Health Metrics and the Spread of Infectious Diseases

Machine Learning Applications and Spatial Modeling Analysis with R

Federica Gazzelloni

12/16/22

Table of contents

Preface

Health metrics and the spread of infectious diseases, with machine learning applications and spatial model analysis are the topics of this book.

Here you will find everything you need to analyze the state of health of a country and compare it with that of other countries. You will also be able to evaluate the best model for predicting future trends.

The author of this book is **Federica Gazzelloni** actuary and a statistician by education and by training. She is also a collaborator of the **Institute for Health Metrics and Evaluation (IHME)** , which inspired this work to serve as a guideline for making health metrics and spatial model analysis for evaluating the state of health of a population and spread of infectious diseases.

All data used in this book are from the **Institute for Health Metrics and Evaluation (IHME)**. **GBD Results**. Seattle, WA: IHME, University of Washington, 2020. Available from https://vizhub.healthdata.org/gbd-results/. (Accessed January 2023) and the **World Health Organization (WHO)**. **Global Health Observatory data repository**. Available from https://apps.who.int/gho/data/.

Introduction

> - What this book is all about
> - How it can be used
> - What is the main take away

Health Metrics and the Spread of Infectious Diseases, featuring machine learning applications and spatial model analysis, serves as a manual and textbook for introductory health data analysis courses. Additionally, it can be a valuable source code for practitioners and data scientists alike.

This book provides a set of tools for analyzing data of various types, but it is specifically designed for health data, such as the number of infections in a population or identifying the health status of a country.

There are techniques that will be evaluated as the most appropriate for a certain type of analysis, while others, on the contrary, will be deprecated. It will be interesting to see which one is the most valuable for making predictions, while another should just be used for evaluating data consistency.

Public health metrics, such as **Years of Life Lost (YLL)** and **Years lived with Disability (YLD)**, are examples of the key metrics discussed and used throughout this book. They are expressed in number of years of life lost or years lived with disabilities, and their sum represents a crucial value named **Disability Adjusted Life Years (DALYs)**. DALYs are generally used for ranking the health status of a population. The book covers the history of the development of health metrics and suggests alternatives, providing insights for health policymakers.

To be more specific, the book compares the metrics used to summarize the health status of a population across different locations. It also tests prediction levels using key models. The initial tools are `{tidymodels}` and `{INLA}` for modeling, but other machine learning packages like `{mlr3}` and `{caret}` are also tested.

In practice, the data will include information about humanity such as age, sex, life expectancy, mortality, and risk levels. The interesting part is related to the identification of the influence of some of the most dangerous infectious disease on the values of health metrics of a population, and this is done to practice the variation to eventually predict through model transfer techniques same pattern procedure on other countries. It's not the only application; more examples of model transfer applications are needed in research. (cit)

A focus on the impact of recent infectious disease outbreaks, such as SARS-Covid19, on the state of health of the population, will be provided along with the most affected locations to compare results of both *deterministic* and *stochastic* (Bayesian) models. Risk factor analysis, also cover an important part of this book and aims at identifying the connection that would lead to an increase on the number of DALYs for specific population and look at providing suggestion for public health policy and practice (Vos et al., 2020).

The book is structured with an alternation of text and chunks of code, primarily in the R programming language; hints for translations in python is in appendix C. This is done to let the reader be a practitioner of real-world case studies on the topic.

The material supports full exploratory data analysis and model data visualization, it contains the code for making some interesting spatial visualization, made using `{ggplot2}`, `{leaflet}`, `{sf}`, `{rgdal}` R packages, plus other main packages for theme user customization. The main reason is to unlock the potentiality of the R language for a wider understanding of both spatial and health metrics.

The book is foreseen for practitioners at early stages and graduated students in STEM, but it will be useful for experts in the field who would love to have all the tools in one place to scan through as needed.

The purpose of this book is to contribute to the scientific development of the field of metrics and evaluation (Murray & Frenk, 2008). It is divided in four main sections, each containing three chapters. The first section **Metrics and Evaluation** is introductory explaining the health metrics definition at first level, the second section **Modeling** goes a bit more inside the topic looking at the tools available, the third section **Data Visualization** allows the user to be able to visualize the results of the analysis, and finally the fourth section makes the things into practice providing **Case Studies**.

In particular, in the `first section`, the first chapter **Health Metrics** dives into the different type of metrics, YLLs, YLDs, and DALYs, along with their history of development, provides a definition of the metrics and their usage. In this section there is also a mention of a fourth metric, the **HALY** or the **Health Adjusted Life Years**, this is a further improvement made in the development of the metrics and evaluation sector.

The second chapter **Metrics Components** looks at the key building blocks of the health metrics as important set of components for assuring consistency and to make sure that the values released are able to picture the real health status of a population. The key components are *life tables*, *life expectancy*, the *mortality level*, and the *weights of disabilities.* This chapter is supported by the **appendix A** where the beyond of the calculations are further explained.

The third chapter evaluates the **Causes and Risks** that are involved with the possibility of the metrics increase in values, even in terms of identifying the most dangerous enemies acting in favor of it. Risks are also considered in order to stimulate alternative solutions for prevention and health policy development.

The `second section` of the book is the the second level dive into the metrics evaluation, it is **Modeling**. A set of tools available for analyzing trends, or phenomenon to catch missing values and attempt filling the missing data to provide an overall picture of the general trend. In the first chapter **Techniques** there is an overview of the different types of models that would be suitable to use, the second chapter **Packages** selects some of the R-packages to use for modeling. Then chapter three shows how to make **Predictions**.

The `third section` of the book is all focused on **Data Visualization**, the first chapter shows the **Application**, the second chapter applies **Spatial Visualization** into modeling results. Then, the third chapter further allows the reader to be able to improve the knowledge acquired with some tailored examples.

The `fourth section` of the book, **Case Studies** presents two main case studies, one is *Covid19* and the other one is the case of *Malaria*. To eventually summarize the variation of the health metrics when these two infectious diseases affect the population.

In **Conclusion**, the technique of modeling transfer learning shows as the foundation of model application how to make prediction and to improve public policies.

Part I

Metrics and Evaluation

Overview

Health metrics are key variables to understand more about the state of health of a population. In this book we'll talk about how to calculate the **numbers of years of life lost (YLLs)** and the **numbers of years lived with disabilities (YLDs)**, to finally obtain the key metric of the **DALYs** which identify the numbers of years of life lost due to death or a disability status, namely **disability adjusted life years**.

The numbers of years of life lost by a population in comparison to other countries or to the Global mean trend, is based on the latest study results relative to how the well being of a country is in terms of the definition of a healthy life. To give an example, let's think about a population whose individuals are living a good life, so defined *healthy life* measured on **life expectancy** established to be 80 years on average, as most of the World population meets this as a deadline.

The part of the population who do not meet this age, but dies earlier, contributes as a building block of the numbers of years of life lost (YLLs), as well as for all that is related with a healthy living, the numbers of years spent dealing with a disability contribute to increasing the numbers of years lived with disabilities for a country's population.

To establish a *healthy life status* for a country, meaning the state of health of a population, the sum of the two values YLLs and YLDs is used to obtain the key metric of DALYs. This metric value is used to quickly identify the level of health of a population compared to a Global review based on the latest findings of the most updated studies.

