The Spread of COVID-19

Azizi, Ilia and Rollini, Maura

26 April, 2020

Contents

Chapter 1

Introduction

1.1 The data

For this project, we use COVID-19 data provided by Johns Hopkins University (updated daily), as well as data from the world bank with demographic information. More specifically, we use daily records of total confirmed infection cases (or cumulative number of cases), total number of fatalities (or cumulative number of deaths) per country starting from 2020-01-22 until 2020-04-05. The dataset, downloaded on 2020-04-06, contains 12,975 observations and 13 variables (country, iso3c, date, confirmed, deaths, population, land_area_skm, pop_density, pop_largest_city, gdp_capita, life_expectancy, region, income).

Special credits: This project has been developed as part of the dsfba course and many thanks to professor Thibault Vatter and his assistants for their contributions.

Furthermore, the repository for this project can be found here.

1.2 A note on Epidemiological Models

Today's epidemiological models are mostly described by so called ***SIR***-like models (see details in Martcheva, 2015, pp.9–12). In this class of models, the population is divided into three groups:

- *(**S**)usceptible* — people, might get infected;
- *(**I**)nfectious* — people, who carry the infection and can infect others;
- *(**R**)ecovered/(**R**)emoved* — people, who have already recovered from the disease and got immunity.

The SIR model is a system of ordinary nonlinear differential equations. In this homework, we focus on the following *logistic* model (see Batista, 2020, pp. 2;

Martcheva, 2015, pp. 35–36):

$$\frac{dC(t)}{dt} = r\, C(t) \cdot \left[1 - \frac{C(t)}{K}\right],$$

where $C(t)$ is the accumulated number of cases at time t, r is the growth rate (or infection rate), and K is the final size of epidemic. Let C_0 be the initial number of cases: in other words, at time $t = 0$, assume that there was C_0 accumulated number of cases. The solution of the *logistic* model is

$$C(t) = \frac{K \cdot C_0}{C_0 + (K - C_0)\, \exp(-r\, t)},$$

which looks like a scaled *logit* model in econometrics.

1.3   This project

Because we only have access to the confirmed cases that are reported, we use those figures as a proxy for the total number of cases, with the understanding that they almost surely underestimates the actual number of interest. In what follows, we do a preliminary exploration of the data. We then use the *logistic* model to analyze the spread of **COVID-19** and try to predict the final number of accumulated confirmed cases for every country. More specifically, we

- start by focusing on modelling the spread in Switzerland;
- then apply the same approach to every country in the dataset.

Chapter 2

Exploratory data analysis

2.1 Description of the state of the spread

We fisrt provide a high level description of the state of the spread. In particular, including the:

- the number of days that have passed since the first confirmed case/death,
- the current stage for confirmed cases/deaths/mortality (i.e., ratio of deaths to confirmed cases).

Note: Please do note that most of the plots in this section, as well as some of the final plots in the the "many models" section, are interactive, hence, in order to see the built-in legend(for each country) you would have to scroll over the lines/bars.

Comment: Most of the current news is based on the absolute values of the confirmed cases and deaths per country, therefore we have decided to start our analysis by plotting them. However, since the sizes of the countries can differ quite a lot, we thought that it was best to have a look at the confirmed cases and deaths per country relative to their population to build a better picture.

We can see that by taking their size into account, the situation for some countries changes, more specifically we can see that the curve for the countries in the region of East Asia and Pacific is flatter when taking into account the size of the population, same for North America, while, on the other hand, the curves for countries in Europe and Central Asia are way steeper. This makes sense as, generally, countries in Europe are smaller, therefore a high absolute value would mean a higher percentage of the population, while in North America and Asia, where the countries have larger populations, the absolute values will correspond to a lower percentage.

What is evident is that for both confirmed cases and deaths, the situation that

is the most worrying is the one in Europe and Central Asia. It is perhaps more interesting to zoom on this area to gain a better understanding.

Comment: Once we focus on Europe and Central Asia, we can notice that the countries with the highest increase are San Marino, Andorra and Luxembourg. As we can see from the table, for the three countries, the population does not exceed a milion people, hence they are fairly small countries. Consequently, as we mentioned already before, this means that the emergence of every new case corresponds to a bigger increase of the percentage of the affected population, which is larger relative to the countries with larger populations.

After looking at Europe and Central Asia, we can move on to analyze the current situation for all the countries in the world.


Top 10 Current Number of Confirmed Cases per Country in Population Percentage
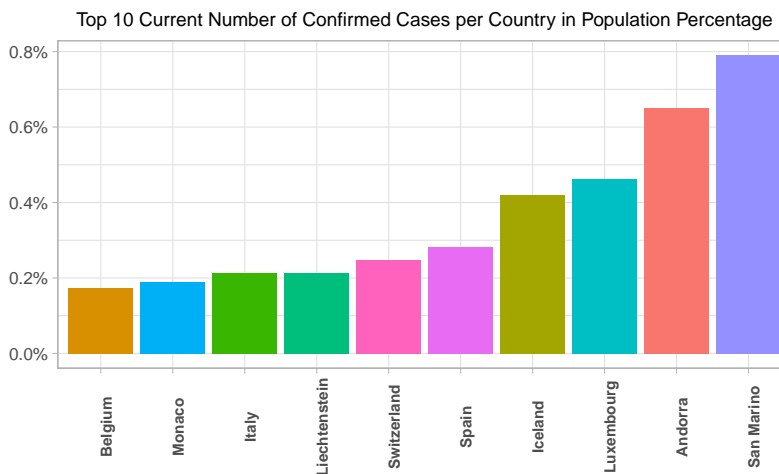
The Population per Country in Europe and Central Asia (from smallest to largest)	
country	population
San Marino	33,671
Liechtenstein	35,994
Monaco	38,682
Andorra	77,006
Iceland	353,574
Luxembourg	607,728
Montenegro	622,345
Cyprus	1,189,265
Estonia	1,320,884
Latvia	1,926,542
Slovenia	2,067,372
North Macedonia	2,082,958
Lithuania	2,789,533
Albania	2,866,376
Armenia	2,951,776
Bosnia and Herzegovina	3,323,929
Moldova	3,545,883
Georgia	3,731,000
Croatia	4,089,400
Ireland	4,853,506
Norway	5,314,336
Slovakia	5,447,011
Finland	5,518,050
Denmark	5,797,446
Kyrgyzstan	6,315,800
Serbia	6,982,084
Syria	6,982,084
Bulgaria	7,024,216
Switzerland	8,516,543
Austria	8,847,037
Belarus	9,485,386
Hungary	9,768,785
Azerbaijan	9,942,334
Sweden	10,183,175
Portugal	10,281,762
Czechia	10,625,695
Greece	10,727,668
Belgium	11,422,068
Netherlands	17,231,017
Kazakhstan	18,276,499
Romania	19,473,936
Uzbekistan	32,955,400
Poland	37,978,548
Ukraine	44,622,516
Spain	46,723,749
Italy	60,431,283
United Kingdom	66,488,991
France	66,987,244
Turkey	82,319,724
Germany	82,927,922
Russia	144,478,050

Confirmed Cases per Country in Absolute Values	
country	confirmed
USA	337,072
Spain	131,646
Italy	128,948
Germany	100,123
France	93,773
China	82,602
Iran	58,226
United Kingdom	48,436
Turkey	27,069
Switzerland	21,100
Belgium	19,691
Netherlands	17,953
Canada	15,756
Austria	12,051
Portugal	11,278
Brazil	11,130
Korea, South	10,237
Israel	8,430
Sweden	6,830
Australia	5,687
Norway	5,687
Russia	5,389
Ireland	4,994
Czechia	4,587
Denmark	4,561
Chile	4,471
Poland	4,102
Romania	3,864
Malaysia	3,662
Ecuador	3,646
India	3,588
Philippines	3,246
Pakistan	3,157
Japan	3,139
Luxembourg	2,804
Saudi Arabia	2,402
Peru	2,281
Indonesia	2,273
Thailand	2,169
Finland	1,927
Serbia	1,908
Mexico	1,890
Panama	1,801
United Arab Emirates	1,799
Dominican Republic	1,745
Greece	1,735
South Africa	1,655
Qatar	1,604
Iceland	1,486
Colombia	1,485
Argentina	1,451
Algeria	1,320

