Week 6 Assignment

Ben Arancibia

July 18, 2015

```r
library(fpp)
```

```
## Loading required package: forecast

## Warning: package 'forecast' was built under R version 3.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.1.3

##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Loading required package: timeDate
## This is forecast 6.1
##
## Loading required package: fma
## Loading required package: tseries
## Loading required package: expsmooth

## Warning: package 'expsmooth' was built under R version 3.1.3

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.1.3
```

HA 8.1 Figure 8.24 shows the ACFs for 36 random numbers, 360 random numbers and for 1,000 random numbers.

a) Explain the differences among these figures. Do they all indicate the data are white noise?

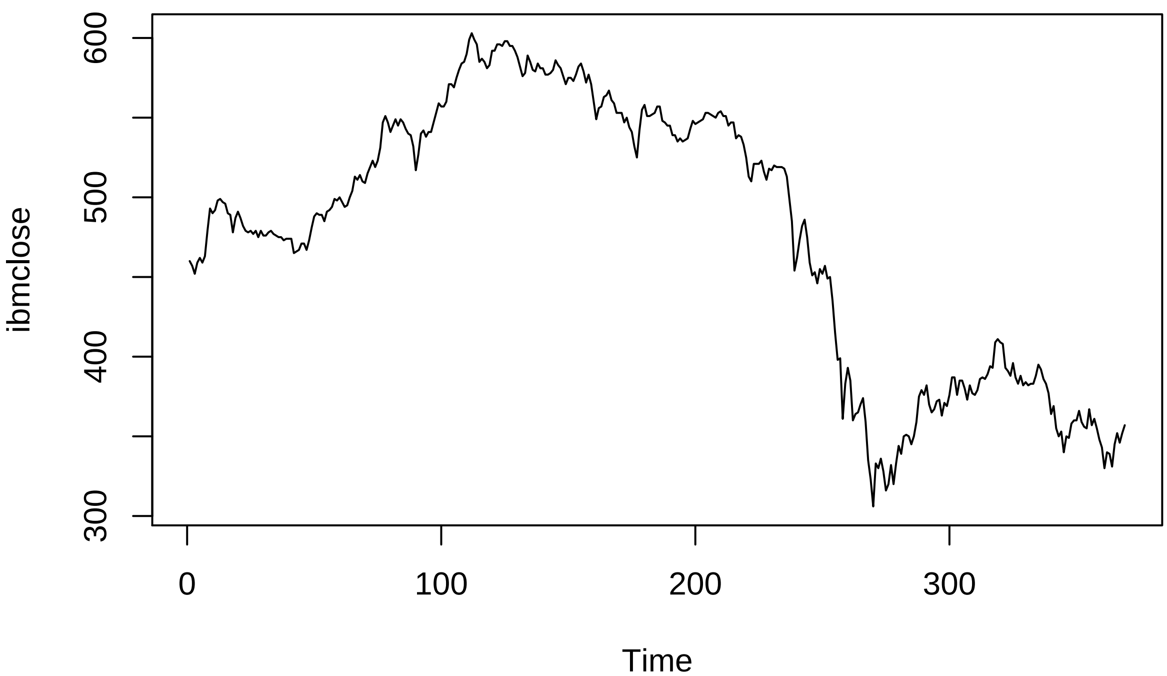
The different among the figures is that the lag increases in each plot. Yes it does appear that the data is white noise because the critical values do not break the threshold.

b) Why are the critical values at different distances from the mean of zero? Why are the autocorrelations different in each figure when they each refer to white noise?

