# UNIVERSITY OF BRITISH COLUMBIA
## Department of Statistics
### Stat 443: Time Series and Forecasting
### Assignment 3

The assignment is due on **Thursday, March 23** at **9:00pm**.

- Submit your assignment online on `canvas.ubc.ca` in the **pdf format** under module "Assignments".
- This assignment should be completed in **RStudio** and written up using **R Markdown**. Display all the R codes used to perform your data analysis.
- Please make sure your submission is clear and neat. The student is responsible for the submitted file being in good order (i.e., not corrupted).
- **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 it on canvas as soon as it becomes possible to make it available for grading.)

1. Let $\{X_1, \cdots, X_n\}$ denote a time series of length $n$ and set $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ as the sample mean. The aim of this question is to explore, through a simulation study, statistical properties of $\bar{X}$ under different temporal dependence scenarios. The results of this study can be used to understand how temporal dependence affects properties of the sample mean as an estimator of the mean of the underlying stochastic process, $\mu$.

   Consider **three** AR(1) processes of the form:

   $$X_t - \mu = \alpha(X_{t-1} - \mu) + Z_t,$$

   where $\mu = 2$, $\alpha \in \{-0.8, 0, 0.8\}$, $\{Z_t\}_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$ with $\sigma^2 = 0.25$ when $\alpha = \pm 0.8$ and $\sigma^2 = 0.25/(1 - 0.8^2)$ when $\alpha = 0$. Note that the choice of $\sigma^2$ ensures that all three processes have the same variance, $\sigma_X^2$.

   For each of the three AR(1) processes, generate $m = 5000$ time series of length $n = 500$ and for each of these series compute its sample mean.

   (a) Provide a summary table giving the values of the empirical mean and standard deviation of $\bar{x}$ values across 5000 replications for the three AR(1) processes considered. Make sure that your table includes a caption.

   (b) On the same plot, display three sampling densities of estimator $\bar{X}$ corresponding to the three values of parameter $\alpha$ based on your simulation output. Add the true value of $\mu$ to the plot. Use this plot and the table in part (a) to compare the stochastic behaviour of $\bar{X}$ (in terms of its bias and variance) as an estimator of the process mean under different temporal dependence scenarios and provide intuition for your observations.

2. In this question, you will revisit the time series of monthly adjusted GDP values for Canada, presented on a 2012 reference year basis, from January 1997 until December 2019. See data file `CanadaGDP.csv`.

   We focus on the year-over-year log-growth rate: $X_t = \ln(S_t/S_{t-12})$, where $S_t$ denotes the monthly adjusted GDP at time $t$.

   (a) Import the data and create a time-series object for the monthly adjusted GDP values. Transform the original series to the year-over-year log-growth rates, and split the series into training and test sets:

   - the training data starts Jan 2010 and ends Dec 2017;
   - the test data starts Jan 2018 and ends Dec 2019.

   Note that we only consider data after the global financial crisis of 2007-2009 to avoid non-stationarity concerns.

   Plot the training data along with its sample acf and pacf. Ensure enough lags can be observed in the acf and pacf. Comment on what you observe and suggest a suitable class of models.

   (b) Using AIC as the model selection criterion and based on your analysis in part (a), select the best fitting model. (You may experience numerical issues when fitting a model for some values of order parameter(s). You can exclude those values from consideration.)

   (c) Perform model diagnostics on the model fitted in part (b) and comment on the goodness of fit.

   (d) Use the fitted model to predict the year-over-year log-growth GDP rates for the period from Jan 2018 until Dec 2019 (the test period). On the same plot, display the test data and your point forecasts along with 95% prediction intervals.

   (e) Now make predictions for the test period using exponential smoothing. Add these point forecasts along with their 95% prediction intervals to the plot in part (d) to allow for comparisons. Make sure you use different line types and/or colours to distinguish lines and add a legend.

   (f) Compare the two forecasting methods (Box-Jenkins versus exponential smoothing) using a variety of criteria including but not limited to the mean squared prediction error (MSPE) and width of prediction intervals. Which method would you recommend and why? Discuss pros and cons of the two approaches.

3. In this question you will predict monthly means of globally averaged $CO_2$ records, which you also explored in Assignment 1, using the Box-Jenkins method. You can download the dataset by following the instructions in Question 1 of Assignment 1.

   (a) Read in the data and create a time-series object for the mean monthly $CO_2$ concentrations. Create training and test datasets. The training dataset should include all observations up to and including December 2019 while the test dataset should include all observations from January 2020 to October 2022.

   (b) **Model formulation:**
   The Box-Jenkins forecasting method is based on a fitted SARIMA$(p, d, q) \times (P, D, Q)_s$ time series model. In this part, you are going to choose values of order parameters (i.e., $p, d, q, P, D, Q$, and $s$).

i. Perform seasonal differencing to remove the seasonal effect. Plot the differenced time series and its correlogram. Comment on what you observe.

ii. If necessary (justify!), difference the series again at lag 1. Plot the new time series and its correlogram, and comment on its features.

iii. Based on the results of steps i. and ii., provide the values of $d$, $D$, and $s$ that you would choose.

iv. Using the sample acf and pacf plots of the suitably differenced series, suggest values of $p, P, q$ and $Q$.

v. Using AIC, select the best model by fixing the values of $p$ and $P$ according to your choice in iv. and selecting $q$ and $Q$ by considering an array of values of $q$ and $Q$ with $q = 0, 1, \ldots, 5$ and $Q = 0, 1, \ldots, 5$.

**Bonus:** You will obtain bonus points if your model has AIC below -682.

(c) **Model fitting:** Fit the model you selected in (b) and print estimates of the model parameters.

(d) **Model diagnostics:** Perform model diagnostics and comment on the goodness of fit of your chosen model.

(e) **Forecasting:** Predict the mean monthly $CO_2$ records for the period from January 2020 to October 2022 based on your model. Plot the test set data along with your point forecasts and corresponding 95% prediction intervals. Comment on how well your forecasting procedure works.