Harvard Extension Data Science

Dynamic Modeling and Forecasting in Big Data

Instructor: William Yu

Assignment 4

Predicting Car Sales

- In this assignment, we would like to build a forecast model from scratch for car sales in the future from using the historical data. Follow my guidelines below.
- Use quantmod library in R (or fredapi library in Python) to import the following series from FRED in which the Total Vehicles Sales will be the dependent variable. The rest of variables will be potential predictors. Feel free to add more variables if you think they will help.
  - *getSymbols("TOTALSA", src="FRED") # Total Vehicle Sales*
  - *getSymbols("PAYEMS", src="FRED") # US payroll jobs*
  - *getSymbols("DFF", src="FRED") # Federal fund rates*
  - *getSymbols("UNRATE", src="FRED") # Unemployment rates*
  - *getSymbols("NASDAQCOM", src="FRED") # Stock market prices*
  - *getSymbols("USSTHPI", src="FRED") # Housing market prices*
  - *getSymbols("CUSR0000SETB01", src="FRED") # Gasoline price in CPI*

- Show the visualizations of these variables.
- Convert the data frequency (there are various frequencies) from daily or monthly to quarterly. (Think about why I suggested this. If you have time, you can try monthly series as a comparison.)

  Note: e.g. you can use "apply.quarterly" function.
- Select the sample period from 1976Q2 to 2023Q2 (June) as the trainset (in-sample).
- Are these time series stationary? If not, transform those non-stationary series to stationary one. Think about why I suggested doing this.

Part A. A Structural Model

- Put all these series together as a data frame and start to build/train a structural model by using a simple OLS model with the trainset data.

  Note: you might need to use "ts" function to make combine these time series from different sources.
- Once you find a model you like, you are ready to forecast the car sales in 2023Q3 to 2024Q2 (testset; out-of-sample).

- Remember what we said in the class in order to use a structural model to forecast in real-time. You will need to forecast all your explanatory variables first. And that is exactly what I wanted you to do. But how?
- Here I suggest a simple way: use ARMA models to forecast all the significant predictors you selected individually. And then you can forecast car sales in 2023Q3 to 2024Q2.
  Note: In this case, we in fact know the values of explanatory variables in the testset. In most real-world situations, we don't know what they will be in the real time and therefore we need to forecast them. Here we just pretend we don't know those values of explanatory variables in the testset.
  In some situation, we could use the values of explanatory variables in the testset as known. That case will be in your next assignment.

Part B. A Reduced-form Model

- Secondly, try a simple reduced form model. Simply run an ARMA model directly on total vehicles sales and forecast it.
- Show these two models' car sales forecast in 2023Q3 to 2024Q2 and compare them to the real one (calculate for forecast errors in terms of number of cars sold).

Part C. A Dynamic Regression Model; Autoregressive Distributed lag model; Mixed-form Model

Part C is optional for students pursuing undergraduate credit.

- Thirdly, try the following model in the trainset and forecast its testset error.
- Cars Sales (t) = a + b1*Car Sales (t-1) + b2*X1 + b3*X2 + …