In addition, this level is used to improve the proportion of countries who are in need of a better health status recognition. To be more specific, the focus on the numbers of years can help identify the areas where most of the years are lost and need for improvement, whether in facilities, research or investments.

An improvement of health at a Global level is reached when the definition of a **healthy file** is met in most of the countries where it wasn't before.

Furthermore, what we want to analyze are a series of data that are produced taking into account the tables of mortality and future life expectancy these are defined as means that these metrics have been considered important for assessing the state of health of one point population

What we refer to are the health metrics and which refer to the number of years lost due to an increase in mortality or in any case to a mortality trend that is above a certain general level which we can consider as the level optimal health globally.

1 Health Metrics

Learning Objectives:

- A closer look at the metrics
- Understand how to use them
- Think about potentiality for improvements

1.1 YLLs, YLDs and DALYs

In this section are shown the methods used for building three key metrics: YLLs, YLDs and DALYs. These will be used throughout the book for making comparisons among the state of health of different countries.

The health metrics, and their components are used to measure the burden of disease and quantify the impact of diseases and injuries on individuals and populations. These metrics can help prioritize public health interventions and evaluate the effectiveness of public health programs.

1.2 YLL (Years of Life Lost)

YLL (Years of Life Lost) measures the number of years a person would have lived if they had not died prematurely due to a disease or injury. YLL is calculated by subtracting the age at death from the expected age at death in a population without the disease or injury.

1.3 YLD (Years Lived with Disability)

YLD (Years Lived with Disability) measures the number of years a person lives with a disability due to a disease or injury. It is calculated by multiplying the prevalence of a condition by the disability weight, which reflects the severity of the disability.

1.4 DALY (Disability-Adjusted Life Year)

DALY (Disability-Adjusted Life Year) is a measure of overall disease burden and is calculated as the sum of years of potential life lost due to premature death (YLL) and years lived with disability (YLD). The DALY takes into account both the quantity and quality of life lost due to disease or injury.

The YLL and YLD are components of the DALY, which are used to assess how diseases and injuries impact populations. As a result, @sec-ch12Components section provides a more comprehensive picture of the overall burden of disease by combining YLLs and YLDs in different life expectancy groups.

1.5 How the metrics are used

The health metrics of DALY, YLL, and YLD can be used in several ways to help prioritize public health interventions, evaluate the impact of diseases and injuries, and inform public health decision-making. Some common uses of these metrics include:

Prioritizing public health interventions: By calculating the overall burden of disease in a population, public health practitioners can prioritize which diseases and injuries to address first. This helps allocate resources and target interventions to the areas of greatest need.

Evaluating the impact of diseases and injuries: These metrics can be used to measure the impact of diseases and injuries on individuals and populations and to track changes over time. This information can help inform public health decision-making and allocate resources more effectively.

Comparing the burden of disease across populations: DALY, YLL, and YLD can be used to compare the burden of disease across populations and between different regions. This information can help identify disparities in health outcomes and inform targeted public health interventions.

Evaluating the effectiveness of public health programs: These metrics can be used to evaluate the impact of public health programs and to assess the effectiveness of public health interventions. This information can help public health practitioners identify areas for improvement and make necessary changes to ensure that programs are achieving their goals.

Monitoring global health trends: DALY, YLL, and YLD can also be used to monitor global health trends and track changes in the burden of disease over time. This information can be used to inform global health policies and allocate resources to address emerging health threats.

Overall, the health metrics of DALY, YLL, and YLD provide valuable information for public health practitioners, researchers, and policy makers to help prioritize and allocate resources, evaluate the impact of diseases and injuries, and inform public health decision-making.

1.6 HALE (Healthy Life Expectancy)

- Description and calculation

The health metric of HALE (Healthy Life Expectancy) is a measure of overall health and well-being that takes into account both quantity and quality of life. It is a composite measure that combines years of life expectancy with a measure of the prevalence and severity of disability in a population. HALE provides a more comprehensive view of health outcomes than traditional measures of life expectancy, which only consider the quantity of life.

The calculation of HALE typically involves estimating the number of years that an individual can expect to live in good health, taking into account the impact of diseases and injuries on quality of life. This information is then used to estimate the overall health status of a population.

HALE is a useful tool for public health practitioners and policy makers, as it provides a more nuanced view of the health outcomes of a population. This information can help inform public health interventions and prioritize resources, as well as help track changes in health outcomes over time. Additionally, HALE can be used to compare the health outcomes of different populations and identify disparities in health outcomes, which can help inform targeted public health interventions.

Overall, the HALE metric provides a valuable perspective on the overall health and well-being of a population, combining information about both quantity and quality of life to provide a comprehensive view of health outcomes.

It adjusts overall life expectancy by the amount of time lived in less than perfect health. This is calculated by subtracting from the life expectancy a figure which is the number of years lived with disability multiplied by a weighting to represent the effect of the disability.[1]

More info [2]

[1] www.healthknowledge.org.uk

[2] about the methodology https://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets/health/health_life_expectancy

2 Metrics components

Learning Objectives:

- Life tables and Life expectancy used in the book
- Mortality level and rates
- Identify the disability weights

This section is dedicated to a closer look at what are the components of the health metrics, how to build them and finally how to use them for making countries comparison. YLLs, YLDs and DALYs can be used for different cause of deaths and disabilities, at different age levels.

2.0.1 YLLs components

The number of years of life lost YLLs is the first of the three metrics that is calculated, and is important for releasing a first look at the status of a population. It is calculated for identifying the area where improvement is required for reducing the loss in health status and clearly reducing the probability of death (Reiner & Hay, 2022)

The components of YLL (Years of Life Lost) include several factors that contribute to the calculation of premature death due to a disease or injury. These components are:

Age at death: The age at which a person died due to a disease or injury is a crucial component of YLL. The earlier the age at death, the greater the impact on potential years of life lost.

Life expectancy: The expected age at death in a population without the disease or injury is an important component of YLL. This value is used to compare the actual age at death with the expected age at death and determine the number of years of life lost.

Standard life expectancy: To make comparisons across populations and over time, YLL is often expressed relative to a standard life expectancy, typically set at an age of 70 years.

Population size: The size of the population affected by a disease or injury is another important component of YLL. A larger population will have a greater impact on overall YLL, regardless of the age at death.

Cause of death: The cause of death is also considered when calculating YLL, as different causes may have different impacts on potential years of life lost.

By taking these factors into account, YLL provides a measure of the potential years of life lost due to premature death caused by a disease or injury. It is an important component of the DALY, which provides a comprehensive view of the overall burden of disease on a population.

2.0.2 YLDs components

The components of YLD (Years Lived with Disability) include several factors that contribute to the calculation of the number of years lived with a disability due to a disease or injury. These components are:

Prevalence: The prevalence of a condition is the number of cases of a particular disease or injury present in a population at a given time. This is an important component of YLD as it determines the number of people who are affected by a disease or injury and the overall impact on the population.

Disability weight: The disability weight reflects the severity of the disability caused by a disease or injury. It is used to quantify the impact of a condition on quality of life, with higher weights assigned to more severe conditions.

Age: The age of the person affected by a disease or injury is also considered when calculating YLD. Conditions that occur at an earlier age will have a greater impact on the number of years lived with disability.

Population size: The size of the population affected by a disease or injury is another important component of YLD. A larger population will have a greater impact on overall YLD, regardless of the age of the affected individuals.

Duration of disability: The duration of disability is also considered when calculating YLD. Conditions that last for a longer period of time will have a greater impact on the number of years lived with disability.