Confirmed Cases per Country in Percentage	
country	% Confirmed
San Marino	0.79%
Andorra	0.65%
Luxembourg	0.46%
Iceland	0.42%
Spain	0.28%
Switzerland	0.25%
Italy	0.21%
Liechtenstein	0.21%
Monaco	0.19%
Belgium	0.17%
Austria	0.14%
France	0.14%
Germany	0.12%
Norway	0.11%
Portugal	0.11%
Ireland	0.10%
Netherlands	0.10%
USA	0.10%
Israel	0.09%
Denmark	0.08%
Estonia	0.08%
Iran	0.07%
Sweden	0.07%
United Kingdom	0.07%
Qatar	0.06%
Malta	0.05%
Slovenia	0.05%
Bahrain	0.04%
Canada	0.04%
Cyprus	0.04%
Czechia	0.04%
Panama	0.04%
Armenia	0.03%
Brunei	0.03%
Croatia	0.03%
Finland	0.03%
Latvia	0.03%
Lithuania	0.03%
Montenegro	0.03%
North Macedonia	0.03%
Serbia	0.03%
Turkey	0.03%
Antigua and Barbuda	0.02%
Australia	0.02%
Barbados	0.02%
Bosnia and Herzegovina	0.02%
Chile	0.02%
Dominican Republic	0.02%
Ecuador	0.02%
Greece	0.02%
Korea, South	0.02%
Mauritius	0.02%

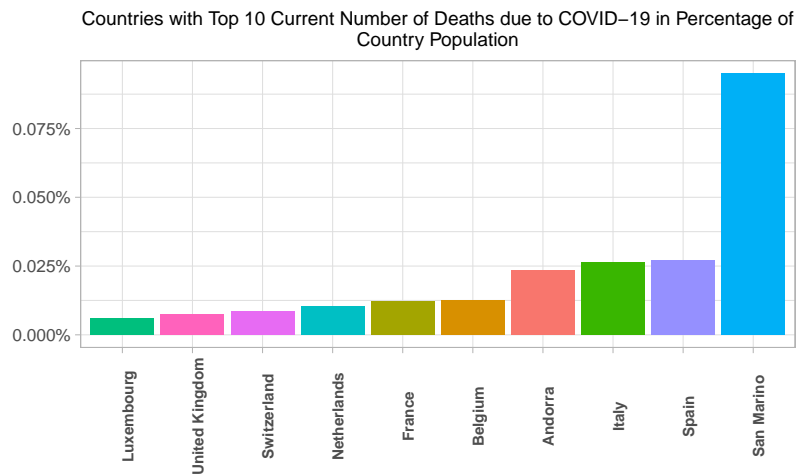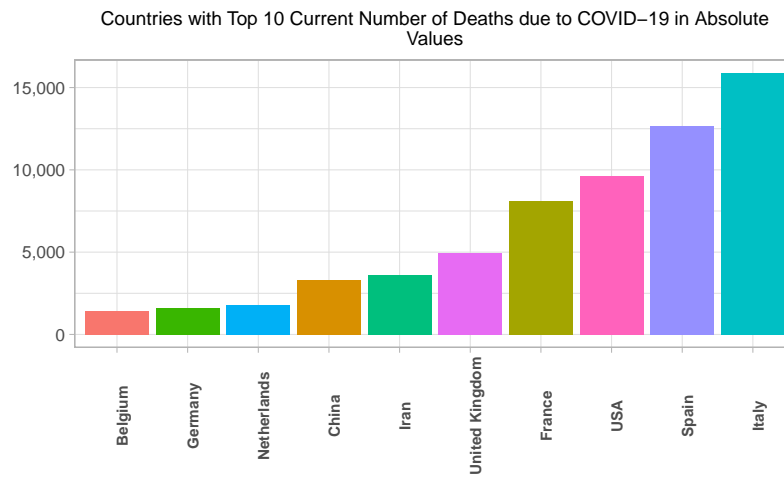Top 10 Current Number of Confirmed Cases per Country in Absolute Values
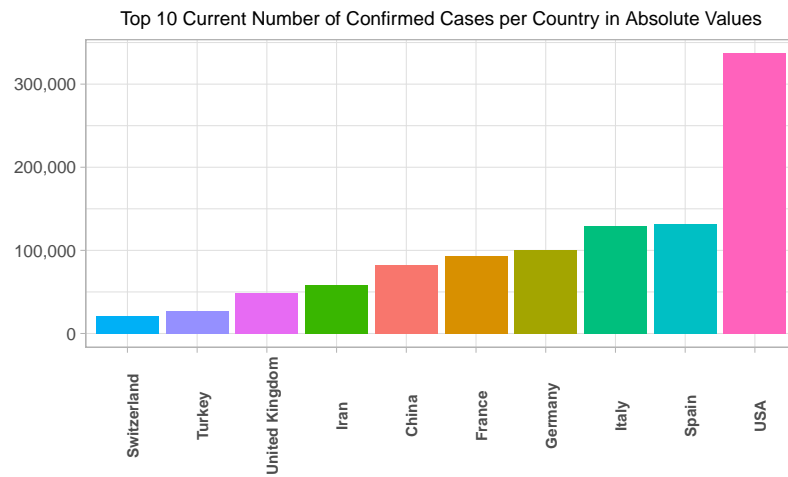
Comment: Thanks to these graphs (showing only the countries with the 10 highest percentage and absolute values in terms of confirmed cases) and the tables (showing all the countries), we can see that actually the situation changes quite a lot if we consider the percentage of the population or the absoulte values. In any case, as to 2020-04-06, we can see that, in absoulte values, the countries with the most confirmed cases are the USA, Spain and Italy, while in percentage the oustanding ones are San Marino and Andorra. Again, this is probably due to the difference in size of these countries. The USA are especially outstanding with regards to the number of confirmed cases, which is somehow a confirm of the news we hear everyday, and the fact that their president didn't take the situation very seriously from the beginning.

Deaths per Country (in descending order)	
country	deaths
Italy	15,887
Spain	12,641
USA	9,619
France	8,093
United Kingdom	4,943
Iran	3,603
China	3,333
Netherlands	1,771
Germany	1,584
Belgium	1,447
Switzerland	715
Turkey	574
Brazil	486
Sweden	401
Portugal	295
Canada	259
Austria	204
Indonesia	198
Korea, South	183
Ecuador	180
Denmark	179
Ireland	158
Algeria	152
Philippines	152
Romania	151
India	99
Poland	94
Peru	83
Dominican Republic	82
Mexico	79
Egypt	78
Japan	77
Greece	73
Norway	71
Morocco	70
Czechia	67
Iraq	61
Malaysia	61
Serbia	51
Israel	49
Pakistan	47
Panama	46
Russia	45
Argentina	44
Ukraine	37
Luxembourg	36
Australia	35
Colombia	35
Chile	34
Hungary	34
Saudi Arabia	34
San Marino	32

Deaths per Country relative to the population	
country	% deaths
San Marino	0.095%
Spain	0.027%
Italy	0.026%
Andorra	0.023%
Belgium	0.013%
France	0.012%
Netherlands	0.010%
Switzerland	0.008%
United Kingdom	0.007%
Luxembourg	0.006%
Iran	0.004%
Sweden	0.004%
Denmark	0.003%
Ireland	0.003%
Liechtenstein	0.003%
Monaco	0.003%
Portugal	0.003%
USA	0.003%
Austria	0.002%
Germany	0.002%
Albania	0.001%
Bahamas	0.001%
Bosnia and Herzegovina	0.001%
Canada	0.001%
Cyprus	0.001%
Czechia	0.001%
Dominican Republic	0.001%
Ecuador	0.001%
Estonia	0.001%
Finland	0.001%
Greece	0.001%
Guyana	0.001%
Iceland	0.001%
Israel	0.001%
Mauritius	0.001%
North Macedonia	0.001%
Norway	0.001%
Panama	0.001%
Romania	0.001%
Serbia	0.001%
Slovenia	0.001%
Trinidad and Tobago	0.001%
Turkey	0.001%
Afghanistan	0.000%
Algeria	0.000%
Angola	0.000%
Antigua and Barbuda	0.000%
Argentina	0.000%
Armenia	0.000%
Australia	0.000%
Azerbaijan	0.000%
Bahrain	0.000%

Countries with Top 10 Current Number of Deaths due to COVID−19 in Absolute Values

Countries with Top 10 Current Number of Deaths due to COVID−19 in Percentage of Country Population

Comment: The situation of the deaths due to COVID-19 is a little bit different from the one of the confirmed cases. The country with the most deaths in absoulte values is Italy, followed by Spain and the USA, while in percentage the highest value is the one of San Marino, followed not very closely by Spain and Italy. This indicates that the fatality of the COVID-19 has been severe in both Italy and Spain which has caused both countries to have very high values boths in absolute terms as well as relative to their populations, while for San Marino

the worrying high value is again mainly due to the fact that the country is really small and it has a small population (33671 inhabitants and 266 deaths due to COVID-19 as to 2020-04-06).

