta S^*}{(\nu + \mu)(\alpha + \mu)}.$$

This notation makes it easier to see the impact an infectious individual has on its number of secondary cases.

Since infectious individuals are the only ones that can spread the disease, we only need one term to describe the number of secondary cases caused by a single infectious individual. The incidence of the infectious class is $\beta S^* I$, so for a single infectious individual this becomes βS^* per unit of time. A freshly infected susceptible becomes exposed first, and the portion of exposed individual that move on to being infected is $\frac{\nu}{\nu+\mu}$. Note how this elegantly integrates the death rate of the exposed compartment into $\mathcal{R}_0$. Finally, an infected individual is infectious for $\frac{1}{\alpha+\mu}$ units of time. All together this gives us exactly the $\mathcal{R}_0$ that we found previously.

6. Conclusion

In this thesis we have seen an overview on the topic of mathematical epidemic modelling. First we introduced the reader to the simple Kermack-McKendrick SIR model. By analysing the behaviour of the differential equations, we concluded that there are two ways the infected compartment, and thus the disease among the population, progresses. To illustrate this we used data from an outbreak of influenza at an English boarding school. With the help of a Python program we could easily determine basic properties of this disease, like how long an individual is contagious.

We then looked at more complex models by introducing new compartments, to show that possibilities are endless when it comes to thinking up a model. We also saw a use for such a complex model by fitting an SIIII model to data on the prevalence of HIV in the UK. The study of complex models quickly raises the question of model selection, which model best fits the data. To this end we introduced the Akaike information criterion, a scale that measures the amount of information loss. Finally, we took a look at the infamous basic reproduction number $\mathcal{R}_0$ and how it should be interpreted. We saw that there are several choices for this number, but its interpretation in all cases is roughly the same; the number of secondary cases generated by a single infectious individual in a population of susceptible individuals. It is important to note however, that $\mathcal{R}_0$ only says something about the inception of the epidemic. As control strategies are applied and people behave differently, this value changes.

6.1. What we can say about COVID-19 (in the Netherlands)

Luckily for us, COVID-19 data is readily available in the Netherlands. Hence, we tried to run the data, specifically on the prevalence of the disease, through the models used in this thesis. Unfortunately, this didn't result in anything conclusive. Of course, this was to be expected. First, our models are vastly too mundane to capture the essence of the situation. For example, a proper COVID-19 model should at least contain a quarantine, a treatment, and an asymptomatic compartment. Our models contained at most one of these compartments. Furthermore, the data itself isn't accurate as well. At the start of the pandemic it was impossible to properly register all cases because we weren't aware of how prevalent it already was, and afterwards all flu-like symptoms were categorised as COVID-19. All this tells us that even if we can deduce some properties of COVID-19 from mathematical models, it should by no means be the leading factor when deciding how best to combat the pandemic. We simply don't know enough.

Bibliography

[1] M. Martcheva, *An Introduction to Mathematical Epidemiology.* Springer-Verlag New York Inc., 2013.

[2] W. O. Kermack and A. G. Mckendrick, "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*, vol. 115, no. 772, pp. 700–721, 1927.

[3] H. J. Rose, "The use of amantadine and influenza vaccine in a type A influenza epidemic in a boarding school," *Journal of the Royal College of General Practitioners*, vol. 30, no. 219, pp. 619–621, 1980.

[4] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B," *ACM Transactions on Mathematical Software*, vol. 23, no. 4, pp. 550–560, 1997.

[5] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[6] E. B. Wilson and J. Worcester, "The law of mass action in epidemiology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 31, no. 1, p. 24, 1945.

[7] V. Isham, "Stochastic models for epidemics with special reference to AIDS," *The Annals of Applied Probability*, pp. 1–27, 1993.

[8] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociological Methods and Research*, vol. 33, no. 2, pp. 261–304, 2004.

