Unit 10 Live Session

Time Series Analysis Lecture 5: Vector Autoregressive (VAR) Models

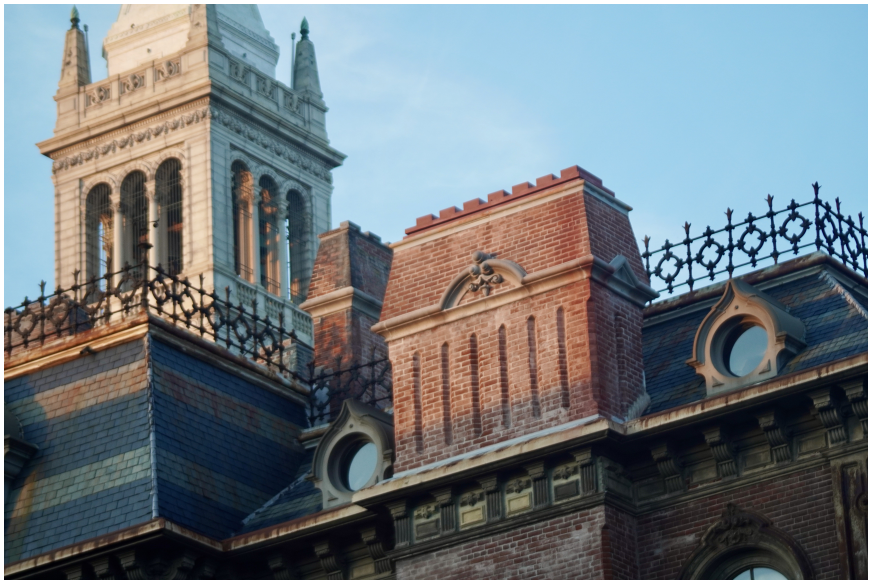

Figure 1: South Hall

Class Announcements

- Lab-2 due in 1 week

Roadmap

Rearview Mirror

- Univariate Time Series Models
  - Estimation, model diagnostics, model identification, model selection, assumption testing, and statistical inference/forecasting

Today

- Regression with multiple trending time series
- Spurious regression
- Cointegration
- Multivariate Time Series Models: Vector Autoregressive (VAR) model
- Notion of cross-correlation

Looking Ahead

- Introduction to panel data
- Using the OLS regression model on panel data
- Exploratory panel data analysis
- First-Difference models
- Distributed Lag models

Start-up Code

```
# Load required libraries
## Load a set of packages inclusing: broom, cli, crayon, dbplyr , dplyr, dtplyr, forcats,
#googledrive, googlesheets4, ggplot2, haven, hms, httr, jsonlite, lubridate , magrittr,
#modelr, pillar, purrr, readr, readxl, reprex, rlang, rstudioapi, rvest, stringr, tibble,
#tidyr, xml2
library(tidyverse)
## To laod All data sets in the book "Forecasting: principles and practice"
#by Rob J Hyndman and George Athanasopoulos
library(fpp3)
# To create and work with tidy temporal data
library(tsibble)
# To work with date-times and time-spans
library(lubridate)
# Provides a collection of commonly used univariate and multivariate time series models
library(fable )
## To interact directly with the Quandl API and download data
library(Quandl)
# For analysing tidy time series data.
library(feasts)
# Provides methods and tools for displaying and analyzing univariate time series forecasts
library(forecast)
# For estimation, lag selection, diagnostic testing, forecasting, and impulse response functions of VAR
library(vars)
#provides tools for statistical calculations
library(stats)
# To assist the quantitative trader in the development,
#testing, and deployment of statistically based trading models.
library(quantmod)
# For statistical analysis
library(car)
## To retrieving and displaying the information returned online by Google Trends
library(gtrendsR)
# To do time series analysis and computational finance.
library(tseries)
```

Multivariate Time Series Models

The goal of a researcher working with time series data does not differ much from that of a researcher working with cross-sectional data. They use regression to explore the relationship between two or more variables.

However, we will face three problems in time series data that we will not encounter using cross-sectional data:

1- A time series variable can be influenced by lags of itself or lags of other variables

2- If the variable is non-stationary, a problem known as spurious regression may arise.

We can address the first problem by following two types of models:

1- The DL(q) (Distributed Lag) Model:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_q X_{t-q} + \epsilon_t$$

Here, the effect of the explanatory variable does not happen all at once but over several periods, and DL(q) incorporates such dynamic effects.

2- ARDL(p,q) (Autoregressive Distributed Lag) Model

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_q X_{t-q} + \epsilon_t$$

Here, The dependent variable Y depends on p lags of itself and q lags of X.

