Assignment

Data Analysis 2 and Coding with R

MS in Business Analytics,
2020/2021 Fall

Your task is to analyse the pattern of association between registered covid-19 cases and registered number of death due to covid-19 on a given day. The aim of this assignment is to guide you through in creating a short report on the pattern of association. You will need to chose your final model, interpret the results and refer to some robustness checks. This assignment is evaluated for both Data Analysis 2: Finding Patterns with Regressions **and** Coding 1: Data Management and Analysis with R.

You need to upload a zip file containing the required file structure and all the files to ceulearning site to Assignment 1. The readme file in the root folder needs to contain your assignment's **public** github repo's url.

Deadline: Sunday, 29 November 2020, 11:55 PM

1 Task

1. Create github repo with proper file structure:
   (a) data folder: with raw and clean sub-folders, also contains variables.xlsx and readme files.
   (b) codes folder: rmd and if want an .R with a readme which tells in 1-2 sentence which file does what
   (c) docs folder: .html and .pdf generated from .rmd. Here there is no need for readme file.
   (d) out folder *if needed*: contains any output, which generated by the codes: e.g. model comparison
2. Download covid-19 data, which is collected by Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Also download population data for 2019 from the World Bank's site (using World Development Indicators)
   (a) See the exact date what you need to download under Section 2.
   (b) Create a separate script, which downloads the data and saving them into the raw data folder.
   (c) You may use my codes, available at the course's github repo. (The aim of this exercise is not to test your data gathering skills, but later you will need it in the last assignment, so you may learn from this!)
   (d) Data sources:
      i. Covid-19: `https://github.com/CSSEGISandData/COVID-19`, use csse_covid_19_data daily reports. (Available here.)
      ii. Population for 2019: use WDI package in R
3. Clean and merge the two data

(a) Create a new script which cleans and merges your data as follows:

(b) Covid data:

   i. Drop variables which are not going to be used: 'FIPS, Admin2, Last_Update, Lat, Long_, Combined_Key, Incidence_Rate, Case.Fatality_Ratio'

   ii. Summarise your observations, thus you only have countries as observations (instead of countries with regions). You need to sum up the cases.

(c) Population data:

   i. Remove all non-country observations, similar to the codes used in Seminar on 11th of November

(d) Merge the two data table with 'full_join' - available in tidyverse package.

   i. After merging you need to correct some country names to match each other fully or partially. (See my codes.)

   ii. In the end you should end up with 7 non-matched observations. These are: a) cruiser ships which stuck in national territory (Diamond Princess, MS Zaandam) b) disputed territories which are accepted by covid statistics but not by world bank (Western Sahara, Taiwan or Kosovo) c) we have no population data on them (Ertirea, Holy See (Vatican))

(e) Finally check the missing values and drop if population, confirmed cases or number of death is missing.

(f) Save your merged file into clean data folder.

(g) You may use my codes, available at the course's github repo. (The aim of this exercise is not to test your data cleaning skills, but later you will need them in the last assignment!)

(h) Do not forget the readme files!

4. **Analysis of the data**

(a) Create a .rmd or .R file for the analysis and carry out the followings:

(b) Check what is your dependent (y) and explanatory (x) variable in Section 2. Note that this is only a 'potential' variable, thus you may transform it with ln transformation, if you decide. But if you have 'number of registered death' then do not transform it to ratio variable of 'number of registered death per capita' and vica versa.

(c) Check all your variables – with the help of histograms, summary statistics and checking extreme values – and make a conscious decision on which observation(s) to drop.

(d) If your task is to analyse 'per capita' measure then create new variables in your dataset. Otherwise skip this step.

(e) Choose a proper scaling for your variable (may divide by a thousand, a million, ect.). Stay consistent during the interpretations with the scaling!

(f) Check and **report** your distributions for y and x variables: use histograms and summary statistics table (mean, median, min, max, standard deviation)

(g) Check the possible different ln transformation for the variables with plotting different scatterplots with lo(w)ess. **Make a substantive and statistical reasoning**, where and when to use ln transformation. You do not need to fit any model here, only use statistical reasoning based on the graphs.

   i. Take care when it is possible to make ln transformation: you may need to drop or change some variables.

(h) Choose your specification for the ln transformation and estimate the following models with graphical visualizations:

   i. Simple linear regression

   ii. Quadratic (linear) regression

iii. Piecewise linear spline regression

iv. Weighted linear regression, using population as weights.

(i) Compare the models and choose your preferred one

i. Use substantive and statistical reasoning for your chosen model.

ii. Show the model results in the report along with the graph.

iii. Report the model comparison (all the estimated model results) in the appendix of your report.

(j) You need to test your β parameter for your chosen model, which interacts with your explanatory variable. (In case of quadratic or piecewise linear spline, test β_1)

i. Carry out the following test: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$.

(k) Finally, using your selected model, analyse the residuals:

i. Find countries who lost (relatively) the most people due to covid using the model result: worst 5 residual.

ii. Find countries who saved (relatively) the most people due to covid using the model result: best 5 residual.

