**Stat 510 Week 9 Homework Solutions**

1. Use datasets “xwk9prob1.dat” and “ywk9prob1.dat” from the Datasets folder. The first file is an x-variable and the second is a y=variable for n = 200 time periods. The task is to identify the connection between the y-variable and the x-variable.

Use Lesson 9.1 and the R-code within Lesson 9.1 for guidance.

A. Determine an ARIMA model for the x-variable. (Hint: differencing is needed.) What is the estimated ARIMA model?

**An ARIMA(1,1,0) works well. An ARIMA (0,1,2) might too, but the ARIMA(1,1,0) will be more convenient for filtering y.**

***Notes* – To identify the model, first do a time series plot of x. It shows an overall upward trend indicating a need for differencing. The ACF and PACF of the differenced data, have a clear AR(1) pattern although an MA(2) is a conceivable interpretation as well.**

**With R’s arima command the estimated AR coefficient = 0.4698.**

**With the sarima command, the estimated AR coefficient = 0.4655 (it included an intercept and arima does not for differenced data.)**

**Minitab might provide a slightly different value, but most likely it rounds to 0.47.**



B. Use the estimated model for x to filter the y-variable for pre-whitening. Explain exactly how you filtered the variable, either by giving an equation or an R command.

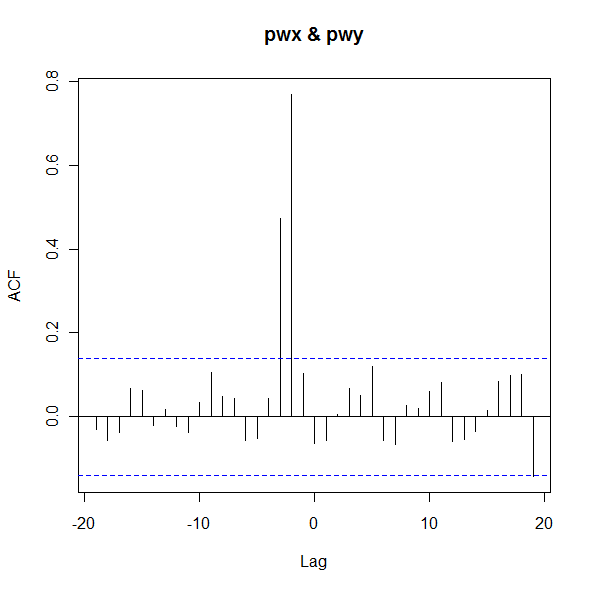
****

**pwy = filter(y, filter = c(1,-1.47,.47), sides =1)**

**The AR(1) polynomial for x and the first differencing are applied to the y-series. Rounding the AR coefficient to 0.47 allows for the use of differing routines to estimate the x model.**

C. Determine the CCF between the residuals from the model for x and the filtered y-values. Give the CCF and write a brief interpretation.

**The CCF given by R is the following. In Minitab this will be flipped over so that the major spikes are on the positive side. (This would also happen in R if you list the pre-whitened y first.) There are major spikes at lags 2 and 3 of x suggesting a regression model in which these lags of x predict y. A conceivable interpretation is that lag 2 of x and lag 1 of y could be used to predict y, but for that we would see more tapering of the ccf values away from lag 2.**



D. Estimate the regression model that is suggested by the CCF in part C. Write the estimated equation.

We’ll stop the analysis here, but if we were to polish it a bit we would add lags of y into the regression to account for the fact that the errors have an AR structure.

**Results given by R lead to predicted yt = 14.8949 + 1.7440xt-2 + 0.9397xt-3**

**Minitab should give the same answer (perhaps different number of decimal places)**

***Note*: R code is something like**

**x = ts(x)**

**y = ts(y)**

**a = cbind(y,lag2x = lag(x,-2), lag3x=lag(x, -3))**

**lm(y ~lag2x+lag3x, data = a, na.action = na.omit)**

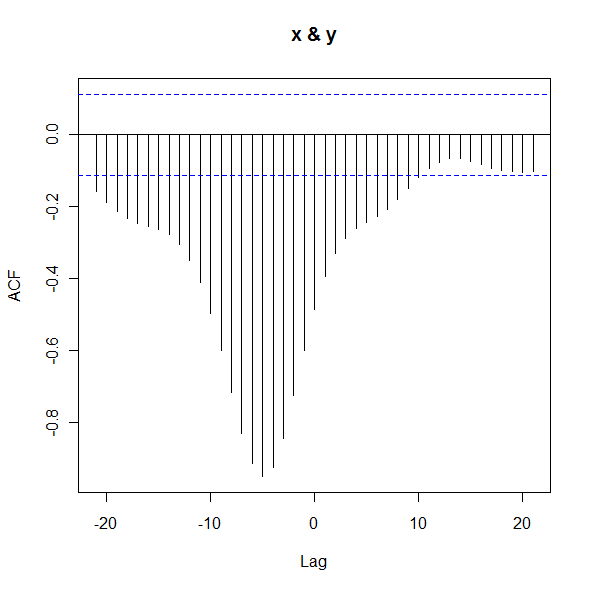
2. Use the datasets “inputrate.dat” and “outputrate.dat” from the Datasets folder. The inputrate.dat file gives an input rate for a chemical added to an industrial process. The outputrate.dat gives the output rate of another chemical as the result of the process.

