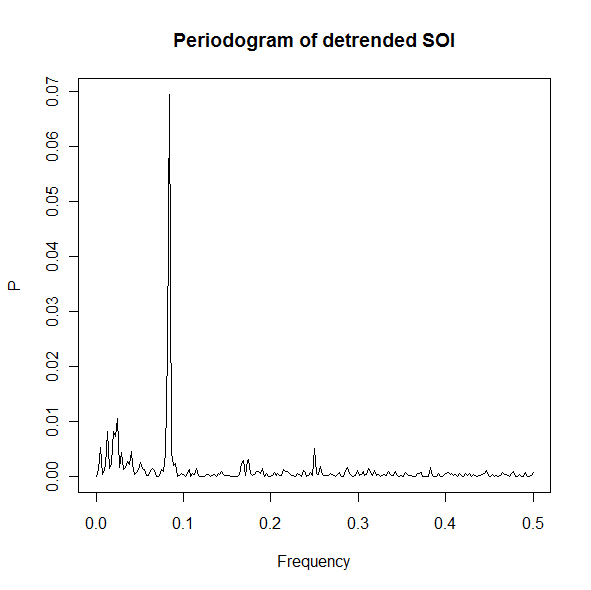
**Stat 510 Week 7 Homework Solutions**

1. Following is the periodogram of the detrended SOI series that we’ve used a few times in the past. SOI = southern oscillation index, a measure of temperature in the southern hemisphere. The series is n = 452 monthly values. The detrended values are the residuals from a linear regression on time (1 to 452).



**a**. Roughly, at what frequency is the dominant peak? The period of a cycle is the reciprocal of the frequency. About what is the period for the frequency at which the dominant peak occurs? (Keep in mind that this is monthly data and this is a weather measure.

**The peak is at a frequency equal to roughly .08 which is a period of 1/.08 = 12.5. We might conjecture that the period is in fact 12 months (for a seasonal effect) and that the peak frequency would then be at 1/12 = .083333…**

**b**. Roughly what is the region of frequency range for the second highest peak(s)? Roughly what would be the period(s) associated with these frequencies? (Hint: The El Nino weather phenomenon is believed to affect things about every 3 to 7 years.)

**It’s hard to judge exactly, but the peak may be a frequency around .02 or so. Thus the period may be about 1/.02 = 50 months, or around 4 years.**

**2.** Do problem 4.1 on page 255 of the textbook. For part a, use a modified version of the R code on page 178. To simply increase the sample size from 100 to 128, you need only change 1:100 to 1:128. For part b, use a modification of the code on page 179 or the code below. To print the periodogram between 0 and .5 only, as opposed to between 0 and 1 in Figure 4.2, you may use the following code:

P = abs(2\*fft(x)/128)^2

f = 0:64/128

plot(f, P[1:65], type="o", xlab="frequency", ylab="periodogram")

The modifications for part b are to change 100 to 128 and change the values 50 and 51 in the second and third lines accordingly. For part c, it’s only necessary to plot the series that is the sum and plot the periodogram of the sum. The command to get the sum will be something like sum = x1+x2+x3+rnorm(100,0,5). The parameters for the rnorm command are sample size, mean, standard deviation. (The notation the authors used in part c gives the mean and the variance.)

