**Stat 510 Week 9 Homework**

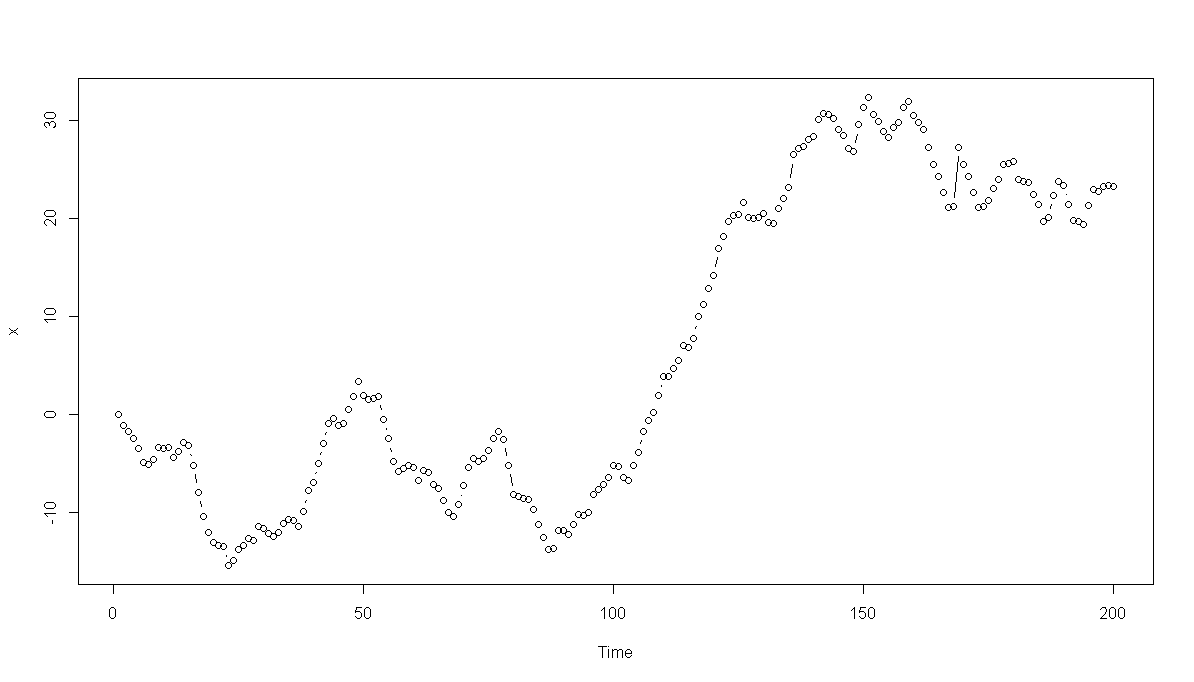
**Submit answers in a Word document or a pdf to the designated ANGEL Drop Box in the Drop Boxes folder.**

1. Use datasets “xwk9prob1.dat” and “ywk9prob1.dat” from the Week 9 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.

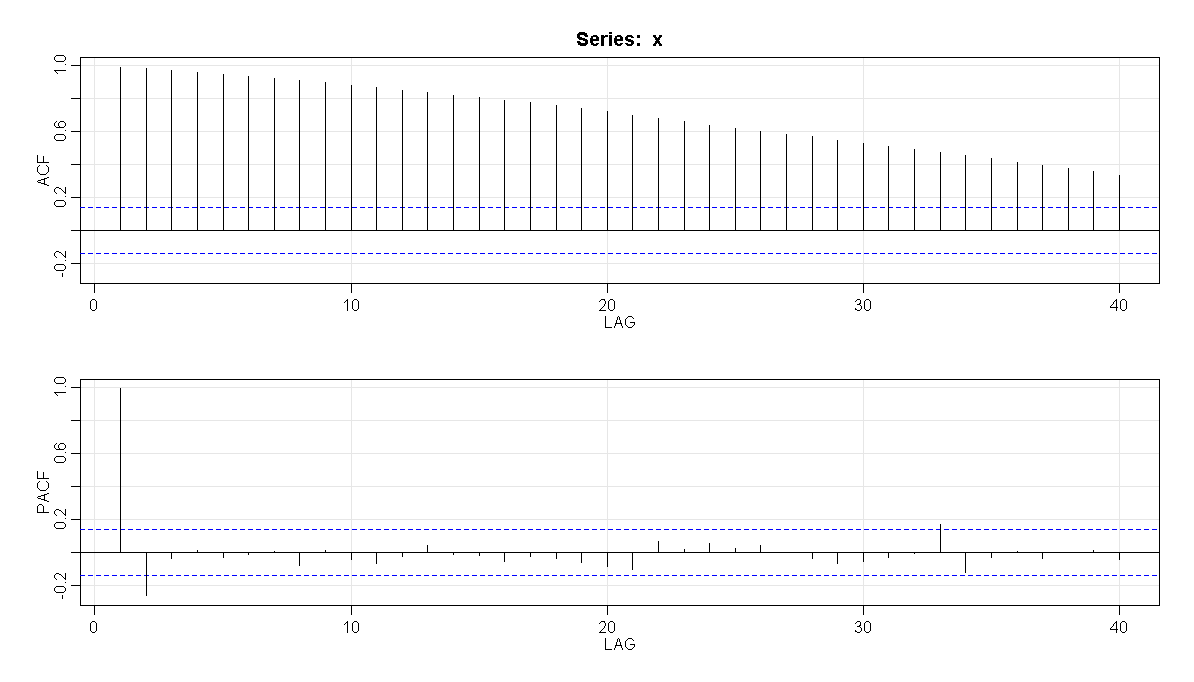
Use Lesson 9.1 and the R-code within Lesson 9.1 for guidance.