This is a dataset where the CCF can be interpreted without pre-whitening (in our opinion).

A. Determine the CCF of the two series. Give the CCF and interpret it. For instance, where is the peak lag? What is the pattern around the peak lag?

**The R CCF is given below. In Minitab, this will be flipped over so that the peak is on the positive lag side.**

**The peak is in the vicinity of lags 4,5, and 6. There is tapering in both directions from there. (By, the way there looks to be negative correlation between x and y.)**



B. Estimate a model for predicting y = output using the first lag of y= output and lags 5 and 6 of the input series. Give the R2 for this model and discuss whether all terms are needed in the model.

**R-squared = .9859 (or 98.59%). All terms are statistically significant. Output (optional for the answer) is:**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.55425 1.27576 1.218 0.224**

**lag1y 0.97117 0.02392 40.603 <2e-16 \*\*\***

**lag5x -1.95197 0.07046 -27.703 <2e-16 \*\*\***

**lag6x 1.88773 0.08754 21.564 <2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 0.3861 on 286 degrees of freedom**

**(12 observations deleted due to missingness)**

**Multiple R-squared: 0.9859, Adjusted R-squared: 0.9858**

**F-statistic: 6666 on 3 and 286 DF, p-value: < 2.2e-16**

Note: You may notice that the residuals from the above regression have an ARMA structure. Following the techniques in Lesson 8, a more proper analysis would be to include an AR(2) model for the residuals as follows:



Use ordinary regression to estimate the model

where the wt are independent and follow a normal distribution of mean 0 and constant variance. For this example, results are

Call:

stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,

Q), period = S), xreg = xreg, optim.control = list(trace = trc, REPORT = 1,

reltol = tol))

Coefficients:

ar1 ar2 intercept outlag1 inlag5

1.3682 -0.4545 31.5202 0.4105 -1.0614

s.e. 0.0803 0.0653 3.8618 0.0719 0.1065

sigma^2 estimated as 0.08929: log likelihood = -62.71, aic = 137.41

$AIC

[1] -1.382805

$AICc

[1] -1.375239

$BIC

[1] -2.321374

3. Use the dataset “highway.dat” from the Week 9 folder. The dataset simulates the number of highway deaths and serious injuries for 139 months - 109 months before a seatbelt law was implemented and then for an additional 30 months after the law.

A. Plot the data. Discuss any noteworthy features. From the plot, can you tell whether the number of highway deaths and serious injuries were lower than expected after the law was enforced?

**There’s a downward trend and perhaps seasonality (looks to be a regular up and down). It’s difficult to judge whether the counts dropped significantly after time 109, but it does look like there was an effect.**



B. Assuming that you called the original series x (you don’t have to), in R, use the commands

before = window (x, 1, 109)

after = window (x, 110, 139)

The before variable now contains the data from before the seatbelt law. The after variable contains the data from after the law.

Estimate an ARIMA(1,0, 0)×(1,0,0)12 model for the before data. Discuss whether this model fits the data.

**The R estimates are**

**Coefficients:**

**ar1 sar1 xmean**

**0.1591 0.5739 1812.1008**

**s.e. 0.0954 0.0777 12.1635**

**sigma^2 estimated as 2684: log likelihood = -587.35, aic = 1182.71**

**The ACF of the residuals and the Ljung-Box statistics both look good. The normal probability plot shows normality. The nonseasonal AR(1) coefficient is not significant and could be dropped from the model. You would need to reassess diagnostics and information criterion to choose between this model and one with just the seasonal term.**



C. Use the ARIMA(1,0, 0)×(1,0,0)12  and the before data to forecast counts for the 30 times in the after period. Then, determine the differences between the actual after data and the forecasts and plot those differences. Give the plot as the answer to this part, and briefly discuss whether you think the law may have led to a generally lower number of fatalities and injuries.

Something like this should work:

afterpred = sarima.for(before, 30,1,0,0,1,0,0,12)

diffs = after - afterpred$pred

plot(diffs)

**For all of the months following the law, the actual incident count was below the predicted value based on the model for the before period. It looks like there was a drop in the count.**



D. Determine the mean difference between actual and predicted for the after period. Something like mean(diffs) should work.

**mean(diffs) = −119.6. On average, the number of fatalities and serious injuries dropped by about 120 incidents after the law.**