Recitation 2

Big Data Science

Friday March 31st , 2023

Assignment Objectives

- Replicating research articles
- Working with financial data
- Fit parametric and non-parametric models
- Evaluate the performance of models

Welch and Goyal (2008)

- Predicting the stock market is a tedious task
- The literature suggests several variables (dividends, earning price, interest rates etc.) but different articles uses different techniques, making it hard to absorb
- Data recency may also change the predictability of a model

Welch and Goyal (2008)

“ The goal of our own article is to comprehensively re-examine the empirical evidence as of early 2006, evaluating each variable using the same methods (mostly, but not only, in linear models), time-periods, and estimation frequencies. The evidence suggests that most models are unstable or even spurious. Most models are no longer significant even in-sample (IS), and the few models that still are usually fail simple regression diagnostics. “

Question 1

Load the dataset

- original dataset (up to 2005)
  - Original dataset
- updated dataset (up to 2021)
  - Updated dataset

Question 2

- Recreate explanatory variables as provided in Table 1, read the paper to understand how some variables are computed.

- **The explanatory variables to be used:** *Default Yield Spread, Inflation, Stock Variance, Dividend Payout ratio,Long Term Yield, Term Spread, Treasury-bill rate, Default Return Spread, Dividend Price Ratio, Dividend Yield, Long Term Return, Earning price ratio, Book to market, Net Equity Expansion*

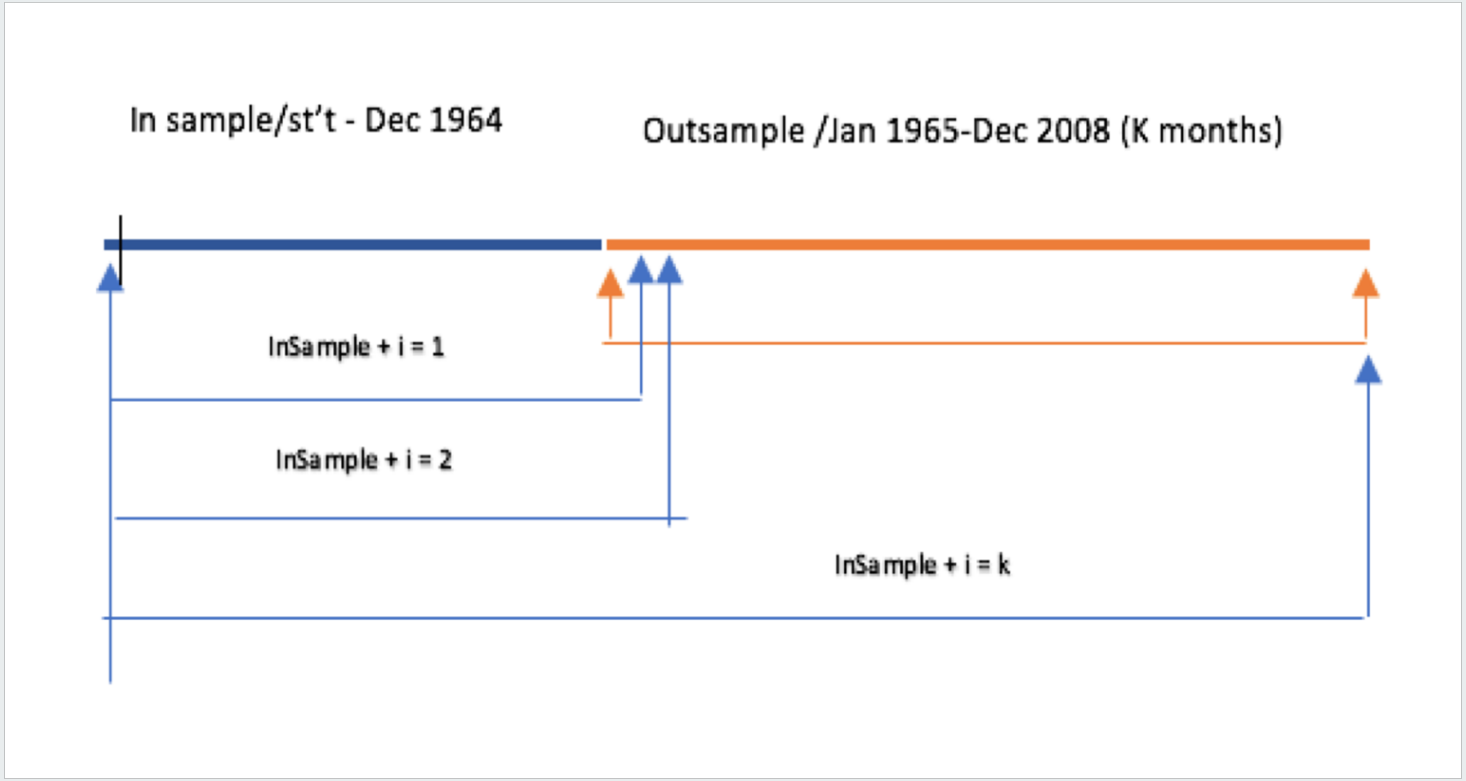
- Plot time series for each of the variables with label

Question 3

- Rolling multiple linear regression model as described in Welch and Goyal (2008)
  Training: **Insample_x**, **Insample_y**
  Testing: Outsample: array of “the next value”

- CRSP_SPvw : total rate of return

- No filling in of missing values.


In sample/st't - Dec 1964
Outsample /Jan 1965-Dec 2008 (K months)
InSample + i = 1
InSample + i = 2
InSample + i = k

Question 4

- Improve on the performance of the model proposed by Welch and Goyal (2008) using any two of the following:

  (a) Economically motivated model restrictions
  (b) Forecast combinations
  (c) Regime switching
  (d) Machine learning (e.g. Lasso, CART, KNN, RNN or SVM)
  (e) Bagging and boosting

Question 4 - cont'd

- Using the same time frames and providing R2, RMSE and MAE values, compare the results of your approach with Welch and Goyal (2008)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This new process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID-BDS-AssignmentNo. For example, mcsharry-BDS-Assignment1, mcsharry-BDS-Assignment2 and mcsharry-BDS-Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

Q&A

Remember to install the ArcGIS software!