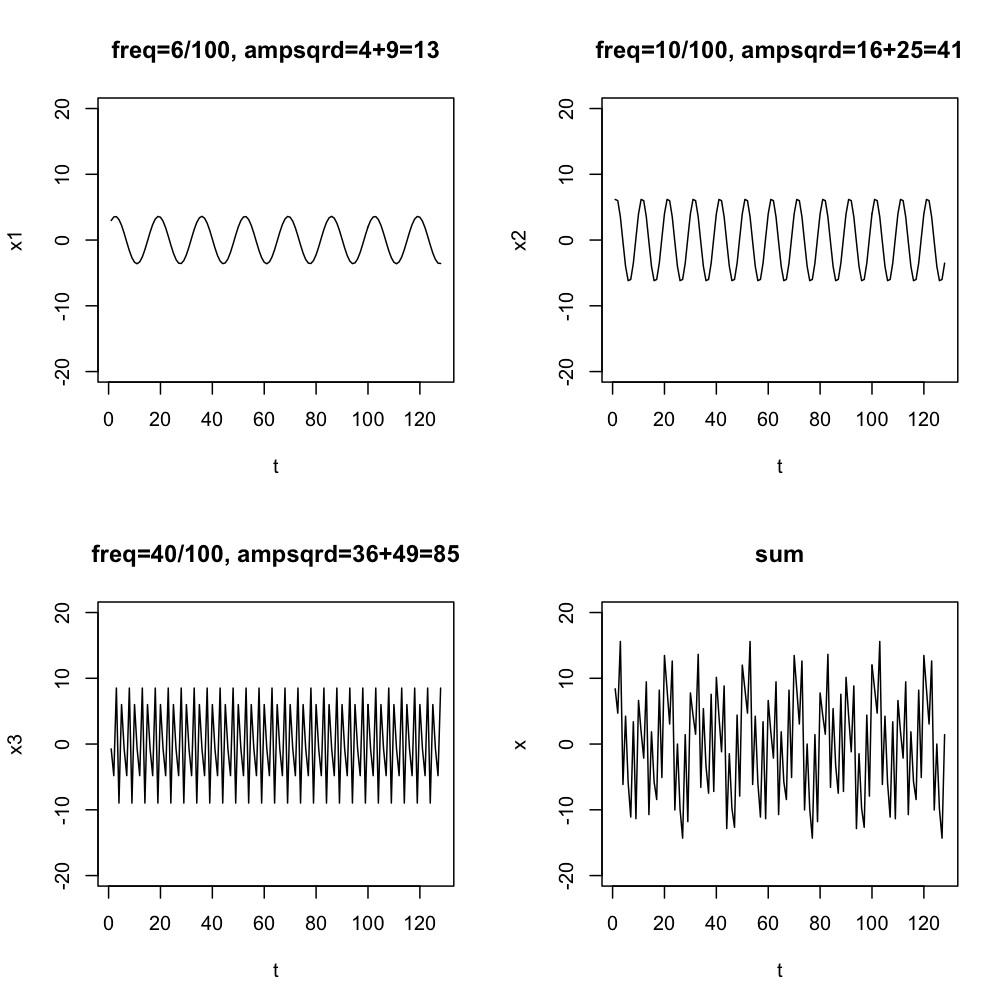
For answers to the book parts–

Part a Give plots of the x1, x2, x3 and the sum. To determine how things differ from Example 4.1, focus on the formula for the series rather than the plots.

Part b. Give the periodogram of the sum and describe the main frequency peaks. What has changed from Figure 4.2?

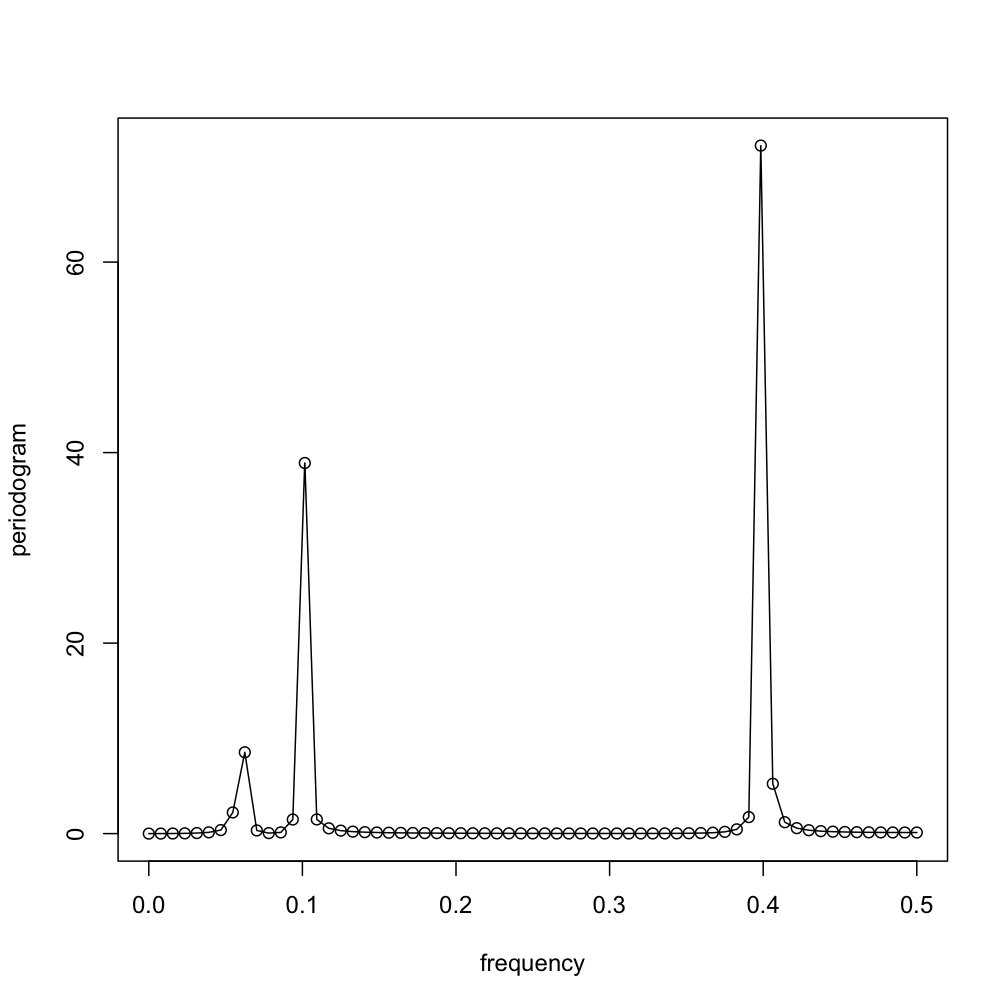
Part c. Give the plot of the sum and its periodogram. And, describe the locations of the main peaks in the periodogram.

