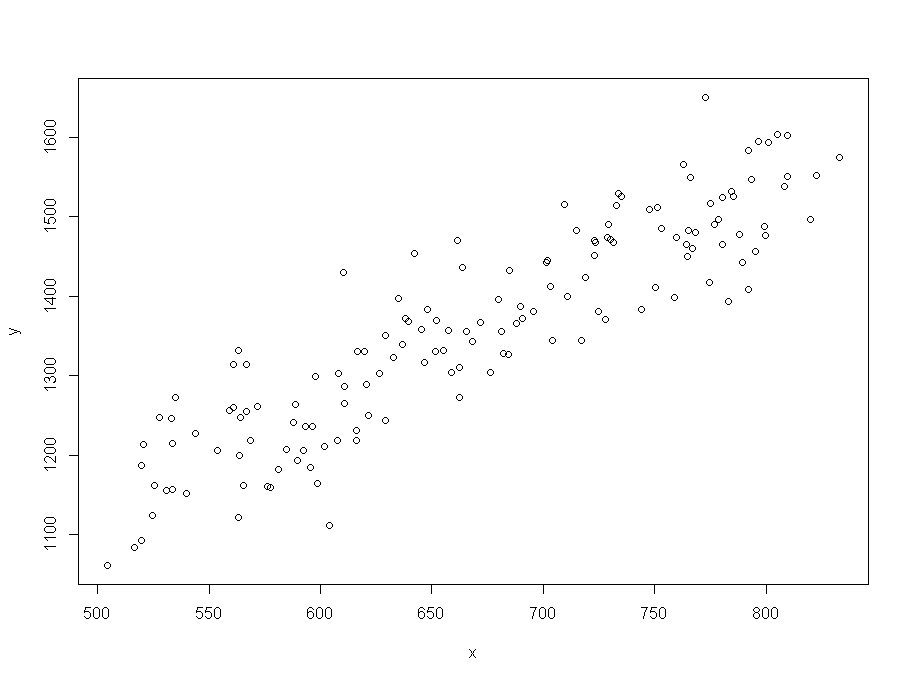
1. Use the e2x.dat and e2y.dat datasets from the Datasets folder. In R, the dataset could be read using something like x=scan("e2x.dat") and y=scan("e2y.dat"). The data are 150 simulated values in sequential order.

(a) Do an ordinary simple linear regression relating e2y to e2x. Give the estimated intercept and slope along with their standard errors as the answer to this part.

The plot for the data:



We have the linear regression model:

Coefficients:

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

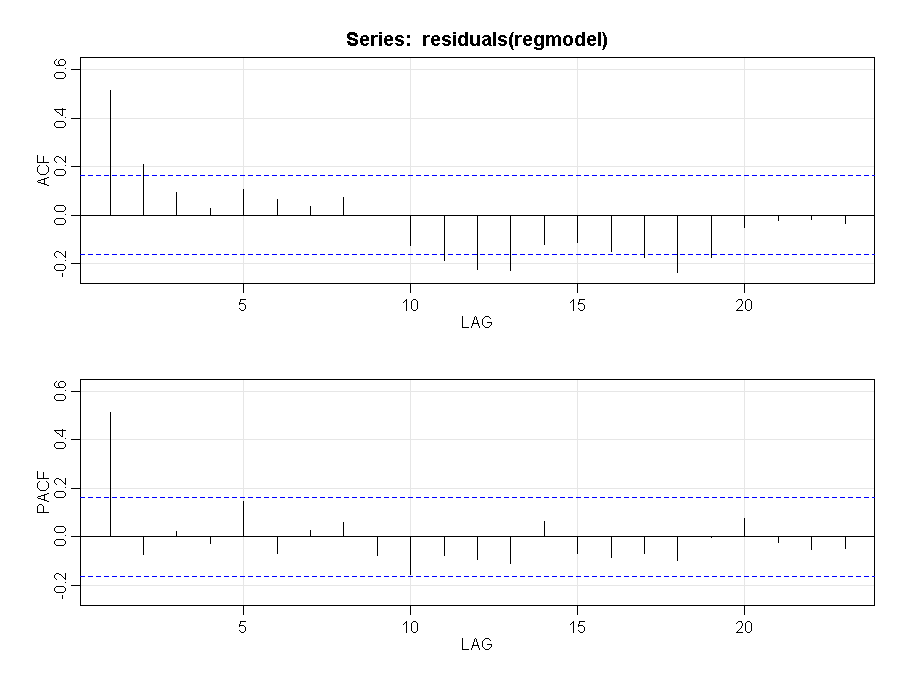
(Intercept) 466.96069 35.59523 13.12 <2e-16 \*\*\*

