Econometria 4
Departament of Economics/Pontifical Catholic University of Rio de Janeiro
Second Assigment
Delivery Date: Aug 15, 2023, 11:59 pm

The second assignment consist of **two empirical questions**.

Your answers to the homework assignments must be completed individually: you may discuss the assignment with your classmates but are not allowed to share spreadsheets, calculations or compare answers.

The following rules apply:

- The answers to the questions can be delivered as Python notebooks, R markdown, Matlab reports, or written in a word processing software.
- The answers must be sent by email in PDF format by the due date and time. Late homeworks will be penalized. Please use the following convention to name your PDF file: LastName_FirstName.pdf

Question	Points	Bonus Points	Score
1	120	0	
2	120	0	
3	220	0	
Total:	460	0	

No not write on the table above.

Good Luck!

1. The first question consists of a factor analysis of a large dataset. We consider monthly close-to-close excess returns from a cross-section of 9,456 firms traded in the New York Stock Exchange. The data starts on November 1991 and runs until December 2018. There are 326 monthly observations in total. In addition to the returns we also consider 16 monthly factors: Market (MKT), Small-minus-Big (SMB), High-minus-Low (HML), Conservative-minus-Aggressive (CMA), Robust-minus-Weak (RMW), earning/price ratio, cash-flow/price ratio, dividend/price ratio, accruals, market beta, net share issues, daily variance, daily idiosyncratic variance, 1-month momentum, and 36-month momentum.

   The dataset is organized as an excel file named `returns.xlsx`.

   (a) (30 points) Compute the principal components of the returns and determine the optimal number of principal factors by one the methods described in Lecture 2. How much of the variance will the factors be able to explain?

   (b) (30 points) Regress the selected factors on the 16 observed "anomaly" factors described above. How do the "principal component factors" relate to the "anomaly factors"?

   (c) (30 points) Now, run a principal component analysis on the 16 "anomaly factors" and select the optimal number of principal components using the same criterion adopted in the first item of the exercise. By inspecting the principal eigenvectors can you identify a dominating "anomaly"?

   (d) (30 points) How do the "anomaly-based principal factors" related to the "return-based principal factors"?

2. The second question consists of an **inflation forecasting** exercise using a large set of monthly macroeconomic variables. We compare four different models: (1) Autoregressive model; (2) Principal Component Regression (PCR); (3) Ridge Regression; and (4) Lasso.

   The data consist of variables from the FRED-MD database:

   https://research.stlouisfed.org/econ/mccracken/fred-databases/,

   which is a large monthly macroeconomic dataset designed for empirical analysis in data-rich macroeconomic environments. The dataset is updated in real time through the FRED database.

   We will use the vintage as of December 2021, the same used for the first assignment. The sample extends from January 1959 to November 2021 (755 observations). You should consider only variables with all observations in the sample period (104 series). The dataset is divided into eight groups: (i) output and income; (ii) labor market; (iii) housing; (iv) consumption, orders and inventories; (v) money and credit; (vi) interest and exchange rates; (vii) prices; and (viii) stock market. Finally, all series should transformed in order to become approximately stationary. Details about the database and the transformations can be found in

   https://research.stlouisfed.org/wp/more/2015-012.

   The data is in CSV format where each column is a different variable. The first row is the variable identifier and the second row indicates with transformation should be used to each variable.

   The dependent variable for this question is **CPIAUCSL** (CPI all items), column 107 in the original CSV file. **Important:** the original data is in price levels. The suggested transformation is to take second differences of the log of the levels. **However, for this exercise you should construct inflation series: $\pi_t = \frac{\Delta y_t}{y_t}$, where y_t is the price level at time t.**

   You should construct one-step-ahead forecasts for inflation according to the following models:

1. **Autoregressive model (`AR`):**

$$\widehat{\pi}_{t+1|t}^{(\texttt{AR})} = \widehat{\phi}_0 + \widehat{\phi}_1 \widehat{\pi}_t + \ldots + \widehat{\phi}_p \widehat{\pi}_{t-p+1},$$

where $\widehat{\phi}_0, \widehat{\phi}_1, \ldots, \widehat{\phi}_p$, $i = 1, \ldots, n$, are OLS estimates. **The order of the AR model must be determined by the BIC criterion.**

2. **AR + Principal Component Regression (`PCR`):**

$$\widehat{\pi}_{t+1|t}^{(\texttt{PCR})} = \widehat{\phi}_0 + \widehat{\phi}_1 \widehat{\pi}_{i,t} + \ldots + \widehat{\phi}_{ip} \widehat{\pi}_{t-p+1} + \widehat{\boldsymbol{\lambda}}' \widehat{\boldsymbol{F}}_t,$$

where $\widehat{\boldsymbol{F}}_t$ is the estimate of the $(k \times 1)$ vector of factors $\boldsymbol{F}_t$ given by principal component analysis of the full dataset. **The number of factors in the model but be determined by one of the criteria learned during Lecture 2.**