Populaire samenvatting

En toen was daar corona. Het is 2020 en wat begon als een onbekende griepachtige infectie in China, groeide uit tot het gesprek van de dag. Elke dag. De ene maatregel na de andere wordt ingevoerd om de situatie onder controle te krijgen, tot ongenoegen van de corona-criticus. Beseft de overheid wel wat de implicaties zijn van een jaar lang de cafés sluiten? Waarom wordt deze onbekende infectie zo angstig benaderd als de symptomen verschrikkelijk veel lijken op die van de griep? Wat is het nut van de gehele bevolking vaccineren als er maar een (klein) deel er langdurig onder lijdt? Wat is die almachtige "R" waar ik steeds over hoor?

Deze eindeloze lijst aan veelgestelde vragen maakt een ding uitdrukkelijk duidelijk: er is nog een heleboel onduidelijk rond corona. Epidemiologie tracht deze onzekerheid te verminderen. Ondanks dat dit veld zich al eeuwen aan het ontwikkelen is, presenteert de recente pandemie toch weer nieuwe uitdagingen. Door de immense hoeveelheden aan gegevens die dagelijks worden verzameld, hebben zowel politici als wetenschappers ontzettend veel informatie om mee te werken. Als wiskundige, was ik voornamelijk gefascineerd door het wiskundige modelleren van zo een pandemie, of in ons geval een epidiemie. En dit brengt ons bij mijn onderzoek.

In deze scriptie ga ik na hoe een simpel model kan worden opgebouwd, uitgebreid en 'gefit' kan worden aan de hand van data. Hierbij spelen differentiaalvergelijkingen een grote rol. We zullen aan de hand van voorbeelden over een griepuitbraak op een engelse kostschool en HIV in het Verenigd Koninkrijk zien hoe dit in de praktijk gaat. Ook gaan we na waarom de beruchte "R" zo een cruciale waarde is.

A. Python Code

A.1. Epidemic Models

A.1.1. SIR

```
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 23 13:24:57 2021

@author: Test
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.interpolate import interp1d
from scipy.optimize import minimize
from AICselect import AIC,AICc

#Import data
bestandsnaam=input('voer bestandsnaam in: ')+'.txt'
f = open(bestandsnaam,'r')
lines = f.readlines()
data=pd.read_csv(bestandsnaam)
tdata=data['t'].values#-data['t'].values[0]
idata=data['q'].values
i0=idata[0]
N=data['N'].values[0]
s0=N-idata[0]
#%%
#initial guesses (a,b) and conditions (z0)
a=0.416666667
b=0.000841081807
x0=np.array([a,b])
z0=np.array([s0,i0])
#%%
#time points
```

```
tlen=len(tdata) #number of steps
delta_t=tdata[1]-tdata[0]
#%%
#interpolate data
ii = interp1d(tdata,idata)
#%%
#Differential equations
def fopdt(z,t,a,b):
    S,I=z
    dSdt=-b*S*I
    dIdt=b*S*I-a*I
    dzdt=np.array([dSdt,dIdt])
    return dzdt
#%%
#Simulate model
def sim_model(x):
    a=x[0]
    b=x[1]
    zm=np.zeros((tlen,2))
    zm[0]=z0
    for i in range(0,tlen-1):
        ts=[tdata[i],tdata[i+1]]
        y1=odeint(fopdt,zm[i],ts,args=(a,b))
        zm[i+1]=y1[-1]
    return zm
#%%
#define objective
def objective(x):
    zm=sim_model(x)
    obj=.0
    for i in range(len(zm)):
        obj += (zm[i,1]-idata[i])**2
    return obj
#%%
#show initial objective
print('Initial SSE Objective: ' + str(objective(x0)))

#optimize a and b
bnds = ((0.0,None),(0.0,None))
sol = minimize(objective,x0,method='L-BFGS-B',bounds=bnds)
x=sol.x
aic = AIC(tlen,len(x),objective(x))
aicc = AICc(tlen,len(x),objective(x))
SSE=objective(x)
```

```
# show final objective
print('Final SSE Objective: ' + str(SSE))
print('AIC : ' + str (aic))
print('AICc : ' + str (aicc))

print('a: ' + str(x[0]))
print('b: ' + str(x[1]))
#%%
# calculate model with updated parameters
z1 = sim_model(x0)
z2 = sim_model(x)
# plot results
plt.figure(figsize=(8, 6),dpi=600)
plt.subplot(2,1,1)
plt.plot(tdata,ii(tdata),'o',linewidth=2,label='Process Data')
plt.plot(tdata,z1[:,1],'b-',linewidth=2,label='Initial Guess')
plt.plot(tdata,z2[:,1],'r--',linewidth=3,label='Optimized Curve')
plt.ylabel('number cases')
plt.xlabel('time')
plt.legend(loc='best')
#%%
plt.subplot(2,1,2)
plt.axhline(y = 0, color = 'r', linestyle = '-')
plt.plot(tdata,idata-z2[:,1],'bx-',linewidth=2)
plt.ylabel('error')
plt.savefig('OPTIMIZATION + ERROR')
```

