Master Thesis Corona - Notes

Mike Weltevrede

March 2020

Contents

1 Model overview

Define:

- $I(t)$: fraction of infected individuals at time t,
- $R(t)$: fraction of recovered individuals at time t,
- $S(t) = 1 - I(t) - R(t)$: fraction of the population that is susceptible at time t,
  - *This assumes that recovery implies immunity*
- a : rate at which new cases develop,
- b : rate of recovery.

Then $R_0 \coloneqq \frac{a}{b}$ is the effective reproduction rate of the virus. If $R_0 < 1$, the virus can be said to be subsiding. Policies are made to halt a.

Standard Inflammatory Response model (SIR)

$$\begin{cases} \frac{dI(t)}{dt} & = aS(t)I(t) - bI(t) \rightarrow \text{daily increase in the fraction of } \textbf{infected} \text{ individuals} \\ \frac{dR(t)}{dt} & = bI(t) \rightarrow \text{daily increase in the fraction of } \textbf{recovered} \text{ individuals} \end{cases}$$

Incidence rate at time t
I.e. the number of new infections in a population P.

$$\begin{aligned} Inc(t) &= \left(\frac{dI(t)}{dt} - \frac{dR(t)}{dt} \right) P \\ &= aS(t)I(t)P \end{aligned}$$

Incidence rate in a region r at time t (Adda, page 922)

$$\begin{aligned} Inc(r,t) = Inc(r, t-lag) \cdot S(r, t-lag) \cdot \sum_k a(k, within) W(k, r, t-lag) \\ + \sum_{k, c \neq r} a(k, between) \widetilde{W}(k, c \neq r, t-lag) \cdot Inc(k, c \neq r, t-lag) \cdot S(k, c \neq r, t-lag) \\ + X(r,t) \cdot d + e(r,t) \end{aligned} \tag{1}$$

where

- $a(k, within) W(k, r, t-lag)$: k within-region spatial weights,
- $a(k, between) \widetilde{W}(k, c \neq r, t-lag)$: k across-region spatial weights for each other region c,
- lag is defined by Adda as the incubation period,

and

- $W(k,t)$: known spatially heterogenous weights for which we will gather data. For connections within Europe, free data is available from EUROSTAT and for the world from the World Bank.
  - EUROSTAT: this data only contains up to 2017 or 2018.
  - EUROSTAT and World Bank: this data only contains per country aggregated data, not a spatial matrix from country to country.
  - WTTC. This has PDFs with data (so not nicely importable) per country with the top 5 inbound and outbound travel in 2019. They say "Note: Data are average shares over the 2015-2017 period. Source: Oxford Economics, national sources and UNWTO"
  - UNWTO, which provides data free of charge to students and academic researchers. So, unfortunately, not open source.
  - NS, partially open source but not in a nice CSV format. Also, this is only available for 2018.
- $a(k, \cdot)$: unknown coefficients to be estimated from the data,
- $X(r,t)$: includes political regimes, development index and population density, but also includes region-time dummies to capture the effects of potentially unobserved characteristics, such as cultural norms and news, medical capacity shortages etc.

2 Our contributions

1. **New data:**

   - **COVID-19 data**: Daily data on new cases, recoveries and deaths from COVID-19 for all countries and provinces within many countries is available at https://github.com/CSSEGISandData/COVID-19 free of charge. However, airline transportation data is only partially available.
   - **Transportation data**: We will construct the spatial weights separately for different transportation means: airline, railway transport and road transport. For connections within Europe, free data is available from EUROSTAT, and for the world from the World Bank. We will carefully assess all freely available data, and while there are private data sources which would lead to more accurate modelling, we refrain from using these, as we want to produce a model that can be used in real-time, free of charge by all researchers and policy-makers in the case of COVID-19 but also in case we experience a new viral disease outbreak in the future.
   - **Other characteristics subsumed in** $X(r,t)$: These are freely available from the World Bank. We will address the issue of whether these characteristics should be included as additional covariates or as spatial weights.

2. **Real-time prediction of the infection rates:**
   The model in equation (1) allows prediction of infection rates many days ahead with and without policy measures.

