Predicting The Incidence Rate And Case Fatality Rate Of COVID-19

by
Mike Weltevrede (SNR 1257560)

A thesis submitted in partial fulfillment of the requirements for the degree of Master in Econometrics and Mathematical Economics.

Tilburg School of Economics and Management
Tilburg University

Supervised by:
dr. Otilia Boldea

Second reader:
dr. George Knox

Date:
May 31, 2020

Contents

1 Acknowledgements

2 Introduction

3 Problem description

In this section, we will elaborate on the methods that we apply in order to model the epidemiological spread of SARS-CoV-2. We are basing our model on specifications as used by Adda (2016). In his paper, Adda investigates the spread of several viral diseases in the past, namely influenza, gastroenteritis, and chickenpox. He starts from the Standard Inflammatory Response (SIR) model, one of the most commonly used, if not the most used, model in epidemiology (Kermack & McKendrick, 1927; Anderson & May, 1992).

The SIR model splits a population into three groups. Let S be the fraction of individuals who are susceptible to being infected, let I be the fraction of individuals who are currently infected, and let R be the fraction of individuals who have recovered but who are still immune. As such, at any point in time, we have that

$$S, I, R \in [0,1] \text{ and } S + I + R = 1.$$

The SIR model is then postulated in continuous time, i.e. the equations in (1) depict the change in the variables S, I, and R for one time period ahead. This type of model is also called a stock-and-flow model because there is a certain stock (for instance the number of infected persons) to which a flow is added or subtracted (for instance the change in the number of infected persons).

$$\begin{aligned}
\frac{dI}{dt} &= \alpha SI - \beta I \\
\frac{dR}{dt} &= \beta I - \lambda R \\
\frac{dS}{dt} &= -\alpha SI + \lambda R
\end{aligned} \tag{1}$$

We can interpret the SIR model as follows: the equation on $\frac{dI}{dt}$ states that the stock of infected people I is increased by a fraction α of the susceptible population, which become infected because of the currently infected people, and is decreased by a fraction β of the infected people, which recover. Note that, therefore, the quantity β^{-1} is the average infectious period (the time that a person stays infected, on average). The equation on $\frac{dR}{dt}$ tells us that the stock of recovered and immune people R is increased by the patients who recover, as described before, and it is decreased by a fraction λ of the recovered patients since these lose their immunity. For instance, Adda mentions that λ is set to 0 for chickenpox as individuals acquire a lifetime immunity while λ will be high for gastroenteritis due to almost no immunity emerging. We will tackle the issue of immunity in the case of COVID-19 later in section 4. Lastly, the equation on $\frac{dS}{dt}$ describes that the stock of susceptible individuals

Explain more later on immunity since this is currently still researched

S is decreased by the fraction of people that become infected. It is increased by the individuals that have lost their immunity. An important addition that Adda makes, is recognizing that there is spatial spillover between regions. That is, there may be infected people in one region that travel to another region and then infect individuals there. As such, the number of new cases would be modelled as $\alpha_{within} SI + \alpha_{between} S\tilde{I}$ where $\tilde{I}$ is the fraction of infected individuals from outside the region of interest who meet susceptible people from within the region. Clearly, this is an important addition to the model and we acknowledge and incorporate this in this thesis.

One of the main measures resulting from the SIR model is the estimation of the basic reproduction number $R_0 \coloneqq \alpha/\beta$. An epidemic is said to develop if $R_0 > 1$. In the same sense, this measure is widely used to indicate that an ongoing epidemic is dying out if R_0 drops below 1. For instance, the Italian health ministry has posted an article on May 9, 2020 stating that the R_0 reproduction rate for COVID-19 is currently below one in Italy, at between 0.5 and 0.7 (Ministero della Salute, 2020), showing that this measure is also used to communicate to citizens.

Later in the paper, Adda presents his full econometric model as follows:

$$
\begin{aligned}
I_{r,t} = {} & I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha^k_{within} W^k_{r,t-lag} \\
& + \sum_{c \neq r} I_{c,t-lag} S_{r,t-lag} \sum_{k=1}^{\tilde{K}} \alpha^k_{between} \widetilde{W}^k_{r,c,t-lag} \\
& + X_{r,t}\delta + \eta_{r,t}
\end{aligned} \tag{2}
$$

There are several notes to place about this model. Firstly, notice the subscript τ. This lag indicates the length of the incubation period, being the period between an infection and the moment that the infected individual starts showing symptoms. For COVID-19, there is an ongoing discussion on this. Secondly, (Adda, 2016) puts weights W and $\widetilde{W}$ on the parameters α_{within} and $\alpha_{between}$. The weighting matrix W consists of region-specific variables that may have an effect on the transmission rate within that region. On the other hand, the weighting matrix $\widetilde{W}$ is made up of variables that may influence the transmission rate between two regions r and c. In section 4 we will explain more about how these matrices are constructed in this thesis.