Now we look at the percentage of deaths over the confirmed cases per country.

_____
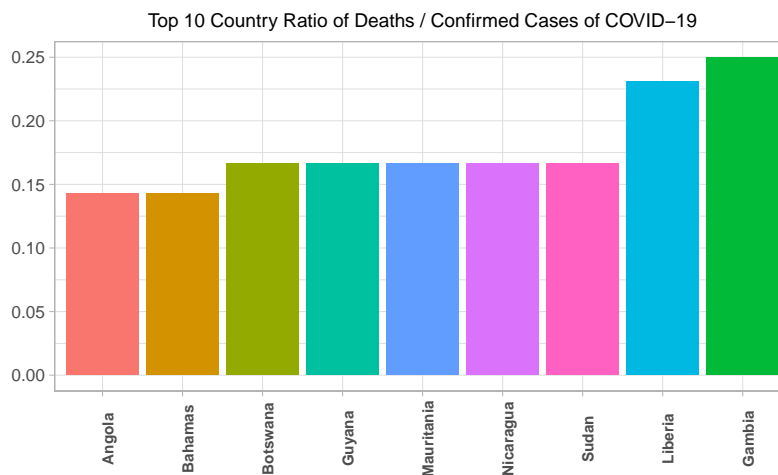
Comment: The countries with the highest percantage are Gambia and Liberia, these values are mainly due to the fact that there are only 4 and 13 cases of COVID-19 respectively, as to 2020-04-06, so, following the same reasoning as before, an increase of one victim here means a high increase in the ration of percentage over the confirmed cases. This opens an interesting discussion of whether this ratio is related to the GDP (a way to measure how rich the country is) or not. That is why we decided to plot the gdp against this ratio to see if there is any kind of correlation between the two variables.

Ratio of Deaths/Cases per Country	
country	ratio
Gambia	0.250
Liberia	0.231
Guyana	0.167
Mauritania	0.167
Nicaragua	0.167
Sudan	0.167
Botswana	0.167
Angola	0.143
Bahamas	0.143
Cabo Verde	0.143
Italy	0.123
San Marino	0.120
Congo (Kinshasa)	0.117
Algeria	0.115
Congo (Brazzaville)	0.111
Zimbabwe	0.111
Mali	0.111
Syria	0.105
Bangladesh	0.102
United Kingdom	0.102
Suriname	0.100
Netherlands	0.099
Spain	0.096
Indonesia	0.087
France	0.086
Honduras	0.082
Belgium	0.073
Morocco	0.069
Togo	0.068
Trinidad and Tobago	0.067
Egypt	0.066
Bolivia	0.064
Iraq	0.063
Iran	0.062
Sweden	0.059
Libya	0.056
Albania	0.055
Niger	0.054
Jamaica	0.052
Ecuador	0.049
Burkina Faso	0.049
El Salvador	0.048
Gabon	0.048
Haiti	0.048
Dominican Republic	0.047
Philippines	0.047
Ethiopia	0.047
Hungary	0.046
Tanzania	0.045
Venezuela	0.044
Brazil	0.044
Greece	0.042

Ratio of Deaths / Confirmed Cases of COVID−19 per Country GDP per Capita
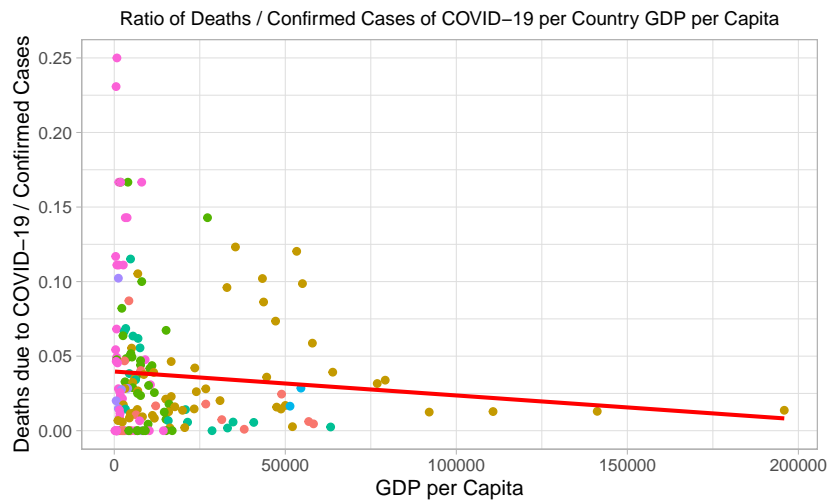
Comment: This economic indicator could be a proxy for the ability of the country to afford the medical emergence and to have a state-of-the-art healthcare system so that they are ready to deal with the alarming situation. However, it is not the only thing to take into account, the size of the country and the regulations taken by the government have also an impact (as we have seen earlier). As we can see here, it is not always the case to have a correlation between GDP and the ratio of deaths due to COVID-19 and the number over the confirmed cases.

We then decided to look at the evolution of the expansion of the virus, by considering the first confirmed case and the first death due to COVID-19 per country.

Note: Please note as the dataset has the first recording for China starting from 22/01/2020, **this is far away from the reality as the first cases were already announced in December**, hence the first Chinese confirmed case is not found the dataset and we have decided for the integrity of our exploratory analysis to ignore it. The same holds for countries such as Japan and Thailand who on the 22nd of January had 2 confirmed cases and not only 1, however, this may be because there were 2 people diagnosed at once or it is an accumulation over time however for simplicity we assumed the latter.

Comment:Considering only the countries for which we have the first confirmed case, the first country to have had a confirmed case of COVID-19 is, surprisingly, the US, the 22nd of January, and the day after Singapore, which is probably due to the high economic and business exchange with the other Asian countries. As we can see, the first countries were mainly in the South-Easth Asia, North America and some countries in Europe. Interestingly, we can notice a sort of
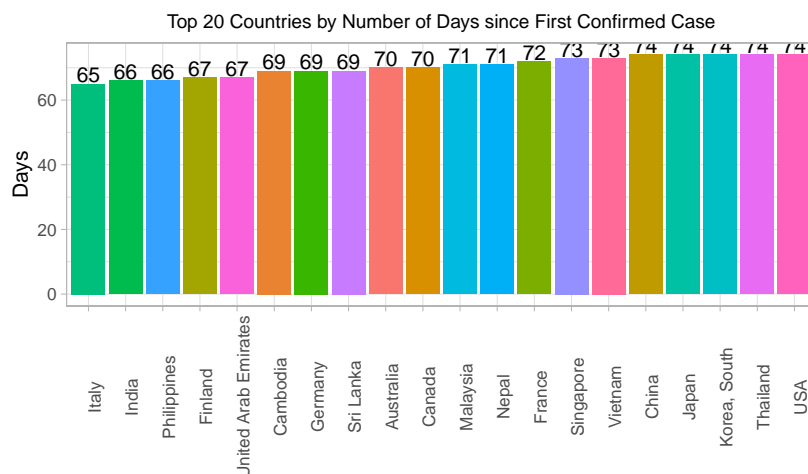
Timeline of First Case per Country		
date	country	region
2020-01-22	Korea, South	East Asia & Pacific
2020-01-23	Singapore	East Asia & Pacific
2020-01-27	Cambodia	East Asia & Pacific
2020-01-30	Philippines	East Asia & Pacific
2020-02-28	New Zealand	East Asia & Pacific
2020-03-09	Brunei	East Asia & Pacific
2020-03-10	Mongolia	East Asia & Pacific
2020-03-19	Fiji	East Asia & Pacific
2020-03-20	Papua New Guinea	East Asia & Pacific
2020-03-22	Timor-Leste	East Asia & Pacific
2020-01-27	Germany	Europe & Central Asia
2020-01-29	Finland	Europe & Central Asia
2020-01-31	Sweden	Europe & Central Asia
2020-02-01	Spain	Europe & Central Asia
2020-02-04	Belgium	Europe & Central Asia
2020-02-25	Croatia	Europe & Central Asia
2020-02-25	Switzerland	Europe & Central Asia
2020-02-26	Georgia	Europe & Central Asia
2020-02-26	Greece	Europe & Central Asia
2020-02-26	North Macedonia	Europe & Central Asia
2020-02-26	Norway	Europe & Central Asia
2020-02-26	Romania	Europe & Central Asia
2020-02-27	Denmark	Europe & Central Asia
2020-02-27	Estonia	Europe & Central Asia
2020-02-27	Netherlands	Europe & Central Asia
2020-02-27	San Marino	Europe & Central Asia
2020-02-28	Belarus	Europe & Central Asia
2020-02-28	Iceland	Europe & Central Asia
2020-02-28	Lithuania	Europe & Central Asia
2020-02-29	Ireland	Europe & Central Asia
2020-02-29	Luxembourg	Europe & Central Asia
2020-02-29	Monaco	Europe & Central Asia
2020-03-01	Armenia	Europe & Central Asia
2020-03-02	Andorra	Europe & Central Asia
2020-03-02	Latvia	Europe & Central Asia
2020-03-03	Ukraine	Europe & Central Asia
2020-03-04	Liechtenstein	Europe & Central Asia
2020-03-04	Poland	Europe & Central Asia
2020-03-06	Serbia	Europe & Central Asia
2020-03-06	Slovakia	Europe & Central Asia
2020-03-08	Moldova	Europe & Central Asia
2020-03-11	Turkey	Europe & Central Asia
2020-03-15	Uzbekistan	Europe & Central Asia
2020-03-22	Syria	Europe & Central Asia
2020-02-26	Brazil	Latin America & Caribbean
2020-02-28	Mexico	Latin America & Caribbean
2020-03-01	Dominican Republic	Latin America & Caribbean
2020-03-03	Argentina	Latin America & Caribbean
2020-03-03	Chile	Latin America & Caribbean
2020-03-06	Colombia	Latin America & Caribbean
2020-03-06	Costa Rica	Latin America & Caribbean
2020-03-06	Peru	Latin America & Caribbean

