**Stat 510 Week 13 Homework**

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

1. Complete problem 5.1 on page 315 of the textbook. The data are in the file fracdiff.dat in the Week 13 folder.

Here are a few pointers and suggestions -

Use Lesson 13.1 for guidance.

First, you’ll have to install the fracdiff package. Either in the R menus, use Install Packages and select fracdiff or use the command install.packages("fracdiff"). After the installation, be sure to then use the command library(fracdiff).

For part c, the command is fracdiff(y, nar=1, nma = 0, M = 30) where y would be the name of the series that you’re analyzing.. The estimates will be given in the output so report them and their significance as the answer. You may compute the test statistic by hand; fracdiff(x,nar=1,nma=0,M=30)$stderror.dpq will provide the standard errors. Or you may use the summary command and R will print the test statistics and corresponding p-values:

mymodel <- fracdiff(x, nar=1, nma=0, M=30)

summary(mymodel)

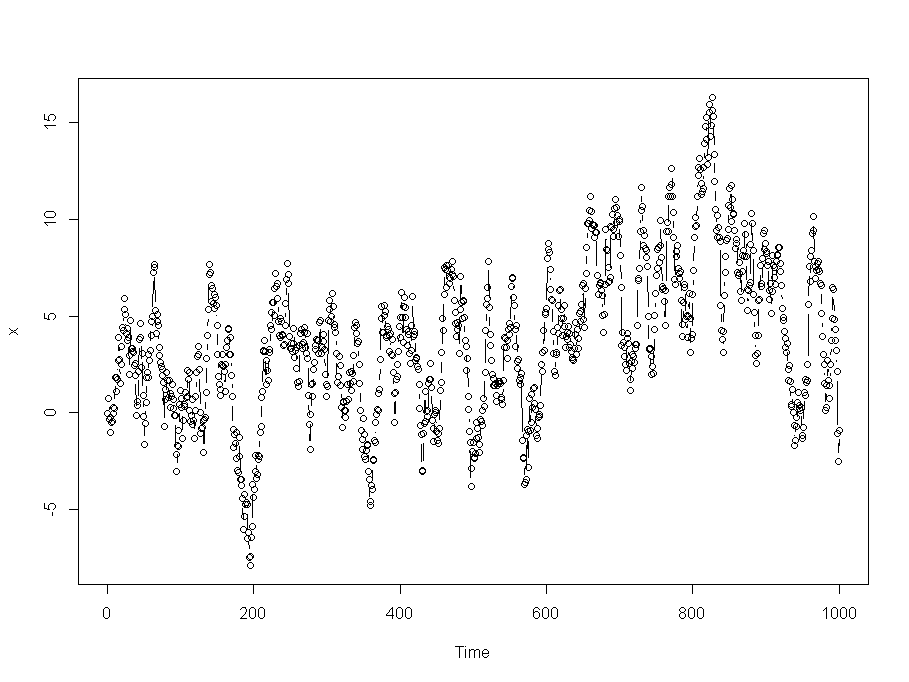
The M = 30 parameter of the command is a bit mysterious. It has to do with the setup of the maximum likelihood estimation of the parameters. It’s what the authors use so we’ll use it too.

In question 2 of the homework, we’ll look at how we would identify this particularly fractionally differenced model in practice.

For parts e and f, you’ll see an ACF and PACF that each have a very modest first order spike. This could be interpreted as either an MA(1) or an AR(1). Try both. For part (f), the “comment” should be a look at diagnostic statistics (from sarima, for example) with a statement about the suitability of the model.

