TIME SERIES ANALYSIS - SPRING 2023

HOME-TAKEN PROJECT

deadline #1: 2023-07-09

deadline #2: 2023-09-10

General information

The home project and its presentation are required for passing the labs section and passing the whole course (for details see the rules of assessment provided in the first lecture).

Home projects should be prepared by **teams of up to two persons**.

Choosing the topic

Each team can choose one out of three projects' topics.

A spreadsheet will be shared with the students for choosing the project's topic on April 26, 2023 at 21:00. There will be teams limit for each topic (maximum 10 teams can choose the same topic). "First come, first served"!

About presentations

Presentations shall contain slightly more than preliminary results not necessarily already finished projects since the deadline is about a month after the presentations.

Presentations schedule

Presentations schedule is:

- group 1: 2022-06-01 starting from 19:30 (after the penultimate labs meeting), up to 6 teams, 10 min. per team
- group 2: 2022-06-05 16:45-18:15 (the penultimate lecture), up to 9 teams, 10 min. per team
- group 3: 2022-06-12 starting from 19:30 (after the last labs meeting), up to 6 teams, 10 min. per team

- group 4: 2022-06-15 16:45-18:15 (the last lecture), up to 9 teams, 10 min. per team

Again, a spreadsheet will be shared with the students for choosing the slots for presentations on April 26, 2023 at 21:00. “First come, first served”!

In case of any questions, please do not hesitate to contact me.

Good luck!

M. Świtała

TOPIC #1

FORECASTING FINANCIAL INSTRUMENTS PRICES

WITH VECM AND ARIMA MODELS

Project description:

The aim of the project is to compare accuracy of forecasts of prices of two cointegrated financial instruments with VECM model and two independent univariate ARIMA models.

In order to accomplish the task you should:

- Download the provided *TSA_2023_project_data_1.csv* file with time series including prices of ten financial instruments.
- Find one cointegrated pair out of ten provided time series. There is more than one cointegrated pair but you are supposed to find just one of them. If you found more than one pair, you can choose any of them for further analysis.
- Build a bivariate VECM model for the identified pair and produce forecasts of prices of two instruments for the *out-of-sample* period.
- Find separately for two instruments the most attractive univariate ARIMA models and produce forecasts for the same *out-of-sample* period.
- Compare accuracy of forecasts of the prices using the *ex-post* forecast error measures.
- Prepare a short report on it.

About the sample:

Please consider the *in-sample* period of 970 observations and *out-of-sample* period of 30 observations.

Detailed grading rules:

The maximum number of points you can get for the project is 40. Detailed grading rules are as follows:

- For checking the cointegration between the chosen time series – 6 pts.
- For estimating VECM, interpreting the estimates along with commenting on model's diagnostics and calculating forecasts – 12 pts.
- For applying the Box-Jenkins procedure in aim of choosing best possible ARIMA models' specifications along with calculating forecasts for both series separately – 12 pts.

- For comparing VECM model's forecasts with ARIMAs alternatives – 5 pts.
- For the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file – 5 pts.

<u>Specific description of the assessment:</u>

To avoid any possible misunderstanding:

- By "checking the cointegration between the chosen time series" I mean exactly the procedure for testing the cointegration presented on lecture #7 and labs #8.
- By "estimating VECM, interpreting the estimates along with commenting on model's diagnostics and calculating forecasts" I mean:
    - performing the Johansen cointegration test and concluding about the number of cointegrating vectors,
    - estimating VECM,
    - interpreting all the parameters of the VECM model,
    - commenting if the Error Correction Mechanism work,
    - commenting about the signs of the estimated parameters,
    - reparametrizing VECM model into VAR model,
    - calculating and interpreting Impulse Response Function for the reparametrized model,
    - calculating and interpreting Forecast Error Variance Decomposition for the reparametrized model,
    - checking serial autocorrelation and normality of the residuals obtained from the reparametrized model,
    - calculating and plotting a forecast using VECM model,
    - calculating *ex-post* forecast errors for both series on the *out-of-sample* period.
- By "applying the Box-Jenkins procedure in aim of choosing best possible ARIMA models' specifications along with calculating forecasts for both series separately" I mean:
    - preparing ARIMA model with an application of the Box-Jenkins procedure i.e. going through the following steps:
        - identification – plotting and interpreting ACF, PACF for the series,
        - estimation – estimating the model and interpreting the results,

        - diagnostics – plotting and interpreting ACF, PACF for residuals as well as performing and interpreting the Ljung-Box test,
        - forecasting – calculating *ex-post* forecast errors on the *out-of-sample* period.
    - estimating many ARIMA models when applying the Box-Jenkins procedure and comparing them with Information Criteria.