x 1.33425 0.05267 25.33 <2e-16 \*\*\*

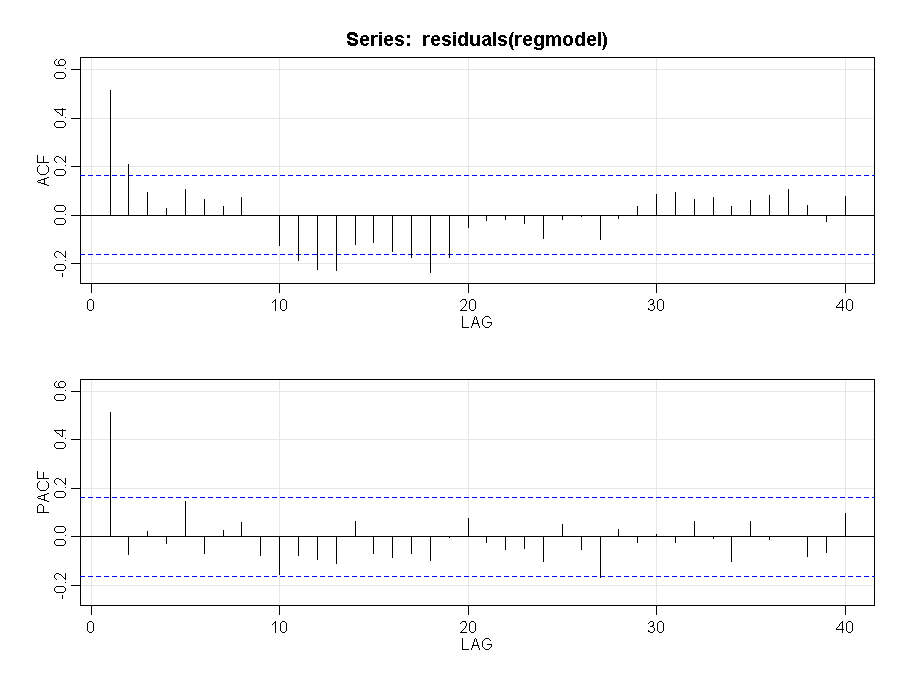
* We can see that the estimates for the intercept and slope are 466.96069 and 1.33425 respectively.
* We can see that the std errors for the intercept and slope are 35.59523 and 0.05267 respectively.

(b) 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 ACF and PACF plots are:



The ACF seems to trail off. However we see some significant values after lag 10. To double check if there is any pattern here, lets expand the ACF for a longer period, we get:



We see the following:

* The PACF peaks at lag1 and then cuts off. This follows the pattern of an AR(1) model
* The ACF seems to trail off to 0 further supporting the AR(1) structure. There are some values between lag 10 and 20 that don’t align well with the expected pattern.

In conclusion an AR(1) ARIMA model may be appropriate for the residuals.

(c) 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.

The coefficients are:

