Applied Statistical Methods
Statistics GU3105 — Fall 2021

Homework 5

The following problems are due on Tuesday, November 30th, 11:59pm.

1. The purpose of this question is to give you some exercise with time series analysis.

   An example is a data set of the number of births per month in New York city, from January 1946 to December 1959, and please familiar yourself with the dataset `Nybirths.csv`.

   (a) Create a time series objective by using `ts(*, frequency=12, start=c(1946,1))`.

   (b) Make a plot of the time series data. From this plot, do you observe some trend of the data? More specifically, do you see a seasonal trend and what is the overall trend?

   (c) A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components: that is, estimating these three components. Use `decompose()` in R and plot the decomposed time series. What can you tell from the results?

   (d) Our time series analysis mainly focused on the irregular component. ARIMA models are defined for stationary time series. The time series of the irregular component appears to be stationary in mean and variance, and so an ARMA(p, q) model is appropriate. The next step is to select the appropriate ARMA model, which means finding the values of most appropriate values of p and q for an ARMA(p, q) model.

   - To do this, you usually need to examine the correlogram and partial correlogram of the stationary time series. Plot the correlogram and partial correlogram of the irregular component using `acf()` and `pacf()` functions in R, respectively. Based on your plots, which lags exceed the significance bounds?
   - Let p equals the largest lag that is still significant in the correlogram , and q for the largest significant lag in the partial correlogram. Then (p, q) is the best candidate for an ARMA model. Please fit the ARMA(p, q) model.
   - Using your fitted model, predict the number of births per month in 1960.

Project Question Series 4

(a) By **November 28th, 11:59 pm**, make a submission to Kaggle that beats the baseline.

(b) Describe the pipeline used for your submission and present your results. Your description should be sufficiently detailed so someone could reproduce your exact pipeline. You must justify the choices you make so that a practitioner could understand why you chose your specific pipeline without having to know much about statistical methods.

(c) Propose concrete and meaningful modifications or extensions to your solution. How would you improve upon your method? Justify your proposal (e.g., through a careful analysis of your current results, ablation studies, preliminary experiments, etc.). If you were to add external sources of information to your pipeline, what would you add and how would it address limitations of your current pipeline?