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:
June 12, 2020

Abstract

TODO

Acknowledgements

TODO

Contents

1 Introduction

2 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 and the disease it causes: COVID-19. We are basing our model on specifications as used by Adda (2016). In the paper, Adda (2016) investigates the spread of several viral diseases in the past, namely influenza, gastroenteritis, and chickenpox. The key additions made are, firstly, that a spatial spillover effect is considered and, secondly, that we allow for some sort of weighting on the parameters on the basis of region specific variables. Adda (2016) starts from the Standard Inflammatory Response (SIR) model, the most commonly used model in epidemiology (Kermack & McKendrick, 1927; Anderson & May, 1992).

The SIR model splits the total population into three groups. S denotes the fraction of individuals who are susceptible to being infected, I denotes the fraction of individuals who are currently infected, also called infectives, and R denotes the fraction of individuals who have been removed from the model, be that because they successfully recovered from the disease or that they have deceased. Adda (2016) defines R to be the group of individuals who have recovered but who are still immune, i.e. the deceased people are not included in R.

Consider that all other sources but Adda look at this group as the removed, i.e. people who overcame the disease but also deaths. How does Adda deal with deceased people?

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 postulated in continuous time, i.e. the equations in (2.1), (2.2), and (2.3) depict the change in the variables S, I, and R, respectively, 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 infectives) to which a flow is added or subtracted.

$$\frac{dS}{dt} = -\alpha SI + \lambda R \tag{2.1}$$

$$\frac{dI}{dt} = \alpha SI - \beta I \tag{2.2}$$

$$\frac{dR}{dt} = \beta I - \lambda R \tag{2.3}$$

It is important to grasp the main assumptions of the SIR model, which also tell us how these equations are constructed. The first assumption that is made, is that the population is constant, meaning that births and deaths are ignored. Next, note that the spread of the virus is determined by the interaction between the infectives and the susceptible population. The second assumption that is made under the SIR model in this light is that there is a constant rate of change in infectives that is proportional to this interaction between the infectives and the susceptible population. This is represented by the term αSI in equations (2.1) and (2.2), which is also called the transmission term (?, ?). The third assumption that the SIR model makes is that there is a constant rate of change at which infectives recover or decease. This relates to the term βI in equations (2.2) and (2.3).

I do not; I take it into account in calculating S. Should see if this matters.

Look into this, if we do use the definition that people who die are included. Reason: the fuller hospitals are, the more people will likely decease

Finally, we assume that there is a constant rate of change at which immune individuals lose their immunity. This is denoted by the term λR in equations (2.1) and (2.3). For instance, Adda (2016) 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. In the case of COVID-19, some studies show

that it is likely that individuals who recovered from COVID-19 may be immune to reinfection, at least temporarily (?, ?). Nonetheless, no definitive results have been shown.

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 $R_0 > 1$. 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 was below 1 in Italy, at between 0.5 and 0.7 (?, ?), showing that this measure is also used communicated to citizens as a way of informing them whether the pandemic is tending to end.

Explain more later on immunity since this is currently still researched.

Explain how this is computed.

3 Methodology

3.1 Model 1: Within-Region Spread

We start with a simple model ignoring effects across regions. First, it is important to understand the concept of an incubation period. This is defined as the period between an infection and the moment that the infected individual starts showing symptoms, at which point the infective is said to be symptomatic. Note that this is not the same as the period between an infection and the moment that the infective is infectious, which is called the latent period. For COVID-19, the latent period is estimated to be approximately 2 days shorter than the incubation period (?, ?). That is, infectives are able to infect others before showing symptoms.

This may be an issue when considering policies such as self-isolation when one is sick, because an infective may have already spread the virus before feeling sick. The incubation period for COVID-19 is estimated to be above 2 and below 11.5 (?, ?), 12.5 (?, ?), and 14 days (?, ?). This is a large range, but this is not rare. For instance, .

