Project on Replication, Regression, Correlation and Monte Carlo

Quantitative Methods in Finance — Fall 2023

Due: Thursday, December 14th, 2023 at noon

This project asks for you to collect and work with data in a realistic setting. Relative to the R exercises you have done to this point in the course, the project is more open ended (a feature meant to resemble more realistic data analysis environments). For this assignment,

1. Write a brief and insightful typed report on your findings using the data.
2. As an appendix to your report, include the relevant R code and output (script files, log files, etc.) you used to perform the calculations you discuss in the report.
3. In the main text of your report, you should have typed tables and figures that help you summarizenicely the relevant computations and graphics from this output (not all output is relevant to a coherent discussion).

Be careful to present your work with proper grammar and professional tone. Your grade will reflect not only your calculations, but the manner in which you interpret them and the form in which you present them. You should submit your final project write up in Canvas by the due date.

John Taylor's Scatter Plot

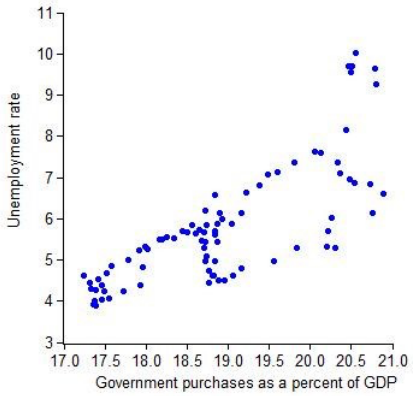
In 2011, an economist named Justin Wolfers wrote a blog post about a striking scatterplot produced by another economist John Taylor on his own blog. A little on the history of this project: I initially wrote it in 2011 and some interesting features kept coming up. Read Taylor's original post,[1] Wolfers' first response,[2] and Wolfers' second response.[3] In the infamous post that generated the controversy, John Taylor actually had two scatter plots. Although the side discussion is interesting controversy, this exercise asks you to produce the other scatter plot relating the unemployment rate to government purchases (presented below). This plot received considerably less attention, but is nonetheless interesting. In reading this exchange, you might be tempted to support one side or the other. Do not let that temptation get in the way of interpreting the data dispassionately (you might be surprised, then surprised again; in fact, if you do the project correctly, that will be the correct feeling). Here is the "other" scatter plot and how Taylor describes it in his post:

[1] https://economicsone.com/2011/01/14/higher-investment-best-way-to-reduce-unemployment-recent-experienceshows/

[2] http://www.freakonomics.com/2011/03/30/how-to-spot-advocacy-science-john-taylor-edition/

[3] http://freakonomics.com/2011/04/01/graph-fight-more-on-taylors-scatterplot/

Some economists argue that the efforts now underway to reduce government spending as a share of GDP will have adverse effects on unemployment. This is not what the data show. Consider this chart which shows the pattern of government purchases as a share of GDP and the unemployment rate over the past two decades. (The data are quarterly seasonally adjusted from 1990Q1 to 2010Q3.) There is no indication that lower government purchases increase unemployment; in fact we see the opposite, and a time-series regression analysis to detect timing shows that the correlation is not due to any reverse causation from high unemployment to more government purchases.



Your task is as follows:

1. Construct the data set that generates this scatter plot. The source is FRED, [4] a great source for data on macroeconomics aggregates. When you download the data from FRED, it is helpful to save in .csv format and read into R using read.csv().[5] Make sure you have obtained the right series of data. You may need to transform data you can download into the variables that Taylor plots. It may be helpful to merge the data into one data frame (merge() is useful for this). Finally, for the later parts of this problem, make sure you download the entire series of observations, not just the ones that Taylor plots. Hint: In the macro aggregate data, "Government Purchases" is called "Government Consumption Expenditures and Gross Investment."

2. Use the data set you constructed to reproduce Taylor's other scatter plot. That is, produce ascatter plot using only the observations from Q1 of 1990 to Q3 of 2010. Because of updates to government statistics, the plot should be very close (but not exactly the same) as the above plot.

   (a) As you do this, try to format the dates as dates (look up how to do this using the as.Date() function). This will help you to produce nice looking plots within R when you want to do time series plots.

[4] http://research.stlouisfed.org/fred2/

[5] There are other ways to download data from FRED – notably, you can install the FRED Excel add-in, and this will be quite effective.

3. Compare the scatter plot you obtain using the full data set to the one that Taylor reports.

   (a) For this comparison, you may want to use different plotting characters and colors, and thepoints() function to produce different plotting characters for different parts of the sample.

   (b) Contrast the correlations across three sample periods: (i) Taylor's sample period 1990Q1:2010Q3,(ii) pre-Taylor 1948:1989, and (iii) post-Taylor 2010Q4:present.

      i. Knowing that Wolfers likely knew the data from pre-Taylor before writing his post, could Wolfers have cherry picked his criticism of Taylor? ii. Would either Taylor or Wolfers know how the data from 2011 to present would look? Whose world view is most consistent with this out-of-sample evidence?

4. Analyze the statistical relationship between seasonally-adjusted unemployment and government expenditures in a way that is informative and insightful. Some questions you should consider as you perform the analysis:

   (a) Is the relationship that Taylor presents in his scatter plot stable over time? In your analysis, you could analyze the relationship by decade, or use multiple regression techniques to learn something similar. In addition to regression output, your analysis should include informative plots that support your main points.

   (b) Per the multiple regression unit, your analysis should use a regression analysis that includes interactions with appropriately defined dummy variables.

   (c) Write up your comments *concisely* in a typed report.

5. Write code – likely using for loops – to simulate the process of "cherry picking" in Wolfers' terminology, and use it to evaluate the seriousness of this part of the criticism. Some suggestions and refinements to guide your coding journey.

   (a) Specifically, the goal of this exercise is to simulate the sampling distribution of "cherrypicked" correlations under the assumption that the full data series is comprised of two entirely uncorrelated random variables.

   (b) For each replication, your code should generate two uncorrelated variables x and y (e.g., from a Normal distribution) over a time series of 256 quarters (this number is meant to match the number of quarters in the data set when Taylor and Wolfers had their blog fight).

   (c) Within each replication, use R to compute the correlation between x and y over the last 60 quarters, over the last 61 quarters, over the last 62 quarters.... all the way until... over the last 256 quarters (the full sample that Taylor and Wolfers fought about).

      i. A researcher who is "cherry picking" the begin date will then select the highest correlation. For each simulated time series of 256 quarters, write code to loop through subsets & compute the correlation for each subset. Then, based on these calculations, store the maximum "cherry picked" correlation among all of the subsets. This is the "cherry picked" correlation. Also, for comparison, you should also store the correlation over the entire sample ("not cherry picked").

(d) To gain a picture of the sampling distribution, repeat this process for 1000 possible "cherrypicked" correlations, and by comparison, "not cherry picked" correlations. Store the output for drawing a concrete comparison later.

   i. Are the cherry picked correlations different than the not cherry picked correlations? Use both visual (plots to compare) and numerical evidence (t-tests).

   ii. Using the 1000 simulated cherry picked sample of correlations (i.e., your approximation to the sampling distribution under the assumption that the data were "cherry picked"), compute a test statistic and a p-value to evaluate whether it is plausible that Taylor's computed correlation came from a cherry picking process like the one you simulated.

(e) Relate the results from your simulation back to Wolfers' criticism of Taylor's selected sample. In particular, is it accurate to describe Taylor's plot as "cherry picking, plain and simple"?