By taking these factors into account, YLD provides a measure of the number of years lived with a disability due to a disease or injury. It is an important component of the DALY, which provides a comprehensive view of the overall burden of disease on a population.

2.0.3 DALYs components

Both YLL and YLD are components of the DALY and are used to provide a comprehensive assessment of the impact of disease and injury on a population. By combining YLL and YLD, the DALY takes into account both premature death and the impact of disease or injury on quality of life, providing a more comprehensive view of the overall burden of disease.

2.1 Build the metrics

2.1.1 Life tables and Life expectancy

Two fundamental components are used for calculating the YLL are:

- life tables
- life expectancy

Both of these elements are key for achieving the highest value of prediction of the state of health of a population.

2.1.1.1 Life tables

The life tables are selected among the most frequently used, more information about how to build a like table can be found in the Appendix A of this book.

```
library(tidyverse)
library(infectious)
```

Glifetables contains five variables:

1. **indicator**:

indicator
Tx - person-years lived above age x
ex - expectation of life at age x
lx - number of people left alive at age x
nLx - person-years lived between ages x and x+n
nMx - age-specific death rate between ages x and x+n
ndx - number of people dying between ages x and x+n
nqx - probability of dying between ages x and x+n

2. **age group**: from 1 to 85+ in 5-year classes

3. **sex:** female, male, and both

4. **value**

```
 #| echo: false
Glifetables %>%
  mutate(indicator = sub(" -.*$", "", indicator)) %>%
  group_by(indicator) %>%
  summarize(value = round(mean(value), 3))
```

```
#> # A tibble: 7 x 2
#>   indicator       value
#>   <chr>           <dbl>
#> 1 Tx        2652370.
#> 2 ex             31.3
#> 3 lx          70145.
#> 4 nLx        313032.
#> 5 nMx             0.041
#> 6 ndx          5263.
#> 7 nqx             0.16
```

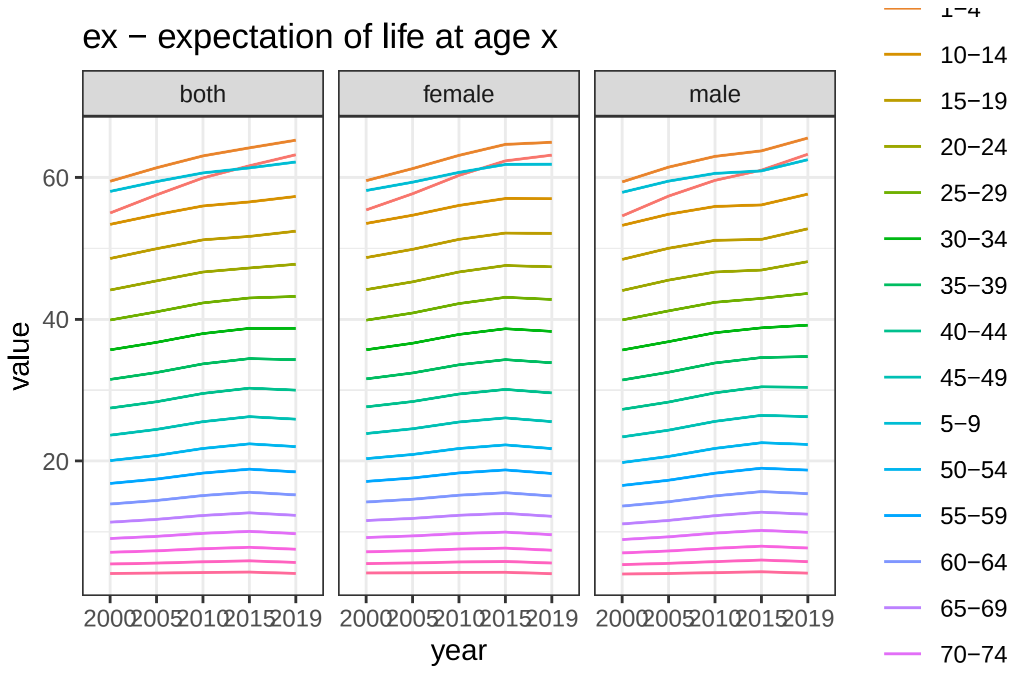
5. **year**: from 2000 to 2019

2.1.1.2 Life expectancy

The life expectancy rates are calculated with consideration of the probability of survival based on key parameter such as age, and deaths probabilities for that age. More info about how to calculate the life expectancy can be found in the Appendix A section of this book.

This visualization of the `ex - expectation of life at age x` shows for each age group its changing level across the years.

```
Glifetables %>%
  filter(indicator == "ex - expectation of life at age x") %>%
  ggplot(aes(year, value, group = age_group, color = age_group)) +
  geom_line() +
  facet_wrap(vars(sex)) +
  labs(title = "ex - expectation of life at age x") +
  theme_bw()
```



2.2 How to build the YLLs: a practical example

In this section a practical calculation of the health metrics is done for the practitioner to be able to replicate this calculation for further analysis based on these key elements.

```
Germany_lungc %>%
  full_join(
    Glifetables %>%
      filter(year == 2019,
             indicator == "ex - expectation of life at age x") %>%
      rename(life_expectancy = value),
    by = c("age_group", "sex")
  ) %>%
  select(-upper, -lower, -year, -indicator) %>%
  group_by(age_group) %>%
  mutate(YLL = val * life_expectancy) %>%
  filter(!is.na(YLL)) %>%
  head()
```

```
#> # A tibble: 6 x 5
#> # Groups:   age_group [2]
#>   sex    age_group   val life_expectancy   YLL
#>   <chr>  <chr>     <dbl>           <dbl> <dbl>
#> 1 male   10-14     0.322            57.7  18.5
#> 2 female 10-14     0.457            57.0  26.1
#> 3 both   10-14     0.779            57.3  44.6
#> 4 male   15-19     1.27             52.8  67.2
#> 5 female 15-19     1.56             52.1  81.1
#> 6 both   15-19     2.83             52.4 148.
```

3 Causes and Risks

Learning Objectives:

- How to use the metrics
- Overview of the causes and risks
- Identify the objective of the research question

Conditions and injuries that are associated with the burden of disease and injury vary according to their specific causes and risks. However, some common causes and risk factors include:

- **Lifestyle choices**: Poor diet, physical inactivity, tobacco use, and excessive alcohol consumption are major risk factors for many chronic diseases and injuries, including heart disease, stroke, cancer, and liver disease.
- **Environmental factors**: Exposure to pollutants, such as air pollution and toxic chemicals, can increase the risk of certain diseases and injuries.
- **Infections**: Many diseases, such as tuberculosis, HIV/AIDS, and malaria, are caused by infectious agents.
- **Poverty**: People living in poverty are often more susceptible to health problems due to limited access to healthcare, healthy food, and safe living conditions.
- **Aging**: As people get older, they are at an increased risk of many health problems, including chronic diseases and disabilities.
- **Genetics**: Some diseases and injuries are caused by genetic factors, such as a genetic predisposition to certain cancers.
- **Injuries**: Injuries, such as falls, road traffic accidents, and violence, can also contribute to the burden of diseases and injuries.

A particular health condition can have multiple causes and risk factors. For example, poverty and lack of access to healthcare can increase the risk of infectious diseases, while poor diet and physical inactivity can increase the risk of chronic diseases. Addressing the underlying causes and risk factors for diseases and injuries is a key component of public health interventions and can help reduce the overall burden of disease.