temporary void before the real boom of the virus, between the first days of February and mid-February. This could be probably due to the fact that the coutries had to develop a way to test the virus, plus, once it started to spread it has been hard to stop, so once some countries started to confirm the first case, the closest countries were quick to follow. The latest to confirm the first case of the virus has been South Sudan on the 2020-04-05.

––––––––––––––––––––

Comment: Regarding the first death due to COVID-19, the first country (again not considering China) announcing the first fatality due to the virus is on the 2020-02-02 in the Philippines, followed more than 10 days after by Japan and then France. Here the void between the countries, especially before March, is even more evident, and we can see that in March there is an exponential increase of the first deaths due to the virus. In general, the countries with the first fatalities are in the East Asia and Pacific area, with the exception of France.

Following, the graph of the number of days since the first confirmed case and first fatality due to COVID-19 per country.

Top 20 Countries by Number of Days since First Confirmed Case

Timeline of First Death due to COVID-19 per Country		
date	country	region
2020-02-02	Philippines	East Asia & Pacific
2020-02-13	Japan	East Asia & Pacific
2020-02-15	France	Europe & Central Asia
2020-02-20	Korea, South	East Asia & Pacific
2020-02-21	Italy	Europe & Central Asia
2020-02-29	USA	North America
2020-03-01	Australia	East Asia & Pacific
2020-03-01	Thailand	East Asia & Pacific
2020-03-03	San Marino	Europe & Central Asia
2020-03-03	Spain	Europe & Central Asia
2020-03-05	Switzerland	Europe & Central Asia
2020-03-05	United Kingdom	Europe & Central Asia
2020-03-06	Netherlands	Europe & Central Asia
2020-03-08	Argentina	Latin America & Caribbean
2020-03-08	Egypt	Middle East & North Africa
2020-03-09	Canada	North America
2020-03-10	Lebanon	Middle East & North Africa
2020-03-10	Morocco	Middle East & North Africa
2020-03-11	Albania	Europe & Central Asia
2020-03-11	Bulgaria	Europe & Central Asia
2020-03-11	Greece	Europe & Central Asia
2020-03-11	India	South Asia
2020-03-11	Indonesia	East Asia & Pacific
2020-03-11	Ireland	Europe & Central Asia
2020-03-11	Panama	Latin America & Caribbean
2020-03-11	Sweden	Europe & Central Asia
2020-03-12	Algeria	Middle East & North Africa
2020-03-12	Austria	Europe & Central Asia
2020-03-12	Guyana	Latin America & Caribbean
2020-03-12	Poland	Europe & Central Asia
2020-03-13	Azerbaijan	Europe & Central Asia
2020-03-13	Sudan	Sub-Saharan Africa
2020-03-13	Ukraine	Europe & Central Asia
2020-03-14	Denmark	Europe & Central Asia
2020-03-14	Luxembourg	Europe & Central Asia
2020-03-14	Slovenia	Europe & Central Asia
2020-03-15	Hungary	Europe & Central Asia
2020-03-16	Bahrain	Middle East & North Africa
2020-03-16	Guatemala	Latin America & Caribbean
2020-03-17	Brazil	Latin America & Caribbean
2020-03-17	Dominican Republic	Latin America & Caribbean
2020-03-17	Iceland	Europe & Central Asia
2020-03-17	Portugal	Europe & Central Asia
2020-03-17	Turkey	Europe & Central Asia
2020-03-18	Bangladesh	South Asia
2020-03-18	Burkina Faso	Sub-Saharan Africa
2020-03-18	Cuba	Latin America & Caribbean
2020-03-18	Moldova	Europe & Central Asia
2020-03-18	Slovakia	Europe & Central Asia
2020-03-19	Costa Rica	Latin America & Caribbean
2020-03-19	Croatia	Europe & Central Asia
2020-03-19	Jamaica	Latin America & Caribbean

Number of days since the first confirmed case	
country	days passed
China	74 days
Japan	74 days
Korea, South	74 days
Thailand	74 days
USA	74 days
Singapore	73 days
Vietnam	73 days
France	72 days
Malaysia	71 days
Nepal	71 days
Australia	70 days
Canada	70 days
Cambodia	69 days
Germany	69 days
Sri Lanka	69 days
Finland	67 days
United Arab Emirates	67 days
India	66 days
Philippines	66 days
Italy	65 days
Russia	65 days
Sweden	65 days
United Kingdom	65 days
Spain	64 days
Belgium	61 days
Egypt	51 days
Iran	46 days
Israel	44 days
Lebanon	44 days
Afghanistan	41 days
Bahrain	41 days
Iraq	41 days
Kuwait	41 days
Oman	41 days
Algeria	40 days
Austria	40 days
Croatia	40 days
Switzerland	40 days
Brazil	39 days
Georgia	39 days
Greece	39 days
North Macedonia	39 days
Norway	39 days
Pakistan	39 days
Romania	39 days
Denmark	38 days
Estonia	38 days
Netherlands	38 days
San Marino	38 days
Belarus	37 days
Iceland	37 days
Lithuania	37 days

Number of Days since First Fatality due to COVID−19 per Country

---
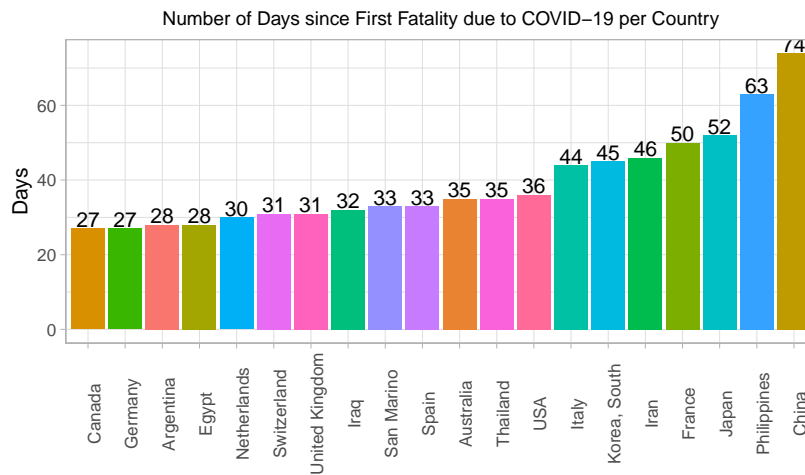
Comment: From the tables above, wan see that the countries that have had the first case of COVID-19 confirmed were the USA and South Korea, and interestingly, as we already saw before, there is a gap between Belgium and Egypt of 10 days and onther one of almost a week between Beligum and Lebanon, and from then on the spread has been increasing exponentially. In general, these graphs and tables are just a confirmation of what we have seen before.

2.2 Worldwide map

We then produce a worlwide map of the **COVID-19** spread at the latest date available in `covid19_data` for each country, and describe what we see.