5.1 The data set arf is 1000 simulated observations from an ARFIMA(1; 1; 0) model with ϕ = .75 and d = .4

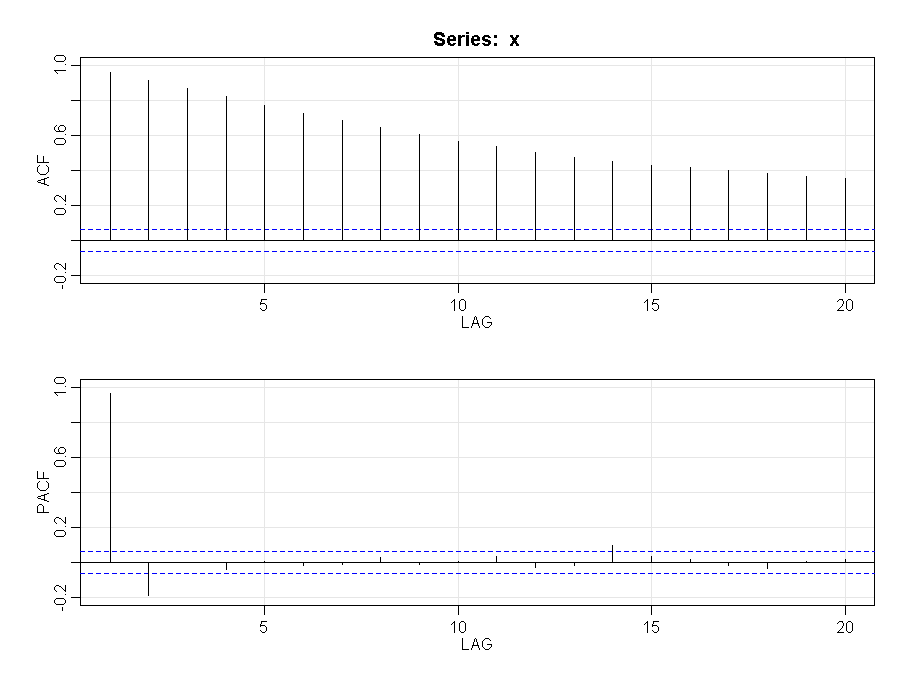
(a) Plot the data and comment.



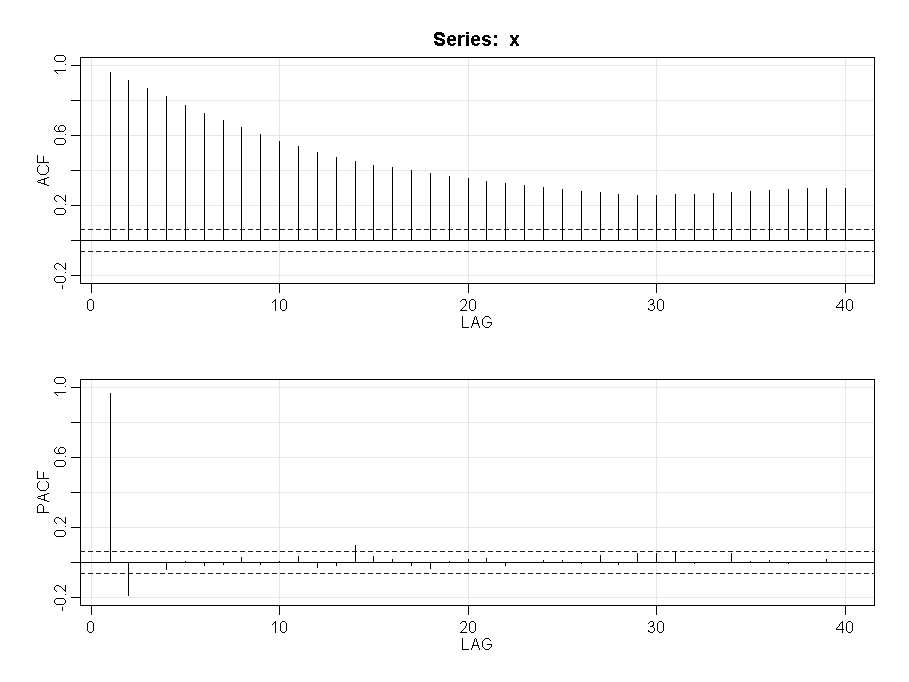
We can see:

* There is no consistent trend and the series wanders up and down.
* There are no obvious outliers.
* The variance appears to be more or less constant.

(b) Plot the ACF and PACF of the data and comment.



Check ACF and PACF for higher number of lags:



We see here:

* The ACF of the data shows a persistent pattern of moderately high values.
* The PACF spikes for lags 1 and 2 and then cuts off indicating AR(2)
  + AR(1) should also be evaluated because the 1st lag PACF is relatively very high and can be seen as cutting off after lag 1.

(c) Estimate the parameters and test for the signicance of the estimates ϕHat and dHat

Coefficients:

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

d 0.252571 0.009824 25.71 <2e-16 \*\*\*

ar 0.875834 0.016154 54.22 <2e-16 \*\*\*

The p-values above show that both ar1 and d are significant. Also with manual computation:

d coeff (z = 0.252571 / 0.009824 = 25.71) is significant since |z| > 1.96

AR(1) coeff (z = 0.875834 / 0.016154 = 54.22) is significant since |z| > 1.96

(d) Explain why, using the results of parts (a) and (b), it would seem reasonable to difference the data prior to the analysis. That is, if xt represents the data, explain why we might choose to fit an ARMA model to 

Often, data for which a first difference is successful will typically have a first lag autocorrelation quite close to 1. We see from the ACF plot above that the lag 1 autocorrelation is very close to 1.