Explain more later since this is currently still researched. We will include this in this section too, as we have tested multiple

BELOW HERE: OLD TEXT

to model the incidence rate $Inc_{r,t}$ for several viruses, being the percentage of the population in a region r who have the virus at a time t:

$$\begin{aligned} Inc_{r,t} = {} & Inc_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} a^k_{within} W^k_{r,t-lag} \\ & + \sum_{c \neq r} Inc_{c,t-lag} S_{r,t-lag} \sum_{k=1}^{\tilde{K}} a^k_{between} \widetilde{W}^k_{r,c,t-lag} \\ & + X_{r,t}\delta + \eta_{r,t} \end{aligned} \tag{3}$$

Adda models the susceptible population as the total population who currently do not have the virus and who are not immune. That is, a certain proportion of immune people lose their immunity and become susceptible again. At this point, we will assume that all recovered patients achieve immunity. This assumption can be challenged because it is currently still unknown whether immunity is always achieved, especially among those who have had only light to medium symptoms. However, it is estimated that COVID-19 antibodies will remain in a patient's system for two to three years, based on what is known about other coronaviruses, but it is too early to know for certain (Leung, 2020). As such, we believe our assumption is generally valid.

TODO: update to our current setting

That is, let S denote the fraction of individuals who are susceptible to contracting the disease, I the fraction of individuals who are infected, and R the fraction of individuals who have recovered but are still immune. Then:

$$\begin{cases} \frac{dI(t)}{dt} = \alpha S(t) I(t) - \beta I(t) \\ \frac{dR(t)}{dt} = \beta I(t) - \lambda R(t) \\ \frac{dS(t)}{dt} = -\alpha S(t) I(t) + \lambda R(t) \end{cases}$$

Notice that Adda also models interaction between regions using the matrix $\widetilde{W}_{r,c}$. At first, we will neglect interactions between regions. The model becomes:

4 Dataset

The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU and the UK (?, ?). In this thesis we focus our attention on Italy, the epicenter of coronavirus cases in Europe. Italy consists of 21 so-called *regioni* (regions), comparable to the Dutch provinces. These constitute the second-level NUTS regions (also called NUTS 2 regions), where the region of *Trentino-Alto Adige* is split into two regions: *Provincia Autonoma di Bolzano/Bozen* and *Provincia Autonoma di Trento*. Italy's first-level NUTS regions are defined as groups of regions, namely *North West*, *North East*, *Centre*, *South*, and *the Islands*. The third-level NUTS regions are 110 provinces, which are subregions of the *regioni*.

Data was gathered from various sources. The specific information on the coronavirus in Italian regions was retrieved from the Ministero della Salute (the Italian ministry of health services), who publish daily reports under a title similar to *Covid-19, i casi in Italia 17 aprile ore 18*, where *17 aprile* would be updated to the relevant date (?, ?). These daily reports were posted with tables containing the following information per region:

- Hospitalized with symptoms *(Ricoverati con sintomi)*
- Active intensive care patients *(Terapia intensiva)*
- Home isolated active cases *(Isolamento domiciliare)*
- Total number of active cases *(Totale attualmente positivi)*
- Dismissed/recovered *(Dimessi/guariti)*
- Deceased *(Deceduti)*
- Total confirmed cases *(Casi totali)*
- Increase in total confirmed cases - compared to the previous day *(Incremento casi totali - rispetto al giorno precedente)*
- Total amount of tests executed *(Tamponi)*

It should be noted that the death statistics for Italy do not include the total amount of coronavirus victims who died outside hospitals, including dozens who died in different nursing homes across the country. Therefore, the official death statistics are considered an underestimate (?, ?). However, we did not model the amount of deaths or the death rate, so this does not impact our analysis. Nonetheless, we do take into account an underestimation of other information. For instance, not all people infected with COVID-19 are tested. These are not only symptomatic patients but also asymptomatic Italians. Moreover, it is unclear how the government collects this information. If regions or provinces submit this information to the government each day, there may be provinces who fail to submit their data for a certain day. Despite this, we assume that this official information is accurate and representative of the region itself.