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

The plot for x:

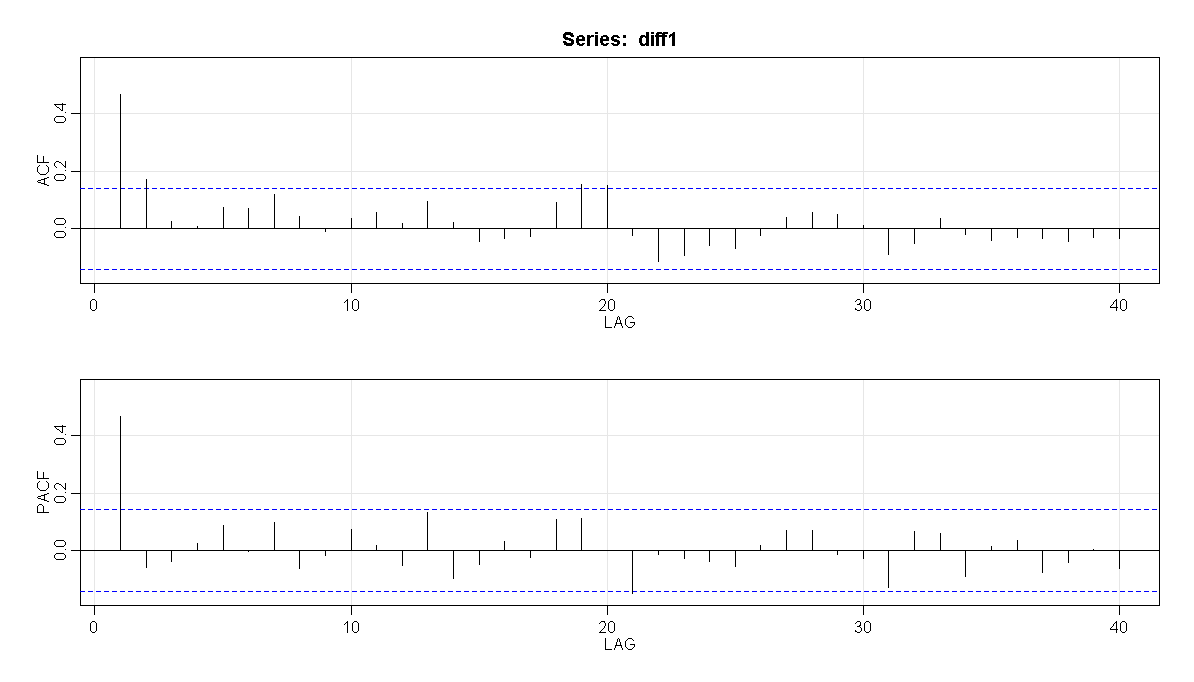


The ACF and PACF are:



We note:

* The ACF of the x-series did not taper to 0 and stayed significant for many lags, suggesting a non-stationary series. Thus a first difference should be tried.