We will use the `ggplot2` package particularly two useful commands of

- `map_data()` to retrieve a map,
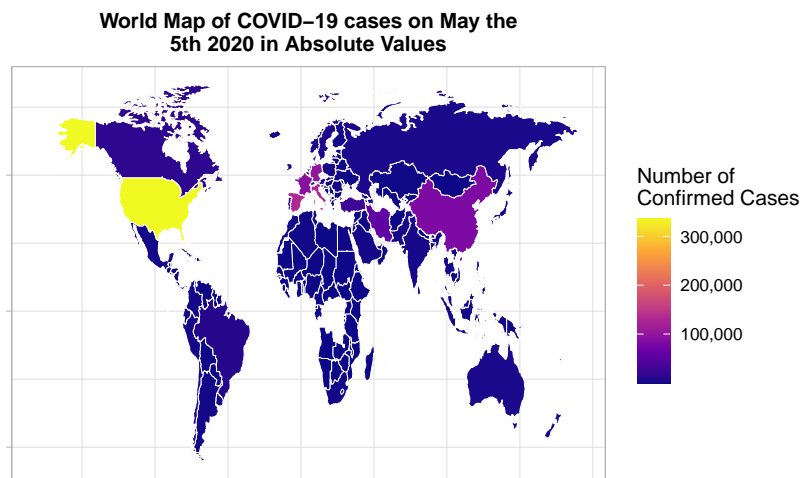- and `geom_map()` to draw a map on a plot.

We then use `expand_limits` to make sure we display the whole map of the world.

```
#> Warning in left_join(end_confirmed, world_map, by = "country"): Each row in `x` is expected to
#> i Row 1 of `x` matches multiple rows.
#> i If multiple matches are expected, set `multiple = "all"` to
#>   silence this warning.
#> Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
#> i Please use `linewidth` instead.
#> This warning is displayed once every 8 hours.
```

Number of days since the first death	
country	days passed
China	74 days
Philippines	63 days
Japan	52 days
France	50 days
Iran	46 days
Korea, South	45 days
Italy	44 days
USA	36 days
Australia	35 days
Thailand	35 days
San Marino	33 days
Spain	33 days
Iraq	32 days
Switzerland	31 days
United Kingdom	31 days
Netherlands	30 days
Argentina	28 days
Egypt	28 days
Canada	27 days
Germany	27 days
Lebanon	26 days
Morocco	26 days
Albania	25 days
Belgium	25 days
Bulgaria	25 days
Greece	25 days
India	25 days
Indonesia	25 days
Ireland	25 days
Panama	25 days
Sweden	25 days
Algeria	24 days
Austria	24 days
Guyana	24 days
Poland	24 days
Azerbaijan	23 days
Sudan	23 days
Ukraine	23 days
Denmark	22 days
Ecuador	22 days
Luxembourg	22 days
Norway	22 days
Slovenia	22 days
Hungary	21 days
Iceland	21 days
Bahrain	20 days
Guatemala	20 days
Brazil	19 days
Dominican Republic	19 days
Malaysia	19 days
Portugal	19 days
Turkey	19 days

```
#> Call `lifecycle::last_lifecycle_warnings()` to see where this
#> warning was generated.
#> Warning in left_join(end_confirmed, world_map, by = "country"): Each row in `x` is expected to
#> i Row 1 of `x` matches multiple rows.
#> i If multiple matches are expected, set `multiple = "all"` to
#>    silence this warning.
```


World Map of COVID–19 cases on May the 5th 2020 in Absolute Values

World Map of COVID–19 cases on May the 5th 2020 as % of Population
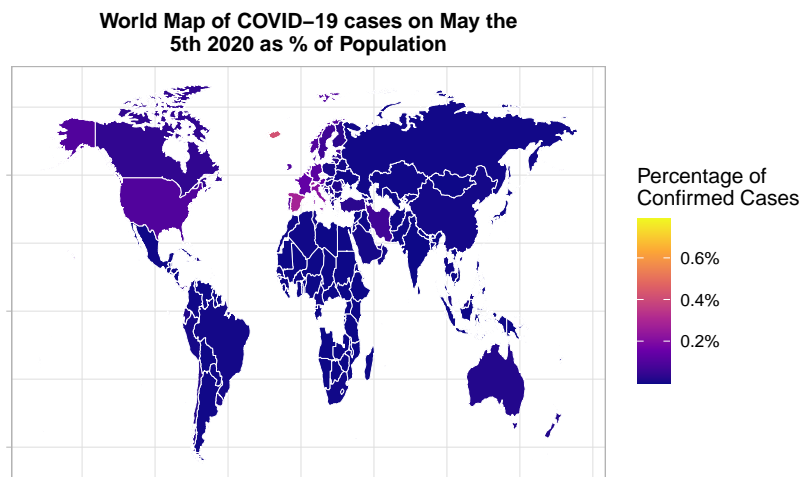
Comment: We can see that the countries with the most cases are the USA, Spain and Italy (warmer colors), in absolute values. In percentage the highest values are still in Europe, (Island, Spain, Switzerland and Italy are the one the warmest colors). This is perfectly in line with what we have seen up to now.

2.3   Data selection and alignment

To compare the speed of the infection spread between countries, we need to "align" the data. In other words, we model the epidemic using equivalent "starting conditions" for every country. To do that, we filter the data so that the number of confirmed cases (in any country) is greater or equal to the maximal number of confirmed cases at the first day of a specific country.

Let $C_{d,i}$ be the number of accumulated confirmed cases on day d for country i. Let country k be the one such that it had the highest number of reported cases on the first day in the dataset, that is $C_{1,k} \geq C_{1,i}$ for any other i. Find k and discuss.

From `covid19_data`, extract a new table `covid19_data_filtered`:

- Select only the countries i, which on some day $d_{0,i}$ have $C_{d_{0,i},i} \geq C_{0,k}$. We will call this time $d_{0,i}$ a day-zero for country i. In the next sections, we model $C_i(t) = C_{d_{0,i}+t,i}$, that is the spread of the epidemic in country i with t representing "event days". Remember, for every country the day-zero is, in general, different. However, when a country entered the epidemic stage, we are only interested in number of days that has passed from the date of entry (i.e. day-zero).
- First, we remove countries who are left with less than two weeks of data, i.e. we keep those countries whose number of days that has passed from the day-zero of this country is 14 or more.
- Then we create a new column called `t` representing the number of days from the day-zero of this country.
- Finally, we create a new column called `confirmed_1e5pop` representing the number of confirmed cases per 100,000 habitants. This is useful in order to compare how the spread of the epidemic differs between countries relative to their population.

Furthermore, we will aim to answer the following three questions:

- Which countries are left in `covid19_data_filtered`?
- What is the state of the spread there?
- What is the relationship between `t` and `confirmed_1e5pop`?

Comment: K is China, with 548 cases, 17 deaths on the 2020-01-22.

Comment: The countries left in `covid19_data_filtered` are Australia, Austria, Belgium, Brazil, Canada, Chile, China, Czechia, Denmark, Ecuador, Finland, France, Germany, Greece, Iceland, Indonesia, Iran, Ireland, Israel, Italy, Japan, Korea, South, Luxembourg, Malaysia, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Saudi Arabia, Spain, Sweden, Switzerland, Thailand, Turkey, United Kingdom, USA.

Comment: We can see which countries were the first ones to reach China's situation; South Korea the 23rd February (almost a month after China), Italy the 27th February and Iran the 29th of February. The evolution of the situation in the

Date in Which Countries Entered an Epidemic Situation	
country	date
China	2020-01-22
Korea, South	2020-02-23
Italy	2020-02-27
Iran	2020-02-29
France	2020-03-06
Germany	2020-03-06
Spain	2020-03-08
USA	2020-03-09
Japan	2020-03-10
Norway	2020-03-11
Switzerland	2020-03-11
Denmark	2020-03-12
Sweden	2020-03-12
Belgium	2020-03-13
Netherlands	2020-03-13
United Kingdom	2020-03-13
Austria	2020-03-14
Malaysia	2020-03-16
Australia	2020-03-18
Canada	2020-03-18
Brazil	2020-03-19
Czechia	2020-03-19
Ireland	2020-03-19
Portugal	2020-03-19
Israel	2020-03-21
Luxembourg	2020-03-21
Pakistan	2020-03-21
Turkey	2020-03-21
Chile	2020-03-22
Ecuador	2020-03-22
Finland	2020-03-22
Greece	2020-03-22
Iceland	2020-03-22
Poland	2020-03-22
Thailand	2020-03-22
Indonesia	2020-03-23
Romania	2020-03-23
Saudi Arabia	2020-03-23

countries that reached the state of China is given by the graph above. One of the biggest concerns is the situation in the US that seems to be growing exponentially and having no intention to slow, while in China the curve has flattened a lot following the regulations introduced by the government. Nevertheless, we do have to mention that the potential lack of transparency by the Chinese government could also explain this flat curve which falls beyond the scope of this analysis.