Two issues that we want to address are missing data and the correction of data. In the official publications that we use, data that was wrongly published on a day $t-1$ is corrected by subtracting the error from the cases from day t. As such, if the error is larger than the number of new cases, the reported amount of new cases is negative. It happened five times that a number was reported to be negative. Since negative numbers are not logical in the context of confirmed cases or deaths, we correct this by subtracting the error from the day before and set the previously negative number to 0. For non-negative corrected numbers, we do not have a way to detect which these are and we cannot reasonably assume how this number should be split up among day t and $t+1$.

Regarding missing data, there are only three cases, namely for Abruzzo on March 10, Puglia on March 16, and Campania on March 18. Given that faulty data is also corrected as described before, we assume that the cases missing on day t are added to those of day $t+1$. This is confirmed by higher values compared to the expected trend, as seen in Table 1. As such, missing data is simply imputed with a value of 0.

	Abruzzo	Puglia	Campania
Day $t-1$	8	64	60
Day $t+1$	46	110	192
Day $t+2$	5	43	97

Table 1: Number of confirmed cases around a day t with missing data

Regressors were obtained from Eurostat, which is the statistical office of the European Union. Statistical data, broken down to the three NUTS levels, are published on their website (?, ?). The data can be freely filtered according to time period, geolocation (being the NUTS regions), and other aspects depending on the data, such as sex, age, or the unit of measure. The specification of the regressors we used can be found in Table 2.

Continue here - fill in the descriptions

Regressor	Description	Unit of measure
air_passengers_arrived	x	Number
air_passengers_departed	x	Number
tourist_arrivals	x	Number
broadband_access	x	Percentage of population
death_rate_diabetes	x	Number per 100,000 inhabitants
death_rate_influenza	x	Number per 100,000 inhabitants
death_rate_chd	x	Number per 100,000 inhabitants
death_rate_cancer	x	Number per 100,000 inhabitants
death_rate_pneumonia	x	Number per 100,000 inhabitants
available_beds	x	Number
maritime_passengers_disembarked	x	Number
maritime_passengers_embarked	x	Number
risk_of_poverty_or_social_exclusion	x	Percentage of population
weekend	x	Binary indicator
weekNumber	x	Number

Table 2: Specification of regressors

We need to make sure that there is no large correlation between regressors. Specifically, we concur that there are people who often have multiple diseases at the same time.

	Diabetes	Respiratory	Hypertension	Cancer	CHD	Pneumonia	TB
Diabetes		0.14	0.07	0.15	-0.23	0.36	0.20
Respiratory	0.14		0.07	0.71	-0.45	0.69	-0.09
Hypertension	0.07	0.07		0.11	0.19	0.02	-0.09
Cancer	0.15	0.71	0.11		-0.02	0.64	0.18
CHD	-0.23	-0.45	0.19	-0.02		-0.40	0.13
Pneumonia	0.36	0.69	0.02	0.64	-0.40		-0.02
TB	0.20	-0.09	-0.09	0.18	0.13	-0.02	

Note, the following is old and is for the specification of Adda. The specification of W will likely be in X for the other models.

The spatial weighting matrix W_r has the following structure:

$$W_r = \begin{bmatrix} V_r & C_r \end{bmatrix},$$

where V_r consists of K_V time-varying regressors and C_r consists of K_C time-constant regressors, so $V_r \in \mathbb{R}^{T \times K_V}$ and $V_r \in \mathbb{R}^{T \times K_C}$. Taking an example:

$$W_r = \begin{bmatrix} V_r^{\text{schools closed}} & V_r^{\text{lockdown started}} & C_r^{\text{hospital beds}} & C_r^{\text{internet access}} \end{bmatrix}.$$

We note that the descriptive data (like demographics and economic data) that we use is assumed to be time-constant during the coronacrisis (due to lack of data). The time-varying information that we use consists binary indicators for whether certain policies (such as closing down schools or instigating a lockdown) were implemented. As such, W_r mostly contains time-constant information.

