Name: ____________________

GSI's name: ____________________

Lab section: ____________________

Statistics 153 (Introduction to Time Series) Homework 5

Due on November 6, 2018

Instructions: Homework is due by 3:50pm in lecture on due date. Please staple your homework when you turn it in. For the computing exercises, be sure to attach all relevant code and plots.

Data analysis and computer exercises

Consider the *birth* dataset in the `astsa` package. The dataset contains monthly live births (adjusted) in thousands for the United States from 1948 to 1979. Our objective is to find a suitable time series model for the data.

1. **Exploratory data analysis**

   a. Make a time series plot of the data (X_t). If stationarity seems like a reasonable assumption, also make a sample ACF plot and a sample PACF plot of the data. Comment.

   (2 point)

   b. Make a time series plot of the differenced data (∇X_t). If stationarity seems like a reasonable assumption for (∇X_t), also make a sample ACF plot and a sample PACF plot of (∇X_t). Comment.

   (2 point)

   c. Make a time series plot of the seasonal difference of the differenced data $(\nabla_{12}\nabla X_t)$. If stationarity seems like a reasonable assumption for $(\nabla_{12}\nabla X_t)$, also make a sample ACF plot and a sample PACF plot of $(\nabla_{12}\nabla X_t)$. Comment.

   (2 point)

2. **Model fitting and diagnostics**

   Consider the following three models:

   a. Model 1: ARIMA$(1,1,1)$;
   b. Model 2: ARIMA$(1,1,1) \times (1,1,1)_{12}$;
   c. Model 3: ARIMA$(2,1,2) \times (1,1,1)_{12}$.

   For each of the model, do the following steps:

   1. Fit the model.
   2. Plot the standardized residuals.
   3. Make an ACF plot of the residuals.
   4. Make a normal probability plot of the standardized residuals.
   5. Plot the p-values of the Ljung-Box statstics.
   6. Comment on the model fit based on Step 2 to Step 5.

   Hint: The command `sarima()` from the `astsa` package is very useful for this problem!

   (6 point; 2 points for each model)

3. **Model selection**

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

   (1 point)

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

   (1 point)

c. 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:

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

where (w_t) is iid $N(0, \sigma^2)$.

Suppose our objective is to predict the data for the next year. Perform the following cross-validation scheme:

i. For each year in $\{1960, 1961, ..., 1978\}$,

   1. Train Models 1 to 4 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, $i = 1, ..., 4$. 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 point)