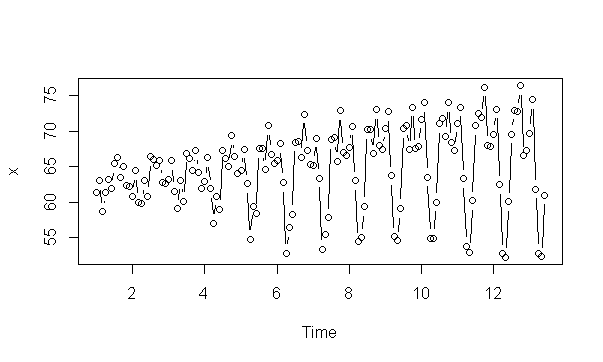
# Analysis for question 1

## Step 1: The time plot of the raw data is:

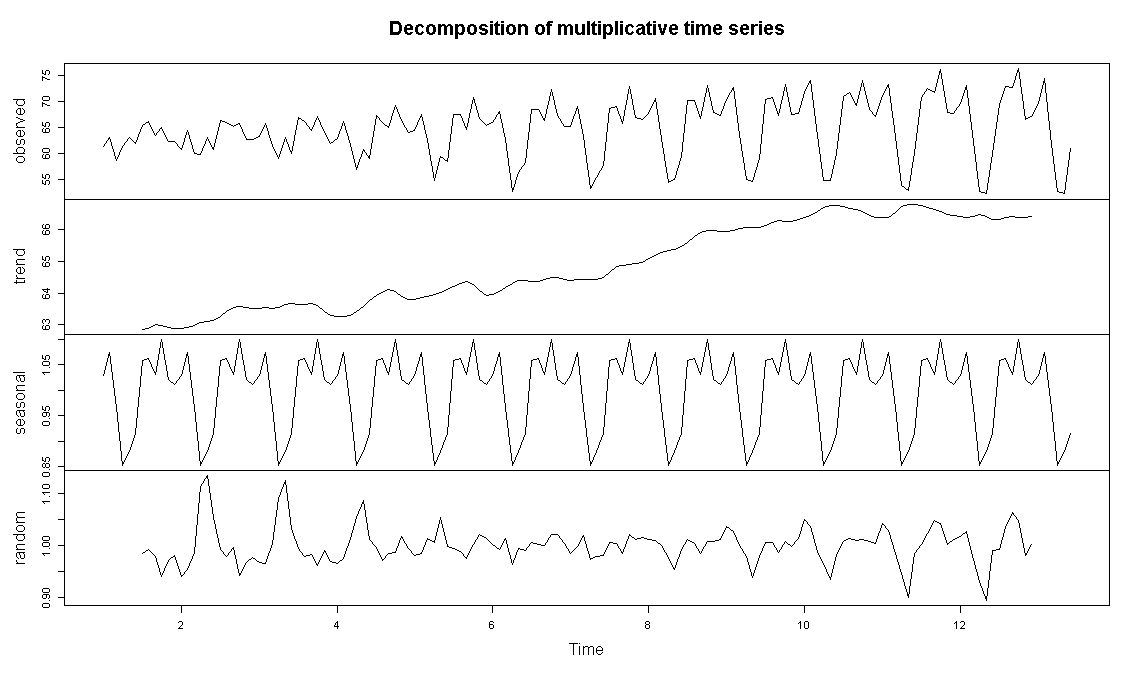


Some features of the plot:

* There is no consistent trend and the series wanders up and down.
* We have monthly data and there appears to be some seasonality.
* There are no obvious outliers.
* The variance is not constant and is increasing over time.

Overall we find that we have monthly data that shows seasonal variation and the variance is increasing over time. This suggests that we may want to check for multiplicative decomposition.

## Step 2: decompose (multiplicative) – we get the following plot:



The seasonal monthly effects are:

Jan Feb Mar Apr May Jun Jul Aug

1.0296780 1.0743008 0.9658527 0.8528570 0.8807737 0.9147313 1.0576098 1.0629899

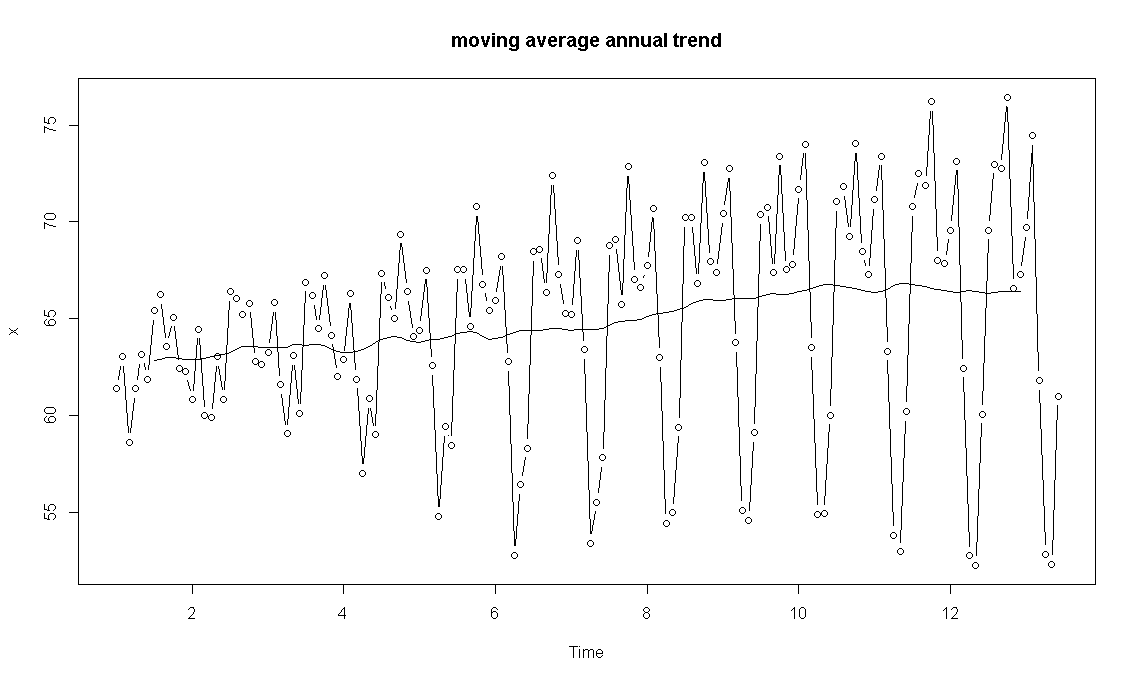
Sep Oct Nov Dec

1.0301841 1.0989718 1.0218366 1.0102143

We observe a trend so let’s try to plot the trend on the raw data.

## Step 3: Taking an annual moving average:

trendpattern =filter(x, filter=c(1/24,1/12,1/12,1/12,1/12,1/12,1/12,1/12,1/12,1/12,1/12,1/12,1/24), sides=2)



From the above analysis we find that there is seasonality and weak trend in our data. **The trend however is not prominent**.

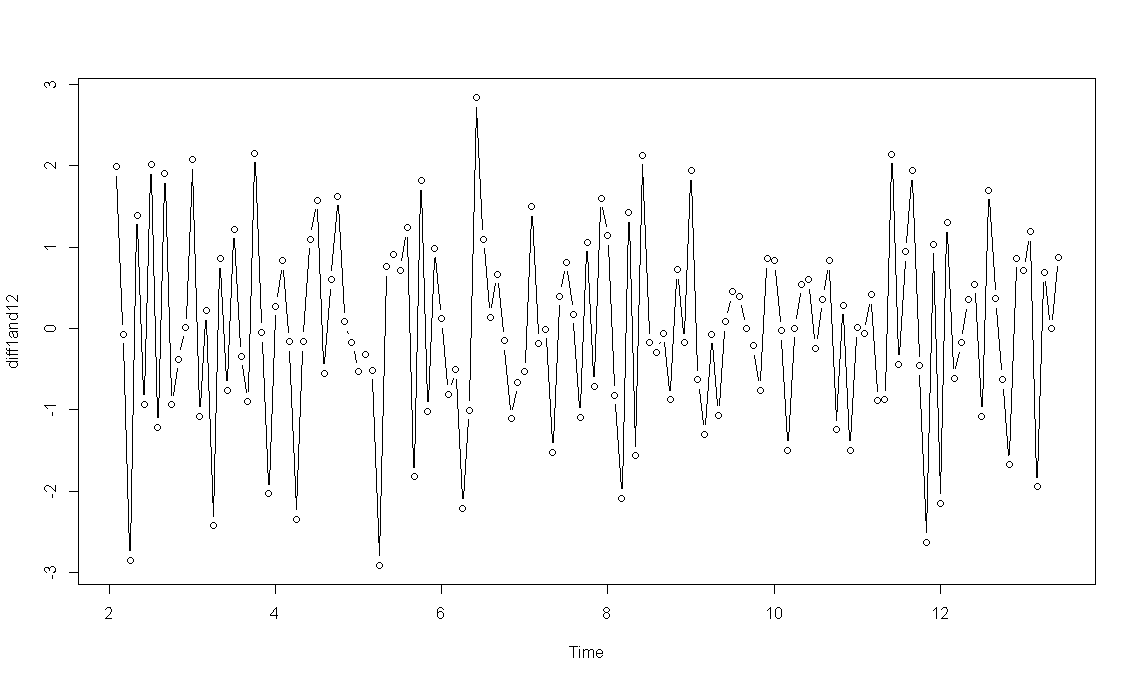
## Step 4: Trying some transformations / differentiations.