Comment: We can see that these two variables are positively correlated with one another, especially for Luxembourg and Iceland, while, on the other hand, for China is not at all the case. However, this make sense, since China is the reference country and it was already in the epidemic situation and was the first to take measurement to stop the expansion of the healthcase crisis.

Chapter 3

Modeling the spread of COVID-19 in a single country

3.1 The logistic model in R

Using the filtered dataset, we study the spread of COVID-19 with the logistic model. Letting $C_i(t) = C_{d_0,i+t,i}$, the model for country i can be expressed as:

$$C_i(t) = \frac{K_i \cdot C_i(0)}{C_i(0) + (K_i - C_i(0)) \exp(-R_i t)}$$

The goal is to find the final number of cases K_i and the infection rate R_i. We implement this in R using the following function:

```
# here, we assume that data is a data frame with two variables
#  - t: the number of event days
#  - confirmed: the number of confirmed cases at time t
logistic_model <- function(data) {
  data <- data %>% arrange(t)
  C_0 <- data$confirmed[1]
  C_max <- data$confirmed[nrow(data)]
  nls(
    formula = confirmed ~ K / (1 + ((K - C_0) / C_0) * exp(-R * t)),
    data = data,
    start = list(K = 2 * C_max, R = 0.5),
    control = nls.control(minFactor = 1e-6, maxiter = 100)
  )
```
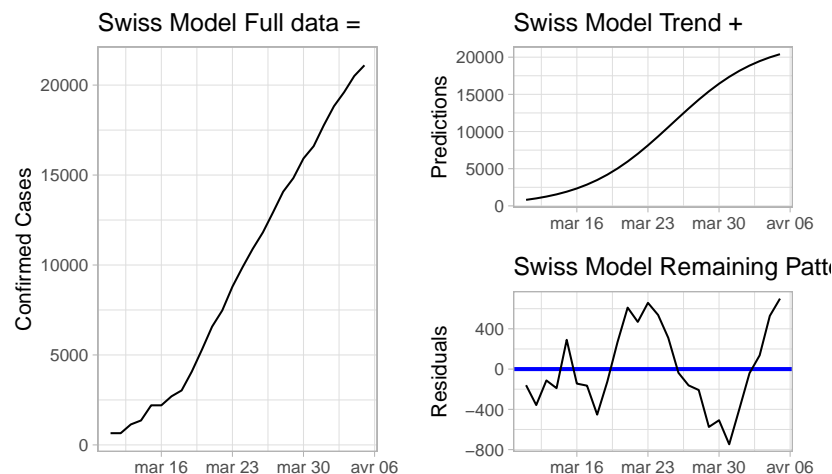
```
}
```

Notice:

- We use the nonlinear least square method `stats::nls` to fit the unknown parameters.
- In R, the formula above is `confirmed ~ K / (1 + ((K - C_0)/C_0) * exp(-R * t))`.
- As starting point (K_0, R_0) for the optimiser, we set $R_0 = 0.5$ and $K_0 = 2\,C(t^*)$, where t^* is the latest information about accumulated confirmed cases.
- We further set the `control` argument as `nls.control(minFactor = 1e-6, maxiter = 100)`.

3.2 The logistic model applied to data from Switzerland

From `covid19_data_filtered`, extract a table `covid19_ch` which corresponds to data for Switzerland. Then:

- We use the function above to fit the logistic model for Switzerland
- We describe its output (fitted parameters, `broom::tidy()` might be useful here)
- We discuss the goodness-of-fit.
- We plot the fitted curve, as well as observed data points.
- And finally, we present the predictions of the model. What is the estimated final size of the epidemic and infection rate in Switzerland?

We will start by presenting the data we own for the model we want to describe.

term	estimate	std.error	statistic	p.value
K	22253.429	345.696	64.4	0
R	0.227	0.002	96.3	0

Comment: In general we can see that there is an increasing trend for the confirmed cases of COVID-19 in Switzerland (which does not come as a surprise). The trend seems to follow quite well the pace of the observations, and the residuals seems to focus around zero, with some variation, especially towards the end of March.

Comment: we can see that the two parameters found are both statistically significant and both positively correlated to the independent variable.