As an example here is shown how the DALY metric can be used for prevention:

Suppose we have data on the number of cases of a particular disease, as well as the average number of years of life lost due to this disease. We can use this information to calculate the total number of DALYs lost due to this disease.

```r
# Load the library 'dplyr'
library(dplyr)

# Create a data frame with the number of cases and average years of life lost
df <- data.frame(YLL = c(5, 10, 15),
                 YLD = c(1,3,4))

# Calculate the number of DALYs lost
df <- df %>% mutate(DALY = YLL + YLD)

# Sum the total number of DALYs lost
total_dalys <- sum(df$DALY)
total_dalys
```

```
#> [1] 38
```

In this example, the number of cases of the disease and the average years of life lost for each case are used to calculate the number of DALYs lost for each case. Finally, the total number of DALYs lost for the entire population.

This information can be used to inform public health interventions to prevent the spread of this disease and reduce the number of DALYs lost. For example, the information could be used to prioritize resources for disease control and prevention activities, such as health education campaigns, vaccination programs, and screening and treatment programs.

Part II

Machine Learning

Overview

This chapter provides a simple explanation on how to model a fast growing phenomenon such as in the case of the spread of infectious diseases, or on the contrary how a fast growing little impact can influence the performance of health metrics for some countries but doesn't affect the Global picture. More specifically, a further look will be provided at the evolution of the infection in some specific countries, and how this influences the overall level of health metrics.

The spread of a virus can be seen as a random process, since the number of individuals who are infected at any given time can change randomly. The exact number of individuals who will be infected in the future cannot be determined with certainty, since it depends on various factors such as the contagiousness of the virus, the behavior of individuals, and the efficacy of mitigation measures.

A deterministic model, on the other hand, could be used to model the spread of a virus under certain conditions, such as a fixed number of individuals, constant contagiousness, and no mitigation measures. This type of model can be used to make predictions about the spread of a virus under certain assumptions, but it will not account for the randomness and uncertainty associated with real-world scenarios.

In general, a **stochastic process**, or random process (Dobrow, 2016) is the type of model which attempt to replicate uncertain outcomes. At opposite, in a **deterministic system** the outcome is obtained from a given input, and for this reason it is reproducible.

Most models used to study the spread of a virus are a **combination of both deterministic and stochastic models**. For example, the SIR (Susceptible-Infected-Recovered) model is a deterministic model that describes the dynamics of the spread of a virus, but it also includes stochastic elements such as random interactions between individuals.

4 Techniques

Learning Objectives:

- How to manipulate data through feature engineering
- Select the most suitable model for your data
- Learn how to apply machine learning algorithms

4.1 Data Analytics

Collecting data to use in a research analysis involves a selection of sources and methods to use for optimizing computational time when downloading and reading the files.

Once data is set and ready to use a further step is required to make the data suitable for the selected model.

In this chapter, an overview of different method of data loading and featuring selection is provided before to get into selecting the best model to use.

The source of data is an important variable. Generally, data can be downloaded by using an API (application programming interface) which allow the user to get access to data directly from source, with the use of specified back-end computations. There are alternatives at using an API; data can be obtained by downloading it directly into the computer, or loaded through library packages.

Usually, available files are provided under various forms such as delimited type of files, .csv, .xls, .json, and other types.

Here is an example of how to use an API for downloading a file directly onto your computer.

```r
library(httr)
url <- ""
httr::GET(url = url)
```

Once data is on your computer available and ready to use, the next step is to have a look at it and decide whether to perform some adjustment to the data to make it suitable for your model.

This step includes:

- data manipulation/wrangling
- featuring engineering
- exploratory data analysis

Let's use the `{HistData}` package for an example on William Farr's Data on `Cholera` in London, 1849. This set of data contains information about the number of deaths due to Cholera in specific districts, the population density, the water provider and other variables.

This is a type of dataset which can be considered ready to use for some type of models such as linear regression models, but it would require some adjustments if a Bayesian approach is desired.

```r
library(tidyverse)
library(HistData)
Cholera <- HistData::Cholera
Cholera%>%head
```

```
#>              district cholera_drate cholera_deaths  popn elevation region
#> 1           Newington           144            907 63074        -2   Kent
#> 2         Rotherhithe           205            352 17208         0   Kent
#> 3          Bermondsey           164            836 50900         0   Kent
#> 4 St George Southwark           161            734 45500         0   Kent
#> 5            St Olave           181            349 19278         2   Kent
#> 6          St Saviour           153            539 35227         2   Kent
#>       water annual_deaths pop_dens persons_house house_valpp poor_rate area
#> 1 Battersea           232      101           5.8       3.788     0.075  624
#> 2 Battersea           277       19           5.8       4.238     0.143  886
#> 3 Battersea           267      180           7.0       3.318     0.089  282
#> 4 Battersea           264       66           6.2       3.077     0.134  688
#> 5 Battersea           281      114           7.9       4.559     0.079  169
#> 6 Battersea           292      141           7.1       5.291     0.076  250
#>   houses house_val
#> 1   9370    207460
#> 2   2420     59072
#> 3   6663    155175
#> 4   5674    107821
#> 5   2523     90583
#> 6   4659    174732
```

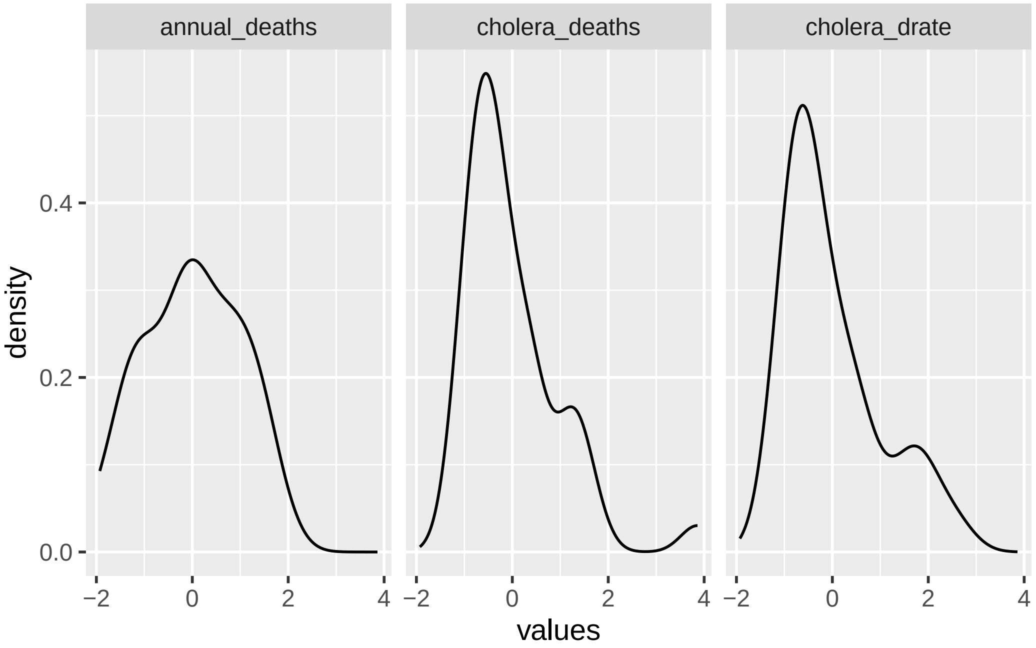
If we are interested in the evolution of mortality due to Cholera. We might want to look at the regional level, how the annual deaths - all causes (`annual_deaths`), the death's rate per