Add to this with example

While the maximum incubation period is not agreed upon by ? (?) and ? (?), their results on the median are similar. ? (?) reports a median incubation period of 5.1 days (95% CI: 4.5 to 5.8 days), while ? (?) reports a median incubation period of 5.2 days (95% CI: 4.1 to 7.0 days). For comparison, ? (?) give the result of a mean incubation period of 5.0 days (95% CI: 4.2 to 6.0 days) when excluding Wuhan residents and 5.6 days (95% CI: 5.0 to 6.3 days) when including Wuhan residents.

Explain more later on how this affects our choice for τ. We will include this in this section too, as we have tested multiple lags.

The within-region model is henceforth defined as:

$$I_{r,t} = \alpha_{within} I_{r,t-\tau} S_{r,t-\tau} + X_{r,t}\delta + \eta_{r,t} \tag{3.1}$$

where the subscript τ is a lag indicating the length of the incubation period.

The matrix X includes fixed effects for regions, as well as weekend and week of the year dummy variables.

Elaborate on the definition of X and why we choose these variables, possibly in the Dataset section.

Lastly, we include an idiosyncratic error term η. The model is estimated by ordinary least squares (OLS). Because fixed effects for regions are included in X, note that this means that running OLS is actually a least-squares dummy variables (LSDV) regression. In general, the main issue with LSDV regression is that there needs to be an indicator variable for each observed individual (in our case, these are regions). However, it is feasible to run an LSDV regression since we consider a relatively small number of regions with a large number of time periods. This will be explained more in section 4.

3.2 Model 2: Weighted Within-Region Spread

In the previous model, it has been assumed that the incidence rate within a certain region is only determined by the previous incidence rates plus some other effects. However, the transmission rate α is likely influenced by other factors as well. These may include policies, such as shutting down restaurants or public transport, but also persistent regional characteristics such as metrics on the quality of hospitals or economic development. In this section, we incorporate these factors in the within-region model (3.1). After defining the between-regions model in section 3.3, we will apply the same methodology to obtain the full weighted model in section 3.4.

Let the tensor W contain K region-specific variables that may influence the transmission rate α. As such, we now allow for α_{within} to differ for these K variables. In section 4, we elaborate on how these variables included in W are specifically defined and selected. For instance, we include the number of rail travellers, which changes over time, but also a measure of the development of health care through the

number of available hospital beds, which does not change over time. We define X and η in the same way as for (3.1). Taking this into account, the weighted within-region model is defined as:

Possibly update later

$$I_{r,t} = I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} + X_{r,t}\delta + \eta_{r,t} \tag{3.2}$$

3.3 Model 3: Within and Between-Region Spread

A key addition made by Adda (2016) is recognizing that there is spatial spillover between regions. That is, there may be infectives in one region that travel to another region and then infect individuals there. As such, the number of new cases would be modeled as $\alpha_{within} SI + \alpha_{between} S\tilde{I}$ where $\tilde{I}$ is the fraction of infectives 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.

Consider the difference in definition in I between the SIR model and Adda. Possibly use the notation from Keeling and Rohani (but this includes X)

The following model is defined:

$$\begin{aligned} I_{r,t} = {} & \alpha_{within} I_{r,t-lag} S_{r,t-lag} \\ & + \alpha_{between} \sum_{c \in R \setminus r} I_{c,t-lag} S_{r,t-lag} \\ & + X_{r,t}\delta + \eta_{r,t} \end{aligned} \tag{3.3}$$

In (3.3), the transmission parameter α is now allowed to be different within and between regions. Adda (2016) estimates (3.3) by OLS and by instrumental variable estimation (IV). Weather episodes, such as the amount of rain and temperature-related instruments, are used as instruments. There is a biological reasoning behind choosing these instruments, for instance that warmer temperatures tend to have a negative effect on the proliferation of some viruses. A social reason is also given, namely that bad weather conditions impact the amount of social interaction between people, meaning that there are less opportunities for viruses to spread. We challenge the choice of these instruments, particularly in the case of SARS-CoV-2. Firstly, we do not have sufficient information on the effect of the weather on the virus. That is, SARS-CoV-2 has only been quite apparent since January 2020 and there has not been enough fluctuation over time in temperatures to show a necessary effect that can be disentangled from, for example, policies being effective in driving the virus back. Secondly, we challenge the social reasons entirely, although not quantitatively. In our view, bad weather conditions in themselves are not likely to be strong enough instruments for the viral spread. That is, even if they are indeed exogenous with respect to the error term and that they are correlated with the viral spread, we expect this to not be quite strong.

