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 14, 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:

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

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):

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

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

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):

$$y = X\beta + u \tag{3}$$

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

Note, this 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.

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

A Tables