Estimate SE t.value p.value

ar1 0.5128 0.0697 7.3591 0

intercept 476.8760 57.0861 8.3536 0

xreg 1.3197 0.0843 15.6470 0

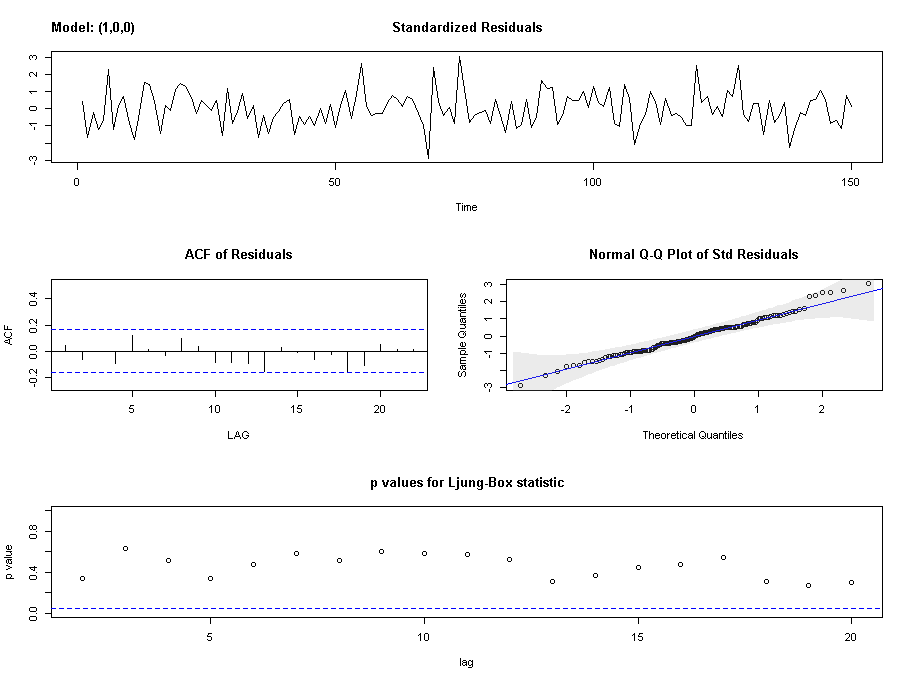
Thus our estimated relationship between yt and xt is:

yt = 476.8760 + 1.3197 xt

The errors have the estimated relationship et = wt + 0.5128 et−1, where wt∼iidN(0,σ2).

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

The diagnostics of the model:

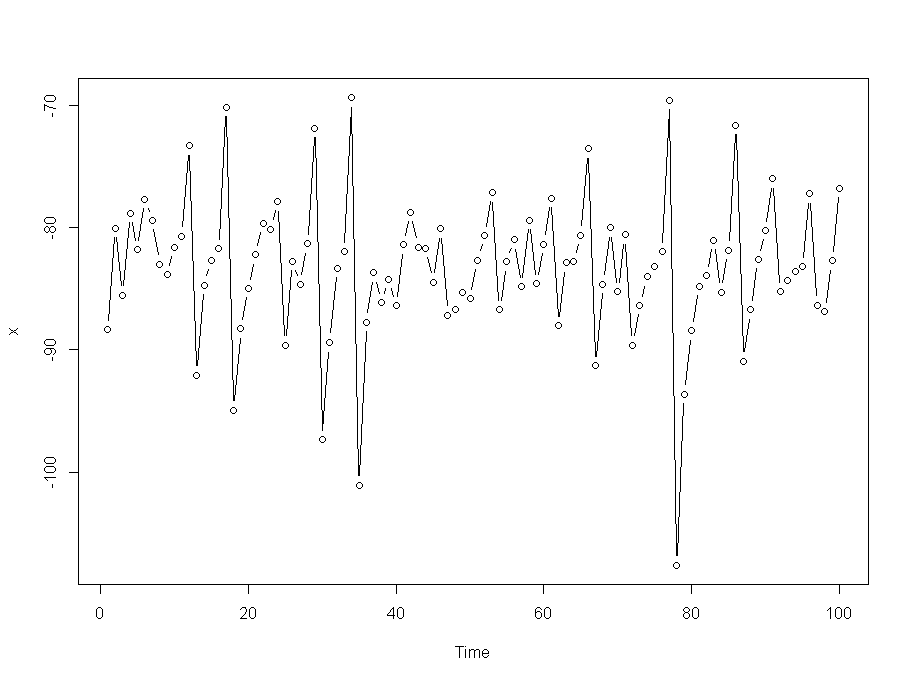


The diagnostics indicate that the model is appropriate because:

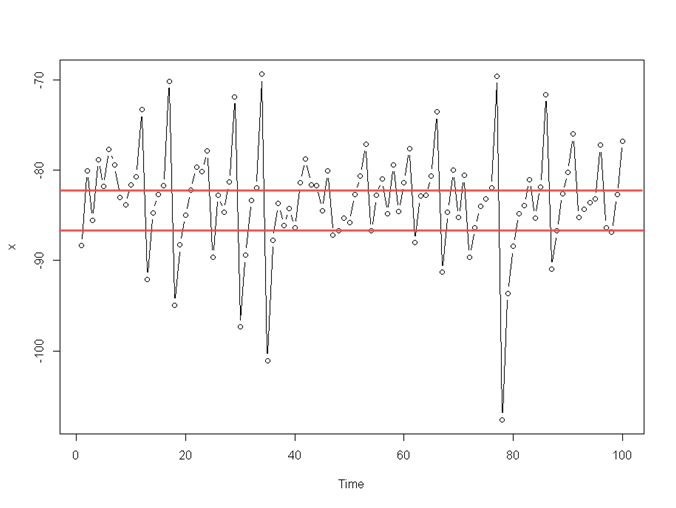
* The time series plot of the standardized residuals mostly indicates that there’s no visible trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations – a good result.
* The Q-Q plot is a normal probability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

