STA457/STA2202 - Assignment 2

Submission instructions:

Submit *three separate files* to A2 on Quercus - the deadline is 11:59PM on Tuesday, June 2.
- A PDF file with your Theory part answers.
- A PDF file with your Practice part report.
- A CSV file with your Practice part forecasts.

Theory

1. Consider two discrete random variables X, Y with joint probabilities given by the contingency table:

$P(X,Y)$	$Y=-1$	$Y=0$	$Y=-1$
$X=-1$	.05	.10	.15
$X=0$	.15	.15	.10
$X=+1$	.15	.00	.15

(a) [2 marks] Find the *Minimum Mean Square Error* (MMSE) predictor of Y given X, i.e. the conditional expectation $g(X) = \mathbb{E}[Y|X]$, and the MSE it achieves, i.e. $\mathbb{E}[(Y - g(X))^2]$.

(b) [2 marks] Find the *Best Linear Predictor* (BLP) of Y given X, i.e. $Y = a + bX$, for the BLP coefficients a, b, and the MSE it achieves.
(Note: This is an example where the MMSE predictor and the BLP are different.)

2. Consider the AR(1) model $X_t = \phi X_{t-1} + W_t, \;\; W_t \sim \text{WN}(0, \sigma_w^2)$.

(a) [3 marks] Find the covariance between the 1- & 2-step-ahead BLP errors, i.e. find

$$\text{Cov}\left[(X_{n+1} - X_{n+1}^n)(X_{n+2} - X_{n+2}^n)\right]$$

as a function of (ϕ, σ_w^2).
(Note: this should be *non-zero*; generally the different-step-ahead forecasts will be correlated.)

(b) [3 marks] Find the covariance between the subsequent 1-step-ahead BLP errors, i.e. find $\text{Cov}\left[(X_n - X_n^{n-1})(X_{n+1} - X_{n+1}^n)\right]$ as a function of (ϕ, σ_w^2).
(Note: These are similar to the model residuals *given perfect knowledge of the parameters*.)

3. [5 marks; **STA2202 (grad) students ONLY**] SS 3.26
(Note: the *estimated* BLPs $\hat{X}_{n+m}^n$ based on the fitted parameters $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}_w^2)$ are less accurate than the *theoretical* BLPs based on the true parameters. This questions shows that for AR(1) 1-step-ahead predictions, their difference is bounded in probability at the usual rate of $1/\sqrt{n}$.)

Practice

Description

It is your first day on the job and your boss, who graduated from the UofT Statistics program in 2013 has given you the task of forecasting a time series. Your forecasts will serve as an input to the firm's budget, so it is critical that they are accurate. Your boss would like you to provide them with forecasts, as well as a description of how you came up with them.

Assignment Structure

You will be given one time series and must produce a forecast for the next twelve observations. You can find your time series in the *Student Data* subfolder of the RStudio Cloud project; the name of your data file is your student number, and the name of the series you are forecasting is on the top row of the file. Your submission will include two files:

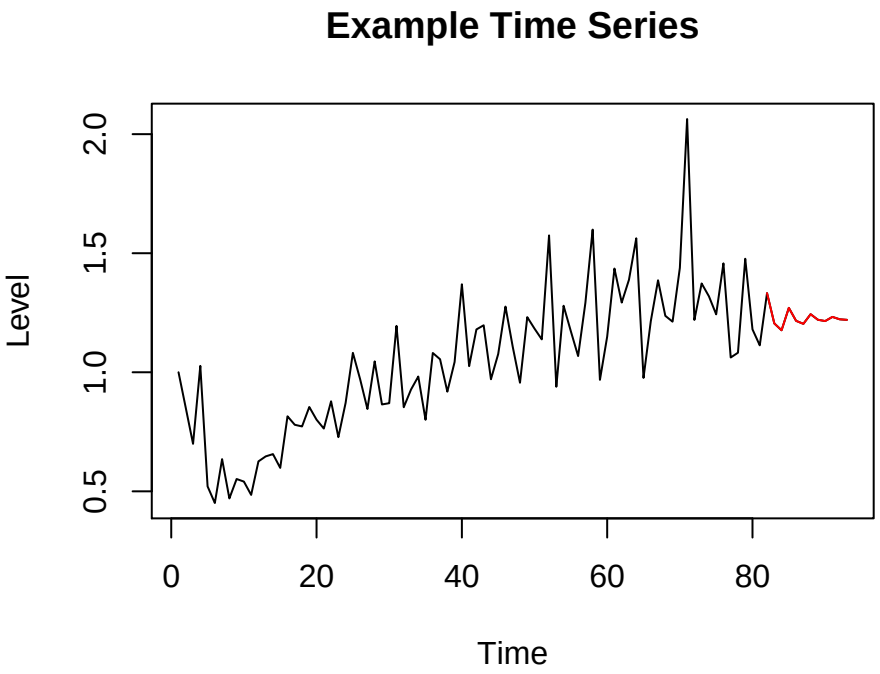
1. A 500-word written report in PDF format, with all your code in the Appendix.
2. A CSV file named `XXXXXXXXXX.csv`, where `XXXXXXXXXX` is your student number. This CSV should include your forecasts for the next twelve values of your series; the first entry should be your one-step-ahead forecast, and the twelfth entry should be your 12-step-ahead forecast (see also the sample file `123456789.csv` in the project's *Examples* subfolder.)