   - **No endogeneity concerns**: Unlike Adda (2016), who was interested in the effect of a transportation strike or school closure on the coefficients $a(k, within)$ and $a(k, between)$, we are interested in forecasting the infection rate, and therefore endogeneity (contemporaneous changes in policy measures and shocks in the new infection rates, such as availability of test kits) is not a problem in our analysis. We can also allow for time changes in the coefficients $a(k, within)$ and $a(k, between)$ as the virus spreads, as long as enough time-series observations are available for a particular country.
   - **Heterogeneity in spatial transmission**: we can allow for the coefficients $a(k, within)$ and $a(k, between)$ to depend on the region, as long as there are enough time-series observations in that region.
   - The data is counts of new infections with many zeros (or missing data) for many regions early on, therefore offering the possibility to model this via a **count maximum likelihood model** for spatial data with truncated observations.

3. **Predicting case fatality rates heterogeneously across countries:** The case fatality rate can be estimated with higher accuracy as the outcomes of the patients can be predicted based on jointly modelling infection and recovery data.

3 Data

3.1 Which country will we use?

- https://github.com/CSSEGISandData/COVID-19 → Only has region-based information for China and the USA (with information split by overseas territory for countries like mainland France and the Netherlands).
- World Bank → For the Netherlands, railway passenger data is confidential.
- https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_Italy#Statistics → has detailed data on Italy by region, sex/age, and date (per region).
- https://en.wikipedia.org/wiki/2020_coronavirus_pandemic_in_the_Netherlands#Statistics → has detailed data on the Netherlands per date (per province).
- Dr. Boldea asked the RIVM for more detailed data.

We have decided to **use data on Italy** (first).

4 Journal

2020-03-18 **Meeting #1 with dr. Boldea**

- Which country shall we take? I.e. which has the most and/or the best data?
  - **Answer: Italy.**
- What actually is the dependent variable?
  - **Answer: Mentioned in the meeting of March 25, we could model the growth rate (#NewCases / #TotalActiveCases).**
- Can we make the link from Adda's model to SIR?

  Answer: I could not find a mathematical link. Adda simply seems to split it up in within-region, across-region, and X.
- We will look into prediction, we can likely not say much, if anything, about causality.

2020-03-22 **Data exploration (R)**

- Tried to automate extracting tables from the Wikipedia page of Italy. This did not work.

- Managed to program extracting the table for the Netherlands. The data cleaning did not work as planned.

2020-03-25 **Meeting #2 with dr. Boldea**

- It is likely that the distribution is **not stationary** over time since policies by the government have an effect.
- How to construct weights on modelling connectivity and policy effectiveness?
    - Within region connectivity: length of railroads, schools, etcetera.
    - Across region connectivity: bordering regions have more weight, use airline data (Alitalia), tourism hotspots, etcetera.
- There are many development indicators available on Eurostat. Perhaps a PCA would be useful.
    - Example: the more hospitals c.q. number of beds there are in a region, the more developed it may be.
- Perhaps we should model growth rate (#NewCases / #TotalActiveCases).
- Adda mentions overestimating and measurement error (which we will also have). Notice that Italy is then good because they (seem to) test a lot. Moreover, then we will receive an upper bound, which may be desirable.
    - **Question: Can we derive how much of an upper bound this is, e.g. at most 10% away from the true answer?**

2020-03-26 **Finding data - I have uploaded these to Github and Google Drive**

- Manually downloaded and cleaned the data for Italy from Wikipedia.
- Looked into the regional data from Eurostat and downloaded 17 possibly interesting datasets, including demographics, internet access, number of hospital beds, transportation statistics, etcetera.

Data exploration

- Used Python to explore the Wikipedia data and found nothing special; all relevant variables exhibit an exponential growth.

2020-03-27 **Data reading**

- Started to work on a data reader for the Eurostat data. It reads data and combines these. Still a work in progress (WIP).

2020-03-30 **Data processing**

- Automated the Excel sheet so that we only need to add new data to the Wide sheet.

– Wrote a Python script to compute the distance between cities given latitude and longitude.