10,000 inhabitants (`cholera_drate`), or is distributed.

```
data <- Cholera%>%
  select(region,cholera_drate,contains("death",ignore.case=T))

data %>%
  head
```

```
#>   region cholera_drate cholera_deaths annual_deaths
#> 1   Kent           144            907           232
#> 2   Kent           205            352           277
#> 3   Kent           164            836           267
#> 4   Kent           161            734           264
#> 5   Kent           181            349           281
#> 6   Kent           153            539           292
```

```
data %>%
  select(-region)%>%
  scale()%>%
  bind_cols(region=data$region)%>%
  pivot_longer(cols = 1:3,names_to = "type",values_to = "values")%>%
  ggplot(aes(values))+
  geom_density()+
  facet_wrap(~type)
```


annual_deaths
cholera_deaths
cholera_drate
density
0.4
0.2
0.0
−2
0
2
4
values

5 Packages and functions

Learning Objectives:

- Overview of the most suitable R-packages
- Combination of packages and functions
- How to find a new R-package

6 Predicting the future

Learning Objectives:

- How to use the Predict function
- Evaluate prediction results
- Improve predictions

Part III

Data Visualizations

Overview

7 Application of the model results

Learning Objectives:

- Overview of the basic plots
- How to customize a plot
- Tell the story with data

8 Spatial data modeling and Visualization

Learning Objectives:

- How to make a map
- Identify the missing pieces
- Map the results

9 Examples of data visualizations

Learning Objectives:

- Learn how to improve the visualization
- Make a custom theme
- Save your plot

Part IV

Infectious Diseases

Overview

Infectious Diseases the invisible enemies

The infective agent begins to thrive and multiply throughout the body (Broemeling, 2021). Its proliferation can be fast or slow depending on the type of organism. Every infectious disease has an incubation period, the length of time the pathogen is established until appearance of symptoms of the disease.

Factors influencing infection

- quantity of invading germs (dose of the infection)
- the virulence of the infection
- the condition of the body's immune system
- contact with source of infection for contagious diseases

Microorganisms adapt far more rapidly than humans as the scene shifts. A bacterial generation ranges from 20-30 minutes, for viruses it's much smaller.

> Who is going to adapt to whom?

Virus means **poisonous substance** ranging from 20 to 400 nm in diameter can be observed only with a electron microscope. Outside of a living cell is a dormant particle of strange shapes. When manage to get inside the cell it starts replicating killing the cell or skewing its functions.

10 Covid19 Outbreaks

Learning Objectives:

- What is Covid19
- How does it spread
- Map Covid19 outbreaks

10.1 Covid19 spread of infection as a cause for increased DALY for selected countries

10.2 The spread of COVID-19

Here is a model example of the spread of COVID-19 using a stochastic process. In this example, we'll use a simple SEIR (Susceptible-Exposed-Infected-Recovered) model to simulate the spread of the virus in a population. The SEIR model divides a population into four compartments based on their status with respect to the virus: susceptible, exposed, infected, and recovered.

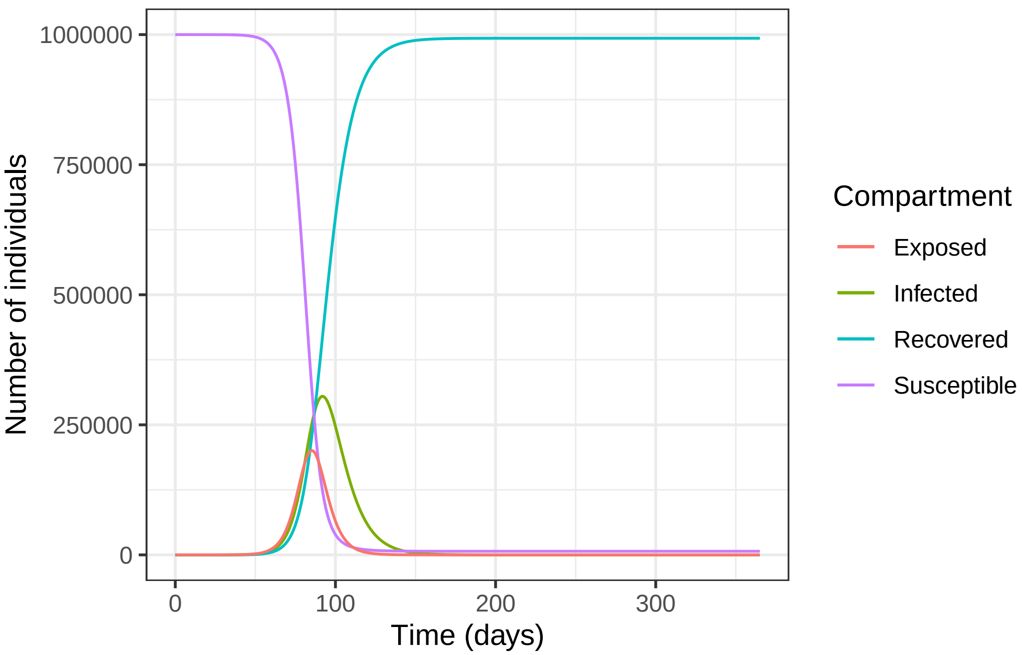
First, we'll load the necessary packages and define some parameters for our model:

```r
library(deSolve)

# Define parameters
N <- 1e6  # Total population
beta <- 0.5  # Transmission rate
gamma <- 0.1  # Recovery rate
t_exp <- 5  # Latent period

# Initial conditions
init <- c(S = N - 1, E = 1, I = 0, R = 0)

# Define the SEIR model
```

```r
seir_model <- function(t, y, parameters) {
  with(as.list(y), {
    dS <- -beta * S * I / N
    dE <- beta * S * I / N - (1 / t_exp) * E
    dI <- (1 / t_exp) * E - gamma * I
    dR <- gamma * I
    return(list(c(dS, dE, dI, dR)))
  })
}
```

Next, we'll use the ode function from the deSolve package to solve the ODEs and simulate the spread of the virus over a period of 365 days:

```r
# Solve the ODEs
times <- seq(0, 365, by = 1)
result <- ode(y = init, times = times, func = seir_model)

# Plot the results
library(ggplot2)
ggplot(as.data.frame(result), aes(time, I, color = "Infected")) +
  geom_line() +
  geom_line(aes(time, R, color = "Recovered")) +
  geom_line(aes(time, S, color = "Susceptible")) +
  geom_line(aes(time, E, color = "Exposed")) +
  scale_color_discrete(name = "Compartment") +
  labs(x = "Time (days)", y = "Number of individuals") +
  theme_bw()
```



In this example, many factors are not taken into account. In reality, the spread of a virus is much more complex and influenced by many factors such as human behavior, government policies, and healthcare systems.

10.2.1 Covid19 modeling

The Bayesian Analysis of Infectious Diseases: COVID-19 and Beyond book is a comprehensive resource that covers the use of Bayesian analysis in the modeling of infectious diseases, including COVID-19.