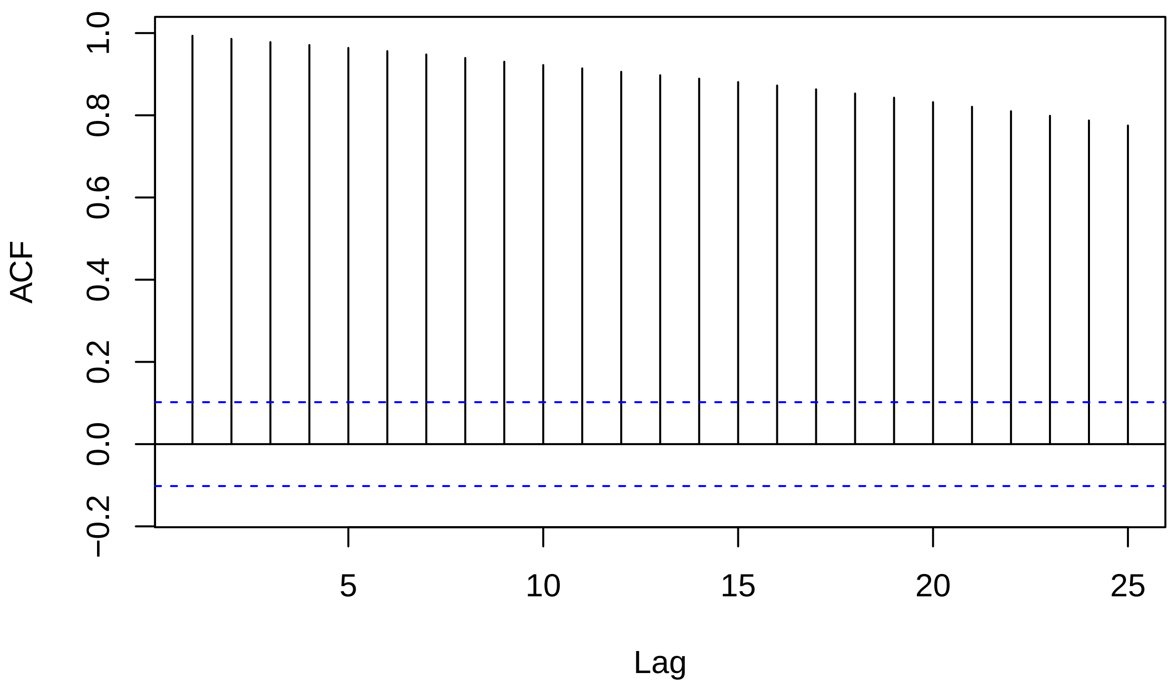
The critical values are at different distances because ACF does not take into account the effect of time lags. ACF plot shows the autocorrelations which measure the relationship between yt and yt-k for different values of k. If yt and yt-1 are correlated, then yt-1 and yt-2 must also be correlated. But then yt and yt−2 might be correlated, simply because they are both connected to yt−1, rather than because of any new information contained in yt−2 that could be used in forecasting yt. PACF removes the effect of time.

HA 8.2 A classic example of a non-stationary series is the daily closing IBM stock prices (data set ibmclose). Use R to plot the daily closing prices for IBM stock and the ACF and PACF. Explain how each plot shows the series is non-stationary and should be differenced.