3. **Ridge Regression (`PCR`):**

$$\widehat{\pi}_{t+1|t}^{(\texttt{Ridge})} = \widehat{\phi}_0 + \widehat{\phi}_1 \widehat{\pi}_{i,t} + \ldots + \widehat{\phi}_{ip} \widehat{\pi}_{t-p+1} + \widehat{\boldsymbol{\beta}}_1' \boldsymbol{X}_t + \cdots + \widehat{\boldsymbol{\beta}}_p' \boldsymbol{X}_{t-p+1},$$

where $\boldsymbol{X}_t$ are the variables in the dataset, except CPIAUCSL. **The parameters of the above model should be estimated by Ridge Regression with the penalty term selected by the BIC criterion.**

4. **LASSO Regression (`LASSO`):**

$$\widehat{\pi}_{t+1|t}^{(\texttt{LASSO})} = \widehat{\phi}_0 + \widehat{\phi}_1 \widehat{\pi}_{i,t} + \ldots + \widehat{\phi}_{ip} \widehat{\pi}_{t-p+1} + \widehat{\boldsymbol{\beta}}_1' \boldsymbol{X}_t + \cdots + \widehat{\boldsymbol{\beta}}_p' \boldsymbol{X}_{t-p+1},$$

where $\boldsymbol{X}_t$ are the variables in the dataset, except CPIAUCSL. **The parameters of the above model should be estimated by LASSO Regression with the penalty term selected by the BIC criterion.**

The forecasts are based on a rolling-window framework of fixed length of 492 observations, starting in January 1959. Therefore, the forecasts start on January 1990. The last forecasts are for November 2021. More specifically, the rolling window forecasting scheme can be described as follows:

1. Run all in-sample analysis and estimation using data from observation a to observation $a + 492 - 1$.
2. Compute the forecast for observation at position $a + 492$.
3. Set $a = a + 1$ and repeat the two steps above.

The exercise consists of the items below.

(a) (60 points) For each forecasting window, compute the squared one-step-ahead forecasting error for the next observation. Plot the cumulative squared errors over the forecasting window. Briefly comment the results with respect to the following points:

   1. Is there a winner model?
   2. Do the results change over the sample, specially after Covid-19?

(b) (60 points) For `PCR`, `Ridge`, and `LASSO`, compute a measure of variable importance. As there will be several estimation windows and a large number of variables, you should proceed as follows:

1. Report results averaged over the estimation windows.
2. Report results grouped in each one of the eight categories plus an additional one representing the lags of inflation.

More specifically, you should, for each forecasting window, compute the variable importance for each variable. Then, you should aggregate the importance of each variable into its respective group. For example, if the variable is a lagged value of inflation, you should include it in the "lag" group; if it is a variable belonging to group "Output and Income", you should include it in this category; so on and so forth. In order to keep things comparable, you should normalize the importance measure. For example, 100 is maximum importance.

Now the question is how to measure importance for the above models. For `Ridge` and `LASSO` this is quite simple: it is just the estimated coefficient multiplied by the variable standard deviation. You should multiply by the standard deviation in order to put all coefficients in the same scale.

For `PCR`, the computation is a bit more complicated. Recall that each factor F_{it} is a linear combination of the original variables: $F_{it} = \boldsymbol{\alpha}_i' \boldsymbol{X}_t$. Therefore, the relative importance of each variable will be given by the product of the respective element of the vector $\boldsymbol{\alpha}$ and the estimated parameter in the `PCR` model. Do not forget to standardize your results.

3. The third question consists of an **inflation forecasting** exercise using a large set of monthly macroeconomic variables and nonlinear models. As in the previous question, the forecasts are based on a rolling-window framework of fixed length of 492 observations, starting in January 1959. Therefore, the forecasts start on January 1990. The last forecasts are for November 2021. More specifically, the rolling window forecasting scheme can be described as follows:

   1. Run all in-sample analysis and estimation using data from observation a to observation $a + 492 - 1$.
   2. Compute the forecast for observation at position $a + 492$.
   3. Set $a = a + 1$ and repeat the two steps above.

   (a) (60 points) Estimate a model based on a neural network specification and compute one step ahead forecasts. You are free to choose between shallow, deep, convolution or LSTM networks. However, you have to motivate your particular choice.

   (b) (20 points) Plot you forecasts against the four linear benchmarks from Question 2. Comment on you results.

   (c) (20 points) Compute the mean squared error of the NN-based model and the benchmarks. Does the NN model outperform the linear alternative?

   (d) (60 points) Estimate a model based on regression trees and compute one step ahead forecasts. You are free to choose between random forests or boosted-trees.

   (e) (20 points) Plot you forecasts against the four linear benchmarks from Question 2. Comment on you results.

   (f) (20 points) Compute the mean squared error of the tree-based model and the benchmarks. Does the tree-based model outperform the linear alternative?

   (g) (20 points) Compare the performance of the tree-based model and the NN specification.