Note: when it comes to the above, please do not use only auto.arima() function! This course is not about running some R functions without any deeper consideration. You need to go carefully through the Box-Jenkins procedure.

- By "comparing VECM model's forecasts with ARIMAs alternatives" I mean calculating forecast *ex-post* error measures considering the *out-of-sample* period and comparing them.
- I guess that "the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file" should be clear.

TOPIC #2

FORECASTING GOOGLE TRENDS

WITH VAR AND ARIMA MODELS

Project description:

The aim of the project is to forecast the number of queries directed to the Google search engine for two chosen phrases with a VAR model and two independent univariate ARIMA models.

In order to accomplish the task you should:

- Come up with an idea for a **possible, theoretically justified and interesting** relationship between the number of queries directed to the Google search engine for two chosen phrases.

For example: number of Google searches of "pregnancy" can cause the number of Google searches of "alimony"; number of Google searches of "weather forecast" can cause the number of Google searches "accommodation by the sea" etc.

Note: it is possible to involve relationships concerning countries of searching. For example: the number of Google searches of "war in Ukraine" can cause the number of Google searches of "helping refugees" in Poland but not in the UK.

- Install **gtrendsR** R package that enables loading Google Trends data directly to R.
- Load some series standing for your initial idea. Use **gtrends() function**.

Note: you can check out the documentation of the gtrendsR package if needed: https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf.

- Check integration order of two chosen time series and examine if there is a Granger causality between them.
- Find a proper specification of VAR model for the considered time series and produce forecasts for the *out-of-sample* period.
- Find separately for two countries the most attractive univariate ARIMA family models and produce forecasts for the same *out-of-sample* period.
- Compare accuracy of forecasts using the *ex-post* forecast error measures.
- Prepare a short report on it.

About the sample:

Data loaded with gtrends() function come from Google Trends, i.e. a Google service providing data on the number of queries directed to the Google search engine.

Function gtrends() enables choosing the country of Google searches origin as well as time horizon. Those parameters are up to you.

To be specific about the data, each series from Google Trends meets values from [0, 100]. It means that a specific transformation of the number of Google searches is considered. Still, the proportions shall be respected.

Detailed grading rules:

The maximum number of points you can get for the project is 40. Detailed grading rules are as follows:

- For checking the integration order of the chosen time series and examining if there is a Granger causality between them – 6 pts.
- For finding a proper VAR specification, estimating the model, commenting on model's diagnostics and calculating forecasts – 12 pts.
- For applying the Box-Jenkins procedure in aim of choosing best possible ARIMA family models' specifications along with calculating forecasts for both series separately – 12 pts.
- For comparing VAR model's forecasts with ARIMAs alternatives – 5 pts.
- For the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file – 5 pts.

Specific description of the assessment:

To avoid any possible misunderstanding:

- By "checking the integration order" I obviously mean testing stationarity. It will be used later when choosing ARIMA family model specification.
- By "examining if there is a Granger causality between them" I mean:
    - formally testing Granger causality with a proper statistical test not assessing it only with a visual analyses of plots,
    - commenting on the type of observed causality.

- By “finding a proper VAR specification, estimating the model, commenting on model’s diagnostics and calculating forecasts” I mean:
    - commenting on the importance of integration orders of the series one uses VAR models on,
    - commenting on the importance of the presence of Granger causality when applying VAR models,
    - justifying the initial choice of the number of lags and other possible model components with Information Criteria,
    - estimating initially chosen VAR model’s specification,
    - analyzing diagnostics of the estimated model:
        - considering original and fitted values plotted,
        - considering residuals via plotting them,
        - examining ACF, PACF for the residuals,
        - testing a multivariate autocorrelation of the residuals formally via Portmanteau test and Breusch-Godfrey test.
    - estimating other, differently specified VAR models (as many as it can be justified by the diagnostics obtained),
    - comparing different VAR models specifications with Information Criteria,
    - examining Impulse Response Functions for the final VAR specification and commenting on them,
    - examining Forecast Error Variance Decomposition for the final VAR specification and commenting on it,
    - calculating and plotting a forecast using the chosen VAR model,
    - calculating *ex-post* forecast errors for both series on the *out-of-sample* period.
- By “applying the Box-Jenkins procedure in aim of choosing best possible ARIMA family models’ specifications along with calculating forecasts for both series separately” I mean:
    - preparing SARIMA model with an application of the Box-Jenkins procedure i.e. going through the following steps:
        - identification – plotting and interpreting ACF, PACF for the series,

    - estimation – estimating the model and interpreting the results,
    - diagnostics – plotting and interpreting ACF, PACF for residuals as well as performing and interpreting the Ljung-Box test,
    - forecasting – calculating *ex-post* forecast errors on the *out-of-sample* period.
- estimating many SARIMA models when applying the Box-Jenkins procedure and comparing them with Information Criteria.