Now, we will discuss the goodness of fit of the model by comparing it to a null model having the confirmed case explained by a constant variable.

```
#>
#> Call:
#> glm(formula = confirmed ~ 1, data = covid19_ch)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#>  -8998   -6868    -314    5999   11450
#>
#> Coefficients:
#>             Estimate Std. Error t value   Pr(>|t|)
#> (Intercept)     9650       1370    7.04 0.00000022 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for gaussian family taken to be 48822367)
#>
#>     Null deviance: 1220559163  on 25  degrees of freedom
#> Residual deviance: 1220559163  on 25  degrees of freedom
#> AIC: 537.1
#>
#> Number of Fisher Scoring iterations: 2
#> Analysis of Variance Table
#>
#> Model 1: confirmed ~ K/(1 + ((K - C_0)/C_0) * exp(-R * t))
#> Model 2: confirmed ~ 1
#>   Res.Df Res.Sum Sq Df      Sum Sq F value              Pr(>F)
#> 1     24    4200305
#> 2     25 1220559163 -1 -1216358858    6950 <0.0000000000000002 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment: The ANOVA results shows that the hypothesis that the two models have the same effect can be reject are 0.1% (since the p-value is lower than 0.001). Moreover, we also calulate the MAPE to measure the accuracy of the prediction of the model. Having a MAPE of 0.08, we can say that the model is fairly good, being the error small.

We then plot the fitted curve and the observations.
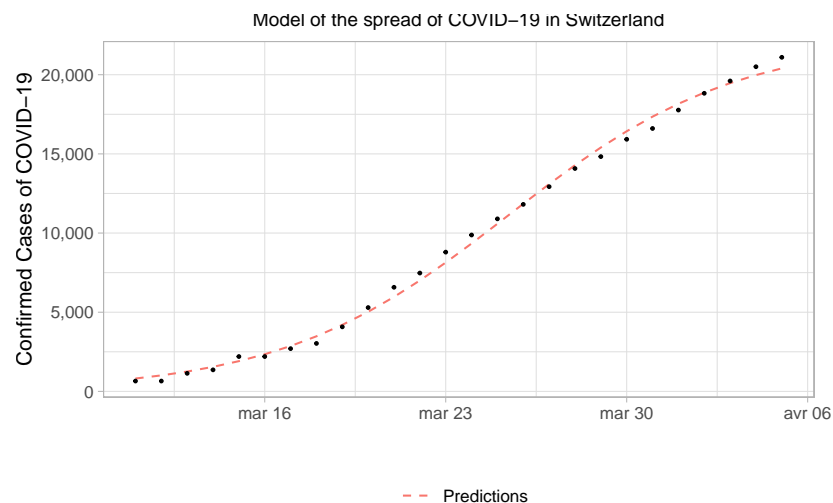
Comment: As we can see the predictions are following quite well the observations, meaning that the model is working quite well. However, it is woth mentioning that the model is fitted on the whole datasets, which could lead to some form of overfitting, not having a separation in training and test set, meaning that once dealing with new data, the prediction ability of the model could decrease quite a lot.

Residulas per Day of logistic_model
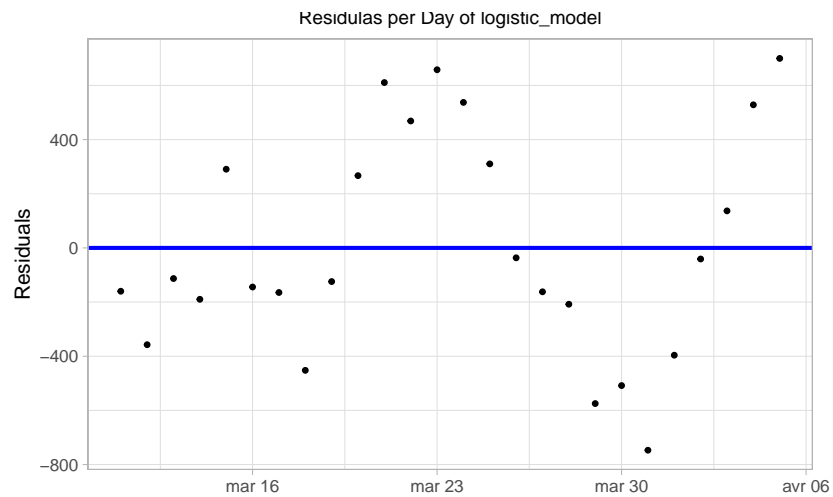
The residuals from the plot tell us that many actual numbers were lower than the predicted ones, however this is not too far off from zero, and it is rather expected that the model overpredicts at this point; the number of cases to be higher as the confirmed does not mean that all the people have been identified and due to the limitations in number of testing kits, perhaps more are infected than our data shows.

Eventually, the prediction the estimated final size of the epidemic and infection rate in Switzerland is 20400.238.

Chapter 4

Modeling the spread of COVID-19 worldwide

In this section, we fit the logistic model to every country in the `covid19_data_filtered` dataset.

## 4.1   Fitting the logistic model to every country

Here we make use of the nested data, list-columns and `logistic_model` to fit the logistic model to every country in the dataset. Because, for some countries, the optimization method might not converge, we will use the `possibly()` function to see which ones fail and which ones succeed. Now one may wonder, for which country does the optimization fail ?

First, let's fit the logistic model to every country and have a look at which one are not converging.

The countries for which the optimization fails are the following:

Denmark, Japan, Pakistan

We will also assess the goodness-of-fit of the logistic model in the various countries. Lets plot the residuals per country to have a look at the general trend of the residual in the model.

Residuals of logistic_model in General
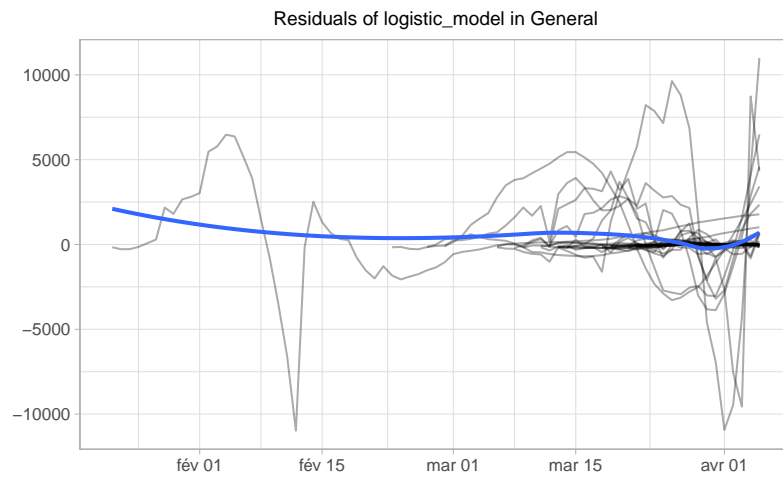
Comment: As we can see, the residuals seem to be quite close to zero, especially towards the end of the timeline (end of March, beginning of April), while at the beginning we can see some more distance from zero (given by the fact that the only country present in the dataset at that time is China, being the only one already in an epidemic situation). From the moment in which more countries join the dataset, we can see that the accuracy of the model increase, hence the residuals decrease and are closer to zero.

Let's plot the residual per country to have a better look, and then we will zoom on the countries with the highest values of the residuals (meaning that they are the ones for which the model has a lower accuracy in the predictions).

Residuals of logistic_model per Country (with Adjusted Scales per Country)
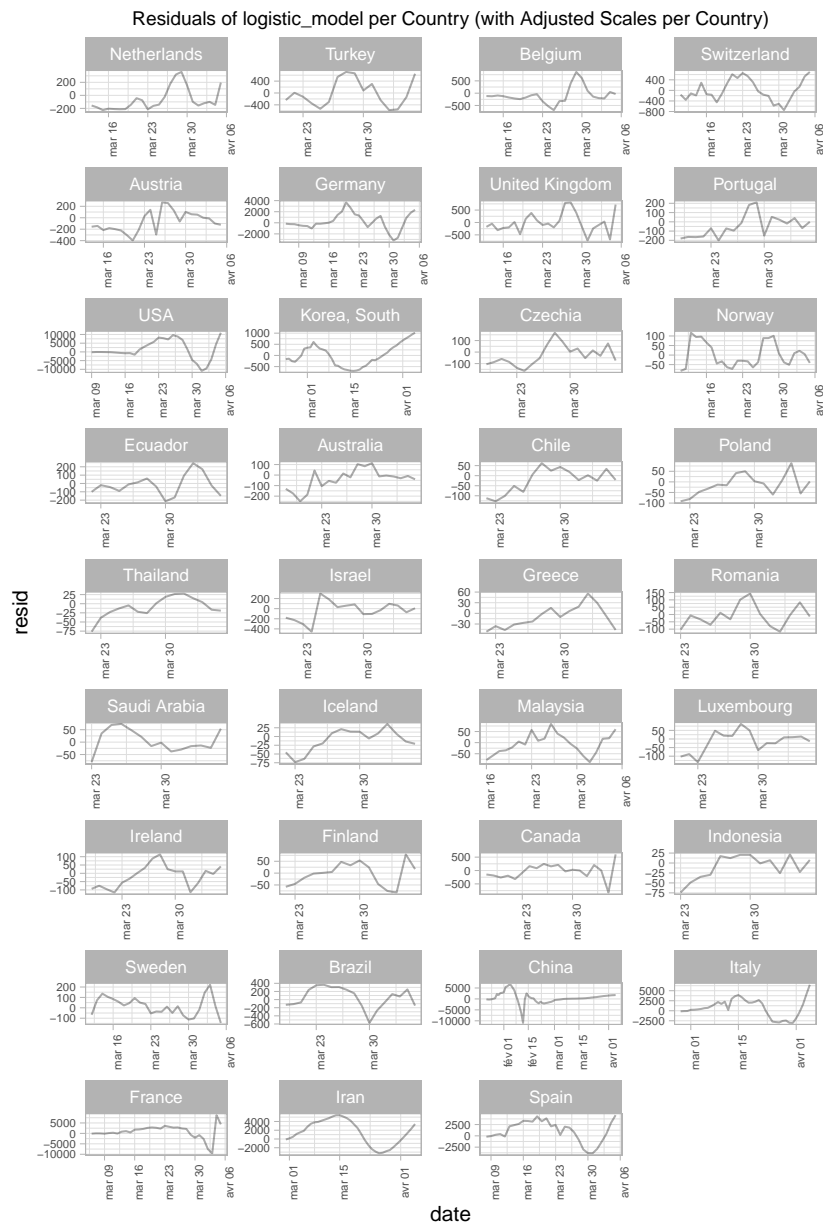
Note: The axis are for this graph `Residual of logistic_model per country` and the one below has been adjusted for each country in order to be able to see better the different distributions.

Zoom on the Countries With Higher Variation in Residuals of logistic_model with Adjusted Axis per Country
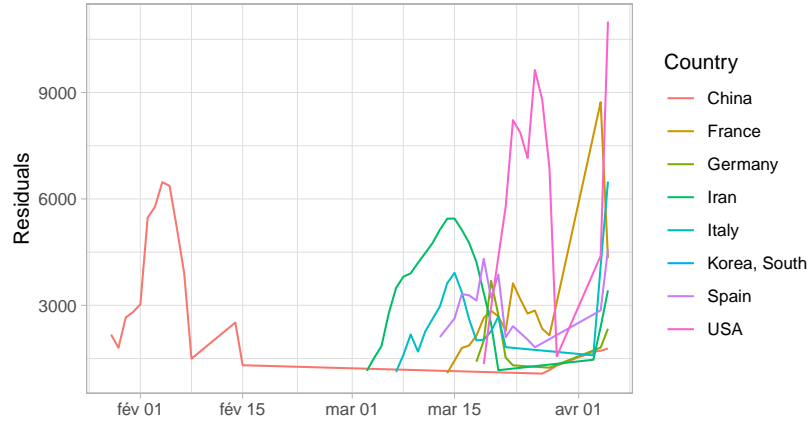
Comment: As we can see the countries for which the model has the highest values for the residuals are China, France, Germany, Iran, Italy, South Korea, Spain and the USA. This does not come as a susprise, actually, since in the EDA part of our analysis, we have seen that they are the countries with the highest exponentiality in terms of absolute values of the confirmed cases (especially the US), while for China, as already mentioned, the residuals are higher at the beginning of the timeline, being the only country in the dataset, since we are considering only the countries in an epidemic situation.

We would like to also have also a score which shows the goodness-of-fit of the model for each country, however according to the (Burnham and Anderson, 2002, pp. 80), it is mentioned that **AIC cannot** be used for models with different number of observations (and also different datasets) as we see in the table below.