Estimation and interpretation of the DL(q) and ARDL(p,q) model depends on whether the series X and Y are stationary or not.

If Y_t and X_t are stationary, the DL(q) and ARDL(p,q) models can be estimated consistently by ordinary least squares.

However, if Y_t and X_t are non-stationary, a problem known as spurious regression may arise.

Spurious Regression Example

- In time series analysis, we have to be particularly careful since an apparent relationship with significant coefficients of the (spurious) regression can be obtained not because the response 'truly' depends on the explanatory variables but because of the deterministic or stochastic time trends "hidden" in these variables.

Case-1: Both Y and X are Trend stationary (TS)

- If Y_t is stationary, or its residuals ϵ_t in the decomposition $Yt = T_t + S_t + \epsilon_t$ are stationary, then Y_t is called a **Trend Stationary (or TS)** series;

Case-2: Both Y and X are Difference Stationery (DS)

- If Y_t has a unit root then its difference $\Delta Y_t = Y_t - Y_{t-1}$ is stationary, and it is called a **Difference Stationary (or DS)** series;

A correct way to distinguish between TS and DS stationary is through a unit root test.

Spurious Regression when both Y and X are Trend stationery (TS)

- Consider the following time series:

$$Y_t = 0.1 + 0.2 \cdot t + W_t$$

$$X_t = 0.3 - 0.1 \cdot t + W_t$$

- W_t is Gaussian white noise with mean zero and standard deviation 1.

a) Simulate 100 realizations of Y_t and X_t and estimate following model. What do you notice?

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

b) Now estimate the following model. What do you expect before estimating the model?

$$Y_t = \beta_0 + \beta_1 t + \beta_2 X_t + \epsilon_t$$

Spurious Regression when both Y and X are Difference Stationery (DS)

- Assume X and Y are two independent random walks without drift:

$$Y_t = Y_{t-1} + W_t$$

$$X_t = X_{t-1} + W_t$$

- W_t is Gaussian white noise with mean zero and standard deviation 1.

a) Randomly draw 100 observations from Y_t and X_T and estimate the following regression:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

b) Estimate a model on variable differences. What do you notice?

$$\Delta Y_t = \beta_0 + \beta_1 \Delta X_t + \epsilon_t$$

Cointegration

- The one time we do not have to worry about the spurious regression problem occurs when X and Y are cointegrated.
- If Y and X have unit roots but some linear combination of them, $\gamma_1 Y_t + \gamma_2 Xt$, is (trend) stationary, then we say that Y and X are cointegrated.
- In other words: $Yt \sim I(1)$ and $X \sim I(1)$ are cointegrated if they share a common trend such that $\gamma_1 Y_t + \gamma_2 X_t \sim I(0)$.
- As mentioned above, if Y and X are cointegrated, then the spurious regression problem does not apply; consequently, we can run an OLS regression of the difference of Y on the difference of X and obtain valid results.

Cointegration test

1- Engle-Granger test

The null hypothesis in the Engle-Granger test is no cointegration, and we conclude cointegration is present only if we reject this hypothesis.

H_0 : No cointegration exists

H_1 : Cointegration exists

This cointegration test involves the following steps:

1- Carry out a (Augmented) Dickey-Fuller test on the null hypothesis that Y and X each have a unit root. If both time series are $I(0)$, standard regression analysis will be valid. If they are both integrated to the same order (usually $I(1)$), proceed to the next step.

2- Run a regression of Y on X and save the residuals;

3- Carry out a unit root test on the residuals (without including a constant or a deterministic trend);

4- If the unit root hypothesis is rejected, then conclude that Y and X are cointegrated. However, if the unit root hypothesis is accepted, then conclude that cointegration does not occur.

Thus, if Y_t and X_t are cointegrated, in $Y_t = \alpha + \beta X_t + \epsilon_t$, the error term is $I(0)$. If not, ϵ_t will be $I(1)$. Hence, one can test for the presence of a cointegration relationship by testing for a unit root in the OLS residuals e_t.

2- The Phillips-Ouliaris cointegration test

The Engle-Ganger test assumes that regression errors are independent with a common variance, which is rarely true in real life. The Philips-Ouliaris test improves the Engle-Ganger test by considering that the errors are not white noise.

H_0 : No cointegration exists

H_1 : Cointegration exists

3- Johansen test

Another improvement over the Engle-Granger test is the test developed by Johansen. This test can detect multiple cointegrating vectors.

