**Stat 510 Week 8 Homework**

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

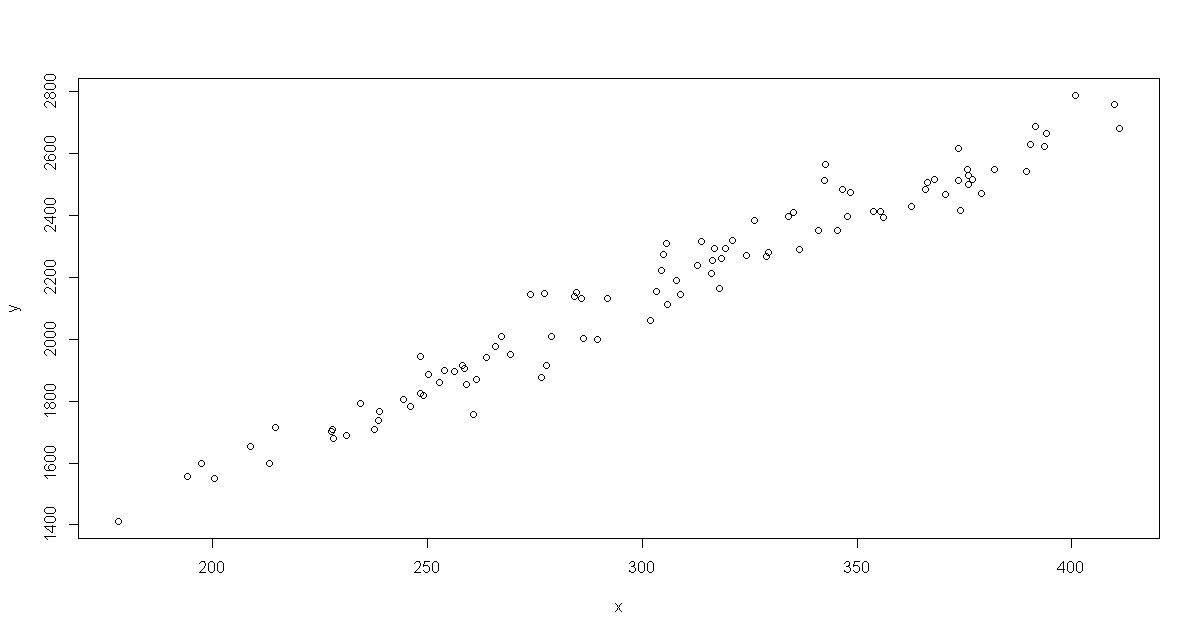
1. Use the xhw8prob1.dat and yhw8prob1.dat datasets from the Week 8 folder. The tasks will be the same that we carried out in Lesson 8.1, so use the code provided there for guidance.

1. Plot the data. Do an ordinary simple linear regression relating  to. Give the estimated intercept and slope along with their standard errors as the answer to this part.

The plots are:

|  |  |
| --- | --- |
| X | Y |
|  |  |

X-Y plot:



Estimate the usual regression model. Results from R are:

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-159.91 -49.90 -11.42 40.54 190.18

Coefficients:

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

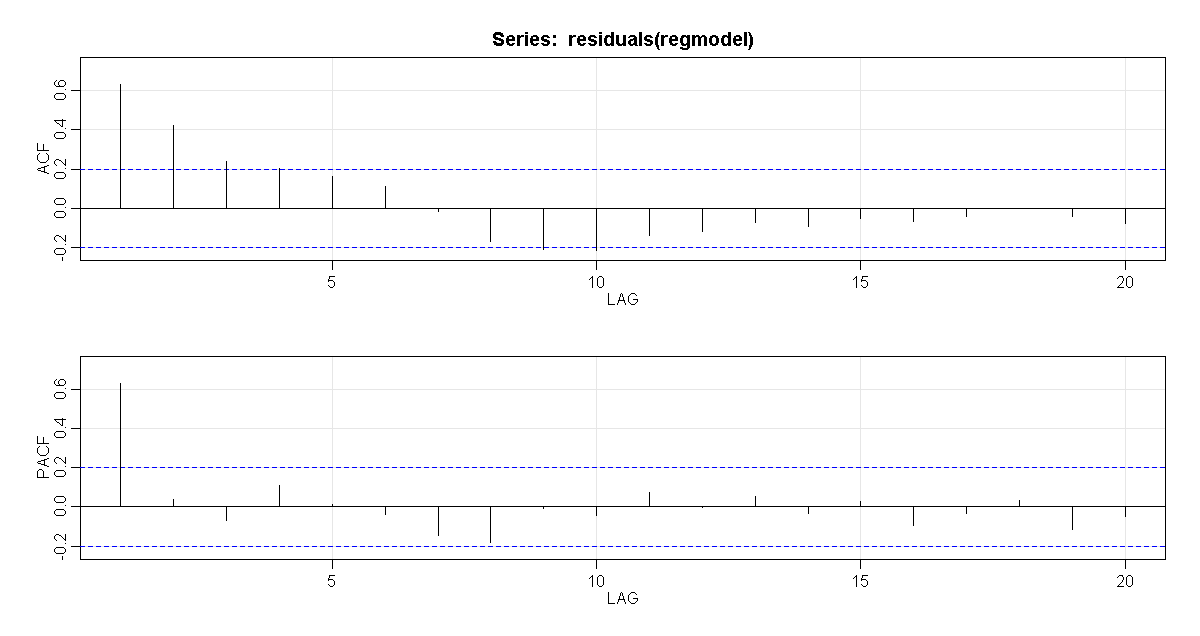
(Intercept) 463.492 36.289 12.77 <2e-16 \*\*\*

x 5.575 0.117 47.64 <2e-16 \*\*\*

---

1. Examine the ACF and PACF of the residuals from the regression in part A. For this part, show the graphs and write a brief explanation of what ARIMA model may be appropriate for the residuals.

The plots are:



These plots indicate:

* + PACF spikes at 1 and after that cuts off and ACF trails off. It looks like the errors have an AR(1) structure.

1. Let’s go through the Cochrane-Orcutt procedure. Estimate the coefficient(s) of the ARIMA model identified in part B. Remember – use the residuals from part A as the variable. Look at Lesson 8.1 for guidance. Give the estimated coefficient(s) as the answer to this part AND discuss whether the model is a suitable fit to the residuals.