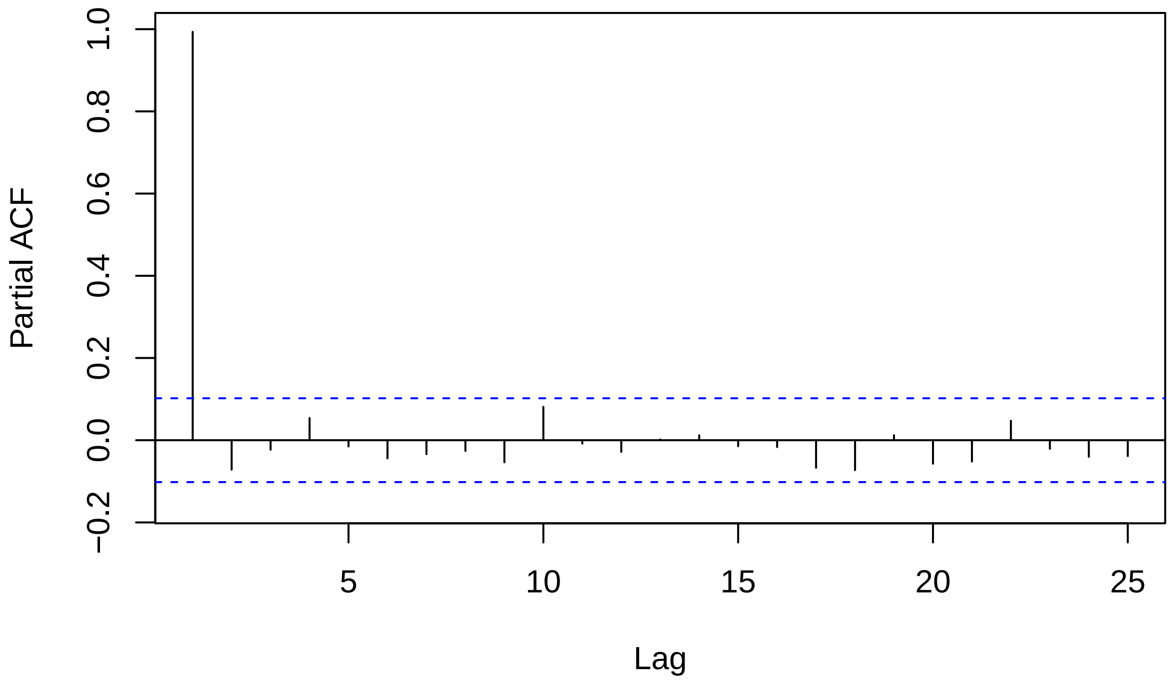
```
plot(ibmclose)
```




```
Acf(ibmclose,main="")
```



```r
Pacf(ibmclose,main="")
```

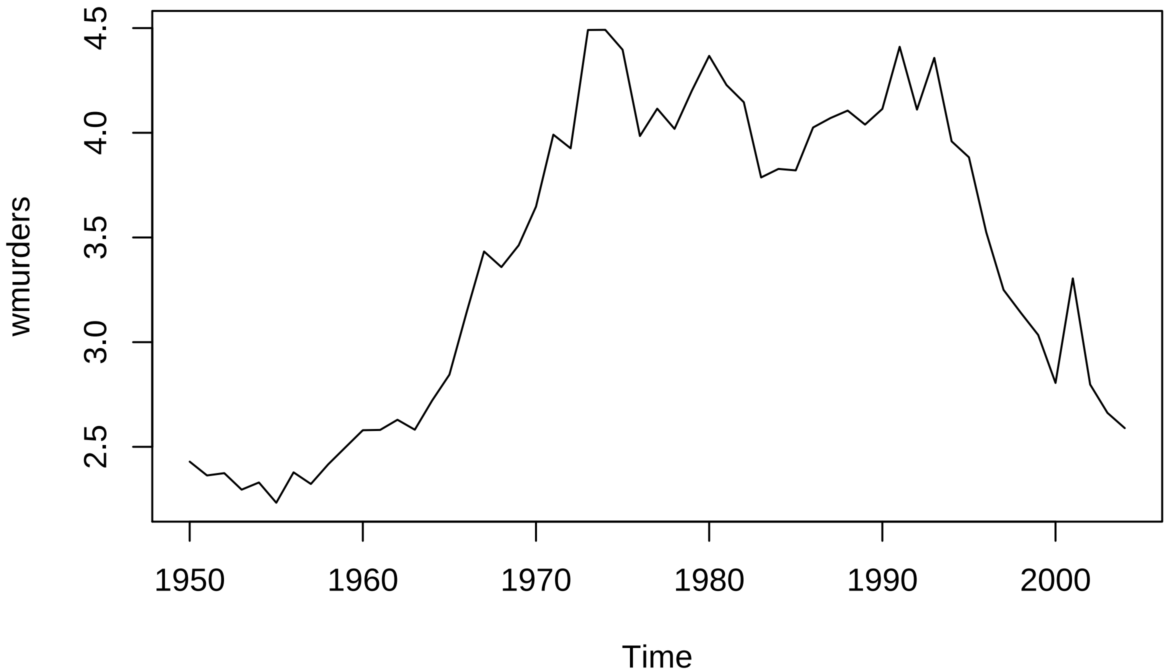


IBM daily closing stock prices is a non-stationary series. If you look at the ACF plot the data is non-stationary because it decreases incremently and does not drop to zero quickly. There is only one value in the PACF pliot that is above the 95% threshold. A random walk model is the best way for this data to be differenced. It is very widely used for non-stationary data, particularly finance and economic data. Random walks typically have long periods of apparent trends up or down and sudden and unpredictable changes in direction.