On trying some of the transformations (**log, sqrt, sqrt of sqrt) we don’t** see that the variance is getting stabilized so we will try a first difference combined with 12th difference:

diff1 = diff(x,1)

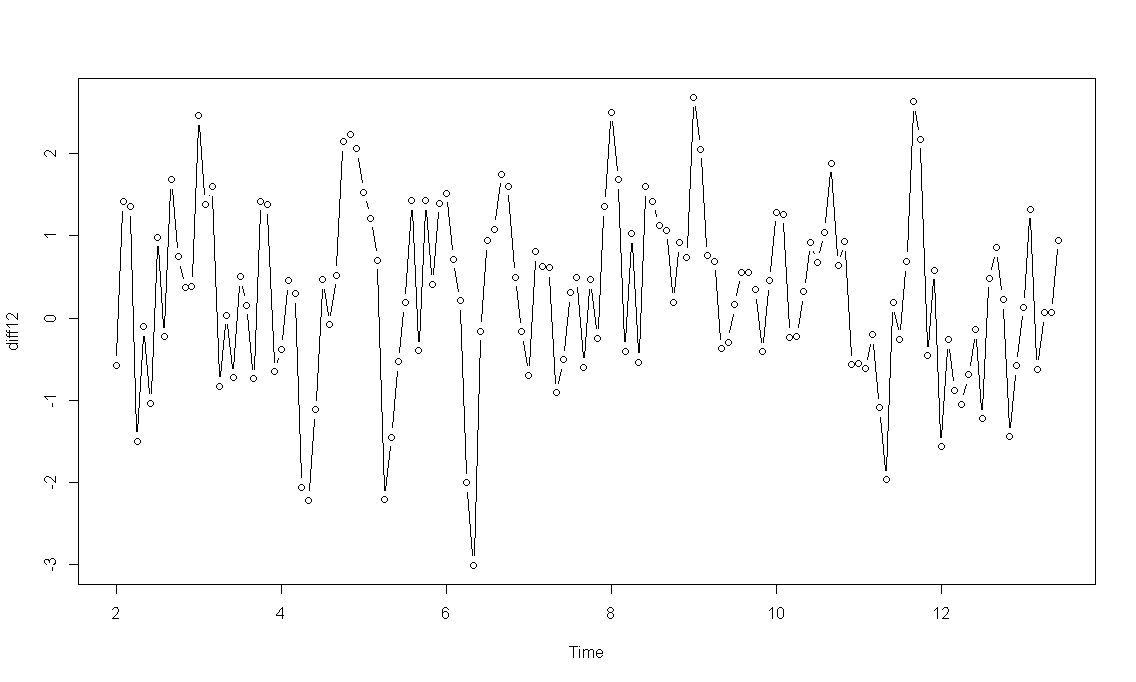
diff1and12=diff(diff1,12)

When we plot diff1and12 we get:



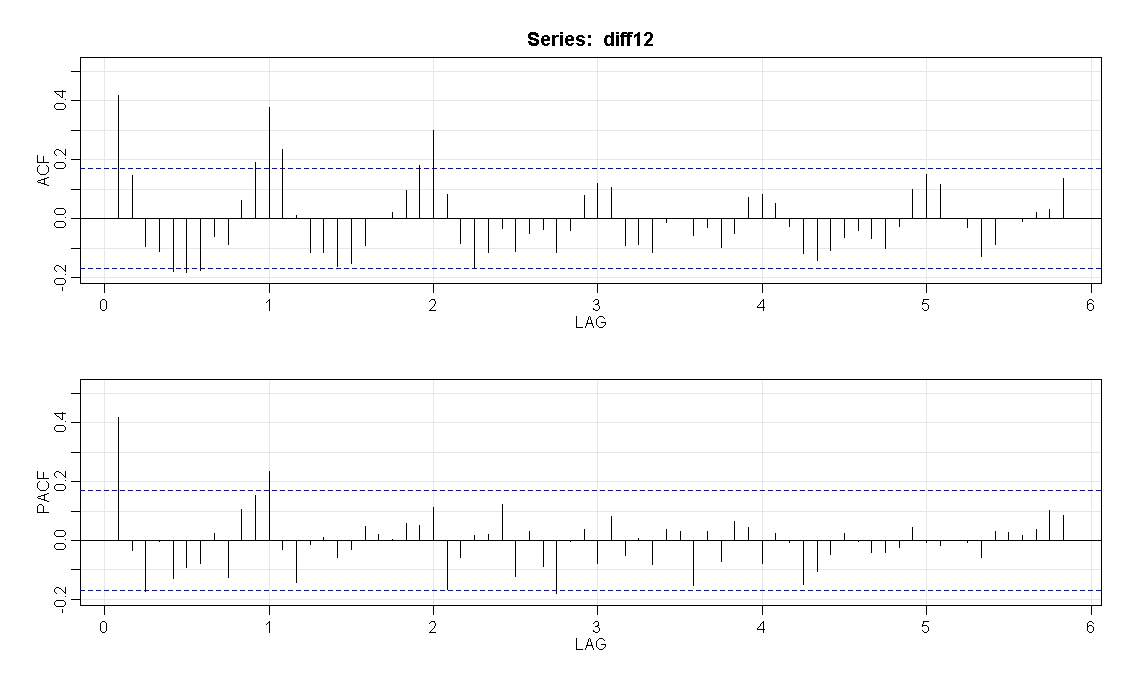
This indicates that the variance has stabilized.