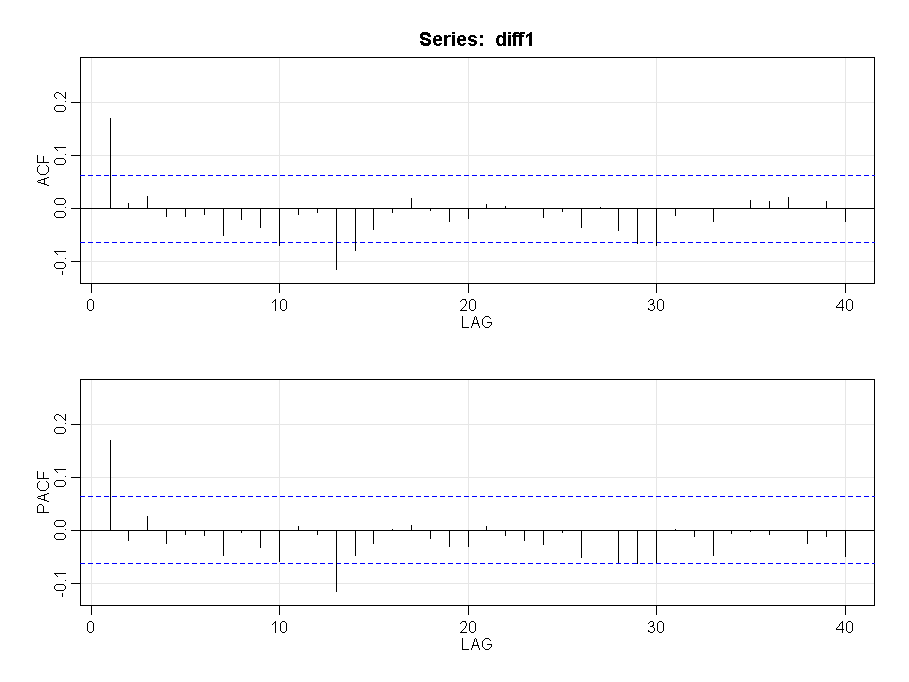
We also see that the PACF supports an AR(1) model. Finally the data plot from part a also indicates a slight trend and a first difference will be useful.

In this instance we also additionally have a significant difference coefficient d indicating applicability of the model:



(e) Plot the ACF and PACF of  and comment.

We get:



Here we see that both ACF and PACF have a modest first order spike followed by a cutoff. We can evaluate AR(1) and MA(1) model for the first difference.

(f) Fit an ARMA model to  and comment.

We evaluate:

sarima (x, 1, 1, 0)

sarima (x, 0, 1, 1)

|  |  |
| --- | --- |
| ARIMA (1, 1, 0) | ARIMA(0, 1, 1) |
| sigma^2 estimated as 0.9994  $ttable  Estimate SE t.value p.value  ar1 0.1698 0.0312 5.4424 0.0000  constant -0.0004 0.0381 -0.0104 0.9917  $AIC  [1] 1.003419  $AICc  [1] 1.005443  $BIC  [1] 0.01323404 | sigma^2 estimated as 0.9988  $ttable  Estimate SE t.value p.value  ma1 0.1741 0.0314 5.5523 0.0000  constant -0.0004 0.0371 -0.0112 0.9911  $AIC  [1] 1.00283  $AICc  [1] 1.004854  $BIC  [1] 0.01264518 |
|  |  |

