Assignment 2

GV903 Advanced Methods – 2020/2021
University of Essex, Department of Government

Instructions

This assignment is due on Wednesday, 23 December at 9:45 am. Please submit on FASER. Instructions can be found in the module handbook on Moodle.

Please create a new document with your answers. Use LaTeX in conjunction with `knitr` via RStudio for document creation, and submit both a PDF file (the compiled document) and the source files necessary for compilation (e. g., with the extensions `.Rnw`, `.bib` etc). Up to five points are given for proper use of technology and proper formatting, as part of the maximum of 100 attainable points.

The number of points you can obtain for each question or task is given in square brackets. The points add up to 100 (including the points for use of technology as described above) and determine your final mark for this assignment. The final mark you earn for this assignment is given by the equation

$$m = t + \sum_{i=1}^{n} p_i, \tag{1}$$

where p_i denotes the number of points you earn for question or task i, n is the number of tasks in this assignment, and t is the number of points you earn for use of technology and formatting. All answers are evaluated on three criteria: how clearly the results are presented; how correct the results are; and how elegant, computationally efficient, or sophisticated the solution is (where applicable). Good luck!

Dataset: Covid-19 Confirmed Infections

In this assignment, you will download a Covid-19 dataset and conduct time series analysis with it. Go to

`https://github.com/GoogleCloudPlatform/covid-19-open-data`

and download the `main.csv` dataset. The direct link to the dataset is

`https://storage.googleapis.com/covid19-open-data/v2/main.csv`.

The file has about 2 GB and may take a while to download.

This CSV file contains variables with a daily time resolution for many countries and regions within countries, including the number of new confirmed cases of Covid-19, population size, weather conditions, the restrictions in place (including lockdowns), and other political and economic variables.

If you have a poor internet connection or limited memory in your computer, you can use a slightly outdated, revised version of this file (without the subnational data for all

countries but the UK), which is about 77 MB large. The file comes with this assignment and has the file name `main_smaller.csv`.

It is well known that the recorded infections deviate in substantial, unknown but systematic, ways from the actual number of infections. Nevertheless, we will use this dataset for educational purposes here. But keep in mind that the validity of the data is limited.

1 Time Series Decomposition

The dataset contains daily time series data for a number of countries, including the UK. In this part of the analysis, focus on new confirmed cases per day in the UK and discard all other data. Focus on the national level, i. e., discard all the different observations for the regions and cities and retain only the data for the United Kingdom overall. There were a few early recorded cases of Covid-19 in January and February 2020, but the virus was spread in the wild only in late February. The last few recorded observations may suffer from incomplete reporting, and the dataset version provided on Moodle is slightly outdated. Therefore, please limit your analysis to everything including and after 20 February 2020 and before and including 1 December 2020. Complete the following tasks:

1. Create the necessary data structures for time series analysis in `R`. Keep in mind the start and end dates and record the observations on a daily basis nested in calendar weeks. Use `print(con, calendard = TRUE)` (where `con` is the number of confirmed cases) to show your data. Plot the raw time series. [7 points]

2. Compute a seasonality component (without any additional packages or specialised functions for this purpose) and plot it. Interpret the seasonality component in up to 50 words substantively—why do these patterns presumably exist? [8 points]

3. Compute the time trend inherent in the new confirmed infections time series (you may use functions we used in the lectures for this purpose) and plot it. Interpret the trend substantively in up to 50 words. What do you see, and why do we presumably observe these patterns? [6 points]

4. Is this an additive or multiplicative time series? Provide details of your reasoning in up to 50 words. You may include diagrams to support your argument if helpful. [4 points]

5. Compute the residual variance and plot it, both in terms of its distribution and in relation to time. Explain how you did it and why, and comment briefly on what you see, in up to 100 words. [7 points]

6. Identify outliers at the $\alpha = 0.10$ significance level, without using any specialised packages or functions for this purpose. Highlight them graphically in the plot of the residual variance (i. e., the fourth plot in this assignment) and in the raw plot of the time series you created above (i. e., the first plot). In up to 50 words, comment on your findings regarding where the outliers are located in the time series. Are you surprised about the result? [7 points]

7. Recompose the time series, then plot again, this time as a new plot. [5 points]

8. Recreate the plot from Task 6 (the original time series with highlighted outliers) using `ggplot2`. [8 points]

2 Time Series Regression and Panel/Multilevel Modelling

1. Use (partial) autocorrelation functions in `R` to examine the autocorrelation inherent in the new confirmed cases in the United Kingdom, Germany, France, Spain, and Italy. For modelling this variable (independent of national context), what kind of lag structure would you recommend in a time series regression model, and is this a stationary or non-stationary time series? Explain in up to 100 words! [8 points]

2. For the five countries (UK, Germany, France, Spain, and Italy) and the duration from 20 February 2020 to 1 December 2020, conduct a joint statistical analysis (i. e., in a single model). In this analysis, test whether stay-at-home orders (i. e., a lockdown) have any effect on the dynamics of new infections, possibly with a lag of some duration you need to choose. Choose appropriate additional covariates. Choose a suitable statistical model for this data structure. Correct for any statistical problems you may diagnose (if and where possible). Choose a good way to take into account the potential temporal dependence in the data. Report the results of your hypothesis test, and justify and explain all your choices in up to 400 words. Your analysis should be as impressive and valid as possible—think of this task as a way to showcase the skills you have acquired so far in this module. [30 points]

3. Present the results of your model, and interpret them substantively in up to 200 words. Include a brief discussion of the goodness of fit of the model in your discussion. [5 points]