This is currently a claim and I have not looked at Adda's quantitative tests for these instruments.

For this reason, we only consider OLS for this model.

3.4 Model 4: Full Model

We now incorporate the between-region effects as well as the weighting of the transmission parameter. In addition to (3.2), we now also put weights on the between-region transmission parameter by some possibly influential variables. Let the tensor $\widetilde{W}$ contain $\tilde{K}$ variables that now can influence the transmission rate α_{within} between two regions r and c.

Possibly consider not following Adda's notation with the tildes and use something like V and L instead of $\widetilde{W}$ and $\tilde{K}$, respectively.

$$
\begin{aligned}
I_{r,t} = I_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} \alpha_{within}^{k} W_{r,t-lag}^{k} \\
+ \sum_{c \in R \setminus r} I_{c,t-lag} S_{r,t-lag} \sum_{k=1}^{\tilde{K}} \alpha_{between}^{k} \widetilde{W}_{r,c,t-lag}^{k} \\
+ X_{r,t}\delta + \eta_{r,t}
\end{aligned}
\tag{3.4}
$$

START OLD TEXT

Adda (2016) 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.

END OLD TEXT

4 Dataset

In this section, we will outline the structure of the data that is used and how it was retrieved. Firstly, we discuss the structure of Italian regions and the reasons why we chose to use Italy as our region of interest. Subsequently, we will look at the data on COVID-19, such as the incidence rate. Here, we also discuss how we tackled possibly errors in the data, as well as missing values. Lastly, we consider the variables that are included in the weighted models in sections 3.2 and 3.4.

Italy has been one of the most intensely struck countries in the world. On June 6th, 2020, it had the seventh highest absolute number of cases, after the United States, Brazil, Russia, the United Kingdom, Spain, and India. Despite dropping in this positioning, Italy reports the highest death-to-cases ratio of 14.47%, followed closely by the United Kingdom, which reports a death-to-cases ratio of 14.18%. The sudden onset of the spread of SARS-CoV-2 put immense pressure on the Italian hospitals, especially in the northern regions such as Lombardy. This forced patients with coronavirus-caused pneumonia to be sent home as well as literal collapses of overworked healthcare workers (?, ?).

Add table

The Italian ministry of Health Services (Ministero della Salute) has posted daily reports containing tables with a detailed numerical overview of new cases, active intensive care (IC) patients, and tests executed, all divided up between the second-level NUTS regions (also called NUTS 2 regions). The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the European Union (EU) and the United Kingdom (?, ?) as used by Eurostat, the statistical office of the EU. Italy consists of 21 so-called *regioni* (regions), comparable to 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 *Nord-Ovest* (North West), *Nord-Est* (North East), *Centro* (Center), *Sud* (South), and *Isole* (Islands). The third-level NUTS regions are 107 provinces, which are subregions of the *regioni*. Ideally, we would want to have coronavirus data on the NUTS 3 regions since many policies are introduced at that level, such as a lockdown put into place on March 7th, 2020 until the strict national lockdown was instated. Unfortunately, these data were not reported. As such, we chose to use the NUTS 2 regions.

4.1 Coronavirus data

As mentioned, the specific information on the coronavirus in Italian regions was retrieved from the Ministero della Salute, who publish daily reports under a title similar to *Covid-19, i casi in Italia: 10 giugno ore 18*, where *10 giugno* (June 10th) 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)*
- Total amount of persons tested *(Casi testati)*
- Increase in total amount of tests executed *(Incremento tamponi)*

The difference between the total amount of tests executed and the total amount of persons tested is that the latter indicates the number of unique persons that were tested. That is, individuals could have been tested more than once. Do note that *tamponi* is a good indication of the *testing capacity* as the number of tests that Italy is able to execute. Henceforth, when the term *testing capacity* is used, this refers to *tamponi*, unless indicated otherwise.