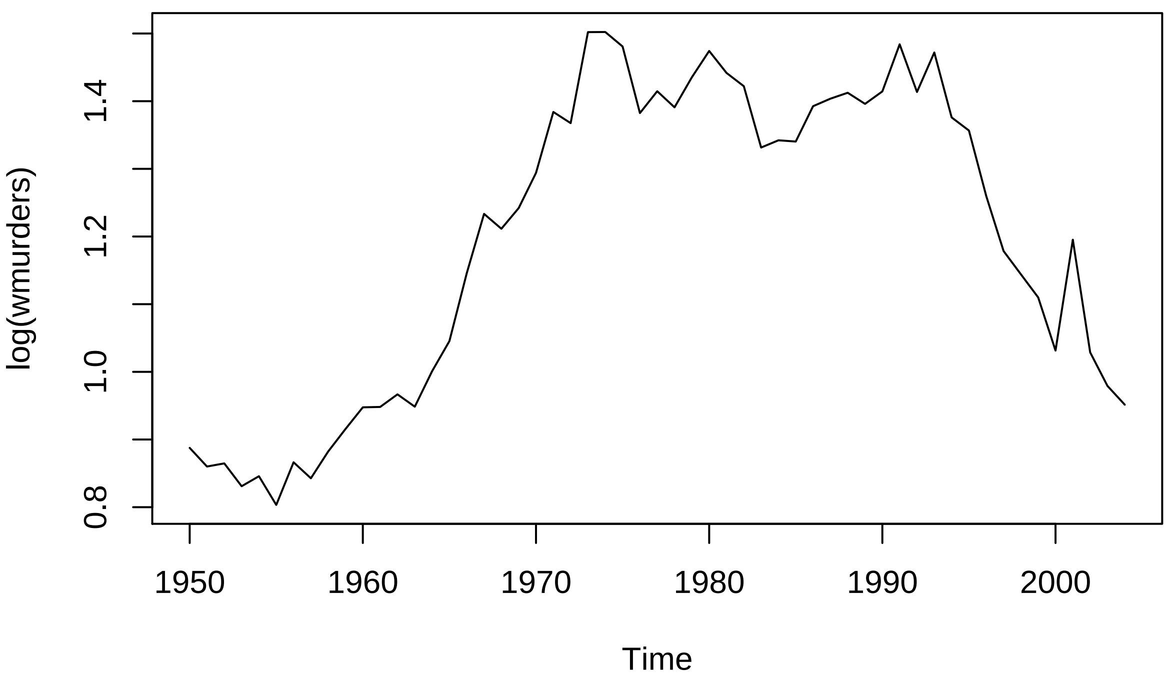
HA 8.6 Consider the number of women murdered each year (per 100,000 standard population) in the United States (data set wmurders).

a)By studying appropriate graphs of the series in R, find an appropriate ARIMA(p,d,q) model for these data.