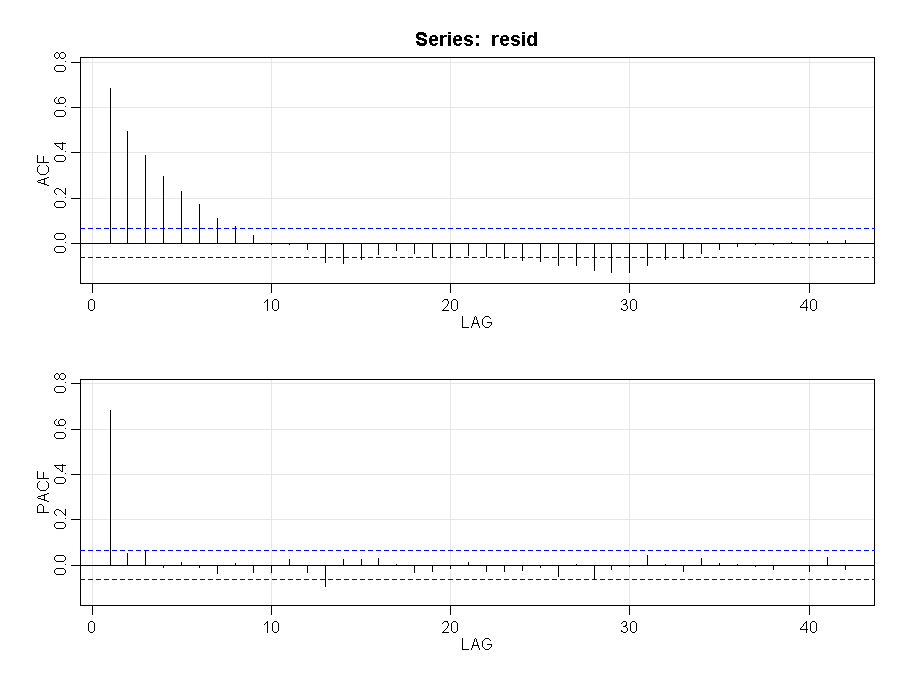
Both the models perform almost the same. The diagnostics of the residuals of both indicate:

* Residuals appear fine with constant variance
* Some ACF are almost significant – **not a desirable result**
* Normality assumption is reasonable
* Ljung statistic is **not desirable** for many lags

2. (View this question as a continuation of problem 1.) For the data in problem 1 (problem 5.1 of the book), use a fractionally differenced model with no ar term (as in Lesson 13.1). The format of the command would be fracdiff(y,nar=0,nma=0,M=30). Determine the residuals and then create the ACF of the residuals. As the answer, give the ACF and PACF and write a brief interpretation.

For programming, use the code in Lesson 13.1.

Teaching note: We were told that this was fractionally differenced ARIMA(1,1,0) so after a model with just the fractional difference we might then expect to see AR(1) residuals.



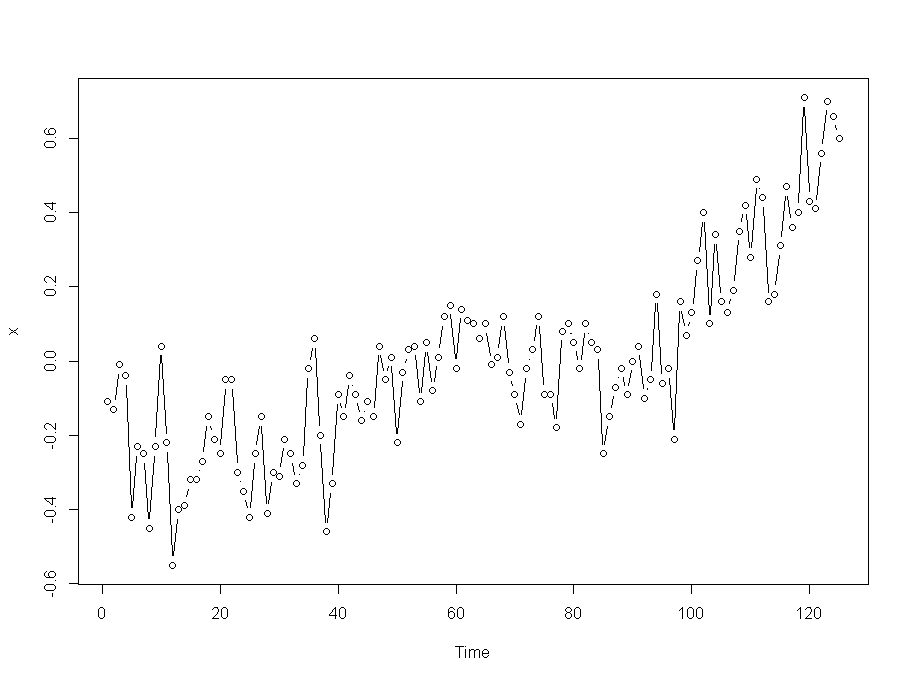
From the ACF / PACF:

* PACF spikes at lag1 and then cutsoff. This is what we expect for an AR(1) model
* Even the ACF indicates an AR(1) behavior in that it trails off gradually

Therefore the residuals indicate an AR(1) model

3. For this problem, use the dataset globaltmp2.dat from the Week 13 folder. Download it to the directory you’ve set up for course datasets. The data are annual global temperature deviations from 1880 to 2004.

a. Do a time series plot of the data. Give the plot.

