Data Science Fundamentals
Assignment for November 1

Johannes Binswanger

Fall 2017

Instructions for this assignment

- You can work together in groups for this assignment. However, every course participant should submit an **individual** solution.
- Please provide the code for your solutions in one single R script file (i.e. a file which ends with `.R`). Name the file `<LastName>_<FirstName>_Nov1.R`.
- Submit your solution on StudyNet via `Abgabe/Assignment Nov 1`. **The deadline is Nov 1, 11.55PM**.
- Make sure to stick to the variable, column and other object names mentioned in the problems below.

Good luck!

Problem 1: Real quarterly changes for Swiss exports

Part 1: Data pre-processing

In the data on exports that we obtained from the Swiss National Bank (SNB), the relative changes in data corrected for changes in the price level ("real"), indicated by the letter `R` in column `D2` of our `trade` data (or indicated as `_R` in the variable name in the dataframe `D`) are growth rates with respect to the previous *year*. (That was a long sentence, did you get it?) In contrast, the growth rates that we calculated for exchanges rates are with respect to the previous *month*! So there is a mismatch. The only way to obtain real growth rates for exports thare are synchronized with growth rates in exchange rates is to download data on the price level, merge them with our dataframe `D` , deflate the respective data series on exports and then calculate synchronized growth rates. In this problem, you are asked to do exactly this.

Although you have seen most of the functions for solving this problem in the course, there may be some issues on which you may read in the book or consult Google. (Googling is a a highly important skill for coders!)

Go the the data portal of the SNB; under "other areas of the economy" choose "prices and salaries/wages". Then, in the left-hand side menu, choose "producer and import prices". Select data from January 2001 to most recent. Then download as "csv (selection)". Put the downloaded file in the `Data` folder where you have the other data from the SNB that we have worked with in the course. Write an R script that solves the below problems. The destination folder for this script is the folder `Programming`, i.e. the same folder where you have the other scripts for the programming part of this script (important for correction). If you have problems with reading the data, close RStudio and then repen it by double-clicking on the respective `.Rproj` item in the `Programming` folder.

1. Read the data into R. Name the data `prices0`.
2. Find out how to identify rows that contain information on the index of producer prices for: 1) the price level in export markets; and 2) the total index for producer prices (this requires some detective skills). Select from `prices0` only the corresponding rows and name the resulting new and smaller dataframe `prices`.

3. Delete all rows in `prices` for which there are no valid observations (i.e. the value is `NA`). Call the result again `prices`.
4. Delete the `D0` column. Call the result again `prices`.
5. Give the entries in `D1` some more telling names and call the resulting dataframe again `prices`. In particular, change the cell entries for the name of the price index related to exports to `p_exp` and for the price index related to the total index for producer prices to `p_tot`.
6. Make the data set "tidy", call the result `X`. You may want to read the respective parts of Chapter 9 in the book.
7. Load the dataframe `D` that we worked on during the class. From `D`, add the columns `Exp_All_WMF`, `EUR`, `USD`, `gEUR`and `gUSD` to `X`, call the result `X`. Be careful that the rows are matched according to the appropriate dates. Code this matching explicitly. The resulting dataframe (`X`) should have about 11 columns (at least not much more than that).
8. Calculate a new column for which the values of the price level `p_tot` are expressed *relative* to the value of that same price level in January 2001 (i.e. divide all values by the value in January 2001). Name the new column `p_tot_rel`. You may get an error (if you don't, then ignore the following remarks). Why do you get this error? Adjust your columns in an appropriate way such that the calculation works.
9. Calculate a new column for which the values of `p_exp` are all expresed relative to the *first available value* for `p_exp`. Name the column `p_exp_rel`.
10. Calculate two new price-adjusted columns for `Exp_All_WMF`. These are defined by dividing that column by the values of `p_tot_rel` and `p_exp_rel` referring to the same time period (i.e. the same month). Name the results `exp_rptot` and `exp_rpexp`.
11. Calculate growth rates for `Exp_All_WMF`, `exp_rptot`, and `exp_pexp`.

Part 2: Plotting and research conclusions

Now we have a number of different measures for exports, we want to inspect whether they make a difference. For this we use the `ggplot` function to create plots.

1. Make a plot that contains all three data series `Exp_All_WMF`, `exp_rptot`, and `exp_pexp` in different colors. There is no need to make the layout particularly nice. The important thing is that the plot allows you to get an answer to the following question: Is there, from purely visual inspection, any evidence that the three series deviate in an important way?
2. Calculate a correlation matrix between all three series (you may have to google about how this works). Since the series deflated with the export price level is only available for later dates, you will have to make a separate analysis for: 1) all dates and only 2 series; and 2) the later dates, with all 3 series. You may still get `NA` values in your correlation matrix. How can you avoid that? What is your conclusion about the three series?
3. Make a plot that shows time in the x dimension and the series for the (monthly) *levels* (i.e. *not* growth rates) of `Exp_All_WMF` and the CHF/EUR exchange rate in the y dimension. Why can you not see much this way? How can you change one of the time series to make the plot more informative (but still with levels and not growth rates)?
4. Also make a plot that shows the series for the *growth rates* of the same two variables (in the y dimension, and time in the x dimension).
5. Now make a scatterplot with the *growth rates* of the same two variables. Put exports on the x-axis. Do you see any interesting relationship between the two variables.
6. Calculate a regression with growth rates of exports as the y variable and growth rates of the exchange rate as the x variable. Do you see any significant correlation between the two variables?

7. Calculate lagged values of monthly growth rates for the CHF/EUR exchange rate for up to 6 months. Then run a regression with the growth rate of exports as y variable, and the current and all the lagged values of growth rates in exchange rates as x variables. What do you conclude?

Problem 2: The WDBC Data

In this problem you are asked to download data that we will use for machine learning (which we haven't covered so far). Some machine learning applications can be ethically quite "charged". For this reason, we want to have some pretty "heavy" data that make us aware of the ethical dimensions involved. So this data is about breast cancer cases. We are interested in developing an algorithm that learns to classify the cases into benign and malign categories. We will do so in class later. For now, you are just asked to download the data and read them into R in an appropriate way. The below instructions are consciously chosen somewhat less detailed. You need to practice your data acquisition skills!

1. Google "breast cancer wisconsin" and find the respective WDBC data by clickling through the respective website. We want to work with the 1996 data. When you look at the data, they are actually just on a website. Save them to your Data folder in an appropriate format. Then read them into R.
2. When you look at the data, you will see that they have no names. So go back to the website where you found the data and engage in some detective work to find out what the names should be. Assign appropriate names to the 32 columns in the data using R code.
3. Save the resulting data with named columns in `.RData` format in your data folder.

Problem 3: Install Latex (not graded)

Install TeX/Latex with the help of the installation instructions that you find in the Dropbox folder. This is required for all Windows users. Among Mac users, only those who could not compile R Markdown do pdf need to install TeX/Latex. Please note that this may take more than 1 hours (in the case of Windows)!

Problem 4 (not yet clear whether graded ;-)

Please put your picture on our "sheet of faces", if you do not mind (and have not yet done so). Here is the link: https://docs.google.com/spreadsheets/d/1QShBGsJZmKn3temsMR862cM0UrPAo294FjgqEVcrXSw/edit#gid=0