2. Use the dataset e2q2.txt in the Exam 2 Module. (You may read in a .txt file just as a .dat file.)

(a) Plot the data, interpret the plot, and comment on two possible threshold values: -82 and -86. Based on the TS plot alone, which do you think is a better choice and why? Would you suggest another value?



With approximate locations of -82 and -86 drawn on the plot:



We can see here that:

* High peaks are as sharp as the low peaks.
* Rate of increase and decrease doesn’t seems to be different
* The values seem to be in a certain range and outside of that range it peaks both upwards and downwards
* Looking at the plot it appears that we need to define 2 different values of thresholds. Threshold of -82 appears to define approximately the level where the value peaks upwards. On the contrary, threshold of -86 appears to define approximately the level where the value peaks downwards.
* Overall though -82 appears to be a better line in that it is more in the middle.
* It doesn’t appear that a single threshold model will be most appropriate for this dataset.

(b) For threshold c = -82, estimate an AR(1) model for the original data in each of the two regions. Provide the model output and discuss significance of terms. Note: Because an AR(1) is requested, you will need to make adjustments to the code provided in Lesson 13.2:

model = ts.intersect(y, lag1y=lag(y,-1))

P = model[,2]

Because P now contains just one column, you may replace P[,1] with P in the rest of your code.

The model is as follows:

##Regression for values below the threshold

Coefficients:

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

(Intercept) -41.96653 4.98863 -8.412 1.43e-11 \*\*\*

P1 0.47661 0.05748 8.291 2.26e-11 \*\*\*

---

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

Residual standard error: 2.049 on 57 degrees of freedom

Multiple R-squared: 0.5467, Adjusted R-squared: 0.5387

F-statistic: 68.74 on 1 and 57 DF, p-value: 2.263e-11

##Regression for values above the threshold

Coefficients:

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

(Intercept) -233.4445 16.1758 -14.432 < 2e-16 \*\*\*

P2 -1.9111 0.2056 -9.293 2.51e-11 \*\*\*

---

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

Residual standard error: 4.797 on 38 degrees of freedom

Multiple R-squared: 0.6944, Adjusted R-squared: 0.6864

F-statistic: 86.36 on 1 and 38 DF, p-value: 2.514e-11

The outputs above indicate that the AR(1) term is significant for both the cases (P1 is significant where variable < threshold and P2 is significant where variable >= threshold)

The R squared results are: below threshold (-82): .5467, above threshold (-82): .6944

(c) For threshold c = -86, estimate an AR(1) model for the original data in each of the two regions. Provide the model output and discuss significance of terms.

##Regression for values below the threshold

Coefficients:

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

(Intercept) -45.19960 6.42813 -7.032 3.64e-07 \*\*\*

P1 0.44247 0.07105 6.227 2.36e-06 \*\*\*

---

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

Residual standard error: 1.815 on 23 degrees of freedom

Multiple R-squared: 0.6277, Adjusted R-squared: 0.6115

F-statistic: 38.78 on 1 and 23 DF, p-value: 2.361e-06

##Regression for values above the threshold

Coefficients:

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

(Intercept) -167.3736 12.3254 -13.580 < 2e-16 \*\*\*

P2 -1.0455 0.1519 -6.883 1.82e-09 \*\*\*

---

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

Residual standard error: 5.069 on 72 degrees of freedom

Multiple R-squared: 0.3969, Adjusted R-squared: 0.3885

F-statistic: 47.37 on 1 and 72 DF, p-value: 1.816e-09

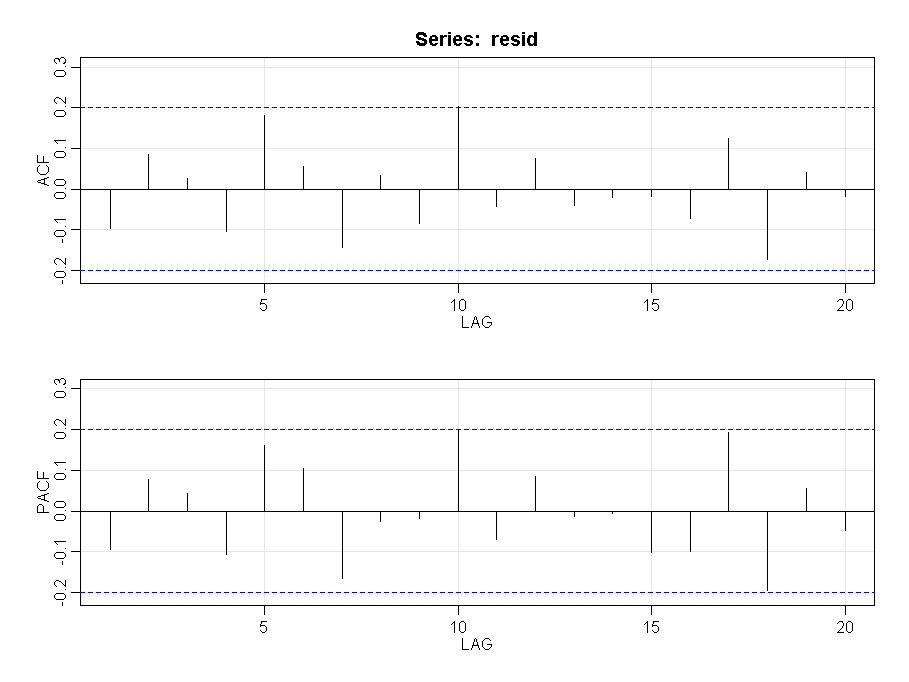
The outputs above indicate that the AR(1) term is significant for both the cases (P1 is significant where variable < threshold and P2 is significant where variable >= threshold)

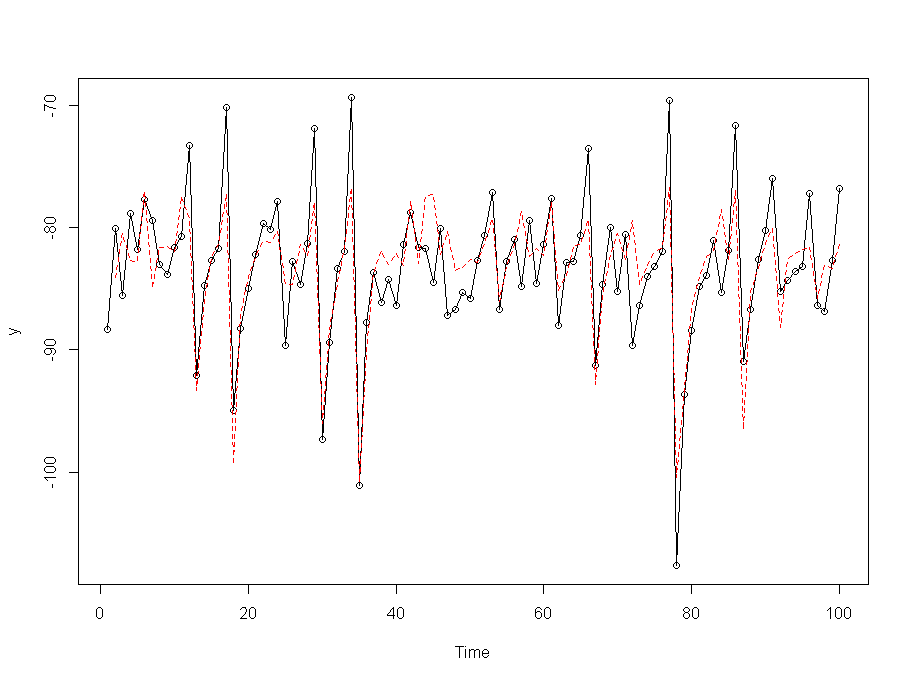
The R squared results are: below threshold (-86): .6277, above threshold (-82): .3969

(d) Comment on the suitability of each model and compare the actual and predicted values. Is either threshold value preferable?