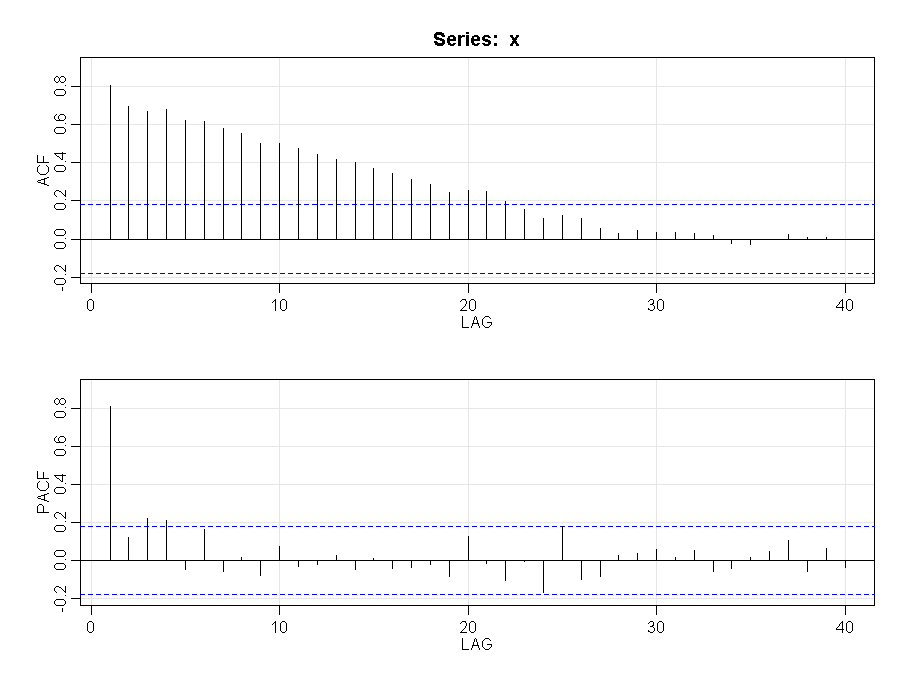


We see:

* Gradual upward trend
* Annual data so no seasonality
* No clear outliers
* Variance appears to be constant

b. Determine the ACF and PACF of the data and write a brief interpretation.

We have:



We see here:

* The ACF of the data shows a persistent pattern of moderately high values. Indicates that we can take difference and/or check fractional differences
* The PACF spikes for lags 1 and then cuts off indicating AR(1)

c. Estimate a fractionally differenced model with no AR or MA terms (just the fractional difference). What is the estimated value of d (the difference parameter)?

We get:

Coefficients:

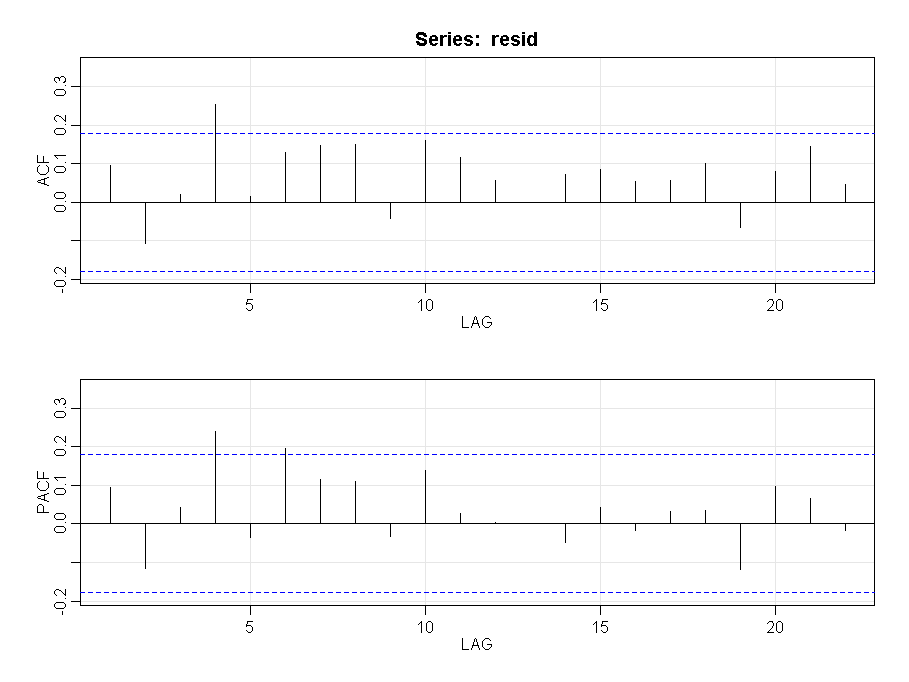
Estimate Std. Error z value Pr(>|z|)