**Option 1 : Using manual procedure, the output is:**

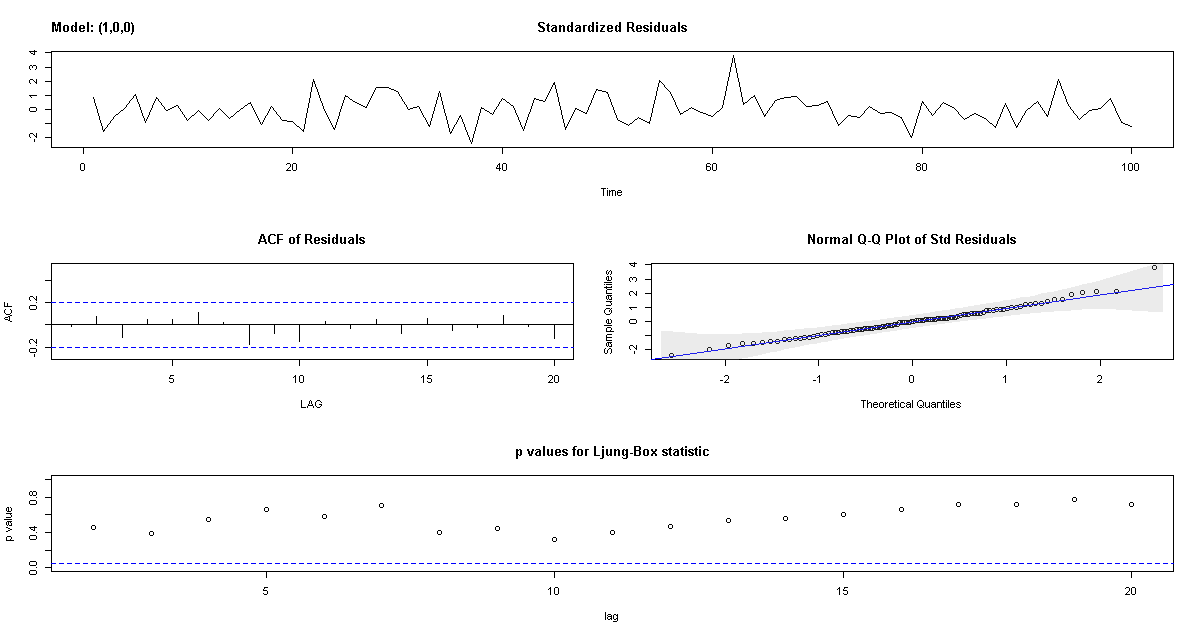
Estimate the AR coefficients

$ttable

Estimate SE t.value p.value

ar1 0.6382 0.0768 8.3111 0

The diagnostics are:



The diagnostics indicate that the model is appropriate because:

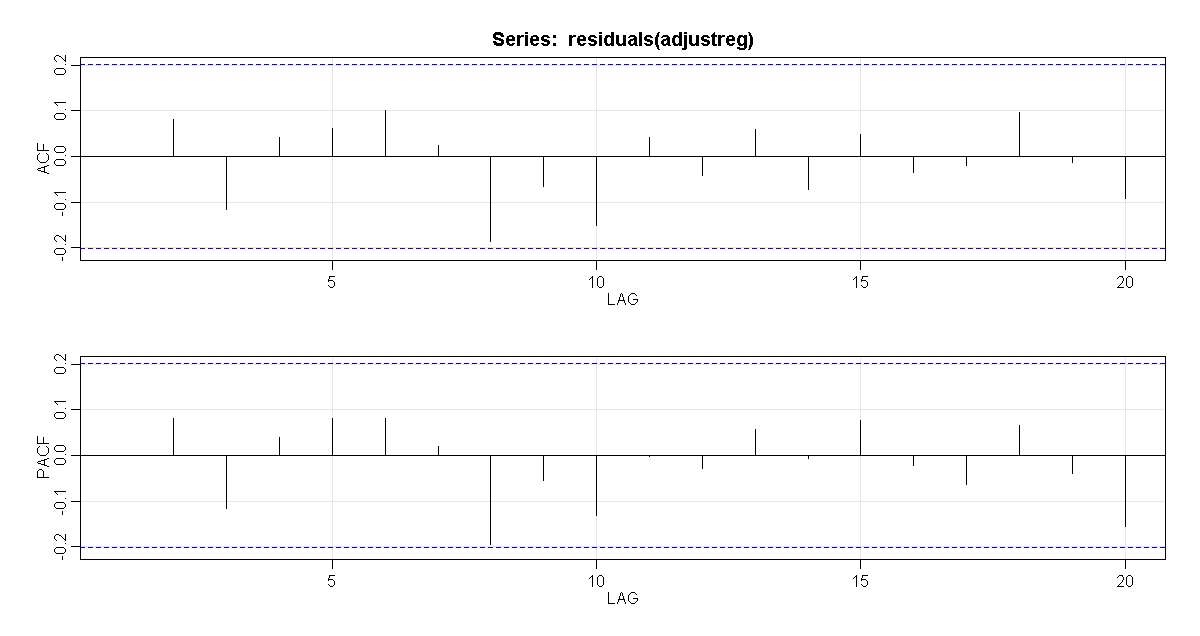
* The time series plot of the standardized residuals mostly indicates that there’s no visible trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations – a good result.
* The Q-Q plot is a normal probability 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.

Calculate variables to use in the adjustment regression:

xt∗ = xt – 0.6382 \* xt−1

yt∗ = yt – 0.6382 \* yt−1

Let’s plot the ACF and PACF (we already have the ACF above though) for the residuals of the adjusted model, we get:



**Option 2 : Using Cochrane-Orcutt procedure, the output is:**

Cochrane-orcutt estimation for first order autocorrelation

Call:

lm(formula = y ~ x)

number of interaction: 5

rho 0.643481

Durbin-Watson statistic

(original): 0.71629 , p-value: 2.255e-11

(transformed): 1.97006 , p-value: 4.304e-01

coefficients:

(Intercept) x

483.617670 5.500303

1. Carry out the necessary “adjustment regression.” Give the adjusted intercept and slope of the regression relationship between and . Also give the standard errors of these estimates. See Step 5 of the first example in Lesson 8.1 for guidance.

Use ordinary regression to estimate the model yt∗ = β0∗ + β1xt∗ + wt. The results are:

Coefficients:

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

(Intercept) 174.801 21.445 8.151 1.28e-12 \*\*\*

xnew 5.502 0.186 29.580 < 2e-16 \*\*\*

---

The slope estimate (5.502) and its standard error (0.186) are the adjusted estimates for the original model.

The adjusted estimate of the intercept of the original model is 174.801 / (1-0.6382) = 483.1426. The estimated standard error of the intercept is 21.445 / (1-0.6382) = 59.273.