Note: when it comes to the above, please do not use only auto.arima() function! This course is not about running some R functions without any deeper consideration. You need to go carefully through the Box-Jenkins procedure.

- By "comparing VAR model's forecasts with ARIMAs alternatives" I mean calculating forecast *ex-post* error measures considering the *out-of-sample* period and comparing them.
- I guess that "the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file" should be clear.

TOPIC #3

ESTIMATING VALUE-AT-RISK OF A PORTFOLIO

WITH GARCH-FAMILY MODELS

Project description:

The aim of the project is to estimate *Value-at-Risk* (VaR) of a portfolio consisting of five financial instruments on the basis of two GARCH-family models.

In order to accomplish the task you should:

- Choose two models from the GARCH family e.g. GARCH and EGARCH.
- Build a portfolio which consists of five financial instruments:
  - list of financial instruments to include: SP&500, DAX, WIG20 and two other series of your choice,
  - weights of each instruments should be equal to 20% for each day of analysis,
  - quotations of equity indices are available on the website www.stooq.com and quotations of cryptocurrencies are available on the website www.coinmarketcap.com, alternatively both can be downloaded with *getSymbols()* function from the R package named *quantmod*.
- Conduct for this portfolio a comparison analysis of:
  - estimates of annualized conditional standard deviation in the in-sample period produced by the two models,
  - estimates of the *Value-at-Risk* produced by the two models in the out-of-sample period.
- Prepare a short report on it.

About the sample:

The *in-sample* period should start on 2018-01-01. The *out-of-sample* period should last for 365 days and should start on 2022-01-01.

Detailed grading rules:

- For computing the returns of the equally-weighted portfolio as well as observing and commenting about the presence of the stylized facts concerning volatility modeling – 5 pts.
- For estimating two chosen GARCH family models – 16 pts.
- For interpreting the diagnostics of estimated models – 8 pts.
- For calculating Value-at-Risk for both *in-sample* and *out-of-sample* periods separately for two chosen models and comparing the quality on the *out-of-sample* period – 6 pts.
- For the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file – 5 pts.

<u>Specific description of the assessment:</u>

- By “computing the returns of the equally-weighted portfolio” I mean:
    - dealing with possible NAs values and non-business days,
    - calculating relevant log-returns for all components of the portfolio separately and all together.
- By “observing and commenting about the presence of the stylized facts concerning volatility modeling” I mean:
    - commenting on a visibility of volatility clusters,
    - commenting on possible leptokurtosis of log-returns,
    - commenting on a possible presence of ARCH effects among log-returns, autocorrelation of squared returns,
- By “estimating two chosen GARCH family models” I mean:
    - estimating many simple GARCH models in aim of finding best parameters specification,
    - choosing best simple GARCH model specification after a deep consideration of at least:
        - Ljung-Box tests for residuals as well as squared residuals,
        - LM Arch test for residuals,
        - Conditional standard deviation,
        - ACF of standardized residuals,
        - ACF of squared standardized residuals.

    - estimating at least one modification of GARCH (e.g. GARCH-in-mean, EGARCH, …) assuming a previously established specification of GARCH.
- By “interpreting the diagnostics of estimated GARCH models” I mean:
    - commenting on parameters interpretation of the chosen specification of GARCH model,
    - commenting on parameters expected signs and values of the chosen specification of GARCH model,
    - interpreting at least:
        - Nyblom stability test,
        - Sign Bias test,
        - Adjusted Pearson Goodness-of-Fit test.
    - interpreting the News Impact Curves obtained for the models.
- By “calculating Value-at-Risk for both *in-sample* and *out-of-sample* periods separately for the chosen models and comparing the quality of estimations on *out-of-sample* period” I mean:
    - calculating Value-at-Risk considering the methodology presented specifically on labs #13 for both periods,
    - comparing at least the shares of situations when the actual values in the *out-of-sample* period were below the Values-at-Risk computed for the chosen models.
- I guess that “the structure, order, transparency and clarity of the provided report as well as for using a form of RMarkdown file” should be clear.