Written Report

Your written report should be able to be understood by your boss, someone who remembers the main ideas from a time series course several years ago, but not the finer details. Be sure to clearly explain what you have done, and if you are using any advanced concepts a sentence or two to refresh your boss on what they are is a good idea. Your written report must include the following:

1. A discussion of the characteristics of the time series (e.g. trend, seasonality, stationarity)
2. An explanation of any data preprocessing you had to do.
3. The model which you used.
4. A graph of the time series, with your forecasts in a different colour (see graph below for an example)
5. A discussion of your model's fit (diagnostics) and limitations.

The list above is what your written report must contain, but not an exhaustive list of all that it can contain. If there are any other topics that are worth discussing related to how you forecasted the data, please include them.



Tips

- This is a report to your boss. Concise & clear is better. They do not want to see single spaced size 6 font with expanded margins. They want to see all important and relevant information neatly organized.
- If you are going to include a code snippet in your written report (this is not required), make sure it is important enough to warrant your boss' attention.
- Make sure the model you choose, and how you fit it, makes sense. The data you are working with may violate some basic time series assumptions.

Assessment (15pts total)

- 1pt Your written report has a clean layout, and includes the requested graph.
- 1pt The text of your report is easy to follow, and conveys ideas effectively.
- 1pt Your CSV file with your predictions is properly formatted.
- 2pt Time Series Characteristics
  - 1/2 Some mention of the important time series characteristics.
  - 2/2 A clear identification of all important time series characteristics.
- 2pt Data preprocessing
  - 1/2 Some vague explanation of how the data has been preprocessed is provided.
  - 2/2 A clear explanation of how the data was preprocessed and the justification for why it was done.
- 2pt Model Explanation
  - 1/2 You have included a model description, but little in the way of explanation.
  - 2/2 You have concisely and clearly explained your model.
- 2pt Model Fit and Limitations
  - 1/2 Give vague description of the model's fit and limitations.
  - 2/2 Give clear and accurate description of the model's fit and limitations.
- 4pt Forecast Accuracy
  - 1pt Your method beats the naive forecast (the entire forecast is equal to the last datapoint)
  - 1pt Your forecast beats the forecast produced by the R code `ts_arima_model = auto.arima(x); forecast(ts_arima_model, h = 12)`
  - 1pt Your forecast beats the forecast produced by the R code `ts_ets_model = ets(x); forecast(ts_ets_model, h = 12)`
  - 1pt Your method beats all of the naive, `auto.arima()`, and `ets()` methods.

The way your forecasts will be judged is via Mean Absolute Percentage Error (MAPE) on the actual subsequent 12 values (not given to you, but known to us). Defining A_t as the actual value at time t, and F_t as your corresponding forecasted values in the submitted csv file, the MAPE for your forecasts will be calculated as

$$MAPE = \frac{1}{12}\sum_{t=1}^{12}\left|\frac{A_t - F_t}{A_t}\right|$$

Your forecast *beats* another forecast if your MAPE is lower.