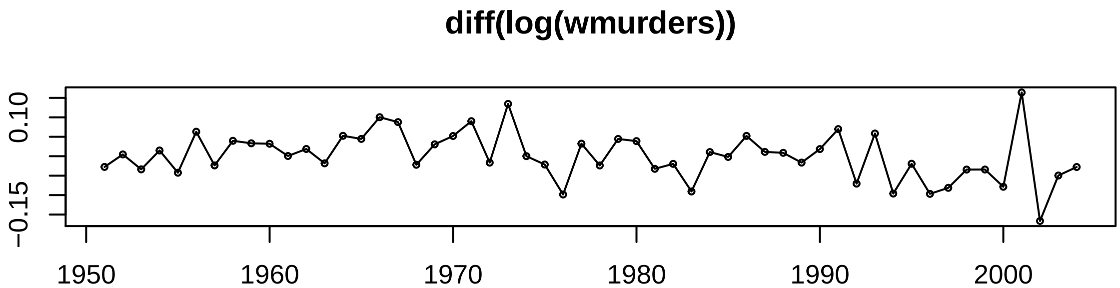
```r
plot(wmurders)
```

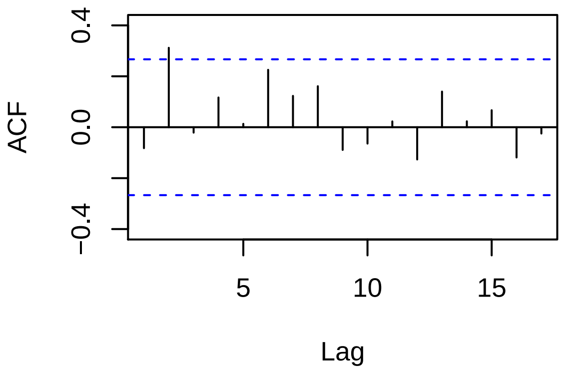


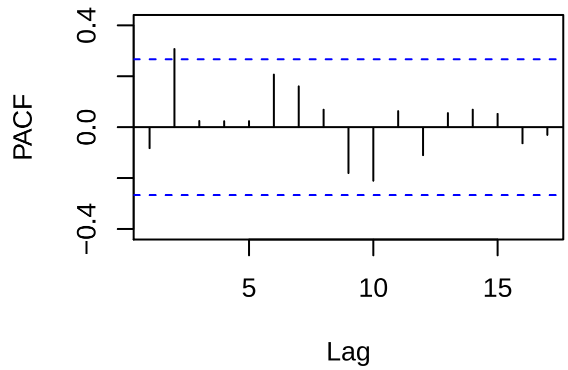

```r
plot(log(wmurders))
```




```r
tsdisplay(diff(log(wmurders)))
```







data seems to be non seasonal. No need for transformation but downward trend could cause negative forecasts so log was used.

```
fit <- arima(wmurders, order=c(1,1,0))
summary(fit)
```

```
##
## Call:
## arima(x = wmurders, order = c(1, 1, 0))
##
## Coefficients:
##           ar1
##       -0.0841
## s.e.   0.1346
##
## sigma^2 estimated as 0.0453:  log likelihood = 6.92,  aic = -9.85
##
## Training set error measures:
##                       ME      RMSE       MAE         MPE    MAPE      MASE
## Training set 0.003312753 0.2108978 0.1600357 -0.06225549 4.65107 0.9841445
##                   ACF1
## Training set 0.0237243
```

```
fit2 <- arima(wmurders, order=c(1,0,1))
summary(fit2)
```

```
##
## Call:
```

```
## arima(x = wmurders, order = c(1, 0, 1))
##
## Coefficients:
##          ar1      ma1  intercept
##       0.9637  -0.0445     2.9596
## s.e.  0.0296   0.1087     0.5646
##
## sigma^2 estimated as 0.04434:  log likelihood = 6.37,  aic = -4.73
##
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.01712634 0.2105752 0.1609971 0.1503576 4.696088 0.9900566
##                     ACF1
## Training set -0.01014992
```

```r
fit3 <- arima(wmurders, order=c(0,1,1))
summary(fit3)
```

```
##
## Call:
## arima(x = wmurders, order = c(0, 1, 1))
##
## Coefficients:
##           ma1
##       -0.0527
## s.e.   0.1070
##
## sigma^2 estimated as 0.04543:  log likelihood = 6.85,  aic = -9.7
##
## Training set error measures:
##                       ME      RMSE       MAE        MPE     MAPE      MASE
## Training set 0.003198768 0.2111909 0.1600467 -0.0627303 4.650887 0.9842122
##                     ACF1
## Training set -0.01729024
```

The above ARIMA model was selected by going through the different models above. I did not include the rest of the models testsed, but I found the model below to give good results. Good results are quantified as having a low aic.

```r
fit4 <- arima(wmurders, order=c(0,1,0))
summary(fit4)
```

```
##
## Call:
## arima(x = wmurders, order = c(0, 1, 0))
##
##
## sigma^2 estimated as 0.04563:  log likelihood = 6.73,  aic = -11.46
##
## Training set error measures:
##                      ME      RMSE       MAE        MPE     MAPE      MASE
## Training set 0.00295268 0.2116718 0.1597016 -0.0635393 4.638393 0.9820898
##                     ACF1
## Training set -0.08564966
```

b)Should you include a constant in the model? Explain.