It should be noted that there is a measurement error in the number of infectives, as is the case in any other country. This is because there is no possibility that every citizen can be tested for COVID-19. For that reason, the actual number of infectives is higher that the official count as reported in the tables of the Ministero della Salute.

Do the functional form thingy here

With respect to the reported death statistics, there is a distinction between Italy and some other European countries. Namely, the Italian numbers include deaths of all patients who were tested for COVID-19 before or after their death, regardless of whether they died inside or outside the hospital. In contrast, other countries may only count deaths in hospitals. French death counts, for instance, only have included deaths at hospitals and clinics caring for patients, excluding people who die at home or in care homes, although the French president Emmanuel Macron did announce that these centers would be tracked from the first week of April onward (?, ?). Moreover, Italian data makes no distinction between people who died because of COVID-19 or simply had the disease but who died from other causes (also referred to as comorbidities). Patients who had pre-existing conditions actually make up around 96% of the total death count in Italy (?, ?). In some other countries, such as Germany, a distinction between these two groups is actually made (?, ?).

We also make the note that it is unclear how the Ministero della Salute collects its information. If regions or provinces submit this information to the government each day, there may be areas that fail to submit their data for a certain day. Despite this, we assume that this official information is accurate and representative of the region for which it has been reported. If this is not the case, the numbers in the report on the next day will compensate for the error on the day before. In the official publications that we use, data that was wrongly published on a day $t-1$ is corrected by subtracting the error from or adding the error to 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. In the case that the error on day t is larger than the number on $t-1$, for instance if a value of -10 is reported on day t while the value for day $t-1$ is less than 10, we propagate the error to multiple lags until this issue no longer occurs. However, this does not happen in our data. 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 $t+1$. As such, these are left as is.

Update accordingly

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 4.1. As such, missing data is simply imputed with a value of 0.

Update accordingly

Table 4.1: Number of confirmed cases around a day t with missing data

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

4.2 Independent variables

Fix title of subsection

Independent variables, or 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 year, geolocation (being the NUTS regions), and other aspects depending on the data, such as sex, age, or the unit of measure. Unfortunately, this data is not available daily and is often not up-to-date. That is, sometimes data is available up to 2016 . For each variable, we kept the most recent data and assumed that this would be representative for the present.

Fix this and look up the actual maximum year per variable

We distinguish three sets of regressors, as mentioned in section 3. Firstly, we have a set of control variables included in the tensor $X_{r,t}$ which are not assumed to have a (large) effect on the transmission parameter α. Secondly, the tensor $W_{r,t}$ consists of variables that are assumed to affect the transmission within regions. Lastly, the matrix $\widetilde{W}_{c,r,t}$ contains variables that are assumed to affect the transmission between regions. The specification of these regressors can be found in Table 4.2.

Insert W_tilde variables and possibly move around variables to X

Table 4.2: Specification of regressors

Matrix	Variable	Description
$X_{r,t}$	weekend	Binary indicator denoting if the day is on the weekend (Saturday or Sunday)
	weekNumber	The calendar week number
$W_{r,t}$	airPassengersArrived	Number of air passengers arrived
	touristArrivals	Number of tourist arrivals
	broadbandAccess	Percentage of population that has access to broadband internet
	deathRateDiabetes	Number of deaths from diabetes per 100,000 inhabitants
	deathRateInfluenza	Number of deaths from influenza per 100,000 inhabitants
	deathRateChd	Number of deaths from coronary heart disease per 100,000 inhabitants
	deathRateCancer	Number of deaths from cancer per 100,000 inhabitants
	deathRatePneumonia	Number of deaths from pneumonia per 100,000 inhabitants
	availableBeds	Number of hospital beds
	riskOfPovertyOrSocialExclusion	Percentage of population at risk of poverty or social exclusion
$\widetilde{W}_{c,r,t}$		

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

likely a large correlation between the various death rates. To investigate this, we consider the correlation matrix in Table 4.3.

Table 4.3: Correlation matrix of the death rates for various comorbidities of COVID-19

	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	

START OLD TEXT

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.

END OLD TEXT

5 Results

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`

Appendices

A Tables