We will also try only 12th difference (since we found the trend to be weak)



Here also we see that the variance has stabilized, we will move forward with only diff12 since we had found the trend to be weak so it makes sense.

## Step 5: Try the ACF and PACF plots for diff12



Non-seasonal - In the first 2/3 lags, we observe:

* The PACF has a spike at (1) and (3) and then cuts off. The ACF is not very clear – it has a spike at (1) and then cuts off to come back .
  + Relying on the PACF though it points to an AR(1) or an AR(3) model
  + Relying on ACF it points to an MA(1) model

Seasonal – Here we observe:

* The PACF has a spike at 1st seasonal lag followed by a cut-off and the ACF tails off gradually to 0. This points to an AR(1) model.

So we have the following potential models to evaluate:

* ARIMA(1, 0, 1) × (1,1,0)12
* ARIMA(3, 0, 1) × (1,1,0)12
* ARIMA(1, 0, 0) × (1,1,0)12

## Step 6 – Get the statistics for the different models

We have now:

|  |  |
| --- | --- |
| **Model** | **Notes** |
| ARIMA(1, 0, 1) × (1,1,0)12 | sigma^2 estimated as 0.8505  $AIC  [1] 0.891417  $AICc  [1] 0.9075281  $BIC  [1] -0.02829937 |
| ARIMA(3, 0, 1) × (1,1,0)12 | sigma^2 estimated as 0.8303  $AIC  [1] 0.8939891  $AICc  [1] 0.9125806  $BIC  [1] 0.0144145 |
| ARIMA(1, 0, 0) × (1,1,0)12 | sigma^2 estimated as 0.8509  $AIC  [1] 0.8785752  $AICc  [1] 0.8937477  $BIC  [1] -0.06121205 |

We find that the statistics of the models is quiet similar. However that said the ARIMA(1, 0, 0) × (1,1,0)12

gives us the best parameters and we choose that as our model of preference.

## Step 7: Details of the selected model

The coefficients are:

$ttable

Estimate SE t.value p.value

ar1 0.3830 0.0795 4.8153 0.0000

sar1 0.3572 0.0826 4.3232 0.0000

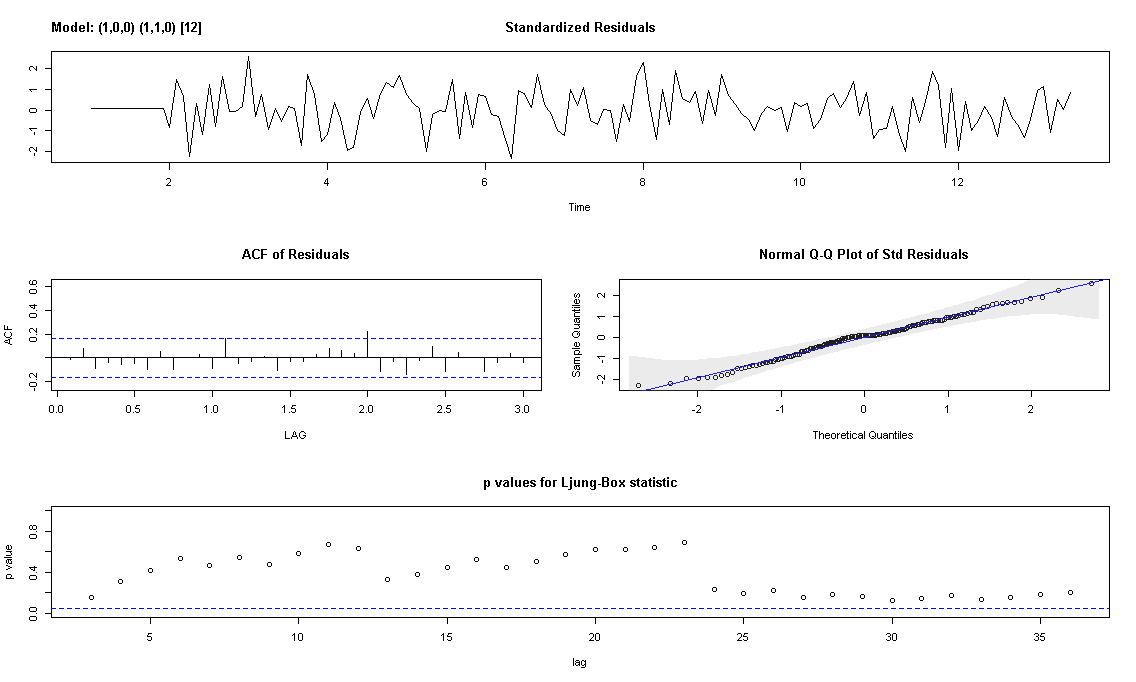
constant 0.0245 0.0157 1.5599 0.1209

Check the significance:

AR(1) coeff (z = 0.3830/0.0795 = 4.8176) is significant since |z| > 1.96

AR(1) coeff (z = 0.3572/0.0826 = 4.3245) is significant since |z| > 1.96

The plots are:



Based on this we deduce the following:

* The time series plot of the standardized residuals mostly indicates that there’s no 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 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 36. All p-values are above the dashed blue line - That’s a good result.