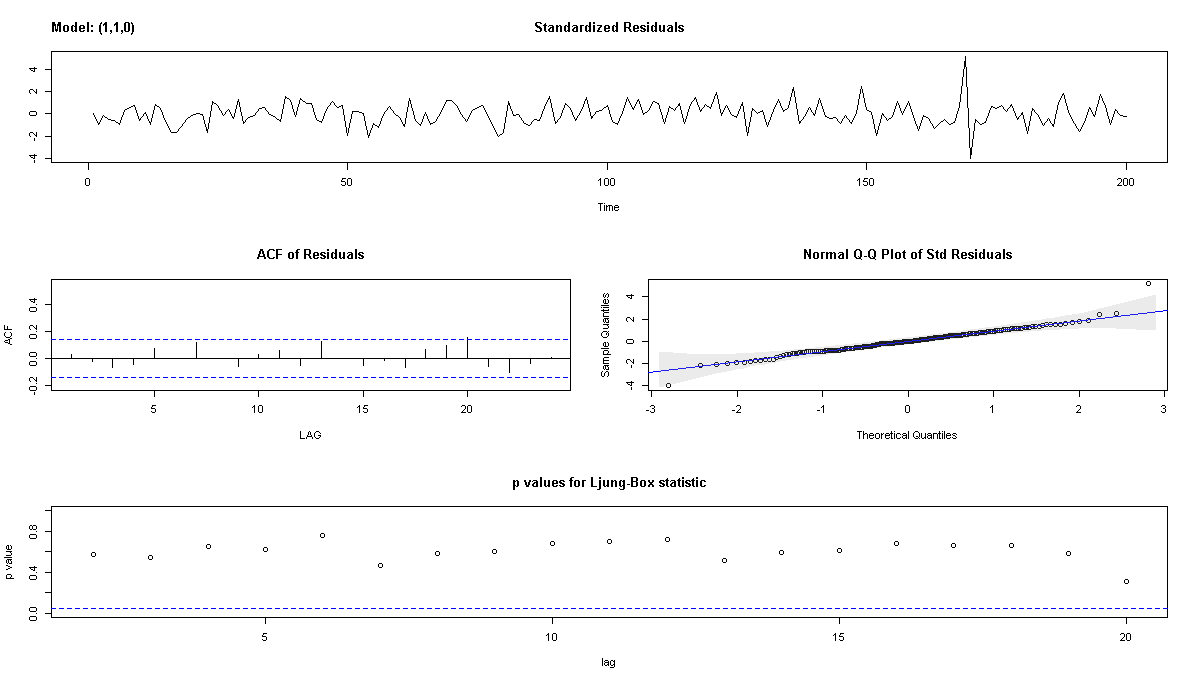


We see:

* The PACF spikes and then cuts off after lag 1. The ACF has a trailing off to 0 (though it does trail off too quickly) – we can use an AR(1), so an ARIMA(1,1,0) should be tried.
* While we do see AR(1) model, we can compare to ARMA(1,1) for the first difference to get better perspective.

In the comparison of AR(1) and ARMA(1,1) we see very similar values for AIC, BIC but get lower values for AR(1)

We will go with ARIMA(1, 1, 0) model for the x series. ARIMA(1,1,0) diagnostics are:



The diagnostics indicate that the model is appropriate because:

* The time series plot of the standardized residuals mostly indicates that there’s no visible trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations – a good result.
* The Q-Q plot is a normal probability plot – The plot looks almost linear and the assumption of normally distributed residuals holds.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. All p-values are above the dashed blue line - That’s a good result.

The model from ARIMA(1,1,0) is:

Call:

arima(x = x, order = c(1, 1, 0))

Coefficients:

ar1

0.4698

s.e. 0.0624

So, for the data, the estimated phi-coefficient for the ARIMA(1,1,0) is ϕ1 = 0.4698

The estimated model can be written as (1 - 0.4698B)(1 - B)(xt - μ) = wt.

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.