2 Date and variables to use

ID	**Date (dd/mm/yyyy)**	**potential Y-variable**	**potential X-variable**
1901587	29/09/2020	Number of registered death	Number of registered case
2000692	23/10/2020	Number of registered death	Number of registered case
2001316	15/09/2020	Number of registered death	Number of registered case
2001385	24/09/2020	Number of registered death per capita	Number of registered case per capita
2001405	22/09/2020	Number of registered death per capita	Number of registered case per capita
2001492	04/11/2020	Number of registered death per capita	Number of registered case per capita
2001531	05/10/2020	Number of registered death per capita	Number of registered case per capita
2001928	11/10/2020	Number of registered death per capita	Number of registered case per capita
2002392	26/10/2020	Number of registered death	Number of registered case
2002824	31/10/2020	Number of registered death	Number of registered case
2003316	13/10/2020	Number of registered death	Number of registered case
2003374	21/09/2020	Number of registered death per capita	Number of registered case per capita
2003382	05/11/2020	Number of registered death per capita	Number of registered case per capita
2003578	17/09/2020	Number of registered death	Number of registered case
2003615	04/09/2020	Number of registered death per capita	Number of registered case per capita
2003617	10/09/2020	Number of registered death	Number of registered case
2003629	17/10/2020	Number of registered death	Number of registered case
2003859	07/10/2020	Number of registered death per capita	Number of registered case per capita
2003870	01/10/2020	Number of registered death per capita	Number of registered case per capita
2003872	25/10/2020	Number of registered death	Number of registered case
2003873	08/09/2020	Number of registered death	Number of registered case
2003877	08/10/2020	Number of registered death per capita	Number of registered case per capita
2003888	20/10/2020	Number of registered death	Number of registered case
2003891	15/10/2020	Number of registered death	Number of registered case
2003898	02/11/2020	Number of registered death	Number of registered case

Table 1: Which date you need to use and which potential variables in the model

3 Evaluation

3.1 DA2

Overall you can earn 10 points. For DA2 only pdf/html is going to be evaluated.

- Introduction: aim of the analysis and introduction of your variables 1p
  - What is your research question?
    - 1 sentence - 0.5/1p
  - What are your variables and how is it measured? What is the population and how your sample relates to that? What are the potential data quality issues?
    - 2-3 sentence - 0.5/1p
- Selecting observations (or dropping) and potential scaling of them. 2 sentence - 0.5p
- Histogram and summary statistics for x and y with 2-3 sentence, explain the main features. - 0.5p
- Investigate the transformation of your variables 1p
  - Substantive reasoning 2-3 sentence - 0.5/1p
  - Statistical reasoning with the help of the graphs, 1-2 sentence - 0.5/1p
    - You can put your graphs in the appendix, but not into the main part.
- Estimating different models 3p
  - Model comparison table with scatter plot visualization in the appendix with explanations what you can see in the table. 5-8 sentence - 2/3p (each worth 0.5 point)
  - State your choice of model: substantive and statistical reasoning. 3-5 sentence in the appendix - 1/3p
- Presentation of your model choice 1.5p
  - State your choice of model in an appropriate format. 1 formula - 0.5/1.5p
  - Interpret the estimated parameters of the model. 2-3 sentence - 1/1.5p
- Hypothesis testing on β (which interacts with x) - 0.5p
  - Test the following hypothesis: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$.
  - Choose a significance level and make your conclusion.
  - Report your results in a table and write 1-3 sentence on the conclusion.
- Analysis of the residuals 1p
  - Interpretation for countries with the largest negative errors. 2-3 sentence - 0.5/1p
  - Interpretation for countries with the largest positive errors. 2-3 sentence - 0.5/1p
- Executive summary - 1 points
  - 3-5 sentence about the main results of the analysis: what is the variable, how the pattern of association looks like, what model you use, what is the main message of your model, what would strengthen your results and what would weaken your results.
- Formatting of your report
  - Your main text for pdf should not exceed 3 pages:
    - Each extra page is penalized by 1 point.
  - Your appendix for pdf can be as long as you wish. You should put the appendix in the same pdf, and indicate it with a new section or title.
  - Your html file should be the same as your pdf
    - Only exception if you make your html interactive.

3.2 Coding in R

Overall you can get 15 points.

- You created the required folder structure with the files as it was asked - overall 2p
  - You created the proper folder structure - 0.5/2p
  - Readme files are easy to read, compact and explain what you can find in the folders' file - 1.5/2p
- Rmarkdown file runs and produces the attached pdf/html file - 2p
- Readability of the code - overall 3p
  - Proper commenting - easy to understand what the code (intended) to do - 2/3p
  - Proper spacing of the code and general outlook - 1/3p
- Code does what it is intended/claimed to do in pdf/html - 3p
- Visualization - how graphs looks and annotated - 2p
- Tables - how table looks and annotated - 1p
- Formatting of pdf/html output - 2p
  - Title, sections, graphs, tables, formulas are well formatted and it looks like a proper analysis that you can give to Head of the Program Director.