A.1.2. SEIR

```
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 23 13:24:57 2021

@author: Test
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.interpolate import interp1d
from scipy.optimize import minimize
from AICselect import AIC,AICc
```

```
#Import data
bestandsnaam=input('voer bestandsnaam in: ')+'.txt'
f = open(bestandsnaam,'r')
lines = f.readlines()
data=pd.read_csv(bestandsnaam)
tdata=data['t'].values#-data['t'].values[0]
idata=data['q'].values
i0=idata[0]
N=data['N'].values[0]
e0=data['E'].values[0]
s0=N-i0-e0
#%%
#initial guesses (a,b) and conditions (z0)
a=0.416666667
b=0.000841081807
n=0.1
x0=np.array([a,b,n])
z0=np.array([s0,e0,i0,0])
#%%
#time points
tlen=len(tdata) #number of steps
delta_t=tdata[1]-tdata[0]
#%%
#interpolate data
ii = interp1d(tdata,idata)
#%%
#Differential equations
def fopdt(z,t,a,b,n):
    S,E,I,R=z
    dSdt=-b*S*I
    dEdt=b*S*I-(n)*E
    dIdt=n*E-(a)*I
    dRdt=a*I
    dzdt=np.array([dSdt,dEdt,dIdt,dRdt])
    return dzdt
#%%
#Simulate model
def sim_model(x):
    a=x[0]
    b=x[1]
    n=x[2]
    zm=np.zeros((tlen,4))
    zm[0]=z0
    for i in range(0,tlen-1):
```

```
        ts=[tdata[i],tdata[i+1]]
        y1=odeint(fopdt,zm[i],ts,args=(a,b,n))
        zm[i+1]=y1[-1]
    return zm
#%%
#define objective
def objective(x):
    zm=sim_model(x)
    obj=.0
    for i in range(len(zm)):
        obj += (zm[i,2]-idata[i])**2
    return obj
#%%
#show initial objective
print('Initial SSE Objective: ' + str(objective(x0)))

#optimize a and b
bnds = ((0.0,0.9),(0.0,1.0),(0.0,1.0))
sol = minimize(objective,x0,bounds=bnds)
x=sol.x
aic = AIC(tlen,len(x),objective(x))
aicc = AICc(tlen,len(x),objective(x))
SSE=objective(x)
# show final objective
print('Final SSE Objective: ' + str(SSE))
print('AIC : ' + str (aic))
print('AICc : ' + str (aicc))
print()


print('a: ' + str(x[0]))
print('b: ' + str(x[1]))
print('n: ' + str(x[2]))
#%%
# calculate model with updated parameters
z1 = sim_model(x0)
z2 = sim_model(x)
# plot results
plt.figure()
plt.figure(figsize=(12, 8),dpi=600)
plt.subplot(2,1,1)
plt.plot(tdata,ii(tdata),'o',linewidth=2,label='Process Data')
plt.plot(tdata,z1[:,2],'b-',linewidth=2,label='Initial Guess')
plt.plot(tdata,z2[:,2],'m--',linewidth=3,label='Optimized Infected curve')
```

```
plt.plot(tdata,z2[:,1],'r--',linewidth=3,label='Optimized Exposed curve')
plt.plot(tdata,z2[:,3],'g--',linewidth=3,label='Optimized Recovered curve')
plt.ylabel('Number cases')
plt.xlabel('time')
plt.legend(loc='best')
plt.ylabel('Output')
plt.legend(loc='best')
plt.savefig('OPTIMIZATION',dpi=800)
#%%
plt.subplot(2,1,2)
plt.axhline(y = 0, color = 'r', linestyle = '-')
plt.ylim([-8, 8])
plt.plot(tdata,idata-z2[:,2],'bx-',linewidth=2)
plt.ylabel('error')
plt.xlabel('time')
plt.savefig('OPTIMIZATION + ERROR')
```