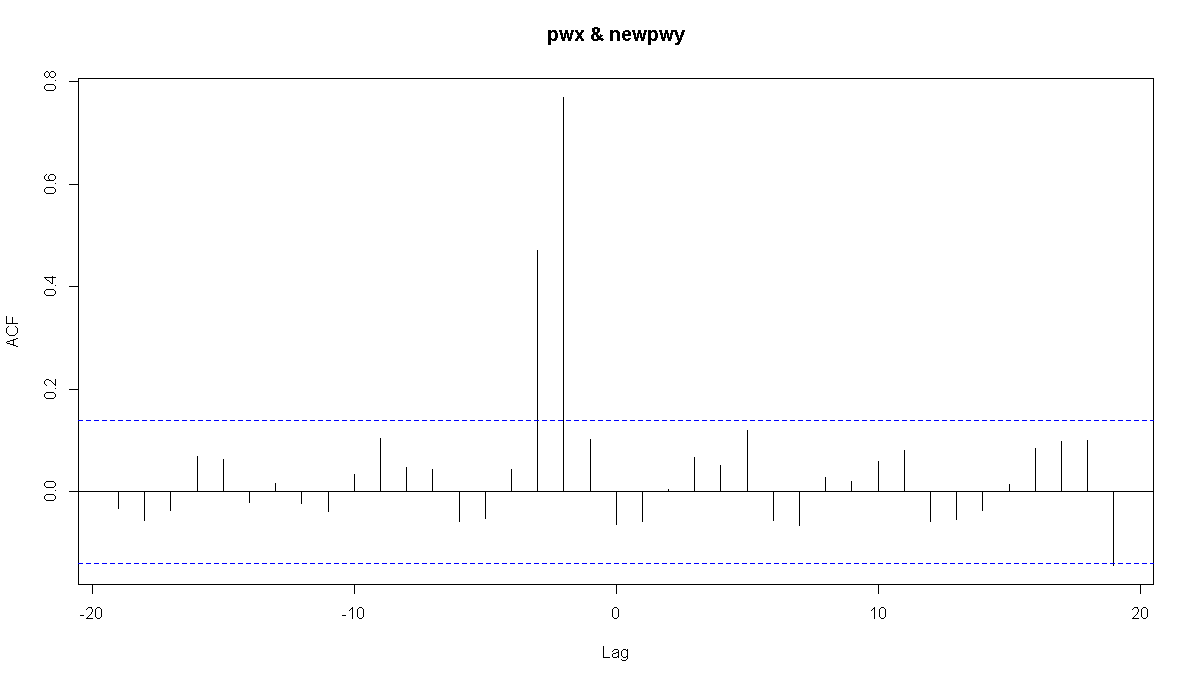
We’ll calculate the filtered y as: (1 - 0.4698B)(1 - B)yt = (1 - 1.4698B + 0.4698B2)yt

An R command that carrys out this operation is:

newpwy = filter(y, filter = c(1, -1.4698, 0.4698), sides =1)

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 plot:



We see clear spikes at lags 2 and 3. Thus xt-2 and xt-3 should be tried as predictors of yt

We can also validate from the ccfvalues:

Autocorrelations of series ‘X’, by lag

-19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9

-0.032 -0.056 -0.037 0.068 0.064 -0.021 0.017 -0.023 -0.038 0.034 0.105

-8 -7 -6 -5 -4 -3 -2 -1 0 1 2

0.048 0.045 -0.057 -0.051 0.044 0.472 0.769 0.102 -0.063 -0.057 0.005

3 4 5 6 7 8 9 10 11 12 13

0.067 0.052 0.120 -0.056 -0.066 0.028 0.020 0.060 0.081 -0.058 -0.054

14 15 16 17 18 19

-0.036 0.015 0.085 0.098 0.101 -0.143

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.

The regression model suggested by the CCF relates y with xt-2 and xt-3

We perform the regression: lm(y~xlag2+xlag3, data = alldata)

Coefficients:

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

(Intercept) 14.89486 0.11162 133.44 <2e-16 \*\*\*

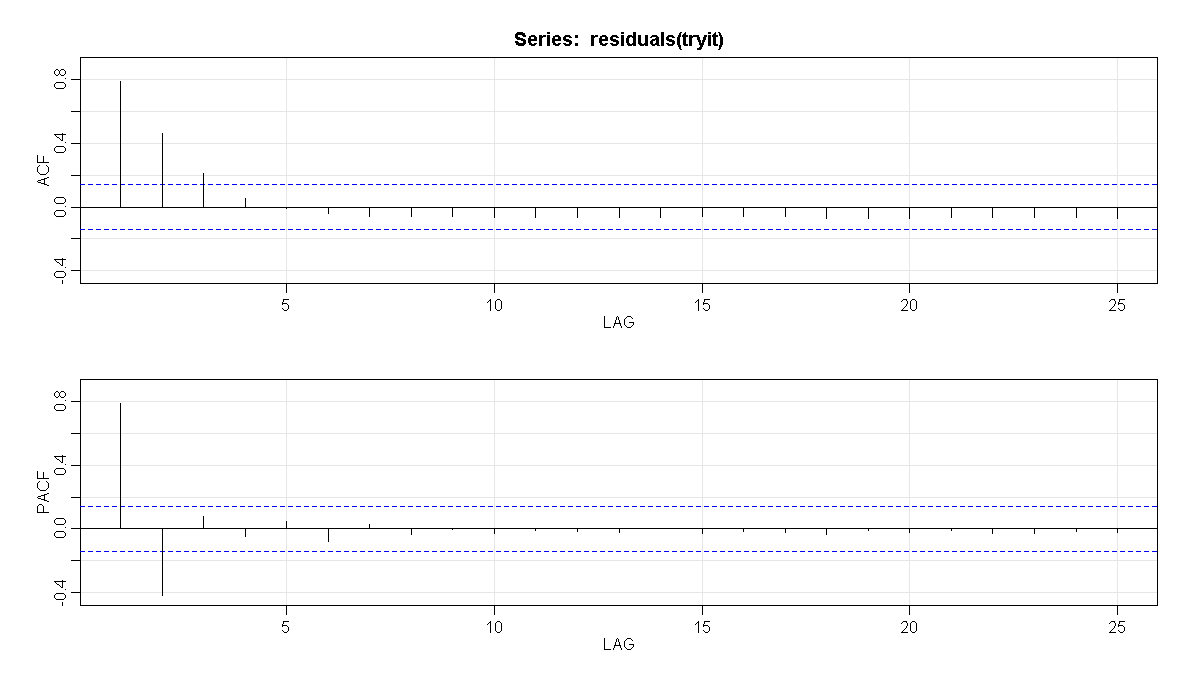
xlag2 1.74400 0.07934 21.98 <2e-16 \*\*\*