We will use the following specifications for the weights and regressors:

- $W_{r,t-lag}$ contains $K \coloneqq K_V + K_C$ region-specific variables that potentially influence the transmission rate of SARS-CoV-2 within a region r. We split these in several categories:

  Economic

  - The amount of freight being transported by plane from and to the region (not available interregionally).
  - The amount of freight being transported by ship from and to the region (not available interregionally).
  - The amount of arrivals at tourist accommodations.
  - The GDP at current market prices per inhabitant.
  - The disposable income per inhabitant.
  - The amount of journeys made for transport of freight by road by loading and unloading region.

  Demographics, social, etcetera

  - The area size.
  - The median age and median age squared.
  - The population number.
  - The percentage of people at risk of poverty or social exclusion.
  - The percentage of people with broadband access.
  - The percentage of people who used internet to contact the public authorities in the last year.

- The percentage of people that attained a certain education level.

Medical

- The average length-of-stay in a hospital.
- The crude death rate for several different diseases.
- The number of health personnel (doctors and nurses).
- The number of hospital beds.

Travelling

- The number of passengers travelling by plane from and to the region (not available interregionally).
- The number of passengers travelling by ship from and to the region (not available interregionally).
- The length of railroads, motorways, navigable rivers, etcetera.

- $X_{r,t}$ contains certain fixed effects to control for, such as a binary indicator whether the day was on a weekend.

When we will also consider interactions between regions, we will define $\widetilde{W}_{r,t-lag}$ to contain $\tilde{K}$ variables that potentially influence the transmission rate of SARS-CoV-2 across regions:

- Amount of passengers that travelled from region c to region r via railroad.
- Amount of freight that travelled from region c to region r via railroad.
- A binary indicator indicating whether the regions border each other.
- The distance between the largest (most populous) cities in the regions.
- The population ratios.
- The log regional GDP ratios.

5 Results

Table 3: Estimates from the Least Square Dummy Variable Regression with standard errors in parentheses.
†: This variable enters the model with a lag and is multiplied by lags of the incidence and susceptibility rate, as described before.

	Lag 1	Lag 2	Lag 5
(Intercept)	6.739×10^{-6}	-3.376×10^{-6}	-9.059×10^{-6}
	(5.836×10^{-6})	(5.106×10^{-6})	(5.772×10^{-6})
weekend1	3.605×10^{-6}	4.438×10^{-6}	4.756×10^{-6}

	Lag 1	Lag 2	Lag 5
	(1.861×10^{-6})	(1.807×10^{-6})	(1.9×10^{-6})
weekNumber	1.109×10^{-6}	1.307×10^{-6}	1.281×10^{-6}
	(2.14×10^{-7})	(2.08×10^{-7})	(2.194×10^{-7})
BAS	-2.968×10^{-5}	-2.514×10^{-5}	-9.136×10^{-6}
	(6.966×10^{-6})	(6.89×10^{-6})	(7.419×10^{-6})
BZ	2.1×10^{-5}	1.066×10^{-5}	1.686×10^{-5}
	(7.07×10^{-6})	(6.31×10^{-6})	(6.806×10^{-6})
CAL	-3.069×10^{-5}	-2.067×10^{-5}	-4.171×10^{-6}
	(7.176×10^{-6})	(6.111×10^{-6})	(7.031×10^{-6})
CAM	-3.533×10^{-5}	-1.361×10^{-5}	-1.035×10^{-5}
	(6.719×10^{-6})	(6.447×10^{-6})	(7.596×10^{-6})
EMR	9.301×10^{-6}	2.795×10^{-5}	1.688×10^{-5}
	(7.312×10^{-6})	(5.929×10^{-6})	(6.913×10^{-6})
FVG	-2.682×10^{-5}	-1.808×10^{-5}	4.857×10^{-6}
	(7.634×10^{-6})	(6.388×10^{-6})	(6.905×10^{-6})
LAZ	-1.259×10^{-5}	-6.242×10^{-6}	-5.802×10^{-6}
	(6.798×10^{-6})	(6.967×10^{-6})	(7.076×10^{-6})
LIG	-1.884×10^{-5}	2.89×10^{-5}	2.268×10^{-5}
	(7.355×10^{-6})	(6.131×10^{-6})	(6.594×10^{-6})
LOM	1.63×10^{-5}	1.333×10^{-5}	2.955×10^{-5}
	(7.829×10^{-6})	(6.665×10^{-6})	(7.059×10^{-6})
MAR	-1.862×10^{-5}	-9.392×10^{-6}	1.541×10^{-6}
	(7.898×10^{-6})	(7.016×10^{-6})	(7.669×10^{-6})
MOL	-2.356×10^{-5}	-3.478×10^{-6}	-5.549×10^{-8}
	(7.436×10^{-6})	(7.236×10^{-6})	(7.106×10^{-6})
PIE	1.728×10^{-5}	-8.308×10^{-7}	2.19×10^{-5}
	(6.833×10^{-6})	(6.67×10^{-6})	(7.344×10^{-6})
PUG	-1.889×10^{-5}	-1.028×10^{-6}	-1.121×10^{-5}
	(7.68×10^{-6})	(6.175×10^{-6})	(7.769×10^{-6})
SAR	-3.666×10^{-5}	-6.784×10^{-6}	-1.514×10^{-5}
	(6.705×10^{-6})	(6.973×10^{-6})	(7.98×10^{-6})
SIC	-3.029×10^{-5}	-2.378×10^{-5}	-1.99×10^{-5}
	(6.655×10^{-6})	(5.977×10^{-6})	(7.24×10^{-6})
TN	1.564×10^{-5}	4.167×10^{-5}	5.44×10^{-5}
	(6.889×10^{-6})	(6.043×10^{-6})	(6.866×10^{-6})
TOS	-7.054×10^{-6}	-1.988×10^{-5}	8.216×10^{-6}
	(7.411×10^{-6})	(6.174×10^{-6})	(7.662×10^{-6})
UMB	-3.252×10^{-5}	-7.968×10^{-6}	-2.448×10^{-5}
	(7.247×10^{-6})	(6.657×10^{-6})	(6.71×10^{-6})
VDA	3.513×10^{-5}	1.987×10^{-6}	5.399×10^{-5}
	(7.017×10^{-6})	(6.642×10^{-6})	(6.797×10^{-6})
VEN	-2.298×10^{-7}	4.538×10^{-6}	-2.134×10^{-5}
	(7.077×10^{-6})	(6.021×10^{-6})	(6.991×10^{-6})
airPassengersArrived†	7.041	5.774	5.279