d 4.946e-01 7.741e-07 638957 <2e-16 \*\*\*

d = 0.4946

d. Determine the residuals. Then get the ACF and PACF of the residuals. Give those plots and write a brief interpretation. Specifically, does the model in part c seem to be suitable?

The ACF and PACF are:



It is clear that both ACF and PACF don’t have any significant values. Therefore the model from part c is suitable.

4. Do problem 5.10 on page 316 of the textbook. The data are in the file sunspots.dat in the Week 13 folder. The problem in the book is unstructured, so we’ll make it structured. Do the following -

a. Use a threshold value of c = 65 and an AR(4) model in each of the two regions.

##Regression for values below the threshold

Coefficients:

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

(Intercept) 4.94410 0.69562 7.107 7.86e-12 \*\*\*

P1[, 1] 1.89372 0.05898 32.108 < 2e-16 \*\*\*

P1[, 2] -1.26002 0.11764 -10.711 < 2e-16 \*\*\*

P1[, 3] 0.53712 0.11653 4.609 5.86e-06 \*\*\*

P1[, 4] -0.26688 0.05880 -4.539 8.05e-06 \*\*\*

##Regression for values above the threshold

Coefficients:

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

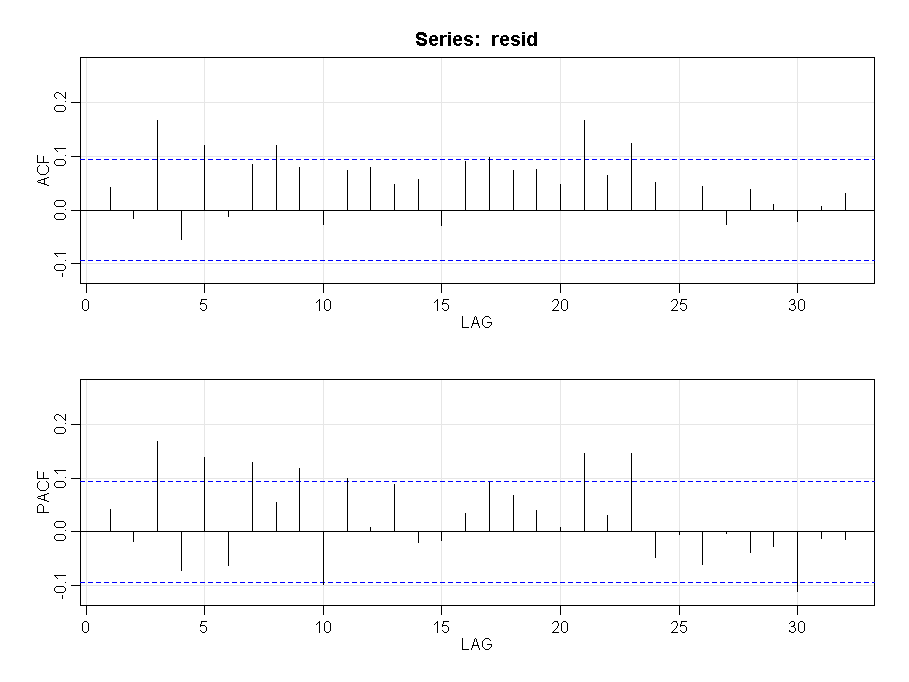
(Intercept) 15.02553 3.17799 4.728 5.89e-06 \*\*\*

