UNIVERSITY OF BRITISH COLUMBIA
Department of Statistics
Stat 443: Time Series and Forecasting
Assignment 1: Exploratory Data Analysis

The assignment is due on **Thursday, February 2** at **9:00pm**.

- Submit your assignment online on `canvas.ubc.ca` in the **pdf** or **html** format under module "Assignments".
- This assignment should be completed in **RStudio** and written up using **R Markdown**. Display all the R code used to perform your data analysis.
- Please make sure your submission is clear and neat. It is the student's responsibility that the submitted file is in good order (i.e., not corrupted).
- Remember to properly label all your plots and have them clearly displayed.
- **Late submission penalty**: 1% per hour or fraction of an hour. (In the event of technical issues with submission, you can email your assignment to the instructor to get a time stamp but submit on canvas as soon as it becomes possible to make it available for grading.)

1. Download the time series of monthly means of globally averaged CO_2 records ('globally averaged marine surface monthly mean data') from the webpage of Global Monitoring Laboratory (https://gml.noaa.gov/ccgg/trends/gl_data.html), from January 1979 to October 2022.

   (a) Read in the data and create a time-series object for the mean monthly CO_2 concentrations. Plot the series and comment on any features of the data that you observe. In particular, address the following points:
   - Does the series have a trend?
   - Is there seasonal variation, and if so would an additive or multiplicative model be suitable? Justify your answers.
   - Is the series stationary? Explain.

   (b) Create training and test datasets. The training dataset should include all observations up to and including December 2019; this dataset will be used to fit ("train") the model. The test dataset should include all observations from January 2020 to October 2022; this dataset will be used to compare forecast accuracy between two estimation methods. You can use the command `window()` on a `ts` object to split the data.

   Using a suitable decomposition model, decompose the training series into trend, seasonal, and error components. Use both the moving average smoothing (`R` function `decompose()`) and the loess method (`R` function `stl()`). Plot the two decompositions.

   (c) Fit a linear model to the trend component under each of the two decomposition methods (you can use `R` function `lm()`).
   - Write down the fitted model for the trend component under each method.

- Does the linear model provide evidence of a trend at the 95% confidence level? Without doing any further analysis, would you use this trend component to make predictions? Justify your answer using the linear model results and the trend component plot.

(d) Predict the mean monthly CO_2 records for the period from January 2020 to October 2022 using both the moving average and loess smoothing decompositions.

Compare your predictions with the test dataset graphically and using the mean squared prediction error (MSPE), defined as the average of the squared distances between the predictions and test data. Which method do you recommend and why?

2. The file `CanadaGDP.csv` contains monthly adjusted GDP values for Canada, presented on a 2012 reference year basis, from January 1997 until December 2019.

(a) Create a time series object for this data and plot it over time. Explain why the time series of the monthly adjusted GDP values is likely non-stationary, according to the plot.

(b) Usually the original series is transformed by taking the logarithmic difference. There are two approaches. Let S_t denote the monthly adjusted GDP at time t. Then the following log-growth rates are defined:

- sequential log-growth rate: $X_t = \ln(S_t/S_{t-1})$;
- year-over-year log-growth rate: $X_t = \ln(S_t/S_{t-12})$.

Plot the log-growth rate series for the two definitions separately and comment on the stationarity of each. Which definition would be preferable? Give a short explanation.

(c) Create the correlogram of the log-growth rates series based on the two approaches above. Comment on what you observe.

3. In this question you will explore the sampling distribution of the sample autocorrelation coefficient for a white noise process through a simulation study. Recall that, for a time series of length n, from a white noise process, the sample autocorrelation coefficient at lag h approximately follows a normal distribution with mean $-1/n$ and variance $1/n$:

$$r_h \sim \mathcal{N}(-1/n, 1/n)$$

for large values of n.

To confirm this theoretical fact, conduct the following simulation study:

(i) Simulate a time series of length $n = 1000$ from a white noise process $\{Z_t\}_{t\in\mathbb{Z}}$ with $Z_t \sim \mathcal{N}(0,1)$ (function `rnorm()`).

(ii) Evaluate r_1, the sample autocorrelation coefficient at lag $h = 1$. Store this value.

(iii) Repeat steps (i) and (ii) $m = 5000$ times; i.e., generate 5000 time series and for each of them compute r_1 (you can use `for` loop).

Plot the histogram for the sample of r_1 values from the simulation study (function `hist()`), add the smoothed version of the histogram (function `density()`) and the theoretical asymptotic normal density (function `dnorm()`). Comment whether there is an agreement between the empirical sampling density and its theoretical approximation.

Make sure your plot is well-presented, including a suitable title, axes labels, curves of different type or colour, and a legend.