Case Study: Cointegration and Pairs trading in finance

Basic Idea of Pairs Trading

- Recall that if two time series are cointegrated, they remain close to each other in the long term. In other words, the spread(OLS residual) between them $z_t = y_{1t} - \beta y_{2t}$ is mean-reverting.
- This mean-reverting property of the spread can be exploited for trading, and it is commonly referred to as "pairs trading" or "statistical arbitrage." The idea behind pairs trading is to:

Assume that spread $z_t = y_{1t} - \beta y_{2t}$ is stationary or mean-reverting with zero mean:

- if spread is low ($z_t < -s_0$), then stock 1 is undervalued and stock 2 overvalued:
  - buy the spread (i.e., buy stock 1 and short-sell stock 2)
  - unwind the positions when it reverts to zero after i time steps $z_{t+i} = 0$
- if spread is high ($z_t > s_0$), then stock 1 is overvalued and stock 2 undervalued:
  - short-sell the spread (i.e., short-sell stock 1 and buy stock 2)
  - unwind the positions when it reverts to zero after i time steps ($z_{t+i} = 0$)
- Here s_0 is some threshold like 3 standard deviations of the historical spread.

The profit from buying low and unwinding at zero is $z_{t+i} - z_t = s_0$.

Design of Pairs Trading

- In practice, pairs trading contains three main steps:

1- Pairs selection: identify stock pairs that could potentially be cointegrated.

2- Cointegration test: test whether the identified stock pairs are indeed cointegrated or not.

3- Trading strategy design: study the spread dynamics and design proper trading rules.

Netflix v.s Amazon

- Let us focus on the NFLX vs. AMZN pair, which are the Netflix and Amazon stocks.

a) Use the following code to get the stock prices using the quantmod package.

```
start <- as.Date("2021-01-01")
end <- as.Date("2022-06-01")
symbols <- c("AMZN", "NFLX")
# The auto.assign parameter allows for the returned object to be stored in a local variable rather than the R session's
amazon<- getSymbols("AMZN", src = "yahoo", from = start, to = end, auto.assign = FALSE, return.class= "ts")
netflix<- getSymbols("NFLX", src = "yahoo", from = start, to = end, auto.assign = FALSE, return.class= "ts")
```

b) Plot NFLX and AMZN closing prices, the ACF/PACF, and examine their stationary.

C) Carry out unit root tests for NFLX and AMZN closing prices. Are they stationary? are they both integrated to the same order?

d) Carry out the Engle-Granger test for cointegration

e) Plot the spread(residual) and discuss a possible trading strategy

Vector Autoregressive (VAR) model

Introduction: Granger Causality

- To motivate why the VAR model is essential, we begin by discussing **Granger causality**.
- VARs can be used to investigate Granger causality
- What is Granger causality?
- In the following regression, we call Y_t the dependent variable and X_t the explanatory variable. In many cases, because the latter 'explained' the former it was reasonable to talk about X 'causing' Y.

$$Y_t = \beta_0 + \beta_1 X_t$$

- However, the causality could run in either direction - or both! Hence, when using the word 'cause' with regression or correlation results, a great deal of caution has to be taken, and common sense has to be used.

In the time series data, we can make slightly stronger statements about causality simply by exploiting the fact that time does not run backward!

That is, if event A happens before event B, then it is possible that A is causing B. However, it is not possible that B is causing A. In other words, events in the past can cause events to happen today. Future events cannot.

- These intuitive ideas can be investigated through regression models incorporating the notion of Granger or regressive causality. The basic idea is that a variable X Granger causes Y if past values of X can help explain Y.
- Of course, if Granger causality holds, this does not guarantee that X causes Y. This is why we say 'Granger causality' rather than just 'causality.' Nevertheless, if past values of X have explanatory power for current Y values, it at least suggests that X might be causing Y. Granger causality is only relevant with time series variables.
- For example, consider Granger causality between two stationary variables (X and Y). Since X and Y are stationary, the following ARDL(q,p) model is appropriate.

$$Y_t = \alpha + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + \beta_1 X_{t-1} + ... + \beta_q X_{t-q} + \epsilon_t$$

- **X does not Granger cause Y if all** $\beta_i = 0$**.**

- In many cases, it is not obvious which way causality should run. In such cases, when causality may be in either direction, we must check for it. If Y and X are the two variables under study, in addition to running a regression of Y on lags of itself and lags of X (as above), you should also run a regression of X on lags of itself and lags of Y.
- In other words, we should work with two separate equations: one with Y being the dependent variable and one with X being the dependent variable. These two equations comprise a VAR. A VAR is an extension of the autoregressive (AR) model to the case where there is more than one variable under study.