	Lag 1	Lag 2	Lag 5
	(1.52)	(1.476)	(1.553)
touristArrivals†	39.31	30.63	13.32
	(4.7)	(4.563)	(4.802)
broadbandAccess†	-0.2785	-0.04844	-0.1411
	(0.02367)	(0.02298)	(0.02417)
dischargeRateDiabetes†	-31.28	-29.85	-7.975
	(4.806)	(4.666)	(4.909)
dischargeRateRespiratory†	-165.9	-119.6	-101.7
	(14.48)	(14.06)	(14.79)
dischargeRateHypertension†	36.23	0.6688	16.28
	(5.398)	(5.241)	(5.514)
dischargeRateCancer†	-45.97	74.58	5.827
	(11.91)	(11.56)	(12.16)
dischargeRateChd†	-6.828	-30.72	-11.71
	(2.563)	(2.489)	(2.618)
dischargeRatePneumonia†	183.8	69.82	102.6
	(14.03)	(13.62)	(14.33)
dischargeRateTB†	32.97	7.312	20.37
	(4.853)	(4.712)	(4.958)
availableBeds†	-16.5	-33.03	5.219
	(2.946)	(2.86)	(3.01)
maritimePassengersDisembarked†	1.79	-0.02792	6.774
	(1.768)	(1.717)	(1.806)
riskOfPovertyOrSocialExclusion†	0.144	0.1581	0.1545
	(0.0107)	(0.01039)	(0.01093)
railTravelers†	3.885	10.95	-9.397
	(1.767)	(1.716)	(1.805)
medianAge†	0.4081	0.04353	0.1563
	(0.04461)	(0.04332)	(0.04557)

6 Conclusion

7 Future research

References

Adda, J. (2016). Economic activity and the spread of viral diseases: Evidence from high frequency data. *The Quarterly Journal of Economics*, *131*(2), 891–941.

Anderson, R. M., & May, R. M. (1992). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, *115*(772), 700–721.

Leung, H. (2020, Apr). *What we know about coronavirus immunity and reinfection.* Time Magazine. Retrieved from `https://time.com/5810454/coronavirus-immunity-reinfection/`

Ministero della Salute. (2020, May). *Coronavirus: Contagion rate r0 below 1. prudence needed in phase two says iss.* Retrieved from `http://www.salute.gov.it/portale/news/p3_2_1_1_1.jsp?lingua=italiano&menu=notizie&p=dalministero&id=4717`

A Tables