## Summary of the model:

We have a model represented as ARIMA(p, d, q) × (P, D, Q)S

ARIMA(1, 0, 0) × (1, 1, 0)12 (non-seasonal AR(1) and seasonal AR(1) model with seasonal first difference)

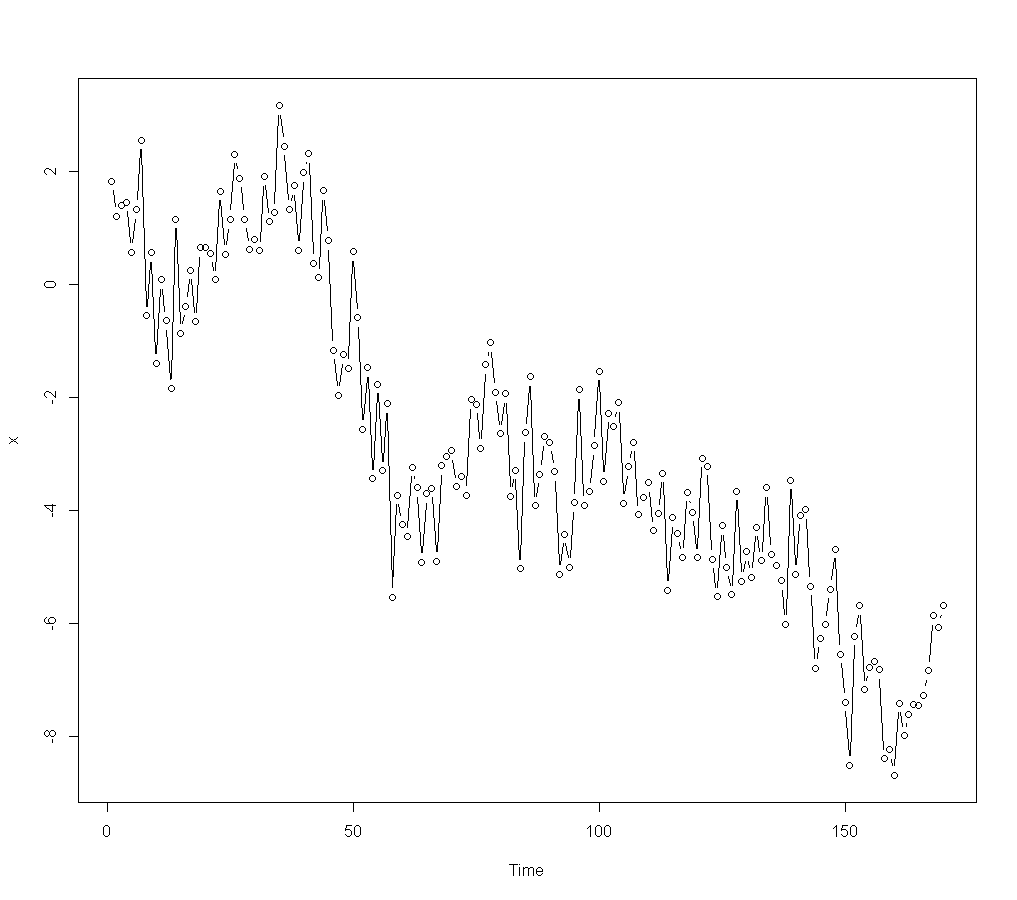
First seasonal difference: ∇12xt = xt − xt−12

The non-seasonal components are: AR(1): φ(B) = 1 - φ1B = 1 – 0.3830 B

The seasonal components are: Seasonal AR(1): Φ(B12) = 1 - Φ1B12 = 1 - 0.3572 B12

# Analysis for question 2

## Step 1: The time plot of the raw data is:



Some features of the plot:

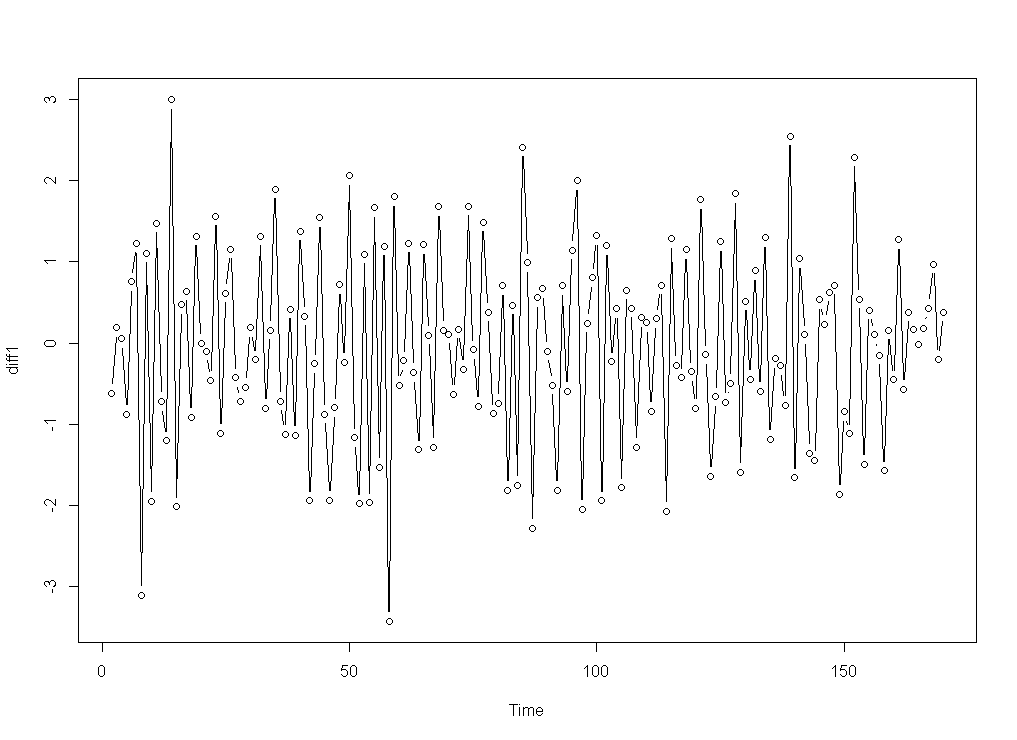
* There is a clear overall downward trend.
* We have annual data and therefore deduce that there is no seasonality
* There are no obvious outliers.
* The variance appears to be constant.

## Step 2: Trying some transformations / differentiations.

Try a first difference and plot:

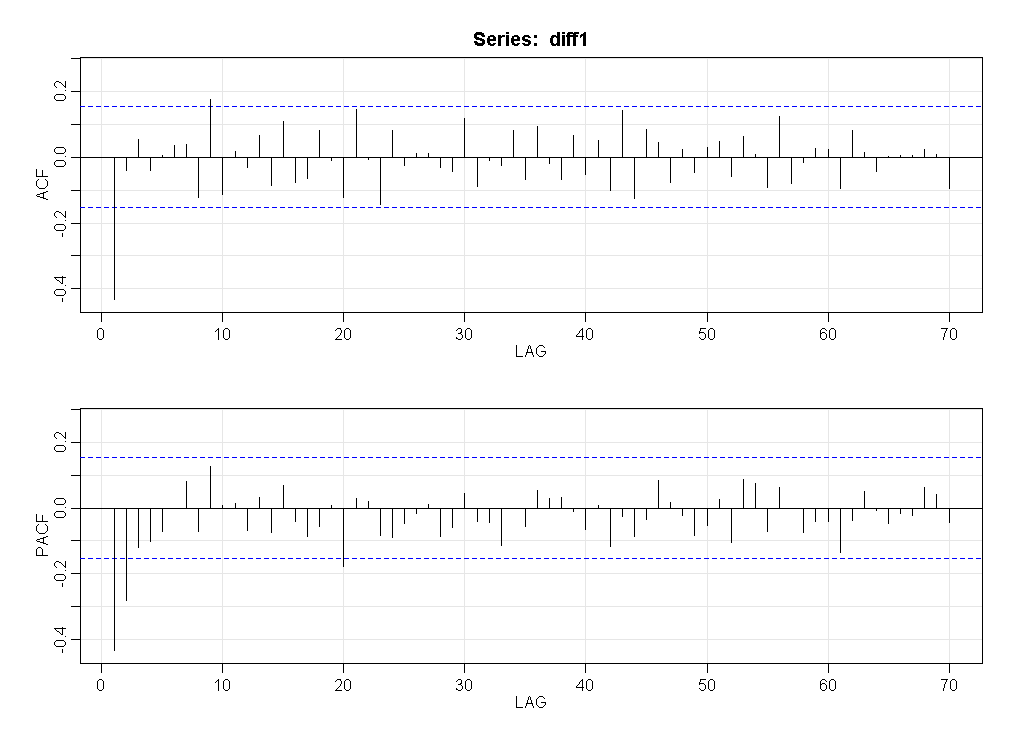
diff1 = diff(x,1)

plot (diff1, type="b")



This indicates that the series has been de-trended.