**The plots are below and besides being longer than the plots in Example 4.1 have not changed. The difference in the series is that each of the frequencies 6/100, 10/100, and 40/100 are no longer fundamental frequencies, that is of the form j/n=j/128.**



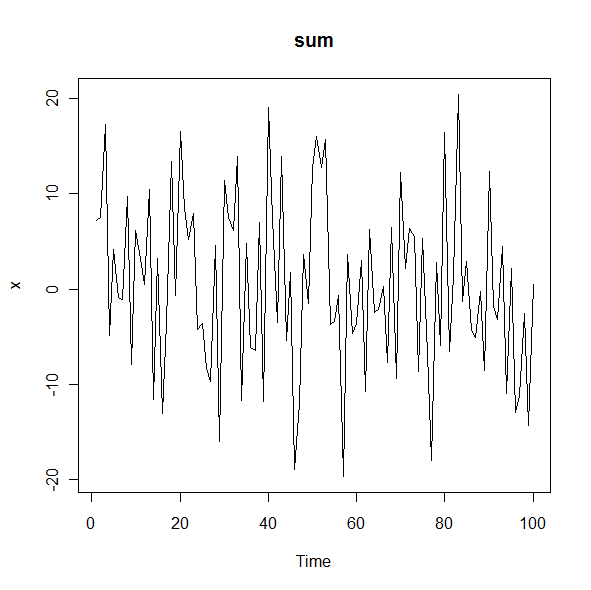
Part b. Give the periodogram of the sum and describe the main frequency peaks.

**The periodogram follows. Because the t are not evaluated at fractions of 128, the frequencies are not fundamental frequencies and the periodogram sees positive values surrounding each frequency as opposed to 0. Also, the value of the periodogram at each frequency is no longer the amplitude squared.**

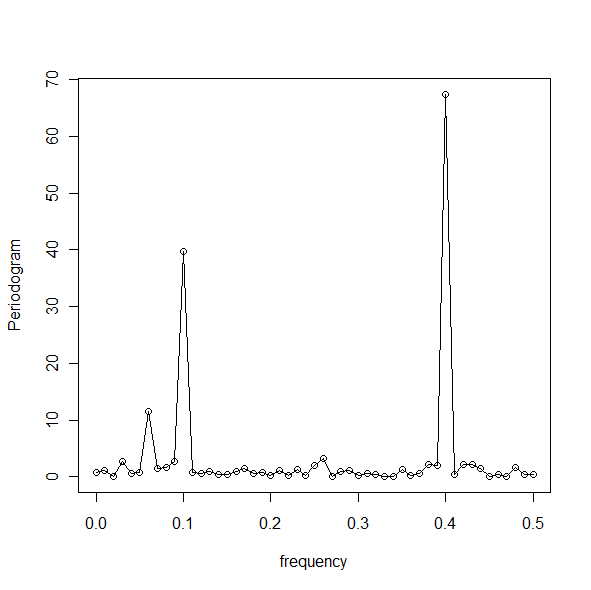


Part c. Give the plot of the sum and its periodogram. And, describe the locations of the main peaks in the periodogram.

**The answer will differ a bit for different students because the random error generation will differ. The plot of the sum that we got is –**



**The periodogram of the sum follows. Again this may differ for different students due to the simulation aspect.**



**The locations of the peaks could conceivably differ slightly from the “theoretical” peaks due to the simulation aspect, but in theory the peaks will be at 6/100 = .06, 10/100 = .10 and 40/100 =.40 (as they are in the above periodogram).**

**3.** Use the dataset “speech.dat” from the Week 6 folder. The data are described and plotted in Example 1.3 on page 5 of the textbook. The series is a sample of recorded speech of the phrase aaa…hhh, sampled for 0.1020 seconds at the rate of 10,000 points per second. Thus there are n = 1020 data values. The authors don’t say what exactly is the response variable, but a likely possibility for the response is a measure of the audio frequency of the sound. (Maybe they held out from saying that to avoid confusion with the frequency of a cosine wave.)

