Job description

The work includes a series of Questions concerning the coronavirus pandemic. The data you will use are given in the files: 1) ECDC-7Days-Testing.xlsx and 2) FullEodyData.xlsx. Both files are from the website https://www.stelios67pi.eu where there is a lot of other stuff you might want to download. The first file has weekly data for the number of cases and tests and mainly for the positivity index (positivity_rate column) for European countries as recorded by the European Center for Disease Control (ECDC). The second file has various data and indicators related to Covid-19 for Greece on a daily basis, such as tests and cases, hospitalizations in simple beds and intensive care units (ICU), deaths, etc.

In the assignment questions that refer to a country in Europe from the 25 European countries given in the file EuropeanCountries.xlsx, you will define that country according to your group number. The country corresponding to the group is the one with a consecutive number, the remainder of the division increased by one of the AEM (if the group consists of two members, of one of the two members) by 25. There is the possibility to use another neighboring European country if it is judged that there are no satisfactory data on the Work Questions from the European country that was initially defined. For example, for AEM=9876 the country sequence number is 2 and corresponds to Belgium. Let's call the country you selected as country A.

The team may also draw information on the Work Questions from other sources if he judges that they are easier to process or more accurate and better updated.

Labor issues

For all Questions at the beginning of each program the relevant data file will be loaded. If it is other than the three data files given for the work, this should also be submitted. There is a free choice in the presentation of the results (graphs, conclusions and results on the command line). If for a country, week and/or day there is no value or it is negative for the index you will use, you can either ignore it (and the sample will have one observation less) or replace it with a value you will find in another source or correct it in a justified way.

It is noted that for each Question you should first present the data you will use in the analysis in a way that allows you to get a first subjective impression of the Question's question (it does not have to be stated in the Question).

1. *What is the distribution of the positivity index in European countries for two specific dates (at week level)? Is there any known parametric distribution that approximates it?* To answer these questions you will use all 25 European countries given in the EuropeanCountries.xlsx file and look at two different dates (weeks) in the ECDC-7Days-Testing.xlsx file: 1) one of the last 6 weeks of 2021 (W45-W50) in which the positive index of country A is maximum and 2) the same for the corresponding 6 weeks of 2020 (W45-W50). You will plot the histograms of the positivity index from the 25 countries for the two dates. You will fit an appropriate known parametric distribution (normal, uniform, exponential, see also fitdist function). You will also investigate whether the two distributions for the two dates can (or cannot) be satisfactorily described by the same parametric distribution. Your answer here can be given simply based on the appropriate plots of the parametric distributions on the respective histograms.

2. *Do the two distributions of the positivity index in Question 1 differ?* To answer this question you will test the equality of two distributions. For the control, use the Kolmogorov-Smirnov statistic which is the maximum difference by rank of the cumulative relative frequencies for the positivity index in the two time periods periods. The cumulative relative frequency estimates the cumulative distribution function FˆX(x) = i/n, where i is the rank of observation x of the t.m. X in the list ascending order of the n observations, and denotes the ratio of the observations in the sample that are less than or equal to x. The Kolmogorov-Smirnov statistic is maxx |F_X(x) - F_Y (x)|, where X is the positivity index in Europe for the former period and Y for the second period. Considering that we don't know anyone known parametric distribution of this statistic you will do a randomization or bootstrap test by randomly selecting (without resampling or with resampling, respectively) from the population sample both periods together (as for testing the equality of two means in Section 3.4.4. in the Notes).

3. *When is Greece's weekly positivity index statistically significantly different from that of the European Union (EU)?* To answer this question first You will create a function that will answer this question for any week. The daily values of the positivity index for Greece can be calculated from the daily tests (rapid and PCR in the AS and AT columns respectively in the FullEodyData.xlsx file) and the daily new cases (in the column B in the same file). The weekly positivity index is the average of the positivity index of the 7 days of the week. You can calculate the weekly EU positivity index from the values of the weekly positivity indices of the 25 countries given in the file ECDC-7Days-Testing.xlsx or more easily by reading the mean values from the corresponding graph on the website https://www.stelios67pi.eu/testing.html. The function will accept the 7 consecutive daily values of the positivity index in Greece for a given week and the corresponding weekly EU positivity index, will calculate the 95% bootstrap confidence interval for the mean index of Greece's positivity in one week (7 days) and the answer to whether it differs with statistical significance will be given by comparing the value of the weekly net positivity of the EU with the confidence interval. At the output the function will it must also give the sign of the difference if there is a difference. In a program you will do the calculations for a period of 12 consecutive weeks (by calling the function 12 times) from the week of the last peak of the positivity index of country A and backwards (the data for country A are in the file ECDC-7Days-Testing.xlsx). The program should generate the diagram for the weekly positivity index for Greece and the EU for the period of interest and statistically significant differences should be noted / shown in some way.