1. Use the Cochrane.orcutt command to iterate this procedure and obtain final estimates.

**Using Cochrane-Orcutt procedure, the output is:**

Cochrane-orcutt estimation for first order autocorrelation

Call:

lm(formula = y ~ x)

number of interaction: 5

rho 0.643481

Durbin-Watson statistic

(original): 0.71629 , p-value: 2.255e-11

(transformed): 1.97006 , p-value: 4.304e-01

coefficients:

(Intercept) x

483.617670 5.500303

Thus our estimated relationship between yt and xt is

yt = 483.6177 + 5.5003 \* xt

The errors have the estimated relationship et=0.6435 \* et−1 + wt

1. Using the arima command with the xreg option, estimate the coefficient(s) of the regression model and the ARIMA model for the residuals identified in part (B) with maximum likelihood. Give the estimated coefficient(s) and their test statistics as the answer to this part.

First lets estimate the usual regression model that includes the trend: 

Results from R are:

$ttable

Estimate SE t.value p.value

ar1 0.6406 0.0766 8.3588 0.000

intercept 501.9920 61.1294 8.2120 0.000

trend 0.1766 0.6846 0.2579 0.797

x 5.4182 0.2620 20.6836 0.000

Because trend is not significant, we will drop it from the model:

$ttable

Estimate SE t.value p.value

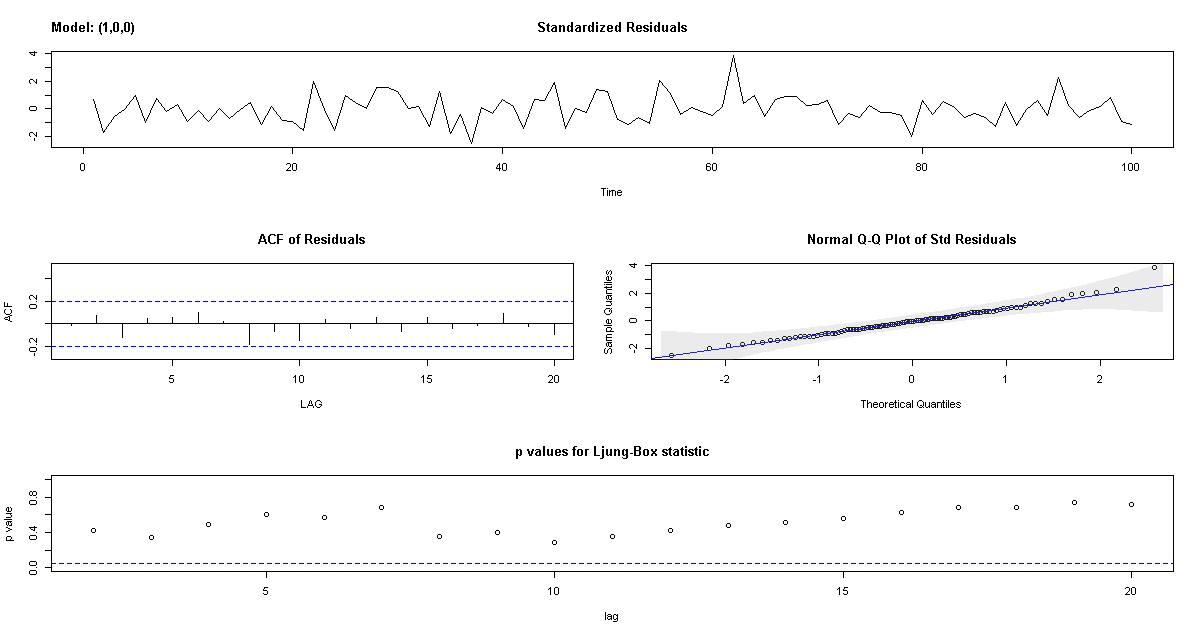
ar1 0.6420 0.0768 8.3628 0

intercept 495.9328 56.7209 8.7434 0

xreg 5.4674 0.1802 30.3344 0

1. Discuss whether the model in part (F) is a suitable fit. If not, repeat step (F) with an amended ARIMA model and refit using the arima command with the xreg option until the fit is suitable. You need only provide output for your final model.

Lets look at the diagnostics of the AR(1) model without trend:



The diagnostics indicate:

* 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.

We also see the coefficients:

$ttable

Estimate SE t.value p.value

ar1 0.6420 0.0768 8.3628 0

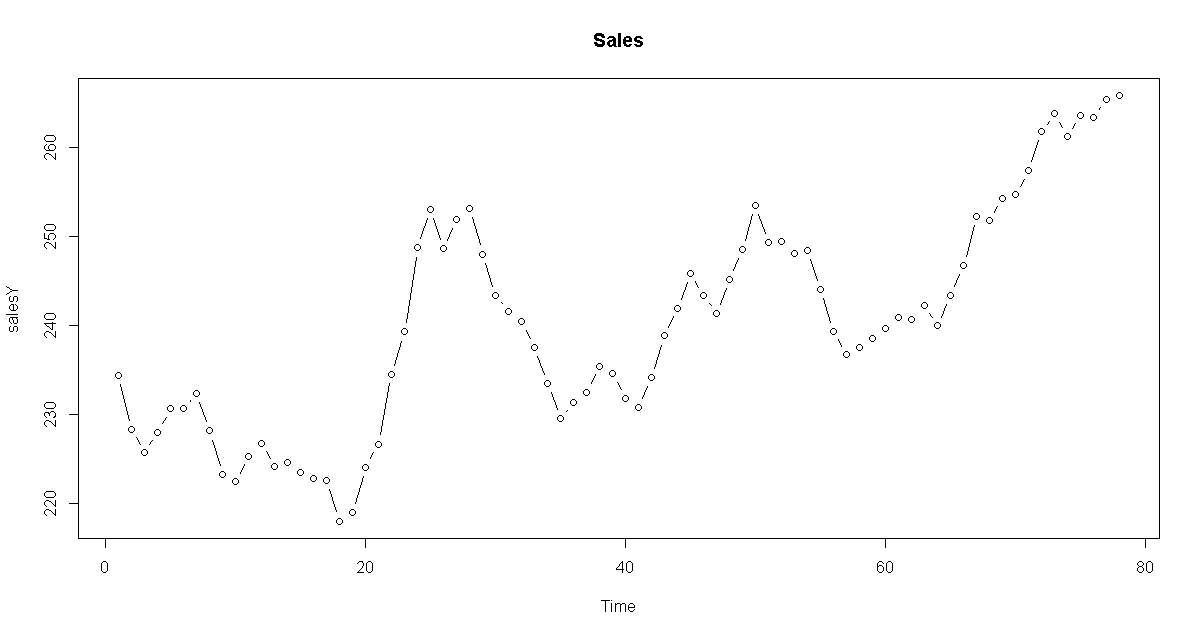
intercept 495.9328 56.7209 8.7434 0

xreg 5.4674 0.1802 30.3344 0

We see that the estimated coefficients are significant. This is a good model.