The data can be read as x = scan("speech.dat") once the data file is in place on your computer.

There’s no obvious trend seen in Figure 1.3 on page 6, but it never hurts to detrend before looking at a periodogram. At worst, detrending does nothing. Detrend in R by doing a linear regression on the time index. The detrended values are the residuals from this regression. The sequence of commands might be something like

t = 1:1020

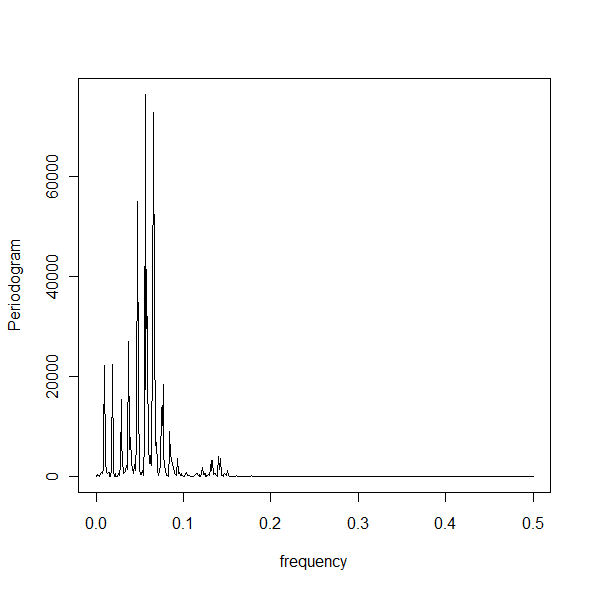
regr = lm(x~t)

y = residuals(regr)

In the above code, y would be the detrended series, the residuals from a linear regression.

Plot the periodogram of the detrended series. Give the plot and describe the pattern of peaks as they relate to frequency (the horixontal axis).

**The scaled periodogram follows. Note that there are several peaks that seem to be spaced at approximately regular intervals. The location of the maximum peak is a frequency of about .056. The second highest peak is at a frequency of approximately .066. The third highest peak is at a frequency of about .047. (In general the frequencies for the peaks differ by about .01.)**



**4.** This problem is a review of the material in Week 3 (non-seasonal ARIMA models). Use the lakesim100.dat dataset in the Week 6 folder. It gives *n* =100 consecutive annual measures of the level of a lake.

a. Plot the time series. Give the plot and an interpretation of the plot as the answer to this part. Please review Lesson 1.1 for what you should interpret in the plot.



**The series series is trending downward without signs of seasonality, outliers, or heteroskedasticity.**

b. Determine the ACF and PACF for the series. (Hint: The most efficient way to do this is to load the astsa library and then use the acf2 command.)

Give the ACF and PACF plots AND write a brief interpretation. Do you think the data are stationary? Do you think you should suggest models from the ACF and PACF?



**Because of the downward trend and somewhat slow tapering ACF, there seems to be an issue with the assumption of stationarity though it is not as pronounced as we have seen in previous examples. Because the lag 1 ACF is close to 1, it would be wise to examine first differences to rule out a random walk. Recall from Lesson 3.1 that you must examine the ACF and PACF together.**

c. Give the ACF and PACF of first differences. Write a brief interpretation that indicates what model(s) you think are suggested.

****

**The ACF and PACF, shown above, have a MA(1) or MA(2) pattern. There are significant values in the ACF at lags 1 and 2 followed by non-significant values. The PACF tapers to 0 as it alternates in sign.**

d. Determine an ARIMA model that you think is suitable for these data. Describe the model and give the estimated coefficients. (You can just give the output that you get from your software.)

**Following is the R output for the MA(2) model. If we let *xt* = Lake level at time t and z*t* = *xt - xt-1*, the forecast model is z*t* = -.0218 + *wt + .*7471*wt-1 +* .8567*wt-2*.**

**Coefficients:**

**ma1 ma2 constant**

**0.7471 0.8567 -0.0218**

**s.e. 0.0694 0.0651 0.0258**

**sigma^2 estimated as 0.009878: log likelihood = 86.3, aic = -164.6**

**$AIC $AICc $BIC**

**[1] -3.557428 [1] -3.533218 [1] -4.479273**

e. For your model, give diagnostic plots/tests for checking the suitability of the model (for example, the diagnostic output from the sarima command). Briefly discuss what the plots indicate about your model.

**The following diagnostics for the MA(2) were given by the “sarima” command. All looks good.**

