Statistics 153 - Introduction to Time Series
Homework 5

Due on Wednesday, November 10, 3 pm

This assignment contains more computer programming/coding work than our typical homework assignment. However, this will be helpful on the project and in real life data analysis.

Data analysis and computer exercises: Consider the time-series dataset contained in the `gas_data.csv` file on bCourses under Files > Datasets. This file contains the Australian monthly gas production for years 1970-1995. (This is a subset of the dataset found in the `forecast` package in R.)

1. **Exploratory data analysis**

   a. Make a time series plot of the raw data Y_t. Is it homoskedastic? If not, propose a variance stabilizing transform $f(\cdot)$. Define $V_t = f(Y_t)$. Plot the time-series for V_t; does it look more homoskedastic?

   (1 point)

   b. Does stationarity seem to hold for V_t? Comment.

   (1 point)

   c. Make a time series plot of the twice differenced data ($\nabla^2 V_t$). If stationarity seems like a reasonable assumption for ($\nabla^2 V_t$), also make a sample ACF plot and a sample PACF plot of ($\nabla^2 V_t$). Comment.

   (2 point)

   d. Come up with your own differenced series, perhaps based on your findings in part (c) and/or considering seasonal differences. Make a sample ACF plot and a sample PACF plot of your differenced series. Comment.

   (2 point)

2. **Model fitting and diagnostics**

   Based on your findings in Q1, come up with 3 different candidates of SARIMA models, which you want to consider for V_t. For each of them do the following steps. Note that steps 2-5 are all contained in the sarima() function's diagnostic plot:

   1. Fit the model. Give the estimated SARIMA parameters and also provide confidence intervals for them.
   2. Plot the standardized SARIMA residuals.
   3. Make an ACF plot of the SARIMA residuals.
   4. Make a normal probability plot of the standardized SARIMA residuals.
   5. Plot the p-values of the Ljung-Box statstics.
   6. Plot the sample ACF and PACF of the respective V_t differenced according to the respective SARIMA model under consideration, together with the ACF/PACF implied by the model (different from the ACF plot on step 3).
   7. Comment on the model fit based on Step 2 to Step 6.
   8. Forecast what happens in the 1996, and give the prediction intervals.

   Hint: The command `sarima()` from the `astsa` package is very useful for this problem! `auto.arima()` may also help in coming with an initial model.

   (6 points; 2 points for each model)

3. **Model selection** Of the three models,

   a. Report the AIC for each model. Based on AIC, which model is the best?

   (1 point)

   b. Report the AICc for each model. Based on AICc, which model is the best?

   (1 point)

   c. Report the BIC for each model. Based on BIC, which model is the best?

   (1 point)

   d. Now suppose we would like to select a model based on forecast performance. One approach is to perform time series cross-validation. To make things more interesting, suppose you don't believe in ARIMA modeling and consider doing a curve fitting with a third-degree polynomial plus nonparametric seasonal components instead. In particular, consider the following model:

   $$\text{Model 4: } V_t = \beta_0 + \beta_1 t + ... + \beta_3 t^3 + \beta_4 I(t \text{ is January}) + ... + \beta_{14} I(t \text{ is November}) + w_t \quad (1)$$

   where (w_t) is iid $N(0, \sigma^2)$. (Note this model is coded up for you below).

   Suppose our objective is to predict the data for the next year. Perform the following cross-validation scheme for the three models you came up with in question 2, as well as model 4 above:

   i. For each year in $\{1985, 1986, ..., 1995\}$,
      1. Train Models 1 to 3 based on all data before the selected year.
      2. For each of the models, generate forecasts for the 12 months in the selected year and compute the sum of squares of errors of the forecasts.

   ii. For each model, average the sum of squares of errors of forecasts over the years considered. Denote these averages CV_i, $i = 1, 2, 3$. These are the cross-validation scores of the models.

   iii. Report the cross-validation scores. Which model yields the smallest cross-validation score?

   Hint: To avoid numerical issues, when fitting the linear regression model (Model 4), you might consider t ranging from 1 *to the number of observations in the training set instead of starting at* 1948.

   (5 points)