No you do not need a constant in the model since it is a nonseasonal model and data looks stationary.

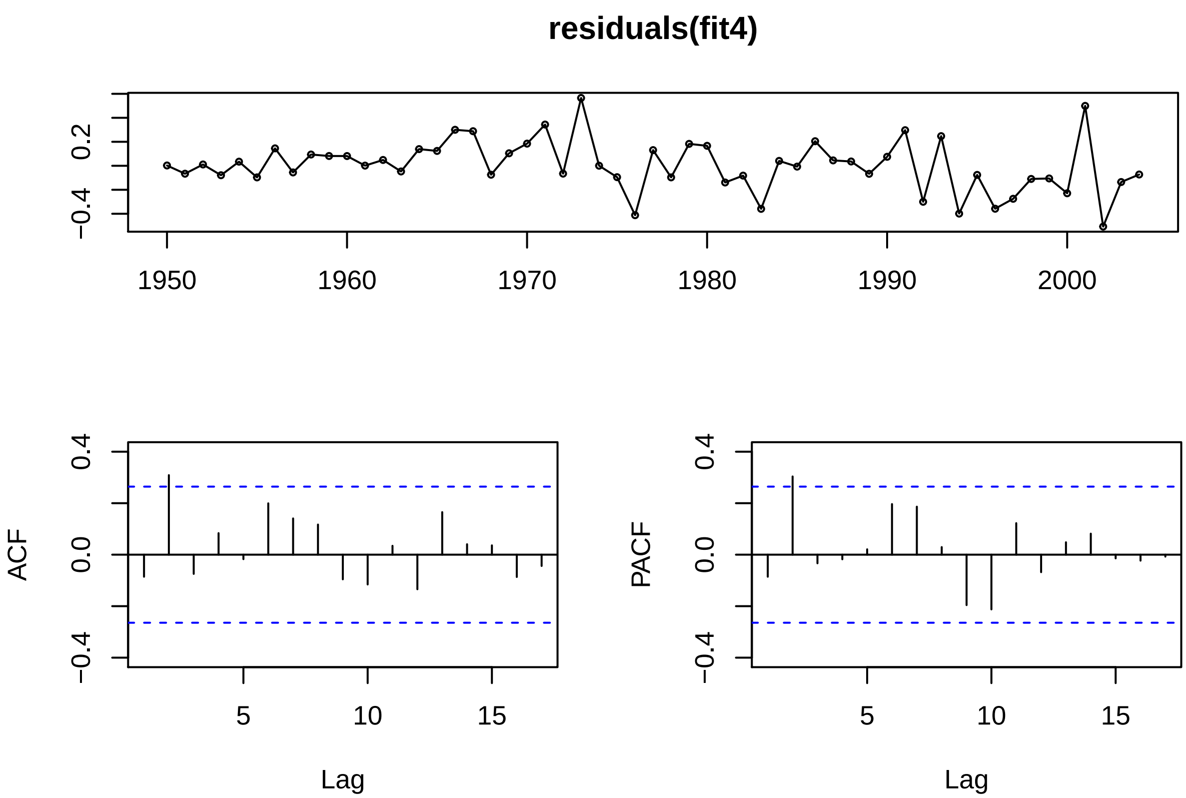
c)Write this model in terms of the backshift operator.