2. Use the sales8.dat and lead8.dat datasets from the Week 8 folder. Each file contains 78 consecutive monthly values. The variable in the sales8.dat file is monthly sales of a product for 78 months. The variable in the lead8.dat file is a company indicator (predictor) of future sales measured in the same 78 months. For this problem, = sales and  = leading indicator. Use Lesson 8.2 for guidance on parts C to G.

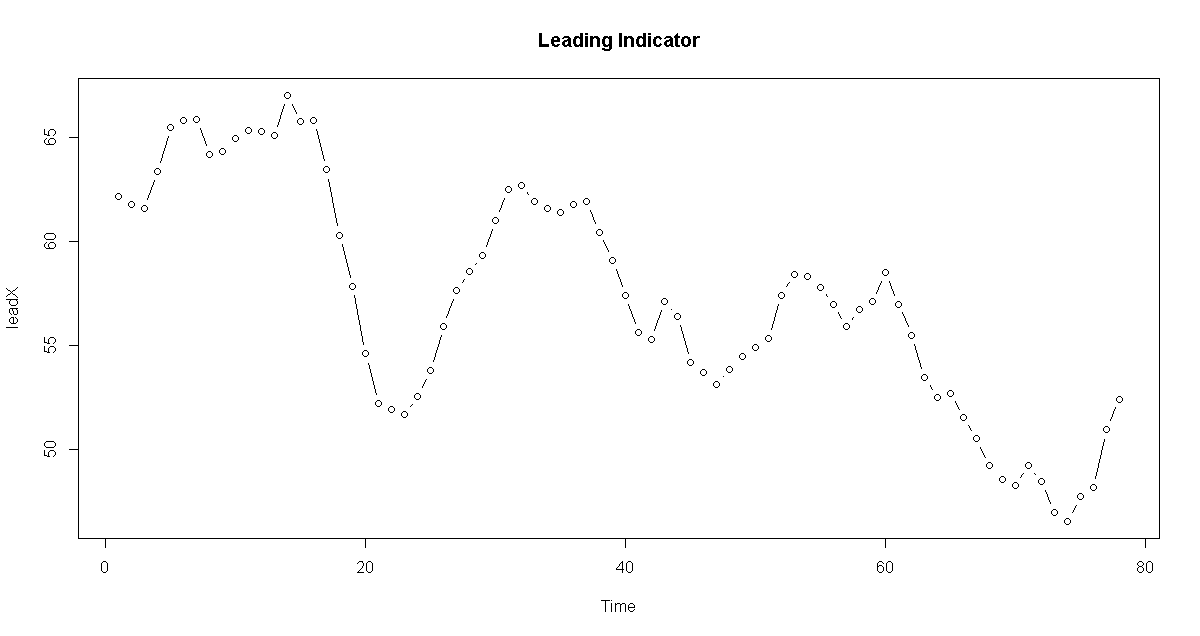
1. Create a time series plot of the sales variable. Describe any noteworthy features.



Some features of the plot:

* + There is some upward trend.
  + There doesn’t appear to be seasonality but we will need decomposition to be sure.
  + There are no obvious outliers except the points between 20 and 30 have a dramatic increase. Interestingly this dramatic increase is closely aligned with the dramatic drop in the leading indicator shown in the next part.
  + The variance appears to be constant though it seems difficult to judge with conviction.

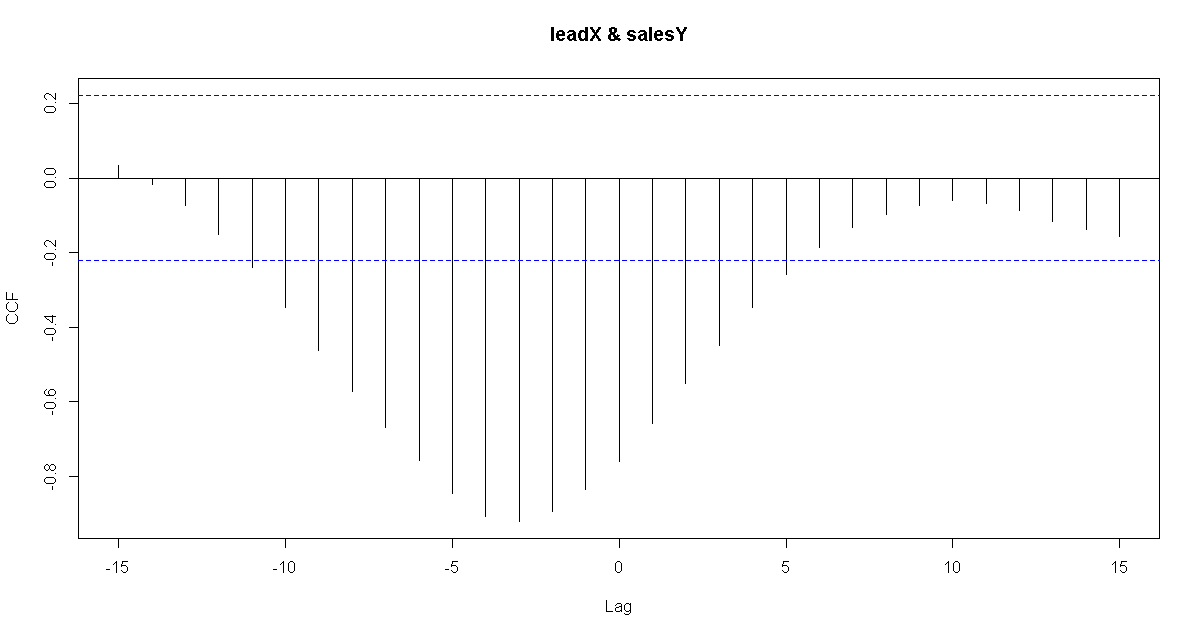
1. Create a time series plot of the leading indicator variable. Describe any noteworthy features.



Some features of the plot:

* + There is some downward trend.
  + There doesn’t appear to be seasonality but we will need decomposition to be sure.
  + There are no obvious outliers except the points between 17 and 25 have a dramatic drop. Interestingly this dramatic drop is closely aligned with the dramatic increase in the sales shown above.
  + The variance appears to be constant though it seems difficult to judge with conviction.

1. Determine the CCF for the relationship between  = leading indicator and = sales indicator. Give the plot as the answer for this part. Notice that it’s a little difficult to identify the main peak(s) because quite a few correlations are relatively large.