A.1.3. SIQR

```
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 23 13:24:57 2021

@author: Test
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.interpolate import interp1d
from scipy.optimize import minimize
from AICselect import AIC,AICc

#Import data
bestandsnaam=input('voer bestandsnaam in: ')+'.txt'
f = open(bestandsnaam,'r')
lines = f.readlines()
data=pd.read_csv(bestandsnaam)
tdata=data['t'].values#-data['t'].values[0]
idata=data['q'].values
i0=idata[0]
q0=0
r0=0
```

```
N=data['N'].values[0]
s0=N-i0-q0
#%%
#initial guesses (a,b) and conditions (z0)
a = 0.416666667
b = 0.000841081807
n = 1
g = 0.7
x0 =np.array([a,b,g,n])
z0=np.array([s0,i0,q0,r0])
#%%
#time points
tlen=len(tdata) #number of steps
delta_t=tdata[1]-tdata[0]
#%%
#interpolate data
ii = interp1d(tdata,idata)
#%%
#Differential equations
def fopdt(z,t,a,b,g,n):
    S,I,Q,R=z

    A = S+I+R

    dSdt=-b*S*I/A
    dIdt=b*S*I/A-(a+g)*I
    dQdt=g*I-n*Q
    dRdt=a*I+n*Q
    dzdt=np.array([dSdt,dIdt,dQdt,dRdt])
    return dzdt
#%%
#Simulate model
def sim_model(x):
    a=x[0]
    b=x[1]
    g=x[2]
    n=x[3]
    zm=np.zeros((tlen,4))
    zm[0]=z0
    for i in range(0,tlen-1):
        ts=[tdata[i],tdata[i+1]]
        y1=odeint(fopdt,zm[i],ts,args=(a,b,g,n))
        zm[i+1]=y1[-1]
    return zm
```

```
#%%
#define objective
def objective(x):
    zm=sim_model(x)
    obj=.0
    for i in range(len(zm)):
        obj += (zm[i,1]-idata[i])**2
    return obj
#%%
#show initial objective
print('Initial SSE Objective: ' + str(objective(x0)))

#optimize a and b
bnds = ((0.0,None),(0.0,None),(0.0,None),(0.0,None))
# bnds=None
sol = minimize(objective,x0,method='L-BFGS-B',bounds=bnds)
x=sol.x
aic = AIC(tlen,len(x),objective(x))
aicc = AICc(tlen,len(x),objective(x))
SSE=objective(x)
# show final objective
print('Final SSE Objective: ' + str(SSE))
print('AIC : ' + str (aic))
print('AICc : ' + str (aicc))
print('a: ' + str(x[0]))
print('b: ' + str(x[1]))
print('g: ' + str(x[2]))
print('n: ' + str(x[3]))
#%%
# calculate model with updated parameters
z1 = sim_model(x0)
z2 = sim_model(x)
# plot results
plt.figure()
plt.figure(figsize=(12, 8),dpi=600)
plt.subplot(2,1,1)
plt.plot(tdata,ii(tdata),'o',linewidth=2,label='Process Data')
plt.plot(tdata,z1[:,1],'b-',linewidth=2,label='Initial Guess')
plt.plot(tdata,z2[:,1],'m--',linewidth=3,label='Optimized Infected curve')
plt.plot(tdata,z2[:,2],'r--',linewidth=3,label='Optimized Quarantine curve')
plt.plot(tdata,z2[:,3],'g--',linewidth=3,label='Optimized Recovered curve')
plt.ylabel('Number cases')
plt.xlabel('time')
plt.legend(loc='best')
```

```
plt.legend(loc='best')
plt.savefig('OPTIMIZATION',dpi=800)
#%%
plt.subplot(2,1,2)
plt.axhline(y = 0, color = 'r', linestyle = '-')
plt.ylim([-8, 8])
plt.plot(tdata,idata-z2[:,1],'bx-',linewidth=2)
plt.ylabel('error')
plt.xlabel('time')
plt.savefig('OPTIMIZATION + ERROR')
```