In the context of modeling COVID-19 in R, the following steps can be taken:

Define the model structure: The first step is to define the model structure, which involves specifying the underlying mechanisms of disease spread and the parameters of interest. For COVID-19, the model structure might include the number of susceptible individuals, the number of infected individuals, the number of recovered individuals, and the rate of disease transmission.

Specify the prior distributions: The next step is to specify the prior distributions for the parameters of interest. This involves defining the prior beliefs about the values of the parameters based on available data and expert knowledge. In the case of COVID-19, this might include prior beliefs about the rate of disease transmission, the incubation period, and the rate of recovery.

Collect data: The next step is to collect relevant data on the spread of the disease. This might include the number of confirmed cases, the number of hospitalizations, and the number of deaths.

Implement the model in R: Once the model structure and prior distributions have been specified, the model can be implemented in R using a variety of packages, including Stan, JAGS, or MCMCpack.

Estimate the parameters: The next step is to use the data and the model to estimate the parameters of interest. This can be done using Bayesian Markov Chain Monte Carlo (MCMC) methods, which involve drawing a large number of samples from the posterior distributions of the parameters.

Evaluate the model: The final step is to evaluate the performance of the model, which involves comparing the model predictions with the observed data and checking the validity of the model assumptions.

This is a general outline of the steps involved in modeling COVID-19 in R using Bayesian analysis. The specific details of the implementation will depend on the particular model structure and data being used.

Suppose we have data on the number of confirmed cases of COVID-19 in a region for a period of time, and we want to model the spread of the disease.

Step 1: Define the model structure

We can use the Susceptible-Infected-Recovered (SIR) model to describe the spread of the disease, where S represents the number of susceptible individuals, I represents the number of infected individuals, and R represents the number of recovered individuals. The model structure can be described as follows:

$$dS/dt = -beta * S * I/N$$

$$dI/dt = beta * S * I/N - gamma * I$$

$$dR/dt = gamma * I$$

where beta is the rate of transmission, gamma is the rate of recovery, and N is the total population.

Step 2: Specify the prior distributions

We can specify the prior distributions for beta and gamma using expert knowledge and available data. For example, we might specify a gamma distribution with a mean of 0.1 and a standard deviation of 0.05 for beta, and a gamma distribution with a mean of 0.05 and a standard deviation of 0.02 for gamma.

Step 3: Collect data

We can collect data on the number of confirmed cases of COVID-19 in the region for a period of time.

Step 4: Implement the model in R

We can use the `{rstan}` package in R to implement the model, as follows:

```r
# Load the rethinking package
library(rethinking)
library(rstan)
library(tidyverse)

# Load the data
source("inst/scripts/covid19_sim_data.R")
data <- out

# Define the model
SIR_model <- function(data) {
  # Define the priors
  beta <- dnorm(0, 0.01)
  gamma <- dnorm(0, 0.01)
  I0 <- dnorm(0, 0.01)

  # Define the model
  S <- numeric(length(data$Time))
  I <- numeric(length(data$Time))
  R <- numeric(length(data$Time))
  S[1] <- 1 - I0
  I[1] <- I0
  for (t in 2:length(data$Time)) {
    S[t] <- S[t - 1] - beta * S[t - 1] * I[t - 1]
    I[t] <- I[t - 1] + beta * S[t - 1] * I[t - 1] - gamma * I[t - 1]
    R[t] <- R[t - 1] + gamma * I[t - 1]
  }

  # Likelihood
  dpois(data$y, lambda = I)
}
```

Step 5: Estimate the parameters

We can use the `sampling` function in the `{rstan}` package to estimate the parameters, as follows:

```
data <- alist(
  time<- data$time,
  S<-data$S,
  I<-data$I,
  R<-data$R
)
```

Step 6: Evaluate the model

We can evaluate the performance of the model by comparing the model predictions with the observed data and checking the validity

```
# Plot the observed data and the model predictions
library(ggplot2)

predictions <- extract(results)$I

data.frame(Time = data$Time,
           Observed = data$y,
           Predicted = predictions) %>%
  ggplot(aes(x = Time, y = Observed)) +
  geom_line(aes(y = Predicted, color = "Predicted")) +
  geom_line(aes(y = Observed, color = "Observed"),
            linetype = "dashed") +
  scale_color_discrete(name = NULL, labels = c("Predicted", "Observed")) +
  labs(x = "Time", y = "Number of Confirmed Cases") +
  theme_bw()

# Calculate the goodness-of-fit measures
rmse <- sqrt(mean((data$y - predictions)^2))
mae <- mean(abs(data$y - predictions))
r2 <- cor(data$y, predictions)^2
cat("RMSE:", rmse, "\n")
cat("MAE:", mae, "\n")
cat("R-squared:", r2, "\n")

# Plot the posterior distribution of the parameters
library(bayesplot)
mcmc_areas(results, pars = c("beta", "gamma", "I0"),
           prob = 0.89, ROPE = c(0.05, 0.1),
           prob_lines = TRUE, prob_args = list(col = "red"),
           rope_args = list(col = "blue"),
             main = "Posterior Distribution of Parameters")
```

11 The case of Malaria

Learning Objectives:

- What is Malaria
- How does it spread
- Map Malaria outbreaks

12 Summary: the state of health

Conclusions

In Conclusion, the technique of modeling transfer learning shows as the foundation of model application how to make prediction and to improve public policies.

References

Broemeling, L. D. (2021). *Bayesian analysis of infectious diseases: COVID-19 and beyond*. New York: Chapman; Hall/CRC. https://doi.org/10.1201/9781003125983

Dobrow, R. P. (2016). *Introduction to Stochastic Processes with R*. John Wiley & Sons.

Kovacheva, Ts. P. (2017). Life tables - key parameters and relationships between them. *International Mathematical Forum*, *12*, 469–479. https://doi.org/10.12988/imf.2017.7225

Life table - an overview | ScienceDirect topics. (n.d.). Retrieved from https://www.sciencedirect.com/topics/medicine-and-dentistry/life-table

Modified logit life table system: Principles, empirical validation, and application: Population studies: Vol 57, no 2. (n.d.). Retrieved from https://www.tandfonline.com/doi/abs/10.1080/0032472032000097083

Murray, C. J., & Frenk, J. (2008). Health metrics and evaluation: Strengthening the science. *The Lancet*, *371*(9619), 1191–1199. https://doi.org/10.1016/S0140-6736(08)60526-7

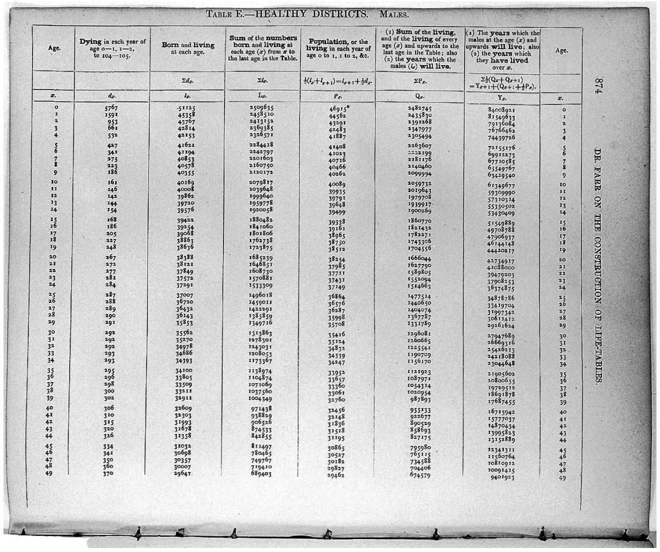
Reiner, R. C., & Hay, S. I. (2022). The overlapping burden of the three leading causes of disability and death in sub-Saharan African children. *Nature Communications*, *13*(1), 7457. https://doi.org/10.1038/s41467-022-34240-6

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., … Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1204–1222. https://doi.org/10.1016/s0140-6736(20)30925-9

A Life tables and Life expectancy

Back in the 1700s the Swiss mathematician and physicist Daniel Bernoulli (1700 - 1782) developed the use of a life table model by differentiating life tables based on specific causes of death *Life Table - an Overview | ScienceDirect Topics* (n.d.).

