R Lab Graded Assignment 2

Stavros Nikolakopoulos*
Department of Statistics, AUEB

Introduction

This assignment is to be done on an individual basis. The scoring will be 0-10, and it will have a weight of 60% of the overall R Labs grade (which counts for 10% of the overall course grade). The data to be used for this assignment can be found in the file `Drugs.txt`. It includes data published by OECD on pharmaceutical drug spending by countries with indicators such as a share of total health spending, in USD per capita (using economy-wide PPPs) and as a share of GDP. Plus, total spending by each countries in the specific year. The variables in the dataset are:

Variable Name	Description
LOCATION	Country code
TIME	Date (Year)
PC_HEALTHXP	% of Health spending that is spent on pharmaceutical drugs
PC_GDP	% of GDP that is spent on pharmaceutical drugs
USD_CAP	USD per capita (using economy-wide PPPs) spent on pharmaceutical drugs
TOTAL_SPEND	Total spending in millions USD

Please complete the tasks below. Make sure you name the Data as indicated below so the script is reproducible by me. You may submit the script as a solution with filename `YOURSURNAME_P/FT.R`, where `P/F` refers to whether you are a Part or Full time student.Assignment is due Friday 06/11/2020 at 23:55.

Questions

1. Load the data in an object named `Drugs`. You will notice there is an additional variable called `FLAG_CODES`. Remove this variable from the dataset (0.5 points)

2. How many countries are there? Make a table with the countries' acronyms and the number of datapoints (years) per country, sorted in ascending order (the country with the least datapoints appears first) Hint: `table()` (0.5 points).

3. Make a selection of countries based on the number of datapoints available. Specifically, select the countries that have a number of points in the top 25% of the distribution of number of datapoints, thus representing the countries with the most information available. (Hint: functions `subset()`, `quantile()` and `%in%` operator might come handy here). (2 points)

4. Using only the data selected in Q3, create a graph with 4 plots (in the same window). Each plot should depict the development of the drug spending for all countries in the reduced dataset, over the available years. Each graph should depict one of the metrics (metrics = variables measuring drug spending in different ways, thus4 metrics= 4 graphs). In each graph, a separate line should represent a country. The legend should show which line represents each country. The main title of each plot should mention "Drug spending in XX (type of metric)". (Hints: `par(mfrow=..)`, define the x-axis limits by the range found in the dataset, that is, smallest year found in the dataset up to largest year found in the dataset.) (2 points)

*e:sknikolak@aueb.gr

5. Your client is a multinational pharmaceutical company. They are interested in the probability that Belgium (BEL) will increase its drug expenditure in at least 4 of the 5 following consecutive years, in order to assess the investing opportunities. Assume that we are at the year following the last record for Belgium. Estimate the yearly probability of drug expenditure increase by the proportion of years where the expenditure was higher than the year before (note: this is not an appropriate way of estimating this probability, only use it for this assignment!). (Hint: functions `diff()` and `dbinom()`). Create a list with the following elements, named accordingly:
`$Data` The data for Belgium
`$Years` The range (in years) of available data points for Belgium, a vector with two elements, the minimum and maximum years
`$Data.points` The number of available data points for Belgium
`$YearlyProbs` The yearly probabilities of increase (probability of increase per year) in expenditure, in all the four metrics available, thus a vector with 4 elements and names according to the metric
`$FiveYeProbs` The requested probabilities (probability of increase in at least 4 out of 5 consecutive years) for all metrics, thus a vector with 4 elements and names according to the metric
(2.5 points)

6. Your client asks for a function that can calculate the above probabilities for a variable amount of years and for any country desired. Create a function, that it will take as arguments (and default values):

   - `DATA=NULL` Data in the same form as the country-specific dataset (without the FLAG_CODES variable). **Attention**: The data input for this function should consist of data concerning one country only (as you did in the previous question with Belgium).
   - `METRIC="pc.gdp"` The metric in which the required probability needs to be reported. Possible values should be "pc.gdp" (% of GDP), "pc.tot" (% of total health expenditure), "per.ca" (absolute expenditure per capita), and "total" (total absolute spending).
   - `nofY=5` The number of following consecutive years that the probability needs to be estimated. Every time the probability estimated should be of the form "The probability of increasing drug expenditure in at least nofY-1 out of the following nofY years".

   The outcome of the function should be a sentence of the form:

   "Based on (XX) datapoints from years (minYear) to (maxYear), the probability that (countrycode) will increase its drug expenditure, in terms of (metric chosen), in at least (nofY-1) years in the period (maxYear+1) to (maxYear+1+nofY) is (estimated probability)".

   If the number of available datapoints for the calculation of the yearly increase probability is less than 10 (thus, less than 11 years of data), the function should return : "Unable to calculate probability (n< 10)", without any other output. Hint: Use `paste()` (2.5 points)