Statistics 153 - Introduction to Time Series
Homework 2

Due on **September 22**, 2021, 3 pm

Computer Exercises

1. Load the data file `Basketball.csv` which contains the google trends data concerning the query *Basketball* from 2004 through most of 2019.

   (a) (5 point) Make a time series plot of (y_t). Is there any trend or seasonality?

   (b) (10 points) Use least squares (i.e. the lm function) to fit the following model

   $$y_t = \beta_0 + \beta_1 t + \sum_{j=1}^{6} \left[\beta_{2j} \cos\left(\frac{2\pi jt}{12}\right) + \beta_{2j+1} \sin\left(\frac{2\pi jt}{12}\right) \right] + w_t, \qquad (1)$$

   where (w_t) is zero mean white noise. Note that $\sin(2\pi(6)/12) = 0$, so you have to only estimate $(\beta_0, ..., \beta_{12})$. Plot (y_t) and the fitted values on the same graph.

   (c) (5 points) Create a periodogram of the data, and note the single dominant sinusoid. Which for which index (value j) does the periodogram achieve the largest value? What frequency and period of sinusoid does this imply? Does this period make sense (annual? monthly? etc.), or does this say something about leakage?

   (d) (10 points) Fit a model like Equation (1) from part b, but only using a single sinusoid of the frequency just found in part c. Plot the fitted values on the data and compare to the other models. Do these new fitted values oscillate at the right frequency with the data? Does this model fit as well as the others? Why not?

   (e) (5 points) Check the periodogram of the residuals from this single sinusoid model. Are there still important sinusoids present even after removing the first dominant sinusoid?

   (f) (10 points) Estimate the trend by smoothing (The `filter` function in `R` might be helpful). Explain reasons behind your choice of the smoothing parameter. Once again, provide a plot of the original data along with the corresponding trend estimate. Also provide a time plot of the residuals. Comment on each of these plots.

   (g) (10 points) Plot the following three versions of (y_t):

      i. (∇y_t): the first difference of the data

ii. $(\nabla_{12} y_t) = (y_t - y_{t-12})$: the seasonal difference of the data (with 12 months being a season)

iii. $(\nabla \nabla_{12} y_t)$: the first difference of the seasonal difference of the data (with 12 months being a season)

Which one looks the most like a stationary process?

(h) (10 points) Now, your turn to pursue stationarity in the residuals! Just do the best you can, it won't be perfect. As a reminder, the goal is for the residuals to look like a stationary process and beware of overfitting. Beware of heteroskedasticity (uneven variance) too, but remember that systematic/deterministic movements in your residuals indicate that your trend/seasonality modeling is not yet complete.

Use what we have seen on the preceding questions, what we have learned in class, and your own knowledge and skill. This will require some trial and error as you find an appropriate model, but your report does not need the travel log of everything you tried. Include only the following:

i. The mathematical expression of your final model, like model (1) above, accept your residuals need only be stationary (X_t) and not white noise (W_t).

ii. A plot that includes both the original basketball time series and your model's fitted values.

iii. A plot of your model's residuals over time.

Theoretical Exercises

2. (10 point) While analyzing their annual sales data (liters of wine sold per year) for the past 30 years, a vineyard found that after taking four successive differences, the resulting data had a mean of 2458 liters and looked like a stationary process (e.g. white noise). If the actual data for the past four years were 2018 - 2134 liters, 2017 - 2403 liters, 2016 - 2076 liters, and 2015 - 2290 liters. What would be a reasonable forecast for sales in 2019? Explain.

3. Let (X_t) be a weakly stationary process with mean μ and autocovariance function $\gamma(k) = \mathrm{Cov}(X_t, X_{t+k})$. Consider a derived series (Y_t) defined as

$$Y_t = \sum_{j=a}^{b} c_j X_{t+j}, \tag{2}$$

where a and b are integers with $a \leq b$, and $(c_a, ..., c_b)$ are all fixed real numbers.

(a) (10 points) Show that $\{Y_t\}$ is a weakly stationary series.

(b) (6 points) Show that order k $(k \geq 1)$ differencing (that is, $\nabla^k X_t$) can be put in the form of Equation (2). Identify the corresponding a, b, and $(c_a, ..., c_b)$.

(c) (5 points) Recall that smoothing via simple averaging with parameter q corresponds to computing for each t

$$\frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}.$$

Show that smoothing via simple averaging with parameter q can be put in the form of Equation (2). Identify the corresponding a, b, and $(c_a, ..., c_b)$.

(d) (4 points) Is the kth differenced version of a weakly stationary process always weakly stationary? Is the smoothed (via simple averaging) version of a weakly stationary process always weakly stationary?