A.1.4. SIIII

```
# -*- coding: utf-8 -*-
"""
Created on Tue Feb 23 13:24:57 2021

@author: Test
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.integrate import odeint
from scipy.interpolate import interp1d
from scipy.optimize import minimize

#Import data
bestandsnaam=input('voer bestandsnaam in: ')+'.txt'
f = open(bestandsnaam,'r')
lines = f.readlines()
data=pd.read_csv(bestandsnaam)
tdata=data['t'].values#-data['t'].values[0]
idata=data['q'].values
i0=idata[0]
N=data['N'].values[0]
s0=N-idata[0]
#%%
#initial guesses (a,b,g) and conditions (z0)
a=.260492
b=0.07
g=0.3399
u=0.01117
A= N*u
```

```
x0=np.array([a,b,g,A,u])
z0=np.array([s0,i0,0,0,0])
#%%
#time points
n=len(tdata) #number of steps
#%%
#interpolate data
ii = interp1d(tdata,idata)
#%%
#Differential equations
def fopdt(z,t,a,b,g,A,u):
    S,I1,I2,I3,I4=z

    I=I1+I2+I3+I4
    N = S + I

    force=b*np.exp(-a*I/N)*I/N

    dSdt=A-force*S-u*S
    dI1dt=force*S-(g+u)*I1
    dI2dt=g*I1-(g+u)*I2
    dI3dt=g*I2-(g+u)*I3
    dI4dtdt=g*I3-(g+u)*I4
    dzdt=np.array([dSdt,dI1dt,dI2dt,dI3dt,dI4dtdt])
    return dzdt
#%%
#Simulate model
def sim_model(x):
    a=x[0]
    b=x[1]
    g=x[2]
    A=x[3]
    u=x[4]
    zm=np.zeros((n,5))
    zm[0]=z0
    for i in range(0,n-1):
        ts=[tdata[i],tdata[i+1]]
        y1=odeint(fopdt,zm[i],ts,args=(a,b,g,A,u))
        zm[i+1]=y1[-1]
    return zm
#%%
#define objective
def objective(x):
    zm=sim_model(x)
```

```
    obj=0.0
    for i in range(len(zm)):
        summ = 0
        for j in range(4):
            summ += zm[i,j+1]
        obj += (summ-idata[i])**2
    return obj
#%%
#show initial objective
print('Initial SSE Objective: ' + str(objective(x0)))

#optimize a and b
sol = minimize(objective,x0)
x=sol.x

# show final objective
print('Final SSE Objective: ' + str(objective(x)))

print('a: ' + str(x[0]))
print('b: ' + str(x[1]))
print('g: ' + str(x[2]))
#%%
# calculate model with updated parameters
z1 = sim_model(x0)
z2 = sim_model(x)
zsum = [0 for i in range(len(z2))]
for i in range(len(z2)):
        zsum[i] += sum(z2[i][1:])

# plot results
plt.figure(figsize=(8, 8),dpi=600)
plt.subplot(2,1,1)
plt.plot(tdata,idata,'o',linewidth=2,label='Process Data')
plt.plot(tdata,z1[:,1],'b-',linewidth=2,label='Initial Guess')
plt.plot(tdata,z2[:,1],'g--',linewidth=1,label='Optimized I1')
plt.plot(tdata,z2[:,2],'y--',linewidth=1,label='Optimized I2')
plt.plot(tdata,z2[:,3],'p--',linewidth=1,label='Optimized I3')
plt.plot(tdata,z2[:,4],'k--',linewidth=1,label='Optimized I4')
plt.plot(tdata,zsum,'m--',linewidth=1,label='Optimized I')
plt.ylabel('Number cases')
plt.xlabel('time')
plt.legend(loc='best')
plt.savefig('OPTIMIZATION')
# %%
```

```
plt.subplot(2,1,2)
plt.axhline(y = 0, color = 'r', linestyle = '-')
plt.plot(tdata,idata-zsum,'bx-',linewidth=2)
plt.ylabel('error')
plt.show()
```