Originally made by the English scientist John Graunt (1620-1674), for the analysis of the mortality of the population of London and the impact of different diseases. Life tables contain fundamental statistics for the calculation of probabilities of deaths and the computation of life expectancy at birth and at different ages.



874 DR. FARR ON THE CONSTRUCTION OF LIFE-TABLES.

TABLE E.—HEALTHY DISTRICTS. MALES.

Age.	**Dying** in each year of age 0—1, 1—2, to 104—105.	**Born** and **living** at each age.	**Sum** of the **numbers born** and **living** at each age (x) from x to the last age in the Table.	**Population,** or the **living** in each year of age 0 to 1, 1 to 2, &c.	(1) **Sum** of the **living,** and of the **living** of every age (x) and upwards to the last age in the Table; also (2) the **years** which the males (l_x) **will live.**	(1) The **years** which the males at the age (x) and upwards **will live**; also (2) the **years** which they **have lived** over x.	Age.
		Σd_x.	Σl_x.	$\frac{1}{2}(l_x+l_{x+1})=l_{x+1}+\frac{1}{2}d_x$.	ΣP_x.	$\Sigma\frac{1}{2}(Q_x+Q_{x+1})$ $=Y_{x+1}+(Q_{x+1}+\frac{1}{2}P_x)$.	
x.	d_x.	l_x.	L_x.	P_x.	Q_x.	Y_x.	x.
0	5767	51125	2509635	46915*	2482745	84008921	0
1	1591	45358	2458510	44562	2435830	81549633	1
2	953	43767	2413152	43291	2391268	79136084	2
3	661	42814	2369385	42483	2347977	76766462	3
4	532	42153	2326571	41887	2305494	74439726	4
5	427	41621	2284418	41408	2263607	72155176	5
6	341	41194	2242797	41023	2222199	69912273	6
7	275	40853	2201603	40716	2181176	67710585	7
8	223	40578	2160750	40466	2140460	65549767	8
9	186	40355	2120172	40262	2099994	63429540	9
10	161	40169	2079817	40089	2059732	61349677	10
11	146	40008	2039648	39935	2019643	59309990	11
12	142	39862	1999640	39791	1979708	57310314	12
13	144	39720	1959778	39648	1939917	55350502	13
14	154	39576	1920058	39499	1900269	53430409	14
15	168	39422	1880482	39338	1860770	51549889	15
16	186	39254	1841060	39161	1821432	49708788	16
17	205	39068	1801806	38965	1782271	47906937	17
18	227	38863	1762738	38750	1743306	46144148	18
19	248	38636	1723875	38512	1704556	44420217	19
20	267	38388	1685239	38254	1666044	42734917	20
21	272	38121	1646851	37985	1627790	41088000	21
22	277	37849	1608730	37711	1589805	39479203	22
23	281	37572	1570881	37431	1552094	37908253	23
24	284	37291	1533309	37149	1514663	36374875	24
25	287	37007	1496018	36864	1477514	34878786	25
26	288	36720	1459011	36576	1440650	33419704	26
27	289	36432	1422291	36287	1404074	31997342	27
28	290	36143	1385859	35998	1367787	30611412	28
29	291	35853	1349716	35708	1331789	29261624	29
30	292	35562	1313863	35416	1296081	27947689	30
31	292	35270	1278301	35124	1260665	26669316	31
32	292	34978	1243031	34832	1225541	25426213	32
33	293	34686	1208053	34539	1190709	24218088	33
34	293	34393	1173367	34247	1156170	23044648	34
35	295	34100	1138974	33952	1121923	21905602	35
36	296	33805	1104874	33657	1087971	20800655	36
37	298	33509	1071069	33360	1054314	19729512	37
38	300	33211	1037560	33061	1020954	18691878	38
39	302	32911	1004349	32760	987893	17687455	39
40	306	32609	971438	32456	955133	16715942	40
41	310	32303	938829	32148	922677	15777037	41
42	315	31993	906526	31836	890529	14870434	42
43	320	31678	874533	31518	858693	13995823	43
44	326	31358	842855	31195	827175	13152889	44
45	334	31032	811497	30865	795980	12341311	45
46	341	30698	780465	30527	765115	11560764	46
47	350	30357	749767	30182	734588	10810912	47
48	360	30007	719410	29827	704406	10091415	48
49	370	29647	689403	29462	674579	9401923	49

Figure A.1: "Life tables", William Farr (England 1859)

A.0.1 Life tables components

More recent life tables are standardized to be used for a population of 100 000 at age 0.

l_x survivors at age x it starts with a value of 100 000

d_x deceased at age x

q_x probability of deaths

p_x probability of survive

The probability of survive is given by:

$$p_x = 1 - q_x$$

Let's start constructing a life table

The **Global Life Tables** are included in the {infectious} package as `Glifetables` dataset. This dataset has been released by the WHO, and contains various indicators.

The construction of the Global Life Tables takes consideration of age-specific mortality patterns, which is the main improvements made on life tables construction since the first set of model life tables published by the United Nations in 1955, see (*Modified Logit Life Table System*, n.d.) for more information about a detailed procedure.

To have a look at the package documentation for this dataset, use:

```r
?infectious::Glifetables
```

```r
library(tidyverse)
infectious::Glifetables %>%
  count(indicator)
```

```
#> # A tibble: 7 x 2
#>   indicator                                             n
#>   <chr>                                             <int>
#> 1 Tx - person-years lived above age x                 285
#> 2 ex - expectation of life at age x                   285
#> 3 lx - number of people left alive at age x           285
#> 4 nLx - person-years lived between ages x and x+n     285
#> 5 nMx - age-specific death rate between ages x and x+n   285
#> 6 ndx - number of people dying between ages x and x+n    285
#> 7 nqx - probability of dying between ages x and x+n      285
```

The indicator of interest for re-building a life table are:

- `lx` - number of people left alive at age x
- `age_group`

These two key elements are crucial for building the life tables.

```
lx <- infectious::Glifetables %>%
  distinct() %>%
  filter(indicator == "lx - number of people left alive at age x",
         year == "2019")
```

```
x <- lx %>%
  filter(sex == "female") %>%
  select(x = age_group)

lx_f <- lx %>%
  filter(sex == "female") %>%
  select(lx = value)

lx_m <- lx %>%
  filter(sex == "male") %>%
  select(lx = value)
```

