INDR 422/522 - Spring 2023 Fikri Karaesmen

**HOMEWORK 2, Due Date: April 14, 2023**

- Please work in groups of two or (individually) and submit one file for each group with all names.
- The first data set is the same as in HW 1: The data is on total monthly average price index of Kenyan tea in commodity markets as reported by the IMF commodity price data portal (data found in the blackboard page). The data starts from January 2013 and ends in December 2022 (included).
- The second data set is the number of daily passengers taking the Istanbul metro starting at Haciosman station (starting on January 1, 2022, 315 consecutive days).
- For the sake of exercise, we sometimes fit models to the entire data set. But a better practice is to separate the data into a training set and a test set. We should use the same training set to train all models (ARIMA, regression etc.) and compare error performance of trained models on the same test set.
- Please perform all computations in python. You are expected to implement your own forecasts (don't use packages/functions from the web). Please submit (upload on blackboard) your commented (with explanations) python notebooks.
- In addition to the python notebook, **submit a short typed summary report** that includes the results (error tables, prediction intervals etc.) of all exercises. Also add a general assessment of the methods (which method is the best, which should be avoided etc.). **The report is part of the overall grade.**

## Exercises

1. (30 points) Forecasting Kenyan Tea Prices (monthly average price index of Kenyan Tea Price in commodity markets as reported by the IMF commodity price data portal). This continues from the first HW.

(a) Plot the data and visually assess whether there is significant trend and seasonality.

(b) Check the ACF and PACF plots after detrending and deasonalizing (if necessary). Check whether there is significant AC/PAC visible after the transformations.

(c) Fit an ARIMA model to the whole data set based on the autocorrelation structure you observe and the patterns (trend, seasonality etc.). Explore the significance of the fitted coefficients, the residual diagnostics and assess its performance by calculating the MAE, MAPE and RMSE.

(d) Experiment with a different ARIMA model for comparison (you can use one or two AR or MA terms after transforming the data). Explore the significance of the fitted coefficients, the residual diagnostics, AIC values and assess its performance by calculating the MAE, MAPE and RMSE. Compare with the previous model.

(e) Separate the data into a training set and a test set (first 70 to 80 % months of the data should be the training set and the remaining data test set). Fit the two models above on the training set and compute the forecast errors for the estimated model (for one month-ahead forecasts) on the test set.

(f) Your report for the exercise should include a table similar to the one below.

2. (40 points) We had looked at the daily metro passengers data from Haciosman Station. We'll investigate whether ARIMA models might be appropriate to make predictions for this data.

(a) Plot the generated observations.

(b) As a benchmark, fit a SARIMA(0,1,0)(0,1,0,7) model to the whole data (i.e. seasonal differencing of 7 days and first order differencing). Explore the significance of the fitted coefficients, the residual diagnostics and the error performance (MAE, MAPE and RMSE). This is a benchmark model for other comparisons.

(c) Detrend and deseasonalize the data. Plot the Auto-Correlation Function and the Partial Auto-Correlation Function of the detrended and deseasonalized data.

(d) Based on the ACF and PACF, fit a SARIMA model to the whole data that has some AR and MA coefficients. Explore the significance of the fitted coefficients, the residual diagnostics and

assess its performance by checking the AIC and calculating the MAE, MAPE and RMSE. Compare with the previous benchmark model.

(e) For the best model you found, find a one-day ahead forecast for days 309 to 315 and report the 95 % prediction intervals for your forecasts.

(f) Provide multi-day look ahead forecasts for days 316 to 329.

(g) Separate the data into a training set and a test set (first 70 to 80 % of the data should be the training set and the remaining data the test set). Fit the two best alternative models above on the training set and compute the one-day ahead forecast errors (MAE, MAPE, RMSE) on the test set.

(h) Your report for the exercise should include a table similar to the one below.

Table 1: Summary of Results for Exercises 1 and 2

| **Method** | **Spec.** | **RMSE (Train)** | **RMSE (Test)** |
|---|---|---|---|
| Benchmark (from HW1) | - | | |
| Model 1 | $\phi_1 =$, $\theta_1 =$, etc. | | |
| Model 2 | | | |
| Model 3 (if any) | | | |

The model specification includes the ARIMA specification (i.e. ARIMA (1,0,2) or SARIMA(1,0,2)(0,1,0,7)) and the fitted coefficients.

3. (30 points) Fit ordinary linear regression models to the Haciosman metro passengers data.

(a) Fit a linear regression model that uses seasonal (daily) dummies as predictors (note that there are seven days and six corresponding seasonal dummies are needed). Check the $R^2$ and the significance of the coefficients and compute the MAE, MAPE and RMSE.

(b) Fit a linear regression model that uses seasonal dummies and a linear trend term as predictors. Check the $R^2$, adjusted $R^2$, AIC (to compare with part a) and the significance of the coefficients and compute the MAE, MAPE and RMSE.

(c) Separate the data into a training set and a test set (first 70 to 80 % of the data should be the training set and the remaining data the test set, please use the same training/test sets as in Q2). Fit the above model on the training set and compute the forecast errors on the test set.

(d) Compute the correlation of your predictions and the observations on the training set (you can use the np.corr() function).

(e) Compute the correlation of your best ARIMA forecasts and regression forecasts on the test set. If the correlation is low (or negative), you can combine both forecasts (by a weighted average) and reduce variance.

(f) Your report for the exercise should include a table similar to the one below.

Table 2: Summary of Results for Exercise 3

| **Method** | **Spec.** | $R^2$ | **RMSE** |
|---|---|---|---|
| Benchmark (from HW1) | - | | |
| Model 1 | $\beta_0 =$, $\beta_1 =$, etc. | | |
| Model 2 | | | |
| Model 3 (if any) | | | |