xlag3 0.93972 0.07953 11.82 <2e-16 \*\*\*

As expected we see that both the lags are significant and give us the model:

y = 1.744 xt-2 + 0.93972 xt-3 + et

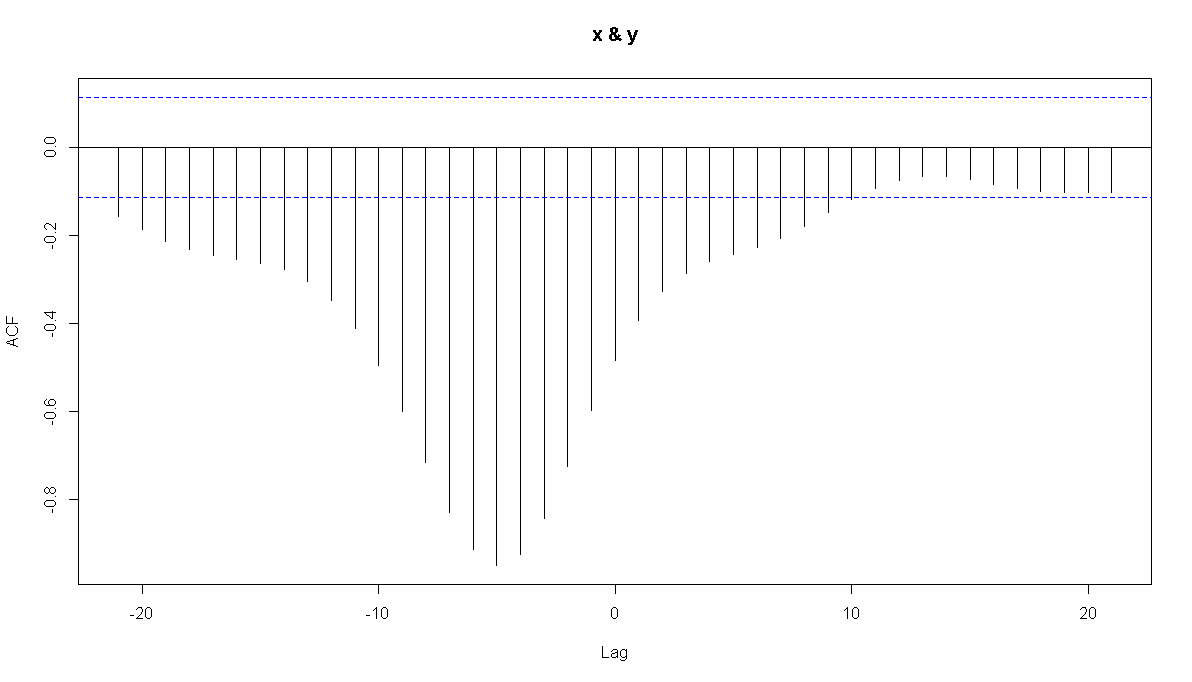
Further the ACF / PACF of the residuals indicate that the errors have a time series structure (AR2):



2. Use the datasets “inputrate.dat” and “outputrate.dat” from the Week 9 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 most dominant cross correlations occur somewhere between h =−12 and about h = 2. It’s difficult to read the lags exactly from the plot, so we list the values:

-21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11

-0.157 -0.187 -0.213 -0.233 -0.246 -0.255 -0.263 -0.278 -0.305 -0.348 -0.411

-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0

-0.495 -0.600 -0.717 -0.829 -0.915 -0.950 -0.925 -0.843 -0.725 -0.598 -0.484

1 2 3 4 5 6 7 8 9 10 11

-0.393 -0.329 -0.286 -0.260 -0.243 -0.227 -0.206 -0.179 -0.149 -0.118 -0.093

12 13 14 15 16 17 18 19 20 21

-0.075 -0.066 -0.066 -0.073 -0.083 -0.093 -0.100 -0.103 -0.103 -0.102

We also see from the plot and values:

* The peak is at lag -5 followed by a tapering pattern is an indicator that lag 1 and possibly lag 2 values of the y-variable may be helpful predictors.

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.

The regression model asked here relates y with yt-1, xt-5 and xt-6

We perform the regression: tryit = lm(y~ylag1+xlag5+xlag6, data = alldata)

The output we get is:

Call:

lm(formula = y ~ ylag1 + xlag5 + xlag6, data = alldata)

Residuals:

Min 1Q Median 3Q Max

-1.31212 -0.22883 -0.01144 0.18897 1.36270

Coefficients:

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

(Intercept) 1.55425 1.27576 1.218 0.224

ylag1 0.97117 0.02392 40.603 <2e-16 \*\*\*

xlag5 -1.95197 0.07046 -27.703 <2e-16 \*\*\*

xlag6 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

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

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

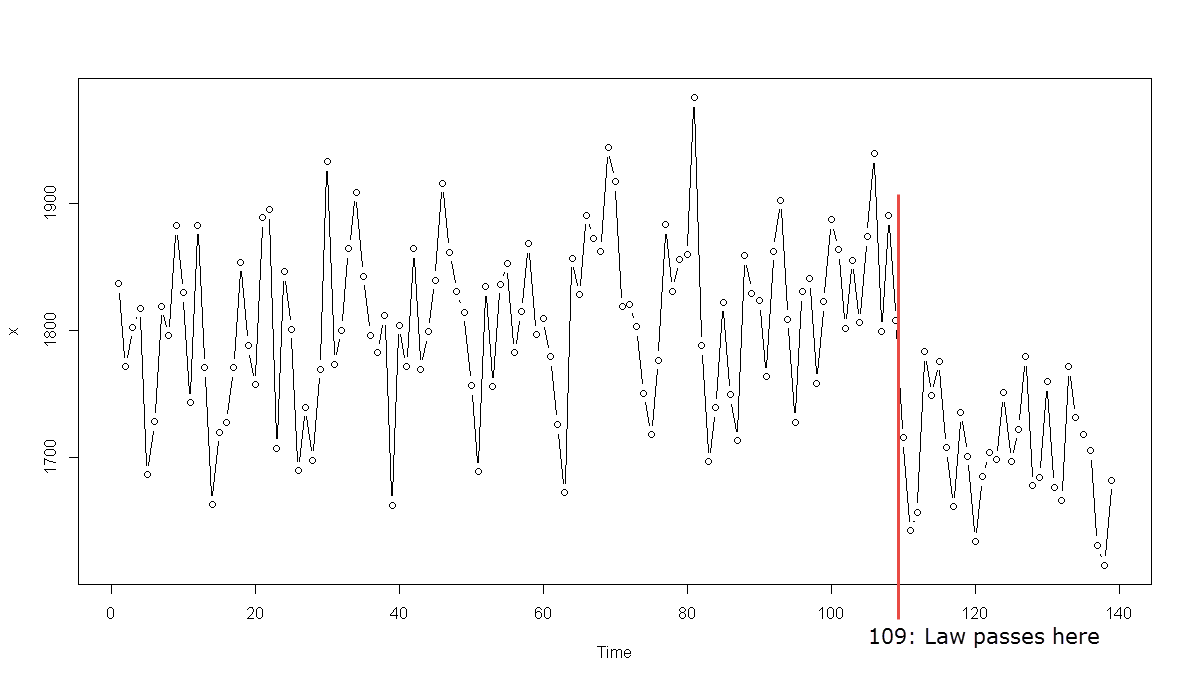
The R-squared value is about 98.6 %

All the lags are indicated to be significant and therefore needed in the model.

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?

The plot is:



We can see:

* The pattern indicates the effect of the intervention caused by the passage of the seatbelt law.
* The law decreased the number of highway deaths and serious injuries.

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.