Data collection

– Official Italian Statistics Navigated a bit. Did not seem to have much more interesting data than Eurostat. They do have **monthly and quarterly data** for some sources as well as **amount of passengers for domestic flights** (arrivals, departures and total but not specifically between which airports)

Flights: DCSC_INDTRAEREO_30032020144404115.csv

– Looked into obtaining weather data for the regions. The ideal goal is day by day historical weather data.

- yr.no: Monthly data (average highest and lowest temperature, as well as average days of precipitation) for around 5 cities per region.
- Il Meteo: This seems to only have it on a city basis and not even for all cities. It is difficult to navigate.

Thesis writing

– Added specification of W, $\widetilde{W}$, and X.

2020-04-01 **Meeting #3 with dr. Boldea**

– We will start by fixing a region, i.e. setting the second line in Adda's model to zero, and then doing analysis for each region.

– Given that there is uncertainty in the regressors and that we have no lag available (most data is annual), we likely need to use a Bayesian approach. How do we do this? Most approaches currently are on uncertainty in parameters.

- Jim Stock paper: he states that, by Bayes Theorem, we can split the symptomatic and asymptomatic effects. Check this out! Moreover, how does he reach a U-shaped R_0?

– Note that for PCA, α is then also not identified up to a rotation, which is tough if we also want to do inference.

– Italy started a total lockdown. Likely, the number of passengers by train then went to 0. Before that, it must slowly decay. How do we decide to decay it? Do we have Italian sources or can we extrapolate trends in other countries to Italy?

– Recall that Italy does not do random testing. They did do this in Germany and Iceland. Can we use this?

– There is no time lag, but do note that policy enacts with a lag. As such, we would have W_t^r and W_{t-lag}^r How can we take this into account? Is there information on this?

- Age may have a nonlinear effect on transmission (younger people likely transmit the disease more), so we can include age^2 too.

2020-04-02 **Reading up**

- Found papers discussing Bayesian approaches to model uncertainty and uncertainty in regressors, most notably (Neff, 1996).
- Started a Coursera course on Bayesian Statistics in R (one-week trial, not continued but Datacamp also has Bayesian courses for R).

2020-04-03 **Reading up**

- Started a Datacamp course on Time Series Analysis in Python.

Data collection

- Redownloaded data and retrieved new data, such as amount of doctors and freight transport.

Thesis writing

- Started specifying my own model.

Data processing

- Improved the Python script to process the Eurostat data and ran this to update the file.

2020-04-06 **Modelling**

- Explored modelling options (see modelling.ipynb).

2020-04-07 **Reading up**

- Explored papers (concerning Bayesian approaches to model uncertainty).

Programming

- Clean Wikipedia data; replace missing values and process negative values (see clean_wide.py).

2020-04-08 **Reading up**

- Started a Datacamp course on Time Series Analysis in R, seeing as we switched languages.

Meeting #4 with dr. Boldea

- Spatio-temporal models are quite different from regular panel data models so we cannot apply a Tobit model. Code and slides will be shared.

- Regarding the concerns on predicting rates due to misspecification of the denominator: we will likely incorporate asymptomatic patients later so this may not be an issue.
- Because we measure with error, we can consider an auxiliary model where some unobserved weighting matrix is a combination of the other weighting measures plus some error:

$$OW = a(b_1 W_1 + \cdots + b_K W_K + error).$$

So, the from the old formulation of

$$Inc_{r,t-lag} S_{r,t-lag} \sum_{k=1}^{K} a^k_{within} W^k_{r,t-lag}$$

we would go to

$$Inc_{r,t-lag} S_{r,t-lag} \cdot aOW.$$

This is akin to a random coefficients model (so look into this) and it is usually estimated with maximum likelihood.

2020-04-09 and 2020-04-10 Did some general things, mostly reading up on spatiotemporal modelling and fixing/experimenting with code.

2020-04-12 **Reading up**

- Looked into the SPLM package in R and the associated paper for theoretical knowledge on spatial panel data models.

2020-04-13 **Coding**

- Created an R script to interpolate the amount of travellers by railroad.

2020-04-14 **Data collection**

- Started to fill out the Google Mobility Report in a spreadsheet. Since no numbers are available for the dates apart from March 29, we have to estimate these from the graphs.

