Statistics 153 - Homework 2

Frequencies and Filters

Due on Wednesday, **February 17**, 2021, 11:59pm

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

   (b) (2 points) Use least squares (i.e. the lm function) to fit the following model

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

   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) (2 points) Use least squares to fit the following model

   $$y_t = \beta_0 + \beta_1 t + \beta_2 I(t \text{ is January}) + ... + \beta_{12} I(t \text{ is November}) + w_t, \tag{2}$$

   where (w_t) is zero mean white noise. The indicator variable $I(A)$ takes values 1 if A and 0 otherwise. Plot (y_t) and the fitted values on the same graph. (Note that the variable `month` has been created for you.)

   (d) (1 point) Compare the fitted values from (b) and those from (c). What do you notice?

   (e) (2 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?

   (f) (2 points) Fit a model like Equation (1) from part b, but only using a single sinusoid of the frequency just found in part e. 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?

(g) (1 points) Check the periodogram of the residuals from this single sinusoid model (perhaps the smoothed periodogram!). Are there still important sinusoids present even after removing the first dominant sinusoid?

(h) (2 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.

(i) (3 points) Recall the definition of a filter, where V_t is the filtered process of raw/original time series Y_t

$$V_t = \sum_{j=-\infty}^{\infty} a_j Y_{t-j}$$

and the filter is defined by coefficients a_j, $j \in \mathbb{R}$. The following filters are known as "differencing". Plot the following three filtered versions of (y_t) (either or both the `filter` and `diff` functions may be helpful):

i. $(\nabla y_t) = y_t - y_{t-1}$ is the first difference of the data, where $a_0 = 1$, $a_1 = -1$, and $a_j = 0$ otherwise.

ii. $(\nabla_{12} y_t) = (y_t - y_{t-12})$: the seasonal difference of the data (with 12 months being a season), where $a_0 = 1$, $a_{12} = -1$, and $a_j = 0$ otherwise.

iii. $(\nabla\nabla_{12} y_t)$: the first difference of the seasonal difference of the data (with 12 months being a season), where $a_0 = 1$, $a_1 = -1$, $a_{12} = -1$, $a_{13} = 1$ and $a_j = 0$ otherwise.

Which one looks the most like a stationary process?

(j) (4 points) Now, it's 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. Don't worry about heteroskedasticity (uneven variance), we won't be able to discuss that in lecture until the day after this is due... 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 models (1) and (2) above, though 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, with the model curve extended past the observed data to demonstrate what your forecast will do through the end of 2020 (this could be used to answer the question, how popular would basketball have been if there was no COVID-19).

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