AAEC 4984/5484: Applied Economic Forecasting

Master Key

Homework #3 - Spring 2024

Instructions: In all cases, please ensure, where necessary, that your graphs and visuals have proper titles and axis labels. Refer to the output, whenever appropriate, when discussing the results. **Lastly, remember that creativity (coupled with relevance) will be rewarded.**

This week, the emphasis will be on pulling real-world data and conducting a more in-depth analysis of the data. The first question examines the Producer Price Index (PPI) series, while the second will explore the Weekly supply of Motor Gasoline in the US.

Question 1: US Producer Price Index

The Federal Reserve Bank of St. Louis provides a wealth of economic data. In this exercise, we will be using the `quantmod` package to import and analyze the `US Producer Price Index (PPI) by Commodity: All Commodities` series.

a. Using the `getSymbols()` function from the `quantmod` package, import the `PPIACO` series from the Federal Reserve Bank of St. Louis. **Store this as `ppi`.** For your convenience, I have called in the `quantmod` library in the `setup` chunk.

Hints: You will find it best to set the `auto.assign` argument to `FALSE` and that the data are returned as a `timeSeries` object. Next, declare as a `ts` object using the `as.ts()` function. Last, declare as a `tsibble` object. This will allow us to use the time series functionalities of our `fpp3` package.

b. You will notice that we just pulled the entire PPI series from January 1913 to January 2024 (at the time of my writing this assignment).

- Use the `filter_index()` function, subset `ppi` to include only the observations from 1972 to the December 2023.
- Next, compute the **logarithmic** growth rate of the PPI series. For clarity, I am asking you to compute the first order logged difference of PPI. Be sure to multiply by 100 to get the percentage change.
- **It is fine to store this back in as `ppi`.**

c. On separate graphs, plot the PPI and its growth rate. Provide descriptions of each plot. It would be helpful to connect the patterns you observed with economic and global events (think COVID, recessions, etc.). Be sure to label axes and give the graphs an appropriate title appropriately. Try using the `grid.arrange` function to help conserve space.

d. Plot the autocorrelation function (ACF) of the PPI growth rate up to four (4) years. What do you observe? Does there appear to be any significant autocorrelation at any lag?

e. Kiddarth, your study partner, is adamant that the PPI growth is a random walk process and entirely random. Therefore, there is no autocorrelation in the growth series. From your reading of the ACF, you are skeptical of this claim and have decided to conduct a formal test. You opted to use the Ljung-Box test to test for autocorrelation up to lag $\ell = 12$, $Q(12)$. What conclusions do you draw at the 5% level of significance?

- Remember to state the null hypothesis and the decision rule of the test.

Question 2: US finished motor gasoline products supplied

**The US Energy Information Administration (EIA) reports data on Weekly US finished motor gasoline products supplied. The most updated data are available here

Our focus will be on estimating and evaluating competing models (albeit the most basic). At the end, you will select an "optimal" model for modeling the future value of weekly supplies. This assignment should bridge the gap between the earlier modules and this current one.

Note: Before you proceed, please install the `rio` package. Do this in your consoles and not within the `.Rmd` file.

a. Using my code below, import the supply data from the EIA into `R`. If you were to download the `.xls` file manually, you would notice that the data is in the second sheet and that the first two rows are not part of the column headings you need.

We will specify this using the `sheet = 2` and `skip = 2` arguments, respectively.

Please feel free to manually download the Excel file from the EIA link above and explore this on your own.

```
# Remember to remove the eval = FALSE argument
gas <- rio::import(
  "https://www.eia.gov/dnav/pet/hist_xls/WGFUPUS2w.xls",
  sheet = -----, skip = -----) #modify here
```

b. Next, let us clean up the data a bit. Your tasks are as follows:

- Using the `mutate()` function, declare the `Date` column as a `yearweek()` object. It is okay to override `Date`.
- Rename the second column to `Supply`.
- Convert `Supply` to `Million Barrels` per day. Keep the column name as `Supply`.
- Declare the dataset as a `tsibble` object using the `as_tsibble()` function.
- Drop all observations after December 31, 2023.

Store the results of this step into a variable called `gas.ts`.

c. Present an `autoplot` and `ACF` of weekly gasoline supply. Briefly comment on your plots. You would find it best to set the maximum lags in the ACF to be at least three (3)years, that is 52×3. Use the `grid.arrange()` function to present your graphs as a 2x1 grid.

Be sure to label your axesproperly.

d. Let us proceed to forecast the `gas.ts` series. We will split the dataset into a training (`train.gas`) and a testing set period (`test.gas`).

Using the `filter_index()` function, assign the observations up to (and including) `"2019 W52"` to `train.gas`. All observations afterward should be assigned to `test.gas`.

e. Confirm that the data is correctly split by using the `autoplot()` and `autolayer()` functions. Be sure to include `gas.ts`, `train.gas` and `test.gas` in this plot.

A title is not necessary. However, please add colors to your lines to highlight each series.

f. Use the `model()` and `forecast()` functions to produce forecasts from the four(4) benchmark models from class. **Ensure that your models are named appropriately.**

- You will find it best to store both the model fits and forecasts for later.
- **Remember that you will need to set the horizon equal to the number of rows in your test set.**

g. Produce an `autoplot` of the forecasts against the *test* dataset. Be sure to turn off the `PI` at this stage. Does any particular forecast method appear to do a better job than others? Does any model appear to consistently over- or under-predict? What, if anything, do you think can explain periods when the forecasts deviate significantly from the actual values?

h. Using the `accuracy()` command, extract the `RMSE`, `MAE`, `ME`, and `MAPE` statistics. Also, I am intrigued to know the `MSE` value. Use the `mutate()` function to create that column.

- Display your results as a table using the `knitr::kable( )` function with `digits = 2`. Add the argument `format.arg = list(big.mark = ",")` to format your table using thousands separator.
- Lastly, include **an appropriate one** by adding and modifying the argument `caption = "Your caption Here"`.
- **Ignoring the `ME`, which model is preferred under the remaining four (4) selection criteria? Be sure to explain how you came to that conclusion.**
- Comment on the bias of each model.

i. As we discussed in class, the "preferred" model might not necessarily satisfy some or all of the necessary assumptions. To illustrate this, use the `gg_tsresiduals()` function, with `type = "innov"`, to comment on the in-sample fit of the model selected by the `MAPE`. **Again, set the maximum lags to 3 years so that you can see more of the dynamics of the `ACF`.**

When commenting, consider the following:

- Do the residuals appear to be normally distributed?
- Do you observe any **potential** patterns (seasonality, trend, cycle) in the `autoplot` and `ACF` of the residuals? From the patterns observed, do you think the residuals are white noise? Why, or why not?
  - Are you surprised by what you found here? Why, or why not?

j. Using the same model as above, employ the `Ljung-Box` test to conduct a hypothesis test of autocorrelation up to lag $\ell = 2 \times m$. **Where m is the frequency of the data.**

- You are required to explicitly state the null hypothesis of the test and how you came to your conclusion. **Think back to the decision rule of the test.**

k. This is a free-form answer. I do not expect you to write any code here. Instead, an intelligent and thoughtful response is required.

- **What are some of the potential reasons why the "optimal" model might have been preferred over the other models? Could our choice of the training vs test period have played a part here?**
- **How would you have gone about improving the model?**

End of Homework 3