EMET3007/8012 Computer Lab 4

Ruitian Lang

Due date: 5 Oct

In this project, we will consider the retail trade data in Australia up to June 2019. The data file can be downloaded from the course Wattle website or from the Data Explorer of the Australian Bureau of Statistics. We only look at the economy total and not the data for each industry group. We use the seasonally adjusted data so you do not need to worry about seasonality.[1]

We will forecast the national total retail trade for the next month, using data until June 2014 as the training set. Therefore, the width of the forecasting window is 1. The benchmark forecast used in this project is the random walk forecast, which uses the observation from the previous month as the forecast.

The data appears to have a deterministic trend:

$$\log(y_t) = f(t) + u_t,$$

where $f(t)$ is polynomial of t and u_t is an ARIMA process.[2] It turns out that the coefficients in the polynomial $f(t)$ can be consistently estimated by OLS even if u_t possesses a unit root. In fact, the ARIMA class in StatsModels allows you to specify the general form of a deterministic trend; see the documentation of ARIMA for details.

We will attempt to forecast future retail trade by an ARIMA model with a deterministic trend. Apart from the order of the ARIMA model, you also need to determine the functional form of the deterministic trend. You should measure the performance of your model by mean absolute percentage error (MAPE): if the data is y_t and your forecast is $\hat{y}_t$, and you perform your forecast between Periods T and $T + h$, then MAPE is

$$\frac{100}{h} \sum_{t=T+1}^{T+h} |(\hat{y}_t / y_t) - 1|,$$

or

$$\frac{100}{h} \sum_{t=T+1}^{T+h} |\log(\hat{y}_t) - \log(y_t)|.$$

The two should be very close to each other as you know from Lab 2.

Besides the performance of your model, you will also be marked on your general design, your model selection approach and your analysis of the results. Please note the following rules.

[1] You may try STL if you want, but you will find no seasonality.

[2] Whether to take the logarithm of the raw numbers is up to you. Even if you do not take the logarithm, the data will still have a deterministic trend.

1. You must consider at least two forecasting models no including the benchmark forecast. Two ARIMA models with different orders count as two different models.

2. If we find that you use the test set (data after June 2014) in your model building and model selection process, you will receive at most 30 marks from this lab. Using the test set in model selection is one of the most serious errors in forecasting and considered cheating in this project.

3. If your forecasting model fails to beat the benchmark forecast, you will receive at most 60 marks from this project.

The hurdle task is to explain to your tutor the set of models you consider and demonstrate how you select the best one from the set.

Good luck!