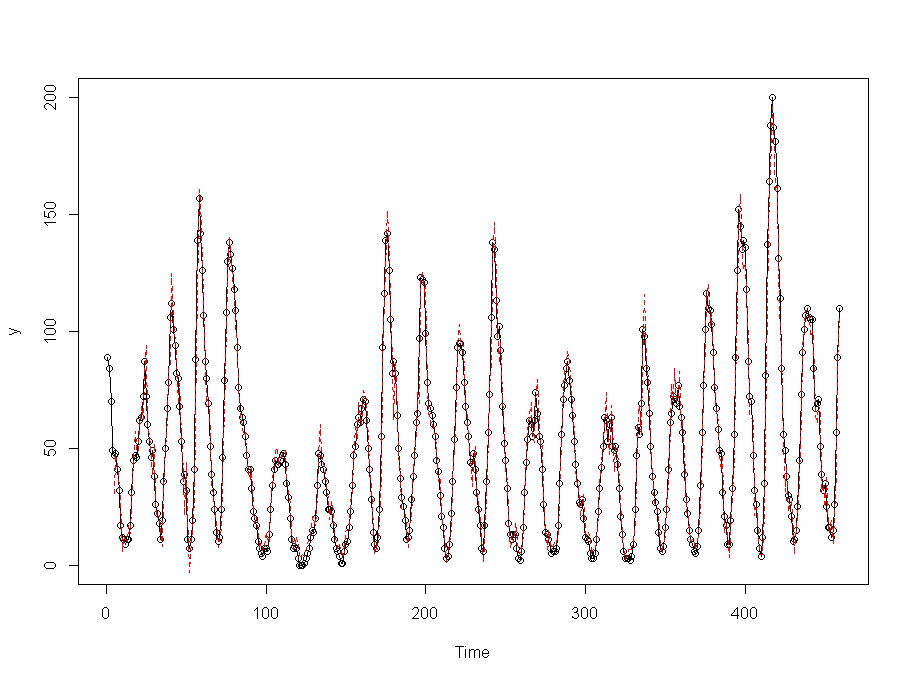
P2[, 1] 1.58245 0.08088 19.565 < 2e-16 \*\*\*

P2[, 2] -1.06810 0.14951 -7.144 6.10e-11 \*\*\*

P2[, 3] 0.65518 0.15061 4.350 2.75e-05 \*\*\*

P2[, 4] -0.37317 0.07613 -4.902 2.82e-06 \*\*\*





b. Use a threshold value of c = 100 and an AR(4) model in each of the two regions.

##Regression for values below the threshold

Coefficients:

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

(Intercept) 4.58915 0.67292 6.820 3.44e-11 \*\*\*

P1[, 1] 1.86373 0.04933 37.779 < 2e-16 \*\*\*

P1[, 2] -1.23749 0.09845 -12.569 < 2e-16 \*\*\*

P1[, 3] 0.61762 0.10055 6.143 1.99e-09 \*\*\*

P1[, 4] -0.31797 0.05182 -6.136 2.07e-09 \*\*\*

##Regression for values above the threshold

Coefficients:

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

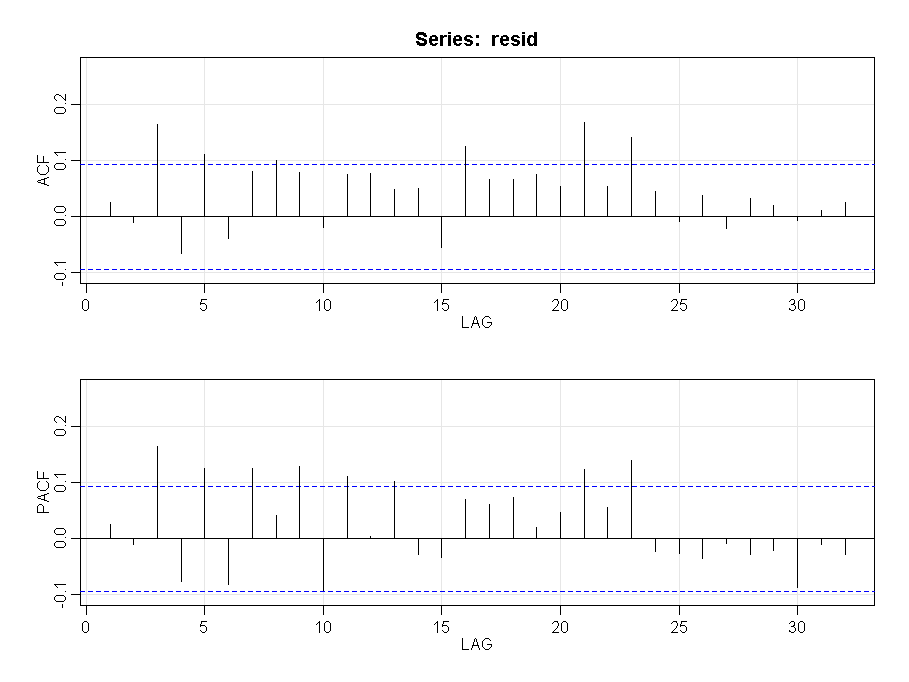
(Intercept) 4.4760 6.7746 0.661 0.511775