We get the following output from sarima():

$ttable

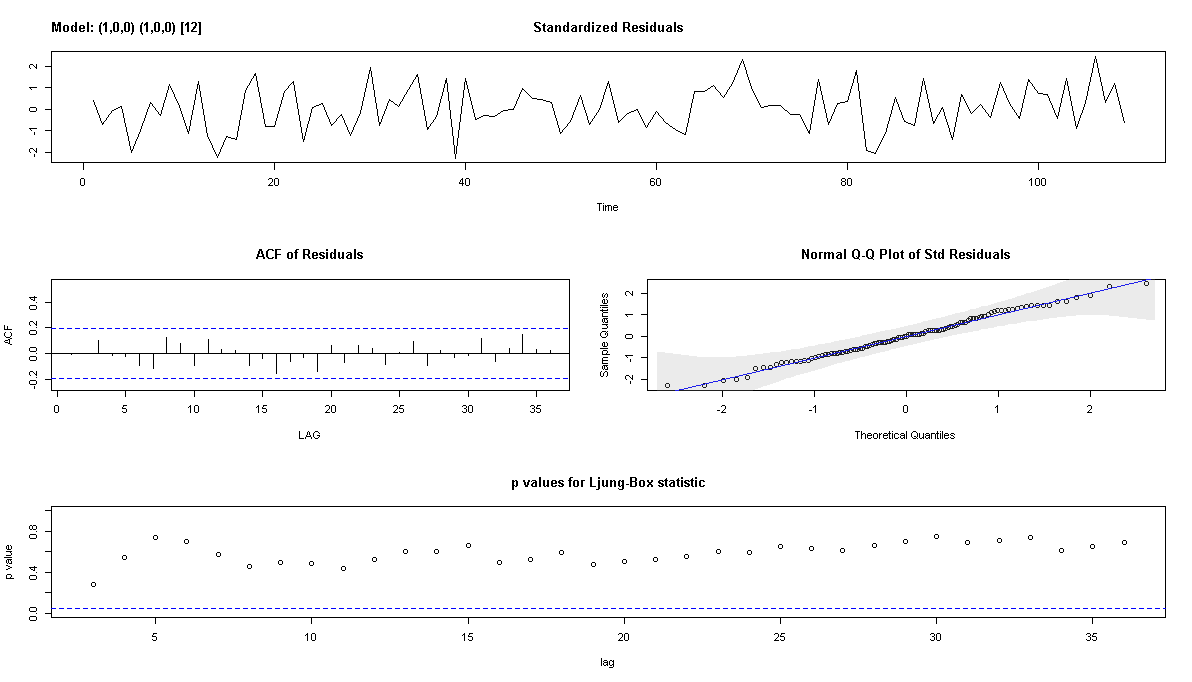
Estimate SE t.value p.value

ar1 0.1591 0.0954 1.6684 0.0982

sar1 0.5739 0.0777 7.3873 0.0000

xmean 1812.1008 12.1635 148.9785 0.0000

The diagnostic output:



The diagnostics indicate that the model is appropriate because:

* The time series plot of the standardized residuals mostly indicates that there’s no visible trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations – a good result.
* The Q-Q plot is a normal probability plot – The plot looks almost linear and the assumption of normally distributed residuals holds.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. All p-values are above the dashed blue line - That’s a good result.

Note that the ar1 term is not significant at alpha=5% but is significance at alpha = 10%

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.