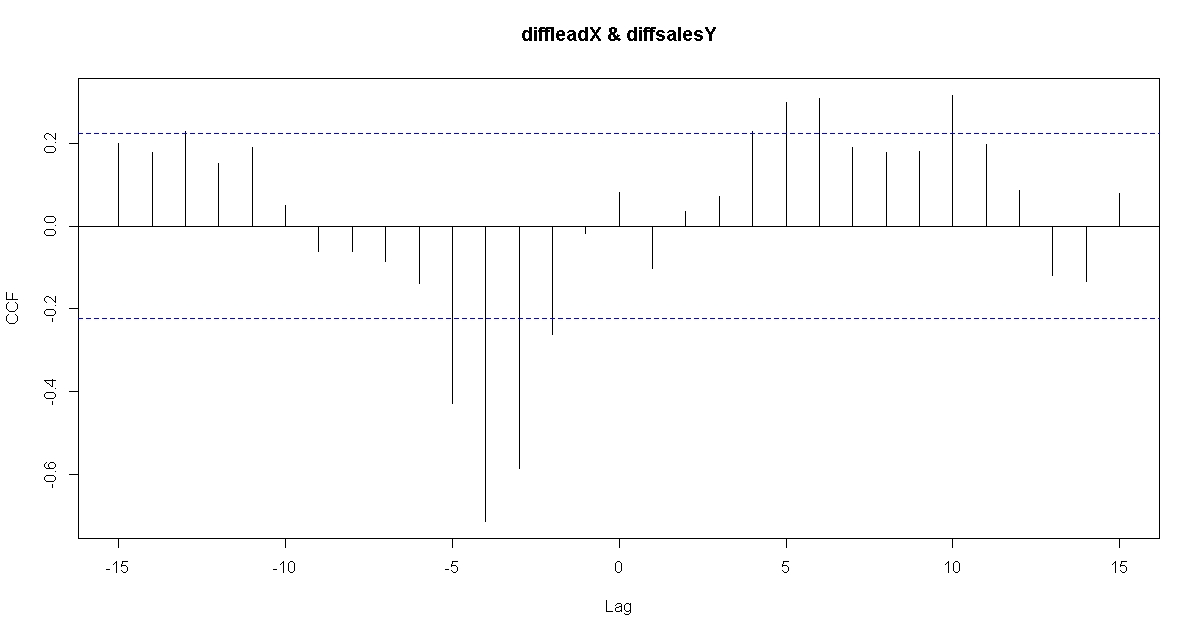
1. The plots in A and B suggest trends, so create variables that are the first differences for the sales variable and the leading indicator variable. Commands will be something like:

diffsales = diff(sales,1)

difflead = diff(lead,1)

Exact commands will depend upon what you called the variables when you read them in.

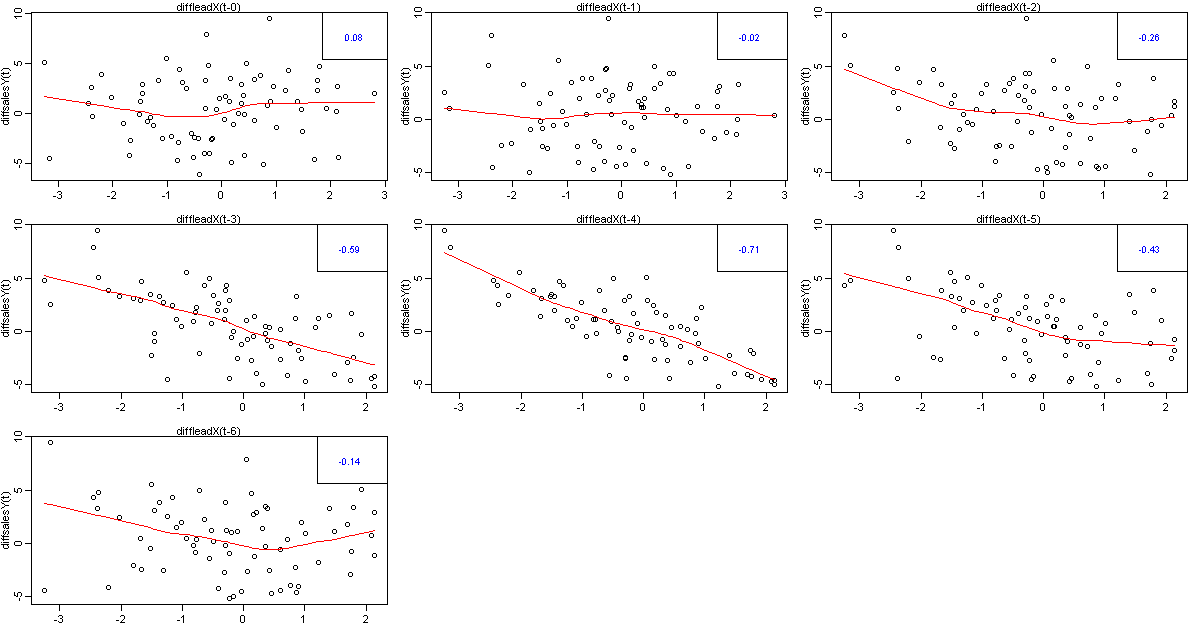
Next, determine the CCF for the relationship between the two sets of differences. The differenced leading indicator variable is the x-variable. Give the plot AND write a brief interpretation.



It is clear from the picture:

* + The most dominant cross correlations occur somewhere between h =−5 and about h = −3.
  + Note that the correlations in this region are negative, indicating that an above average value of first difference of leading indicator is likely to lead to a below average value of first difference of sales about 4 months later. And, a below average of first difference of leading indicator is likely to lead to above average value of first difference of sales about 4 months later.

1. Use the lag2.plot command to create scatter plots of the differenced sales data (as y) and the differenced leading indicator data (as x) for 6 lags. Describe noteworthy features. See Lesson 8.2 for guidance.



The noteworthy features and interpretation are:

* + Note that the correlations are significant and negative for -2, -3, -4 and -5, indicating that an above average value of first difference of leading indicator is likely to lead to a below average value of first difference of sales about 4 months later. And, a below average of first difference of leading indicator is likely to lead to above average value of first difference of sales about 4 months later.

1. Carry out a regression with the original undifferenced sales variable as the y-variable and lags 3 and 4 of the original undifferenced leading indicator variable as the predictor variables. Give the regression results and for each of the predictor variables, indicate whether the variable has a statistically significant relationship with the y-variable.

*Teaching Note*: In parts D and E we used differenced data to identify a possible relationship. This is a variation of pre-whitening, a technique that we’ll discuss further next week. Pre-whitening removes common elements in the two series and that allows us to see the relationship more clearly. Usually, the relationship seen in the pre-whitened data holds for the original data. Thus we’re using the undifferenced data in part F.