Threshold = -82

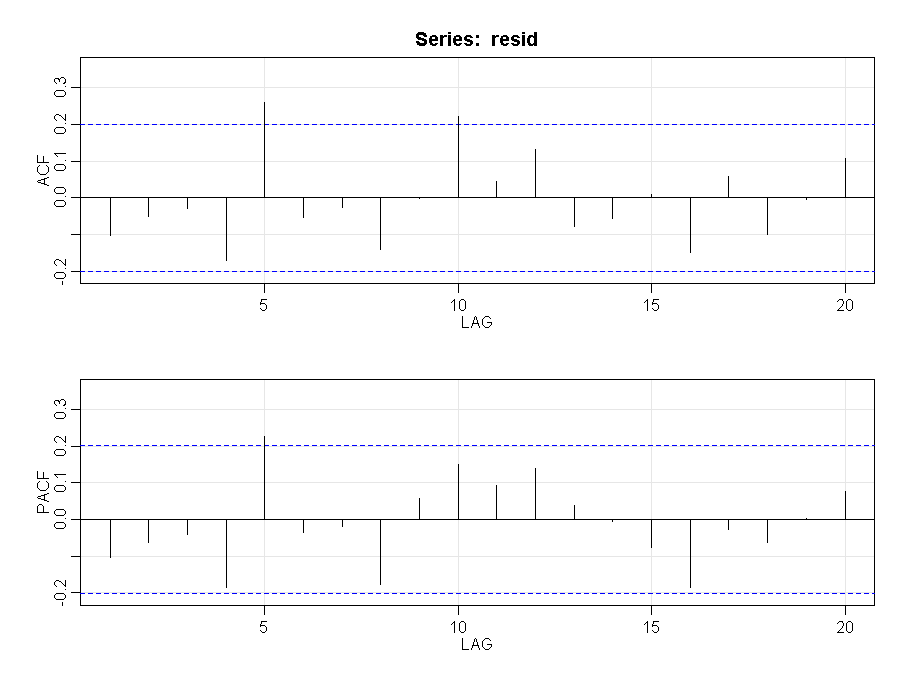
The plot for ACF / PACF of the residuals:

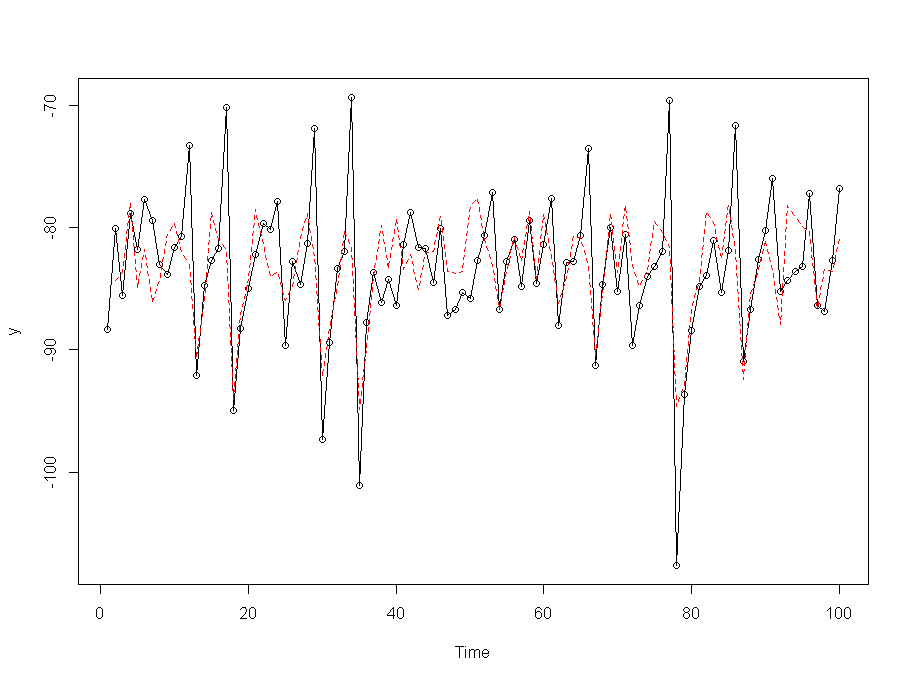




Threshold = -86

The plot for ACF / PACF of the residuals:





The observation from the above are:

* While the ACF / PACF results are better for threshold = -82, both the models are deficient. The poor prediction results indicate that both the thresholds are insufficient.
* Overall though -82 has higher levels of R squared and the prediction results look slightly better than -86 threshold.