$$Y_t = \alpha_1 + \phi_{11}Y_{t-1} + ... + \phi_{1p}Y_{t-p} + \beta_{11}X_{t-1} + ... + \beta_{1q}X_{t-q} + \epsilon_{1t}$$

$$X_t = \alpha_2 + \phi_{21}Y_{t-1} + ... + \phi_{2p}Y_{t-p} + \beta_{21}X_{t-1} + ... + \beta_{2q}X_{t-q} + \epsilon_{2t}$$

- The first of these equations test whether X Granger causes Y; the second, whether Y Granger causes X. Note that now the coefficients have subscripts indicating which equation they are in. The errors now have subscripts to denote that they will differ in the two equations.
- The VAR model can be extended to the case of many variables, and we could include more than two variables in a VAR model.
- The variable in VAR should be stationary. If the original variables have unit roots, then we assume that differences have been taken such that the model includes the changes in the original variables (which do not have unit roots).
- If the original variables have unit roots but are cointegrated, then we should work with a vector error correction model (VECM) involving these variables, which is beyond the scope of this course.

The bottom line - if X Granger causes Y, this does not mean that X causes Y; it only means that X improves Y's predictability (i.e., reduces residuals of the model).

VAR: Estimation

Building a VAR model involves three steps:

1- Use some information criterion to identify the order.

2- Estimate the specified model by using the least-squares method and, if necessary, re-estimate the model by removing statistically insignificant parameters.

3- Use the Portmanteau test statistic of the residuals to check the adequacy of a fitted model (this is a multivariate analog of the Ljung-Box Q-stat in an ARIMA model and is to test for autocorrelation and cross-correlation in residuals). If the fitted model is adequate, then it can be used to obtain forecasts.

Analytical Exercise

- Consider the following bivariate VAR

$$Y_t = 0.3Y_{t-1} + 0.8X_{t-1} + \epsilon_{1t}$$
$$X_t = 0.9Y_{t-1} + 0.4X_{t-1} + \epsilon_{2t}$$

- Is this system covariance stationary?

Empirical Exercise: Bitcoin price and public attention

In this exercise, we will examine the linkage between the bitcoin prices and public attention, proxied by Google Trends data. Nowadays, many investors gather market information mainly through the internet, and Google searches signal investors' attention. Google Trends allows analysts to see how often specific terms are searched.

a) Use the code below to pull the following time series from the Quandl and Google trends API.

1- Weakly Bitcoin price from Quandl API since 01/01/2020.

2- Weakly Google search volume for four main words associated with bitcoin, including "Bitcoin," "bitcoin," "BTC," and "btc" from Google trend API since 01/01/2020

```
### qundl
Quandl.api_key("mbGCKg2ifLUx_DmxzbGv")
bitcoin_weekly = Quandl("BCHAIN/MKPRU", start_date="2020-01-01", collapse = "weekly")
head(bitcoin_weekly)
```

```
##         Date    Value
## 1 2022-07-10 20545.61
## 2 2022-07-03 19224.75
## 3 2022-06-26 21481.38
## 4 2022-06-19 18977.51
## 5 2022-06-12 28344.50
## 6 2022-06-05 29845.23
```

```
# The gtrends default method performs a Google Trends query for the 'query' argument and session
# define search keyword
keyword <- c("Bitcoin", "bitcoin","BTC", "btc")
# define the location
geo <- "all"
#define the channels "web","news", "image", "youtube"
grop = c("web")
#define the time window
time <- "all"
#extract trend
google <- gtrends(keyword, geo = "", grop, time = "2020-01-01 2022-06-01")

df <- data.frame(google[1]) %>%
  rename (Date = interest_over_time.date) %>%
  mutate(Date = as.Date(Date)) %>%
  group_by(Date) %>%
```

```
  summarise(google = sum(as.double(interest_over_time.hits)))
## join google trend and bitcoin data
df <- df %>%
  left_join(bitcoin_weekly, by = "Date") %>%
  rename (bitcoin = Value)

## Import data from csv file
#df <- read_csv("./data/google_bitcoin.csv")
```

b) Plot the time-series of bitcoin prices and google trends. Do they look stationary?

c) Plot ACF/PACF the Perform the unit root test on the bitcoin prices and google trends and report the results. Do you reject the null of unit root for them?

d) Now calculate the first difference for the log of bitcoin prices and google search volume. Are they stationary? Test using the unit root tests.

e) Determine the lag length of the VAR using the information criteria. Estimate the VAR and comment on the fit.

f) Test for Granger-causality. Does google trend Granger-cause bitcoin prices? Does bitcoin prices Granger-cause google trend?

g) Do diagnostic checking of the VAR model.

h) We finally conduct a 3-step ahead forecast:

Reminders

- Before the next live session:
  1. Complete and turn in the Lab-2
  2. Complete all videos and reading for unit 11