## Step 3: Try the ACF and PACF plots for diff1



* The ACF has a spike at (1) and then cuts off. The PACF is tailing off to 0. This is a pattern for MA(1). Here ACF spike at 9 is not explained

While MA(1) provides the best explanation we will get a couple of more candidates to assess how our choice works out:

* We can also evaluate and AR(2) model given that the PACF has spikes at lags 1 and 2 and then cuts off.
* Since both PACF and ACF are not clearly in sync we can also test ARMA(2, 1) and ARMA(1, 1) models

So we have the following potential models to evaluate:

* ARIMA(0, 1, 1)
* ARIMA(2, 1, 0)
* ARIMA(1, 1, 1)
* ARIMA(2, 1, 1)

## Step 4 – Get the statistics for the different models

We have now:

|  |  |
| --- | --- |
| **Model** | **Notes** |
| ARIMA(0, 1, 1) | sigma^2 estimated as 1.007  $AIC  [1] 1.030264  $AICc  [1] 1.042879  $BIC  [1] 0.06715543 |
| ARIMA(2, 1, 0) | sigma^2 estimated as 1.033  $AIC  [1] 1.067519  $AICc  [1] 1.08071  $BIC  [1] 0.1228565 |
| ARIMA(1, 1, 1) | sigma^2 estimated as 1.007  $ttable  Estimate SE t.value p.value  ar1 -0.0067 0.1196 -0.0560 0.9554  ma1 -0.6012 0.0929 -6.4712 0.0000  constant -0.0457 0.0309 -1.4823 0.1402  $AIC  [1] 1.042007  $AICc  [1] 1.055197  $BIC  [1] 0.0973442  We find ar(1) is not significant – leads back to MA(1) |
| ARIMA(2, 1, 1) | sigma^2 estimated as 1.004  $ttable  Estimate SE t.value p.value  ar1 -0.0788 0.1754 -0.4495 0.6536  ar2 -0.0728 0.1187 -0.6134 0.5404  ma1 -0.5290 0.1630 -3.2458 0.0014  constant -0.0454 0.0318 -1.4293 0.1548  $AIC  [1] 1.051537  $AICc  [1] 1.065454  $BIC  [1] 0.1253205  We find ar(1) and ar(2) is not significant – leads back to MA(1) |

ARIMA(0, 1, 1) gives us the best parameters and we choose that as our model of preference.

## Step 5: Details of the selected model

The coefficients are:

$ttable

Estimate SE t.value p.value

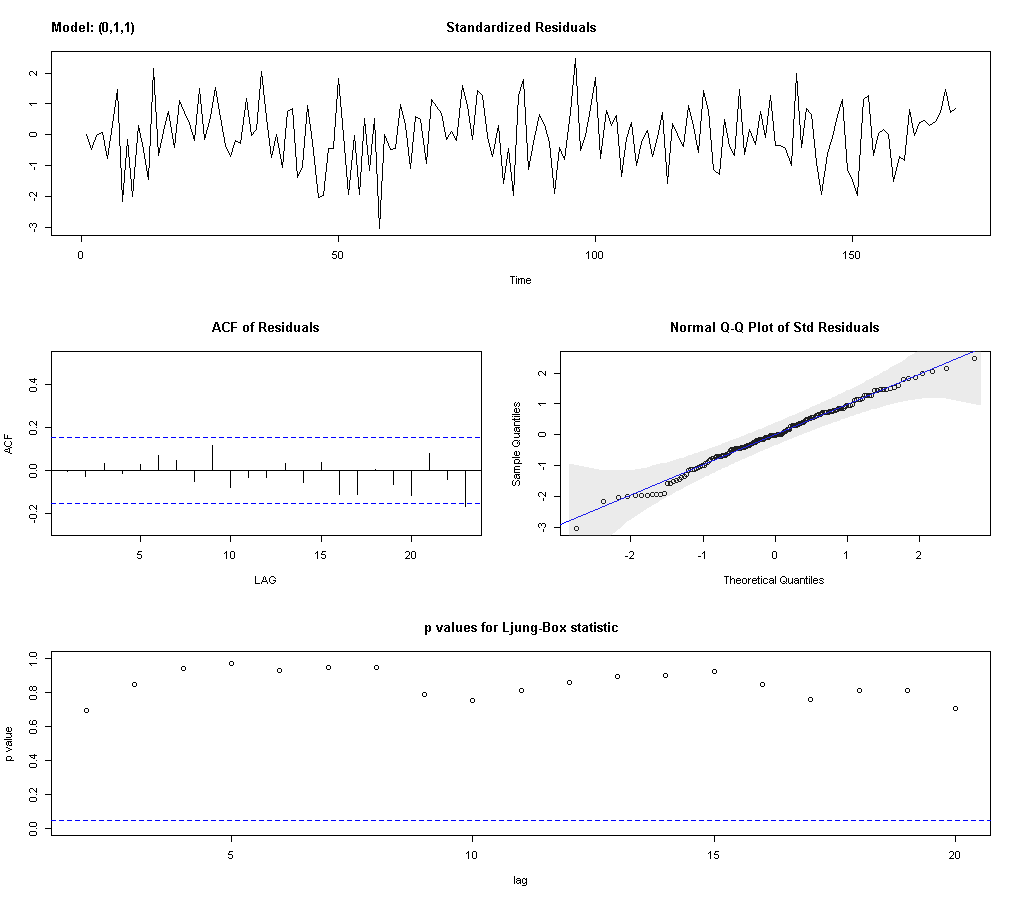
ma1 -0.6051 0.0594 -10.1898 0.0000

constant -0.0458 0.0308 -1.4881 0.1386

Check the significance:

MA(1) coeff (z = -0.6051/0.0594 = -10.1869) is significant since |z| > 1.96

The plots are:



Based on this we deduce the following:

* The time series plot of the standardized residuals mostly indicates that there’s no trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations (except a significant value at lag 23) – a good result overall.
* The Q-Q plot is a normal probability plot – The plot looks almost linear and the assumption of normally distributed residuals generally 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.