The probability of survival is calculated as follow:

$$p_x = \frac{l_x}{l_{x+1}}$$

```
px = lx_f$lx / lag(lx_f$lx)
```

```
data.frame(
  x,
  lx = round(lx_f),
  dx = round(c(-diff(lx_f$lx), 0)),
  px,
  qx = 1 - lead(px),
  Lx = c((lx_f$lx[1] + (lx_f$lx[2])) / 2,
         5 * (lx_f$lx[-1] + lead(lx_f$lx[-1])) / 2)
) %>% head
```

```
#>         x     lx   dx        px          qx        Lx
#> 1       1 100000 4278        NA 0.042783388  97860.83
#> 2     1-4  95722 1346 0.9572166 0.014064138 475242.70
#> 3     5-9  94375  214 0.9859359 0.002264391 471342.84
#> 4   10-14  94162  166 0.9977356 0.001758090 470394.72
#> 5   15-19  93996  553 0.9982419 0.005885446 468597.83
#> 6 20-24  93443  833 0.9941146 0.008917104 465131.71
```

```
infectious::Glifetables %>%
  distinct() %>%
  filter(
    indicator == "nLx - person-years lived between ages x and x+n",
    year == "2019",
    sex == "female"
  )
```

```
#> # A tibble: 19 x 5
#>    indicator                                       age_group sex    value year
#>    <chr>                                           <chr>     <chr>  <dbl> <chr>
#>  1 nLx - person-years lived between ages x and x+n 1         female 9.70e4 2019
#>  2 nLx - person-years lived between ages x and x+n 1-4       female 3.80e5 2019
#>  3 nLx - person-years lived between ages x and x+n 5-9       female 4.71e5 2019
#>  4 nLx - person-years lived between ages x and x+n 10-14     female 4.70e5 2019
#>  5 nLx - person-years lived between ages x and x+n 15-19     female 4.69e5 2019
#>  6 nLx - person-years lived between ages x and x+n 20-24     female 4.65e5 2019
#>  7 nLx - person-years lived between ages x and x+n 25-29     female 4.60e5 2019
#>  8 nLx - person-years lived between ages x and x+n 30-34     female 4.54e5 2019
#>  9 nLx - person-years lived between ages x and x+n 35-39     female 4.45e5 2019
#> 10 nLx - person-years lived between ages x and x+n 40-44     female 4.32e5 2019
#> 11 nLx - person-years lived between ages x and x+n 45-49     female 4.14e5 2019
#> 12 nLx - person-years lived between ages x and x+n 50-54     female 3.90e5 2019
#> 13 nLx - person-years lived between ages x and x+n 55-59     female 3.56e5 2019
#> 14 nLx - person-years lived between ages x and x+n 60-64     female 3.12e5 2019
#> 15 nLx - person-years lived between ages x and x+n 65-69     female 2.59e5 2019
#> 16 nLx - person-years lived between ages x and x+n 70-74     female 1.98e5 2019
#> 17 nLx - person-years lived between ages x and x+n 75-79     female 1.32e5 2019
#> 18 nLx - person-years lived between ages x and x+n 80-84     female 7.34e4 2019
#> 19 nLx - person-years lived between ages x and x+n 85+       female 3.84e4 2019
```

A.1 Life expectancy

Life expectancy is the expected number of years a person will live, based on current age and prevailing mortality rates. There are several methods to calculate life expectancy, but one common approach is to use the actuarial life table, which is a statistical table that provides the mortality rates for a population at different ages. The following steps can be used to calculate life expectancy using a life table:

Identify the relevant mortality rates for the population and time period of interest. Calculate the probability of surviving to each age, given the mortality rates. Multiply the probability of surviving to each age by the remaining life expectancy at that age to obtain the expected number of years of life remaining at each age. Sum the expected number of years of life remaining at each age to obtain the total life expectancy. Note that life expectancy is a statistical estimate and can be influenced by many factors, such as lifestyle, health, and environmental factors, so actual individual life expectancies can vary widely.

Here are some key references for calculating life expectancy:

1. United Nations World Population Prospects - The UN provides detailed life tables and population data, including life expectancy, for countries and regions around the world.

2. Centers for Disease Control and Prevention (CDC) - The CDC provides life tables for the United States, as well as information on how life expectancy is calculated and factors that affect it.

3. World Health Organization (WHO) - The WHO provides information on global health and life expectancy, including data and reports on trends in life expectancy and mortality.

4. Actuarial Science textbooks - Books such as “Actuarial Mathematics” by Bowers, Gerber, Hickman, Jones, and Nesbitt, or “An Introduction to Actuarial Mathematics” by Michel Millar, provide comprehensive coverage of the methods and mathematics used in calculating life expectancy.

5. Journal articles - Articles in actuarial and demographic journals, such as the North American Actuarial Journal or Demographic Research, often provide in-depth coverage of the latest research and methods for calculating life expectancy.

B Tools used to make this book

To set up the environment for replicating the code used in this book the R language is needed as well as R and Rstudio IDE environments. The following sections contain the directions for installing **R** and **RStudio**, how to set up a book with **quarto** and how to use **GitHub** as a version saver source.

B.1 RStudio installation

Download and install R: Download and install RStudio IDE:

B.2 Info on how to setup this project in quarto

quarto is the new version of **Rmarkdown**, it can be used for making notes, presentations, websites, books and more.

In this project the book has been made in quarto and version saved on github.

Quarto publishing

B.2.1 GitHub useful commands

You can do the same using the command line. In RStudio create a new project on a new directory, add git, and select `quarto book project`

The automated process will create a `_quarto.yml` file, the top of the file will look like this one:

```
project:
  type: book
```

On terminal type: `quarto preview`

It creates a folder `_book`

B.2.1.1 Add Github later:

existing-github-last to connect with github create a github repo with the same name of your project then type

```
usethis::use_git()
```

It asks you to commit all files in the RStudio project. This pushes all files in R to a remote folder designed to head to the github repo.

In terminal connect with the github repo:

```
git init
git remote add origin
```

https://github.com/Fgazzelloni/infectious.git

```
git branch -M main
git push -u origin main
```

B.2.2 Publish your book on github pages

change the `quarto.yml` file into:

```
project:
  type: book
  output-dir: docs
```

add a .nojekyll file, type in terminal:

```
touch .nojekyll
```

then type

```
quarto render
```

some issues might arise if more than one calculation is made inside a single chunks split the chunks!

quarto render creates a folder `docs`

B.3 Add a package

```
devtools::create("yourpkg")
```

Now that you have your book done, you might need to add some customized data to use within your analysis.

In order to do that, you'll need to just add data in the way that is usually done when inside a package.

B.3.1 Building blocks of the package inside the quarto book

B.3.1.1 ADD DATA TO PACKAGE

set the data from the original source and tidy appropriately

```
usethis::use_data_raw()
```

it creates a .R script where to put the steps for get the data ready

once data are ready

```
usethis::use_data(yourdata)
```

It creates the R folder where the .R scripts of data documentation is located

```
usethis::use_r("yourdataset")
```

read how to document your work

```
vignette("rd-other") # for datasets
vignette("rd")

devtools::document()
devtools::load_all(".")
```

B.3.1.2 ADD PACKAGE INFO

so the ? (NAMESPACE) works

```
usethis::use_package_doc()
devtools::document()
```

B.3.1.3 ADD DATA PACKAGE INFO

```
usethis::use_r("yourdataset")
devtools::document()
```

when filling the data .R script for explaining what's inside the dataset a specific structure needs to be used (see examples)

when new data are added

Build tab and —> More —> clean and install (this is not there anymore whae you start with `devtools::create("yourpkg")`) to add the new data to NAMESPACE

C Hints for translating into Python