STAT 4181 - Project 2 - Due May 4, Wednesday

Part 1 - R Coding (20 points)

In this part, you are asked to update your R code from Project 1 so that the final code will take a time series named `x` and will:

- *Print the ACF and PACF plots of the time series named* `x`.
- *Fit a series of AR models with* $p = 0, 1, 2 \ldots, 10$ *and store the AIC of each fit.*
- *Fit a series of MA models with* $q = 1, 2 \ldots, 10$ *and store the AIC of each fit.*
- *Fit a series of ARMA models with* $1 \leq p \leq 10, 1 \leq q \leq 10$*, and store the AIC of each fit. Note that this requires fitting 100 (10 × 10) different models and might take a little more time than the other steps.*
- **Fit a series of ARIMA models with** $0 \leq p \leq 10, 0 \leq q \leq 10, 1 \leq d \leq 3$**, and store the AIC of each fit. Note that this requires fitting 363 (11 × 11 × 3) different models.**
- **Find the best among all possible 484 models by using AIC, and then finally fit the best model and store it as `best.model`.**
- *Print the ACF and the PACF plots of the residuals of* `best.model`.

For this part, only submit your R code. Make sure it is commented properly; the code should include instructions on what it is doing at each stage. As in:

```
#acf and pacf plots
acf(x)
pacf(x)
```

Part 2 - Applications (10 points)

In this part, you will be using your R code from Part 1 to analyze the same tv show you have analyzed in Project 1, Part 3.

- Using the `tv.R` function from Project 1, obtain the IMDB scores for the tv series you have chosen to analyze in Project 1, Part 3.
- Use your code from Part 1 and find *"the best model"* for this tv series. If at least one of the best 3 fits is an ARIMA model (with $d \neq 0$), write a short report that analyzes the output of your R code. This should include acf, pacf plots and the parameters of the all three best models.

Part 3 - GLS with ARIMA (70 points)

For this problem, you can use your own dataset instead of using the given one. If you would like to use your own dataset please carefully read the "How to choose a dataset" section, and then check with me.

How to choose a dataset

- The dataset should contain at least 4 related time series (e.g. unemployment, year on year GDP change, inflation and average salary would be a nice quadruple). Obviously, the observation points (and the sample sizes) of these time series should coincide.
- One of the time series will be treated as *the dependent variable* (response, or y, in statistical terminology) and you will be building a series of models to understand how other time series (call them X) relate to the response.

Here are a couple of resources with a lot of useful economic datasets:
- *Federal Reserve Board: Data Download*
- *The World Bank Data Bank*

Please note that any of those datasets may require some data cleaning before they are exported to R.

What to do if you don't want to use your own dataset

- Install and load the `vars` package
- Load the dataset named `Canada`.
- Save the columns of Canada as separate `ts()` variables. Name them `C.e` (employment), `C.prod` (labour productivity), `C.rw` (real wage) and `C.U` (unemployment rate).
- `C.e` will be your target variable (i.e. the response, y), and others will be your independent variables (i.e. the covariates, X).

Analysis

1. For all of your time series, difference them with the `diff()` function and store the differences in a separate variable.
2. Plot the `acf` and the `pacf` of all of your time series (including the differenced series).
3. Plot the `ccf` between y and all other covariates. Further, plot the `ccf` between y and the differenced versions of the other time series. If you had 4 time series in the beginning (3 covariates + your response), you are asked to plot all $6 = 3 \times 2$ possible combinations for the `ccf`. Check if there are any significant lags in the `ccf` plots (these will look like mountains (or valleys) in the ccf curve and the tip of the

mountain (or the valley) will be significantly above (or below) the dotted blue lines). Make note of any significant lags and record them for future reference.

4. Using the results from the previous stage, come up with a reasonable linear model to predict y from X and `diff(X)`. For example: if you identified a significant cross-correlation between `C.u` and `C.e` at lag=4 and at lag =0 and found no no other correlations, your model will be `C.u ~ lag(C.e,4) + lag(C.e,0)`.
5. Use the `lag` function to create lagged versions of your time series.
6. Fit the `lm` model from *(4.)* using the lagged variables you created in *(5.)*.
7. Plot the `acf` and the `pacf` of the residuals.
8. Using your code from Part 1, find the 'best' ARIMA model for the residuals. If the best model has an integrated component (i.e. $d \geq 1$), then difference your response and all of your variables one more time and repeat stages 6 and 7 until the best model for the residuals is an ARMA model.
9. ~~Fit a `gls` model using your ARMA model from stage 8, and comment on the estimated coefficients and the p-values~~.
10. ~~Repeat stages 4-9 until you obtain a reasonable model in which all coefficients are significant ($p <$ your favorite critical level)~~.

Write up a short report with all of the figures and a summary of your results. The report should also include details on your thinking&model building process (e.g. "x1 and y are correlated, furthermore the Z theory of economics suggests that they should be related and I'm including x1 in the model"). All R code should be included in the appendix and the main part of the report should not contain any code.