4.2 Fitted parameters and long-term predictions

We then describe the fitted parameters (i.e., the final size and the infection rates), both on a per-country basis and some aggregate numbers (e.g., total size of the epidemic over all considered countries). Furthermore, we study the evolution (say for t from 0 to 50) of the predictions of the number of confirmed cases from our models. Similarly as was discussed in the last sub-section of the exploratory data analysis, the number of confirmed cases per 100,000 habitants is also important to understand how specific countries are managing the spread of the epidemic. Thus, we predict the evolution of this number (i.e., by dividing our predictions for confirmed cases by the population size) and discuss.

We will do the aforementioned using the following functions:

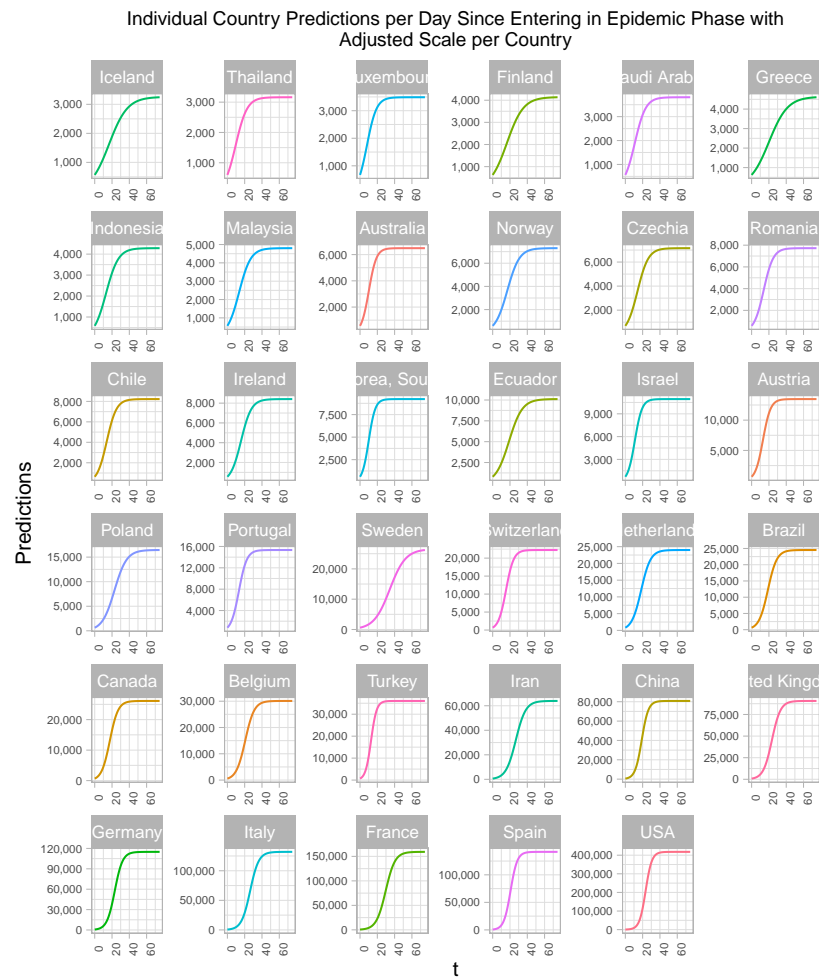| Number of observations per country	
country	observations
Australia	19
Austria	23
Belgium	24
Brazil	18
Canada	19
Chile	15
China	75
Czechia	18
Ecuador	15
Finland	15
France	31
Germany	31
Greece	15
Iceland	15
Indonesia	14
Iran	37
Ireland	18
Israel	16
Italy	39
Korea, South	43
Luxembourg	16
Malaysia	21
Netherlands	24
Norway	26
Poland	15
Portugal	18
Romania	14
Saudi Arabia	14
Spain	29
Sweden	25
Switzerland	26
Thailand	15
Turkey	16
USA	28
United Kingdom	24

- Format the fitted parameters using `broom::tidy()`.
- For the long-term predictions, we use `data = data.frame(t = 0:50)` in `add_predictions()`.

First we can see the parameters of the various models for each countries.

Note: Please do note that we have rounded our results to 2 decimal places.

Comment: It doesn't come as a surprise the fact that they vary quite a lot.

Now let's look at the prediction per country.

Individual Country Predictions per Day Since Entering in Epidemic Phase with Adjusted Scale per Country
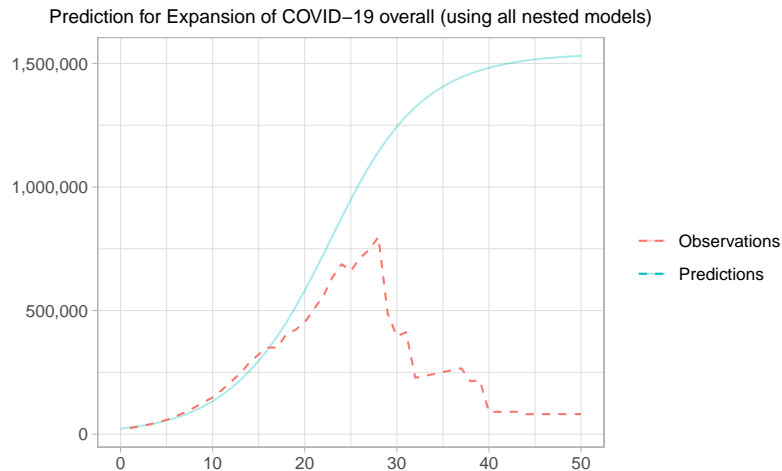
Comment: We can see that the countries with the highest absolute values are the US, Spain, Germany, France and Italy (over a 100'000 of final confirmed

Fitted parameter per Country country	K	R
Australia	6490	0.229
Austria	13475	0.227
Belgium	30106	0.192
Brazil	24528	0.194
Canada	26125	0.209
Chile	8226	0.178
China	80822	0.263
Czechia	7156	0.159
Ecuador	10116	0.131
Finland	4143	0.105
France	159261	0.185
Germany	114953	0.222
Greece	4636	0.093
Iceland	3258	0.094
Indonesia	4284	0.141
Iran	63936	0.175
Ireland	8411	0.167
Israel	10943	0.242
Italy	131962	0.201
Korea, South	9220	0.265
Luxembourg	3488	0.179
Malaysia	4802	0.148
Netherlands	24000	0.184
Norway	7300	0.143
Poland	16456	0.141
Portugal	15335	0.219
Romania	7725	0.180
Saudi Arabia	3811	0.159
Spain	141690	0.259
Sweden	26528	0.109
Switzerland	22253	0.227
Thailand	3162	0.151
Turkey	36048	0.311
United Kingdom	90683	0.201
USA	417895	0.280

cases predicted). This is in line with what we have found in the EDA part of our analysis.

In order to look at the aggregated sum, we can either use the logistic model on the filtered data and calculate new coefficients as we did with the swiss model but now applied to all countries and the second approach is to do it by region to use the fitted parameters of each country (nested) which makes more sense.

Prediction for Expansion of COVID−19 overall (using all nested models)

Comment: We can see that up to the 28th period the aggregated model predicts well however after that we do not have observations for most of the countries so it does not make sense to look at the observed values.

Note: Please do note that that the highest "t" belongs to China which is about 75 periods and therefore we have decided to extend our model to also include the all these dates. However, please do keep in mind that the 50th period the observation is mainly representative of predictions for China.

We can also do the aggregated sum for all countries only for the first 50 periods (because afterwards we do not have observed data points for almost all countries hence the observations naturally goes down). This model is a better one because it takes into account all the different coefficients rather than assigning the same one to all the countries.

Note: In the graph below, Please feel free to scroll over the country to see which one is contributing the most.

Comment: In terms of regions, we see the highest increase for North America due to the predicted increase for the US, followed by Europe & Central Asia, Italy, Spain and France among many and lastly, East Asian & Pacific with the most predicted cases for China.

Furthermore, referring back to the per-country predictions, we can also calculate the same for cases per 100,000 habitants displayed by the interactive plot below.

Comment: We can see that Iceland (green line) and Luxembourg (blue line) will have the highest number of confirmed cases per 100,000 habitants which is same as what we saw previously in the section of exploratory data analysis. This is due to their small populations and their infection numbers will be far larger than the countries that follow like Spain, Switzerland Belgium.

Bibliography

Batista, M. (2020). Estimation of the final size of coronavirus epidemic by the logistic model.

Burnham, K. P. and Anderson, D. R., editors (2002). *Information and Likelihood Theory: A Basis for Model Selection and Inference*, pages 49–97. Springer New York, New York, NY.

Martcheva, M. (2015). *An Introduction to Mathematical Epidemiology.* Springer, Boston, MA.