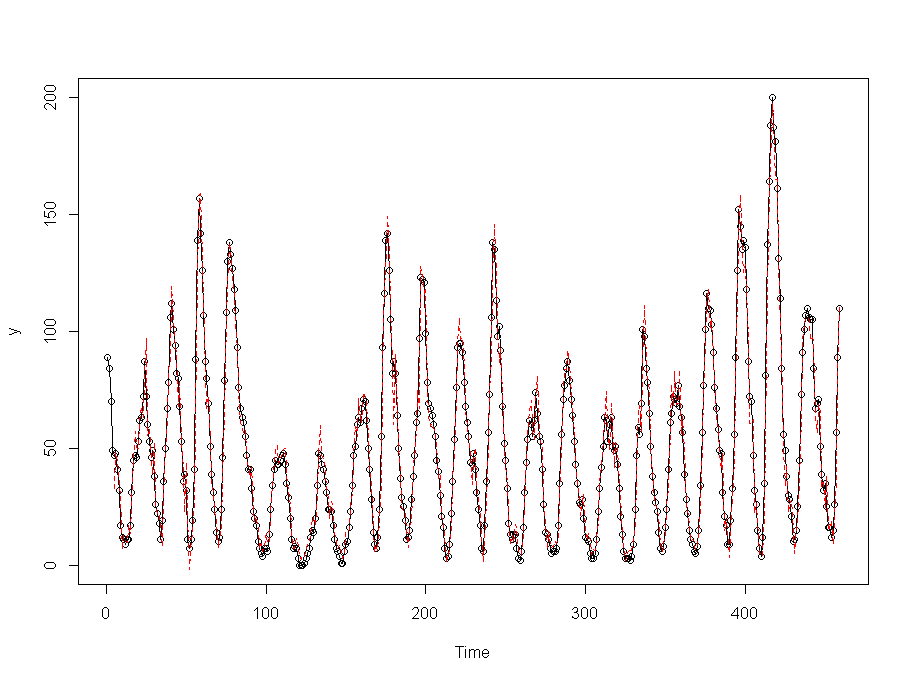
P2[, 1] 1.5455 0.1258 12.280 < 2e-16 \*\*\*

P2[, 2] -0.8479 0.2252 -3.765 0.000431 \*\*\*

P2[, 3] 0.4791 0.2159 2.219 0.030931 \*

P2[, 4] -0.3000 0.1031 -2.908 0.005367 \*\*





For each “attempt” give the regression output, the ACF and PACF of the residuals and a plot showing the actual and predicted values. Comment on the suitability of each model.

Don’t difference the data. Use Lesson 13.2 and the code in those notes for guidance.

c. Comment on which of the two threshold values seems to be better. (Compare the results of parts a and b).

From the above results we do get comparable results. Let’s compare the values

|  |  |
| --- | --- |
| c = 65 | c = 100 |
| ##Regression for values below the threshold  Residual standard error: 6.168 on 317 degrees of freedom  Multiple R-squared: 0.915,  Adjusted R-squared: 0.9139  MSE: 37.45725  ##Regression for values above the threshold  Residual standard error: 9.477 on 128 degrees of freedom  Multiple R-squared: 0.9146,  Adjusted R-squared: 0.912  MSE: 86.43731 | ##Regression for values below the threshold  Residual standard error: 7.042 on 394 degrees of freedom  Multiple R-squared: 0.9389,  Adjusted R-squared: 0.9383  MSE: 48.97506  ##Regression for values below the threshold  Residual standard error: 8.915 on 51 degrees of freedom  Multiple R-squared: 0.904,  Adjusted R-squared: 0.8965  MSE: 72.38612 |

Most adjusted R-squared values are comparable except the regression for values above c=65 threshold is higher than c=100

The total MSE is comparable.

Teaching comment: An AR(5) may be better than an AR(4) but we’ll take it easy on you so that you can (almost) use the code in Lesson 13.2 as is.