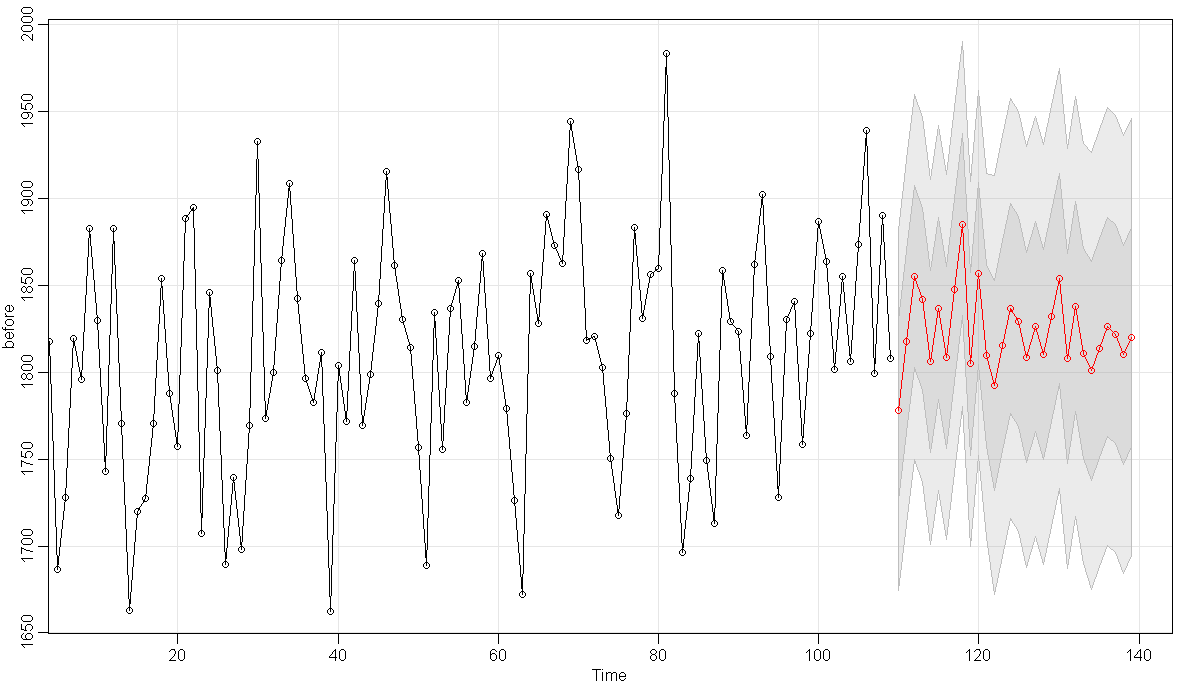
Something like this should work:

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

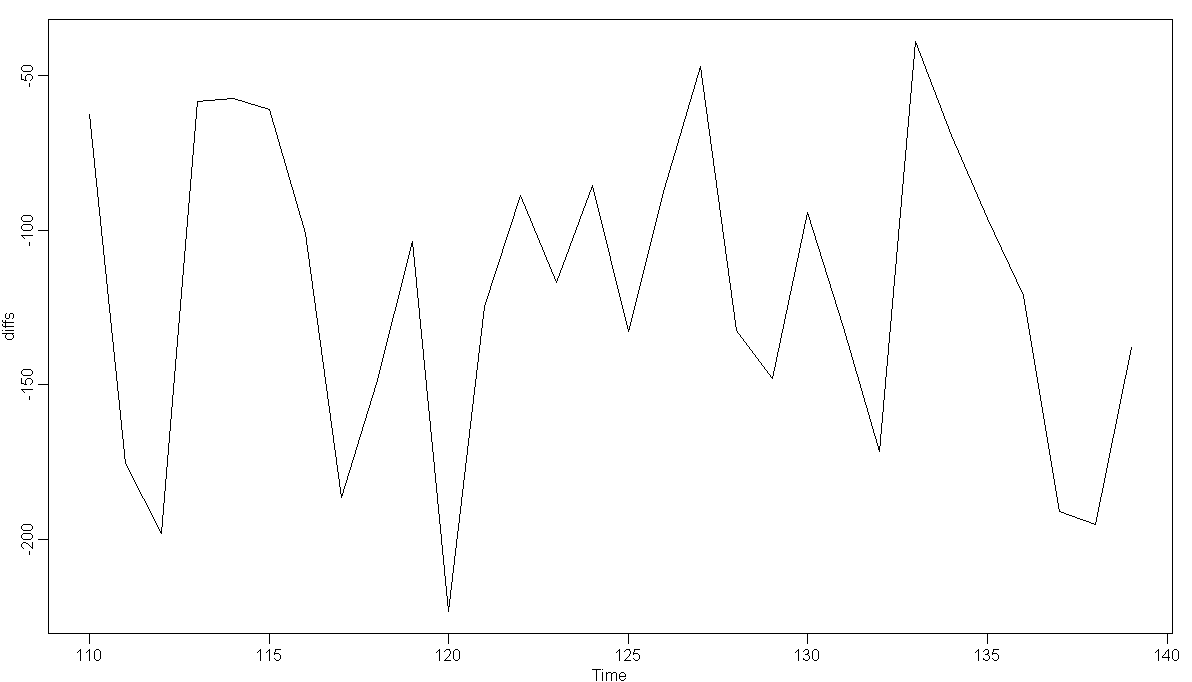
diffs = after - afterpred$pred

plot(diffs)

The predictions for the after period are plotted:



The plot of the difference is:



We see here that the diffs are negative for the entire after duration (and therefore have a negative mean) clearly indicating that the law has led to a generally lower number of fatalities and injuries.

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

The mean is: -119.6088

We again validate here that the mean diffs are negative clearly indicating that the law has led to a generally lower number of fatalities and injuries