NYC DATA SCIENCE ACADEMY

NYC Data Science Bootcamp Fall 2015

Time Series Analysis

Question #1: ARIMA Modelling

For each of the listed datasets, follow the procedure below to fit an ARIMA model.

- The `hare` dataset in the `TSA` library (load by calling `data(hare)`); this dataset describes the annual number of Canadian hare near the Hudson bay area.
- The `WWWusage` dataset; this dataset describes internet usage by logging the number of users connected to a server every minute.
- The `nhtemp` dataset; this dataset describes the average yearly temperatures (℉) in New Haven, Connecticut.
- The `brobs.csv` dataset; this dataset describes the monthly armed robberies in Boston.
    - **Hint #1:** Use the following commands to read the dataset correctly and create a time series object:

```
brobs = read.csv("07brobs.csv")
brobs = ts(brobs[-119, 2], start = 1966, frequency = 12)
```

    - **Hint #2:** Comment on why you might want to log transform this data before proceeding with the ARIMA procedure.

The ARIMA procedure:

1. Plot the original time series.
2. Determine the estimated number of differences necessary to make the time series stationary. If non-stationary, plot the differenced time series.
3. Run the Augmented Dicky-Fuller test and interpret the results.
    a. If necessary, return to part 2.
4. Plot the autocorrelation and partial autocorrelation functions to determine initial values of p & q.

5. Fit an initial ARIMA model.
6. Follow the process of overfitting the initial ARIMA model you fit in part 5. At each step, examine the Wald test statistics for the coefficient estimates. (**NB:** If you seem to have already overfit your model, try taking away terms by reversing the overfitting process).
7. Evaluate the family of models you created in part 6 and select the best (examine the AIC, BIC, and RMSE).
8. Loosely interpret the coefficient estimates of the model you selected in part 7.
9. Perform diagnostic tests for the best model by assessing:
    a. The constant variance assumption (graphically).
    b. The normality assumption (graphically).
    c. The independence assumption (graphically & statistically).
10. Forecast the next few observations and display this graphically.
11. Answer the research questions:
    a. How will the Canadian hare population change over time?
    b. How will server traffic change over the next few minutes?
    c. How will the weather change in New Haven Connecticut over the next few years?
    d. How will the Boston crime scene change over the next few years?