Programming

- Generalised the interpolation script for all 6 subjects in the Google Mobility Report.

2020-04-15 **Meeting #5 with dr. Boldea**

- Spatial spillover happens with a time lag (e.g. if people travel). As such, spatial lag is not applicable in our model.
- We can simply define our lagged regressors in X, e.g. $w_{t-1} y_{t-1}$.

- We have $T > 57$ time periods, which is large enough, and $N = 21$ regions. For panel data models, we don't need both N and T to be large enough. The problems regarding consistency only apply for small T.
- Random Effects: we can use GLS; the random effects do not change over time. As such, there is no effect on the mean, just the variance.
- Fixed Effects: we can use (pooled) OLS. We do not need to do within-transformations as these are done to correct for a small (fixed) amount of time periods T! To add fixed region effects, simply add a dummy for each region.
- Adda's definitions are: $Inc = \frac{\#\text{new cases}}{\#\text{population}}$ and $S = \frac{\#\text{susceptible people}}{\#\text{population}}$.
- If we use the same regressors, we can test whether estimating GLS is the same as estimating OLS equation-by-equation. They should yield the same results as long as we have homoskedastic errors that are uncorrelated across regions.
  1. Estimate OLS;
  2. Compute the residuals;
  3. Construct the covariance matrix of residuals;
  4. Estimate GLS with this;
  5. Do a Hausman test.
- With $T \approx 56$, we cannot have too many regressors. We will look into regressors that deal with the transport of passengers, health care (e.g. death rate for comorbidities of COVID-19 and available (ICU) beds, and, if available, a measure of increase in testing capacity). We will use ratios of the total instead of the absolute number (NOT standardised). We can later apply Ridge or LASSO.
- To look into: policy is an endogenous effect since it depends on the past. Perhaps we can look into what would have happened in the short term and long term is the policy was not enacted.

2020-04-19 **Programming**

- Progress on the data cleaning process. I added the calculation of the susceptible and incidence rate.

Journal not kept between this day and the next

2020-04-23 **Meeting #6 with dr. Boldea**

- Consider the email concerning the consistency of the FE estimator for dynamic models. This is a bit different since we do not only have a lag of the dependent variable but a product with various variables. Usually, α depends on $(1 - \alpha)^T \to 0$ (in the backwards substitution part). Dr. Boldea will look into a proof and/or possible conditions on α.

- As $T \to \infty$, endogeneity disappears.
- Regarding IV/GMM: Arellano-Bond is likely to have weak instruments if the spread is rapid so results are imprecise. Prof. Cizek is looking into how to get around this, but GMM is likely less applicable.
- Regarding variables:
  * Broadband access is more a measure of development. Perhaps we can consider other development indices. Look into these and correlations.
  * Instead of death rate from comorbidities, the discharge number is likely more of a good indicator of the incidence rate. Possibly, we should not split this into multiple comorbidities but rather use just one variable.
  * Notice that air passengers arrived and departed have nearly the same coefficient with opposite signs. Most arrived passengers likely leave so these are highly collinear. Same with maritime passengers. We should only use one of the two. Moreover, can we set these to nearly 0 at the moment of lockdown?
- Can we find some way of finding information on vaccination for tubercolosis? Even nationwide would be good.
- We should use Information Criteria to select our model and avoid multiple testing. Which variables should stay or go?
- Should we go for out-of-sample or in-sample analysis? For policy analysis, we should consider whether we have to deal with exogeneity (as lags are long enough) or endogeneity. For simulations, we would assume that the model specification is good and then estimate α. One thing to consider too: can we use α_W instead of αW or quadratic effects: $\alpha W + \beta W^2$?
- Next things to do:
  * Which variables should be kept in?
  * Use recovery rate over death or combine them.
  * Add crossregion variables.
  * Can we use the Stock & Watson paper to augment scenarios for tested/infected? This depends on clustering.
  * Are the residuals stationary?
  * Only then, simulate policy and do IC for in-sample model selection.

Journal not kept between this day and the next

2020-04-30 **Meeting #7 with dr. Boldea**

- X