Call:

lm(formula = salesY ~ leadXlag3 + leadXlag4, data = alldata)

Residuals:

Min 1Q Median 3Q Max

-5.2823 -0.8735 0.1474 1.1128 3.6117

Coefficients:

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

(Intercept) 372.3614 2.1439 173.687 <2e-16 \*\*\*

leadXlag3 -0.3734 0.1542 -2.422 0.018 \*

leadXlag4 -1.9105 0.1570 -12.171 <2e-16 \*\*\*

We see from above that both lag3 and lag4 are significant

Lag3 significance: z = -0.3734 / 0.1542 = -2.4215 is significant since |z| > 1.96 (also p-value < 0.05)

Lag4 significance: z = -1.9105/ 0.1570 = -12.1688 is significant since |z| > 1.96 (also p-value < 0.05)

Before moving on, let’s also try the model where we have trend, lag3 and lag4 of leading indicator. We get:

Call:

lm(formula = salesY ~ trend + leadXlag3 + leadXlag4, data = alldata)

Residuals:

Min 1Q Median 3Q Max

-4.9642 -0.8977 0.1700 1.1256 3.5577

Coefficients:

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

(Intercept) 370.260128 3.876782 95.507 <2e-16 \*\*\*

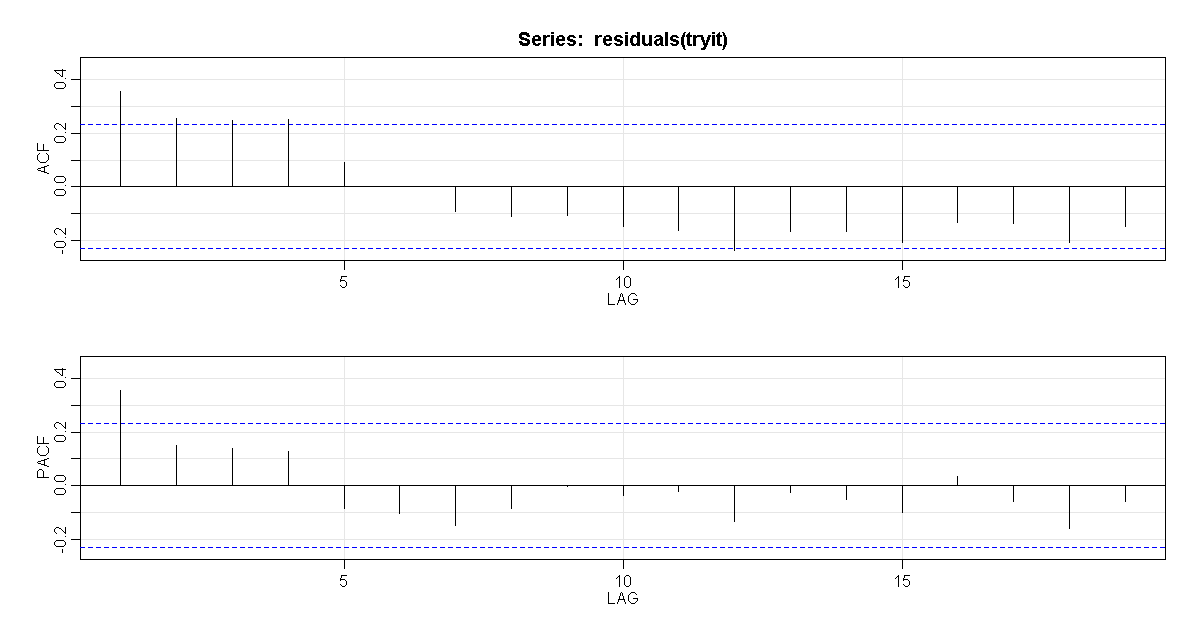
trend 0.009501 0.014579 0.652 0.5167

leadXlag3 -0.352569 0.158085 -2.230 0.0289 \*

leadXlag4 -1.901675 0.158189 -12.022 <2e-16 \*\*\*

We see here that the trend is not significant. On removing trend we get back to the same model as above with only lag3 and lag4.

1. For the residuals from the model in part F, determine the ACF and PACF (of the residuals). Discuss a possible model for the residuals.



We see here:

* + PACF cuts-off after lag1 and ACF is trailing off to 0 (though the trailing is not very clear). These suggest an AR(1) model.

1. Using the arima command with the xreg option, estimate the coefficient(s) of the regression model and the ARIMA model for the residuals identified in part G with maximum likelihood. Give the estimated coefficient(s) and their test statistics as the answer to this part. Because you cannot supply a data set to the arima command, you must access the columns of alldata as follows:

trend=time(alldata[,1])

adjreg = arima (alldata[,1], order = c(p,d,q), xreg = cbind(trend, alldata[,2], alldata[,3]))

*The brackets retrieve values from a matrix. The first entry corresponds to rows and the second to columns. Leaving the first entry blank tells R to retrieve all rows in the corresponding column. “Sales” is the first column in my alldata matrix. You may include more or fewer external regressors in the xreg option as you see necessary.*

Call:

arima(x = alldata[, 1], order = c(1, 0, 0), xreg = cbind(trend, alldata[, 2],

alldata[, 3]))

Coefficients:

ar1 intercept trend alldata[, 2] alldata[, 3]

0.3794 370.5980 0.0077 -0.5222 -1.7371

s.e. 0.1130 5.3404 0.0204 0.1800 0.1788

sigma^2 estimated as 2.332: log likelihood = -136.41, aic = 284.82

Let’s check what is significant:

ar1 significance: z = -0. 3794 / 0.1130 = -3.3575 is significant since |z| > 1.96 (also p-value < 0.05)

trend significance: z = 0.0077/ 0.0204 = 0.3775 is **not** significant since |z| < 1.96 (also p-value < 0.05)

lag3 significance: z = -0.5222/ 0.1800 = -2.9011 is significant since |z| > 1.96 (also p-value < 0.05)

lag4 significance: z = -1.7371/ 0.1788 = -9.7153 is significant since |z| > 1.96 (also p-value < 0.05)

We can also see this from sarima output:

$ttable

Estimate SE t.value p.value

ar1 0.3794 0.1130 3.3584 0.0013

intercept 370.5980 5.3404 69.3951 0.0000

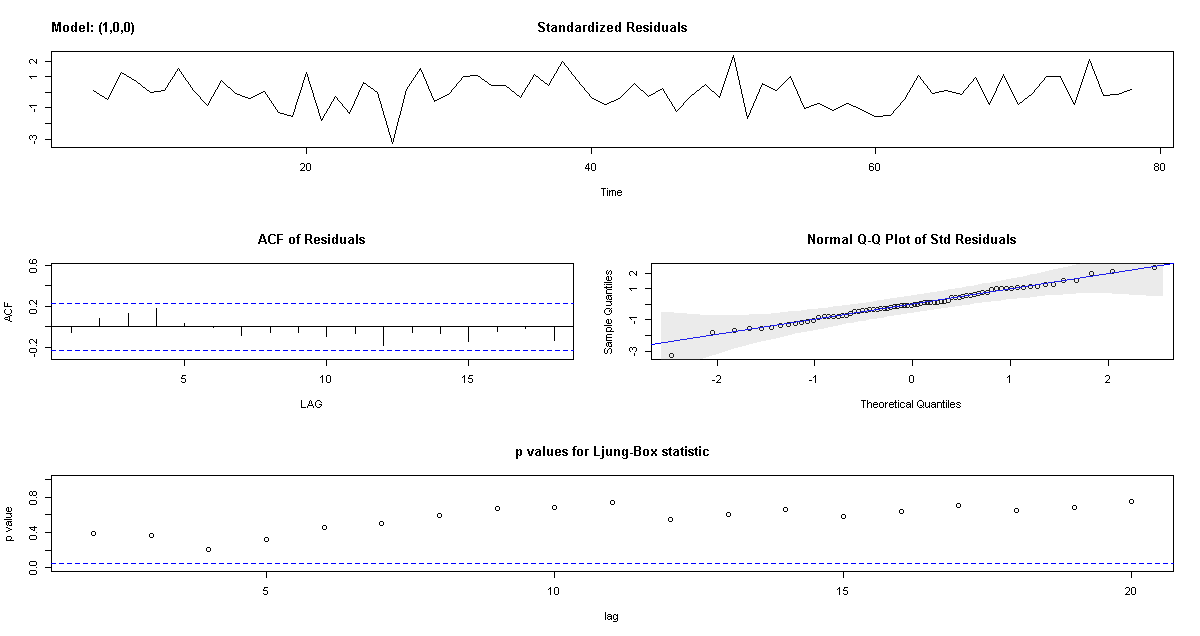
trend 0.0077 0.0204 0.3770 0.7073

alldata[, 2] -0.5222 0.1800 -2.9014 0.0050

alldata[, 3] -1.7371 0.1788 -9.7127 0.0000

1. Discuss whether the model in part H is a suitable fit (e.g. acf/pacf, Ljung-Box, QQ-plot). If not, repeat step H and refit using the arima command with the xreg option until the fit is suitable.

The diagnostics output is:



The diagnostics indicate:

* 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 in part H is not completely suitable because we see that the trend element is not significant. We do the same analysis after removing the trend:

sarima (alldata[,1], 1,0,0, xreg = cbind(alldata[,2], alldata[,3])) #AR(1) for residuals

We get the following:

$ttable

Estimate SE t.value p.value

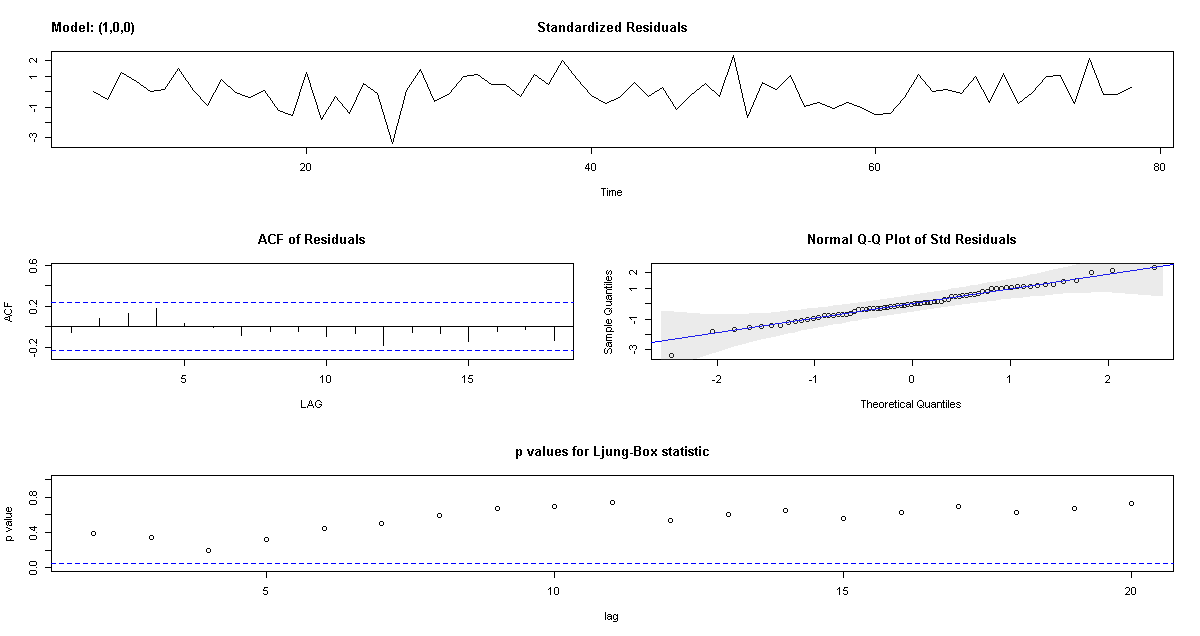
ar1 0.3835 0.1124 3.4119 0.0011

intercept 372.2613 3.0282 122.9317 0.0000

alldata[, 2] -0.5383 0.1755 -3.0673 0.0031

alldata[, 3] -1.7444 0.1774 -9.8312 0.0000

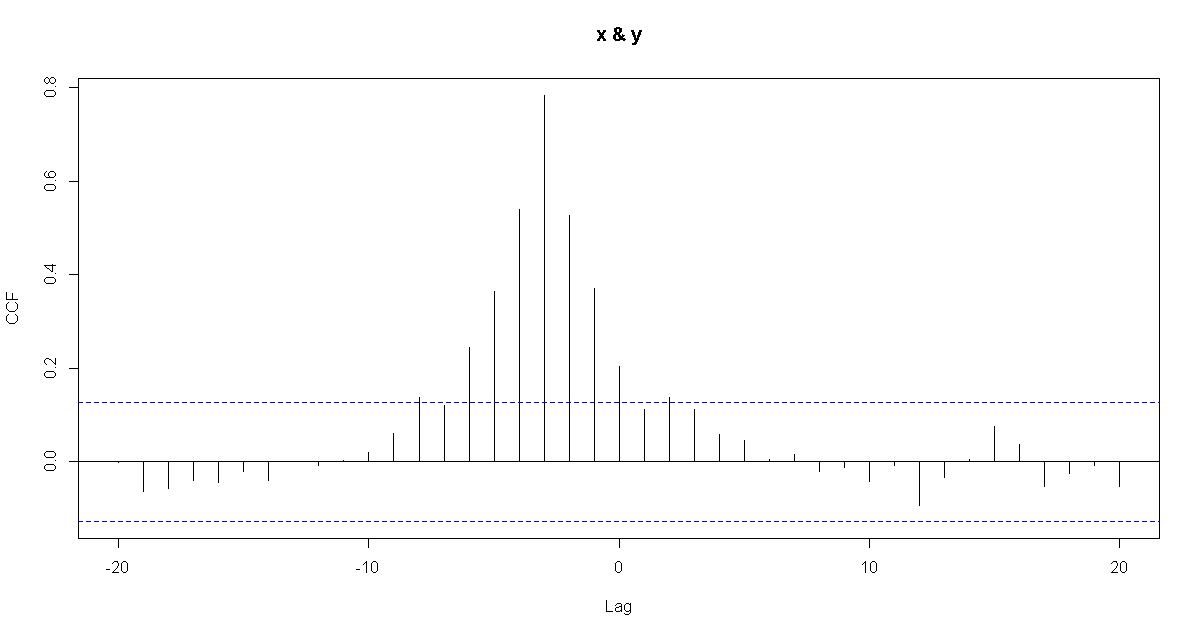
All the components are significant and the diagnostics are:



3. For this problem, use the files xvar-week8prob3.dat and yvar-week8prob3.dat from the Week 8 folder. The files contain n = 240 simulated observations for a relationship between two time series. The *x*-variable is in xvar-week8prob3.dat. The *y*-variable is in yvar-week8prob3. Use Lesson 8.2 of this week for guidance on this problem.

A. Determine the CCF for the relationship between x and y. Write a brief interpretation – for instance, where is the peak value? What’s happening in the neighborhood of the peak value?

The plot is:



The most dominant cross correlations occur somewhere between h = -8 and about h = 0. The peak is at h = -3

Autocorrelations of series ‘X’, by lag

-20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9

-0.001 -0.064 -0.057 -0.040 -0.044 -0.021 -0.040 0.001 -0.008 0.003 0.021 0.061

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

0.138 0.121 0.244 0.364 0.539 0.782 0.527 0.371 0.203 0.111 0.138 0.111

4 5 6 7 8 9 10 11 12 13 14 15

0.059 0.047 0.005 0.015 -0.022 -0.011 -0.043 -0.007 -0.093 -0.033 0.006 0.075

16 17 18 19 20

0.037 -0.053 -0.026 -0.007 -0.053

We also see that we have a peak at h=-3 in a CCF followed by a tapering pattern – this is an indicator that lag 1 and possibly lag 2 values of the y-variable may be helpful predictors.

B. Carry out the regression. That is use lag 1 of y and lag 3 of x to predict y. For each of the two predictors, indicate whether the variable has a statistically significant relationship with the y-variable.

We get the following output:

Coefficients:

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

(Intercept) 0.28178 0.88636 0.318 0.751

xlag3 1.56627 0.09788 16.001 <2e-16 \*\*\*

ylag1 0.05484 0.04738 1.157 0.248

---

We find here that xlag3 is significant while ylag1 is not. Also manually calculated as:

xlag3 significance: z = 1.56627 / 0.09788 = 16.0019 is significant since |z| > 1.96 (also p-value < 0.05)

ylag1 significance: z = 0.05484 / 0.04738 = 1.1575 is **not** significant since |z| < 1.96 (also p-value < 0.05)

C. Now leave the lag 1 of y variable out of the model and carry out a regression with only lag 3 of x as a predictor. For the residuals from this model, determine the ACF and PACF (of the residuals), and interpret them.

We do regression for the following:

Coefficients:

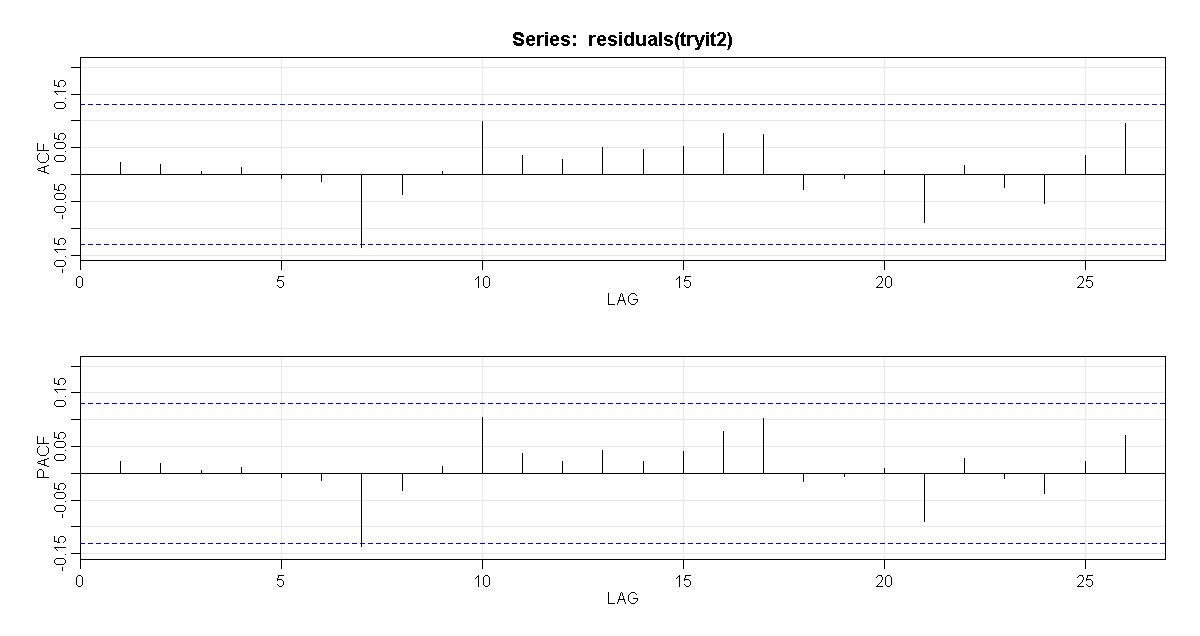
Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.60881 0.84073 0.724 0.47

xlag3 1.62614 0.08316 19.554 <2e-16 \*\*\*

---

The ACF / PACF for the residuals:



The ACF and PACF of the residuals looks pretty good and pointing to an absence of the time series structure of the residuals. There’s a barely significant residual autocorrelation at lag 7 which we may not want to worry about.