4. *There are significant differences in the positivity index in Europe in the last two months with the equivalent of 2020?* To answer this question for country A you will compare in the two 2-month periods, i.e. W42-W50 for 2021 and 2020, the average weekly positivity index (over a two-month period) in the first and second periods. The comparison will be made with appropriate parametric testing and bootstrap or random testing

poetry. You will repeat the same checks for another 4 countries, which are alphabetically adjacent to country A in the list of 25 European countries. Is there agreement in the results of the comparisons in the 5 countries?

5. *With which of the 5 European countries does the course of the weekly positivity index of Greece correlate in the last quarter?* To answer this question you will take the same 5 countries that you used in Question 4 and the 3 month period, i.e. W38-W50 for 2021, and form the path of the 5 weekly positivity indicators over the period of interest. You will calculate the correlation coefficient
Pearson for the Pair of Greece's weekly positivity index with each of the 5 countries. You will also perform a significance test of the correlation coefficient, parametric and randomization, for each of the 5 Pairs of countries at a significance level of ÿ=0.05 and ÿ=0.01. Is there a statistically significant correlation and with which country is it greater?

6. *Is the weekly positivity index of Greece significantly more strongly correlated with any of the 5 European countries in the last quarter?* As a continuation of Question 5, we want to check if, for the two countries whose weekly positivity index is more related to the Greek index, the difference in the respective correlation coefficients is statistically significant. Let the maximum value of the correlation coefficient be for the pair of countries (A,B) and the second highest value for the pair (A,C). To test whether the correlation coefficient of (A,B) is significantly different from that of (A,C) you will perform either a bootstrap or randomization correlation coefficient test, randomly selecting (without replacement or with replacement, respectively) from the common sample of Pairs of observations of (A,B) and (A,C) (as for checking the equality of two mean values).

7. *Can I predict a country's coronavirus deaths from the previous week's weekly positivity rate?* You will answer this question for country A. You should find for which week lag of the weekly positivity rate with respect to the week in which deaths are reported the linear simple regression model of the weekly number of deaths (per million population) is best fit as to the weekly positivity index of some week before. You can read the weekly number of deaths (per million inhabitants) for country A in the relevant graph on the website https://www.stelios67pi.eu/testing.html. You will test the models for lag up to 5 weeks in advance. You should choose two different periods of 4 months (16 weeks, free choice) and apply the procedure by fitting the models to the data for each of the two periods separately. Do the conclusions for the week lag of the positivity index that gives the best prediction of deaths in the two periods seem to agree?

8. *For Greece, can I predict daily deaths due to coronavirus knowing the positivity rates for many days before?* You should find the best multiple linear regression model that predicts the daily number of deaths for Greece from the positivity index in previous days, going up to 30

days back (30 independent variables). You should compare the fit of the full model with the 30 independent variables to models applying dimensionality reduction (free choice) to the data. To compare the models you should use the adjusted coefficient of multiple determination. You should choose two different periods of 3 months (12 weeks, free choice) and apply the procedure by fitting the models to the data for each of the two periods separately. Compare the best models you selected in the two periods in terms of their structure and fit their.

9. *How can I check if my model is suitable for predictions?* You will answer this question in Problem 8 and use cross-validation to calculate a prediction error statistic or the (adjusted) coefficient of multiple determination. Specifically, you will divide the set of data into 5 equal parts, e.g. if the total of observations is 56 you will divide into 5 parts of 11 observations (the last one can have 12 to contain all the observations). You will remove one of the 5 parts and fit the model to the remaining 4 parts. Then you will predict the part you will remove. You will repeat this process 5 times (one for each part) and finally collect the predictions to calculate the prediction error statistic or (adjusted) coefficient of multiple determination. You will repeat the model comparison process as in Question 8 with the cross-validation approach. Comment on the results with the simple fit in Question 8 and the cross-validation.

10. *For Greece, can I predict daily deaths due to coronavirus knowing other relevant indicators and for many days in advance?* While in Question 8 we investigated the dependence of the number of daily deaths in Greece on the positivity index in the previous days, here we want to include in the model other (or even other) indicators that may have predictive power. Such indicators (such as indicators of hospitalization in single beds, intubated vaccinated and non-vaccinated), can be found in the file FullEodyData.xlsx (free choice). You can also consider a lag of days for each indicator as in Question 8 for the positivity indicator (free choice of lags). You can freely choose the period in which you will adapt and evaluate the models but it must extend until 27/12/2021. Having chosen the appropriate model, you should make predictions of daily deaths at least for 12/28/2021. You can also extend the forecasts to later days by getting more recent data for your indicators from the updated FullEodyData.xlsx file given at https://www.stelios67pi.eu/eody.html.