B. Akaike Information Criterion

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jun  9 18:49:23 2021

@author: Test
"""
import numpy as np

# prompt1 = input('Do u want to calculate (1) AIC or (2) AICc? : ')

# choice = 'AIC' if prompt1 == 1 else 'AICc'

# prompt2 = input('Do u want to calculate the '+choice+' for a single model (

# single = 1 if prompt2== 1 else 0

# n = input('Input number of data points in data set: ')
# k = input('Input number of paramaters: ')
# SSE = input('Input the least-squares error: ')
# k=para+1

# values = [n,k,SSE]

def AIC(n,k,SSE):
    k+=1
    return n * (np.log(SSE/n))+2*k

def AICc(n,k,SSE):
    k+=1
    return n * (np.log(SSE/n)) + 2*k + 2*k*(k+1)/(n-k-1)

# ============================================================================
# def AW ():
#     J = input('How many models do you have: ')
#     sum = 0
#     for i in range (J):
# ============================================================================
```

C. Additional Plots

C.1. SEIR

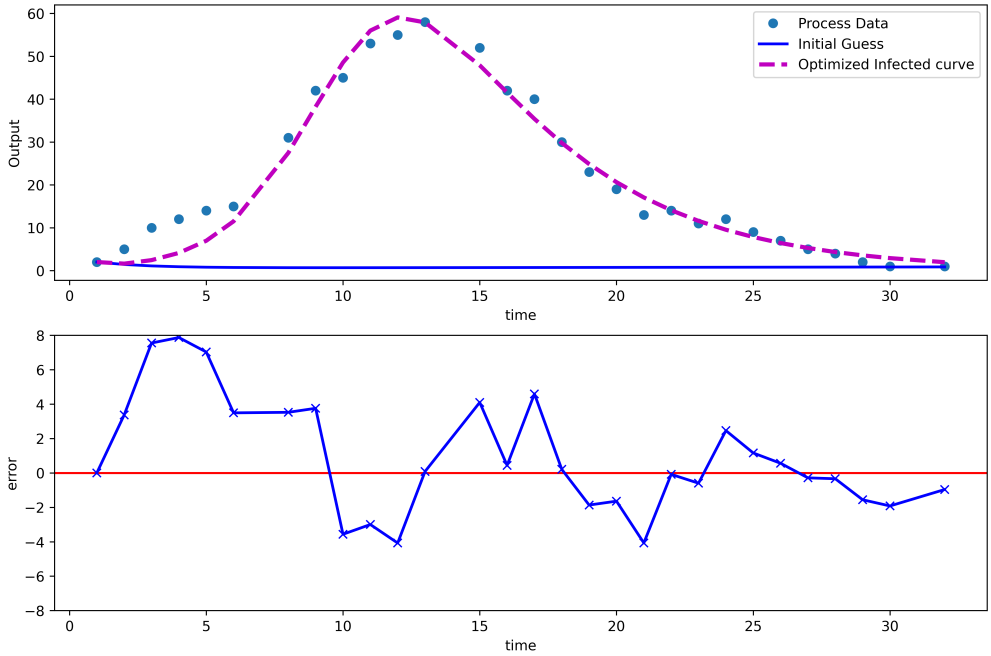


Figure C.1.: Optimized plot and error for fitting influenza data to SEIR model.

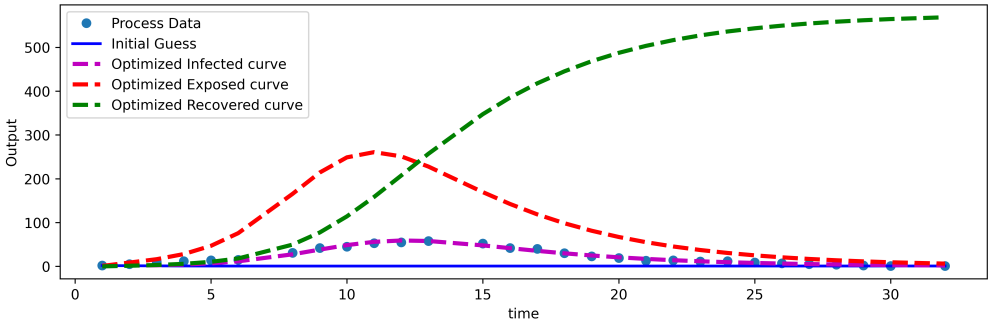


Figure C.2.: Optimized plot for fitting influenza data to SEIR model. All compartments

C.2. SIQR

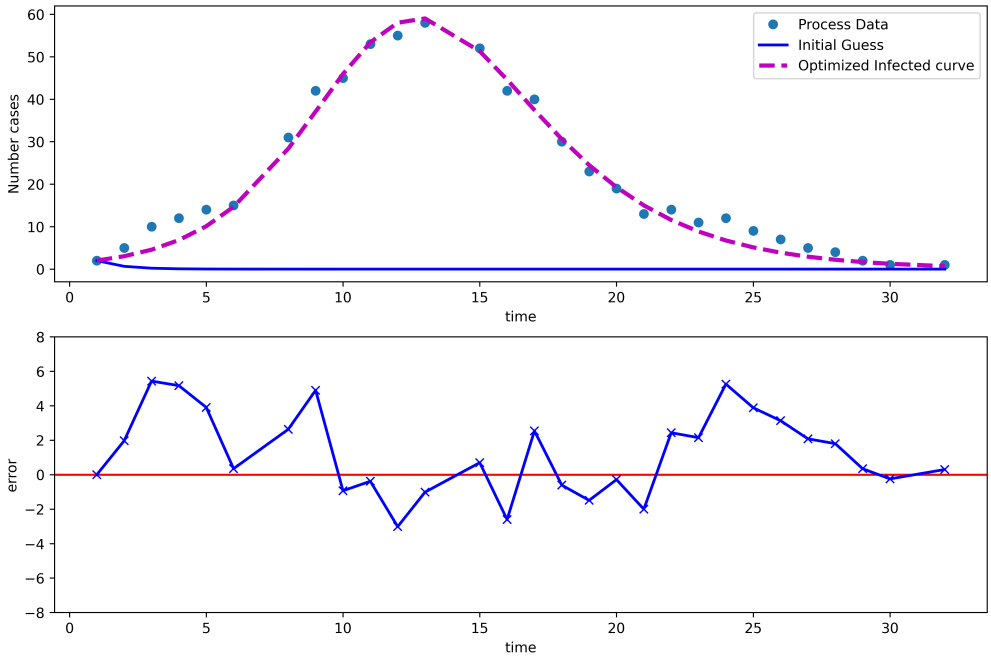


Figure C.3.: Optimized plot and error for fitting influenza data to SIQR model.

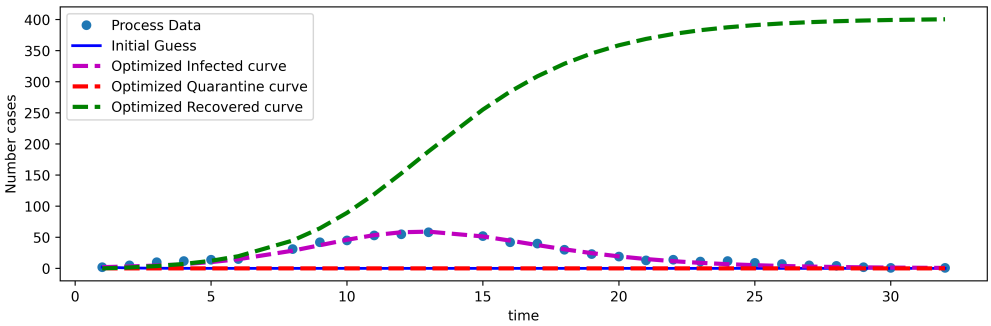


Figure C.4.: Optimized plot for fitting influenza data to SIQR model. All compartments