Predicting The Incidence Rate And Case Fatality Rate Of The Novel Coronavirus SARS-CoV-2

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 (Tilburg University)
dr. George Knox (Tilburg University)

Date:
April 19, 2020

Contents

1 Acknowledgements

2 Management summary

3 Introduction

4 Problem description

The following specification is used by Adda 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:

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

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.

The following spatial panel data random effects specification with a spatial lag of the dependent variable is used for the `splm` package in `R` (Millo & Piras, 2012):

$$y = \rho \left(I_t \otimes W_N \right) y + X\beta + u \qquad (2)$$

We may consider fixed effects later. This may be less applicable due to many time-constant regressors.

Note that we write ρ as the spatial lag parameter, instead of λ as Millo and Piras do, to avoid confusion later on in the thesis when using λ. Excluding the spatial lag means that ρ is set to 0. This model then becomes (Baltagi, Song, & Koh, 2003):

$$y = X\beta + u \qquad (3)$$

In these models, $y \in \mathbb{R}^{RT \times 1}$ is a vector where y_{rt} is the observation on region r at time t, $X \in \mathbb{R}^{RT \times K}$ is the matrix of K regressors where X_{rtk} is the observation on regressor k for region r at time t, and $u \in \mathbb{R}^{RT \times 1}$ is the error vector. We assume that X is of full column rank, meaning that its columns are linearly independent, and that its elements are assumed to be asymptotically bounded in absolute value. We assume that the error vector u can be decomposed in a part for random region effects as well as spatially autocorrelated residual disturbances (Anselin, 2013):

Should I specify here why this is true? It seems more appropriate in the data section.

$$u_t = \mu + \epsilon_t.$$

Here, $\mu \in \mathbb{R}^{R \times 1}$ is the vector of random region effects. We assume that $\mu_r \sim IIN\left(0, \sigma_\mu^2\right)^1$. The spatially autocorrelated residual disturbances are given by $\epsilon_t = \lambda W \epsilon_t + \nu_t$, where λ is a spatial autoregressive coefficient with $|\lambda| < 1$ and $W \in \mathbb{R}^{R \times R}$ is a known spatial weighting matrix, with diagonal elements equal to zero, such that $(I_N - \lambda W)$ is nonsingular for all $|\lambda| < 1$. We define $\nu \in \mathbb{R}^{RT \times 1}$ and assume that $\nu_{rt} \sim IIN(0, \sigma_\nu^2)$ as well as that ν_{rt} is independent of μ_r for all r and t.

We will specify the weighting matrix later. Should I do this in the data section?

Define $B = I_N - \lambda W$. Then, we can rewrite ϵ_t as

$$\epsilon_t = \left(I_N - \lambda W\right)^{-1} \nu_t = B^{-1} \nu_t,$$

so that

$$u = \left(\iota_T \otimes I_N\right) \mu + \left(I_T \otimes B^{-1}\right) \nu,$$

where $\iota_T \in \mathbb{R}^{R \times 1}$ is a vector of ones and $\otimes$ denotes the Kronecker matrix product. The Kronecker matrix product for two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$ is defined as:

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

[1] IIN: independently and identically normally distributed

5 Materials

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 (*NUTS - Nomenclature Of Territorial Units For Statistics - Background*, n.d.). 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 (*Elenco pagine per tag: Coronavirus*, n.d.). 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)*

Deaths statistics for Italy include coronavirus victims who died in hospital, as well as those who died outside of hospitals and were tested before or after dying. Indeed, post-mortem tests are routinely carried out, and there is no distinction between people who died "with" vs "because of" coronavirus, including deaths of patients with pre-existing conditions.[459][460] This is in contrast with some other European countries, where this distinction is made, thus often excluding deaths of people with pre-existing conditions, and post-mortem tests are not routinely carried out.[461][462] In addition to this, some countries only report deaths in hospitals, which represent but a fraction of the total number of deaths.[463][464] For these reasons, the Italian coronavirus death toll is not comparable to that of other European countries. It should be said, however, that in specific towns in regions where the healthcare system has been overwhelmed by the pandemic (i.e. Lombardy), official death statistics may have missed a portion of deaths outside hospitals.[465]

This is directly copied from Wikipedia. We should still research the sources and write this in our own words. The words below should be amended accordingly.

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 (Stancati & Sylvers, 2020). 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 (*Eurostat Regions Database*, n.d.). 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

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

Table 2: Specification of regressors

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.

6 Results

7 Conclusion

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.

Anselin, L. (2013). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Baltagi, B. H., Song, S. H., & Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of econometrics*, *117*(1), 123–150.

Elenco pagine per tag: Coronavirus. (n.d.). Ministero della Salute. Retrieved from `http://www.salute.gov.it/portale/news/p3\_2\_0.jsp?lingua=italiano\&id=1285`

Eurostat regions database. (n.d.). Eurostat. Retrieved from `https://ec.europa.eu/eurostat/web/regions/data/database`

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

Millo, G., & Piras, G. (2012). splm: Spatial panel data models in r. *Journal of Statistical Software*, *47*(1). Retrieved from `https://www.jstatsoft.org/article/view/v047i01` doi: 10.18637/jss.v047.i01

Nuts - nomenclature of territorial units for statistics - background. (n.d.). Eurostat. Retrieved from `https://ec.europa.eu/eurostat/web/nuts/background`

Stancati, M., & Sylvers, E. (2020, Apr). *Italy's coronavirus death toll is far higher than reported.* Dow Jones & Company. Retrieved from `https://www.wsj.com/articles/italys-coronavirus-death-toll-is-far-higher-than-reported-11585767179`

A Tables