## Summary of the model:

We have a model represented as ARIMA(p, d, q)

ARIMA(0, 1, 1)

First difference: ∇xt = xt − xt−1

The components are: MA(1): A MA(1) model is defined as xt = μ + wt + θ1wt−1 and could be written as xt = μ + (1+θ1B) wt

Also represented as (xt−μ) = Θ(B) wt

where,

Θ(B) = 1+θ1B = 1 - 0.6051 B

# Other questions



* This is an ARMA model



* This is an AR model

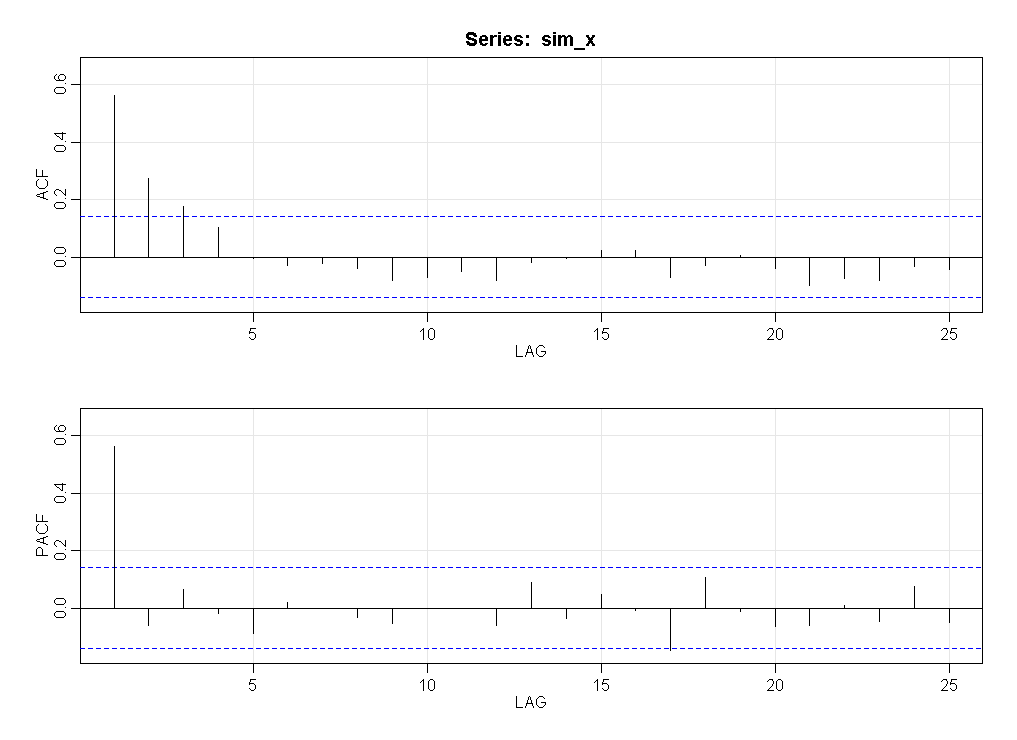


The first lag autocorrelation: we will use ρh=ϕ1h

First autocorrelation = ρ1 = 0.58

Second autocorrelation = ρ1^2 = 0.58 = 0.3364







The ACF and PACF support the AR(1) model since:

* For an AR model, the theoretical PACF “shuts off” past the order of the model i.e. in theory the partial autocorrelations are equal to 0 beyond that point. In the image above we see a spike and significant PACF for lag 1 followed by small and insignificant values of PACF. This is what is expected from an AR(1) model.
* The ACF for an AR(1) model: The correlation between observations h time periods apart is ρh=ϕ1h

This implies that we should see an exponential decrease (tailing off) of the ACF values. The image above is in accordance with this pattern.



xt = 2 + 0.58xt-1 + wt

x201 = 2 + 0.58x200 + w200 = 2 + 0.58 \* 1.22 + 0 = 2.7076

The standard error of the forecast error at time 201 is = sqrt(0.7 \* 1) = 0.8367

The 95% prediction interval for the value at time 201 is 2.7076 ± 0.8367(1.96), which is 1.0677 to 4.3475. We are therefore 95% confident that the observation at time 201 will be between 1.0677 to 4.3475. If we repeated this exact process, then 95% of the computed prediction intervals would contain the true value of x at time 201.