Arima (0,1,0)

$(1 - B)Yt = u + (1 - thetaB)/(1 - phiB)$

d)Fit the model using R and examine the residuals. Is the model satisfactory?

```
tsdisplay(residuals(fit4))
```




```
Box.test(residuals(fit4), lag=10, fitdf=0, type="L")
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit4)
## X-squared = 13.1653, df = 10, p-value = 0.2146
```

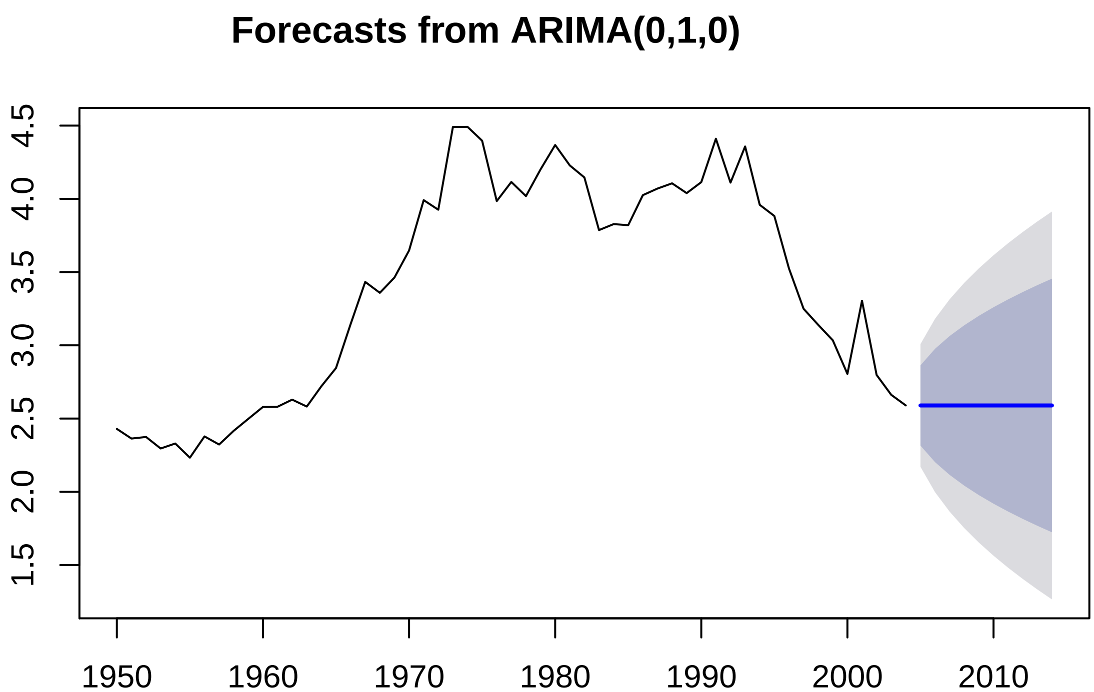
The model is not good, the pvalue is quite high.

e)Forecast three times ahead. Check your forecasts by hand to make sure you know how they have been calculated.

```
fc1 <- forecast(fit4)
```

f)Create a plot of the series with forecasts and prediction intervals for the next three periods shown.

```r
plot(fc1)
```



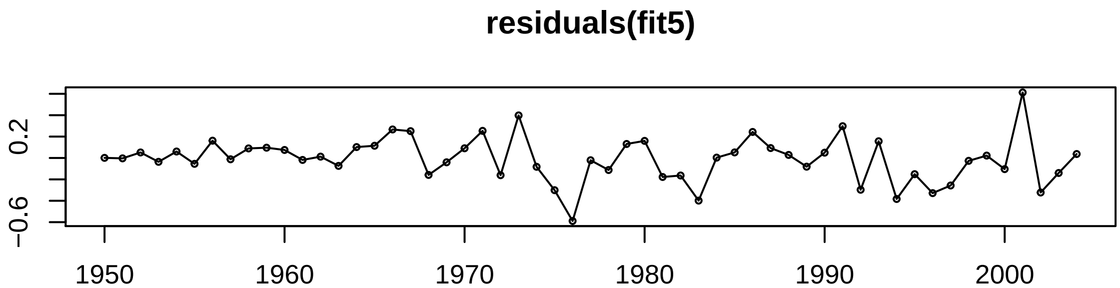
g)Does auto.arima give the same model you have chosen? If not, which model do you think is better?

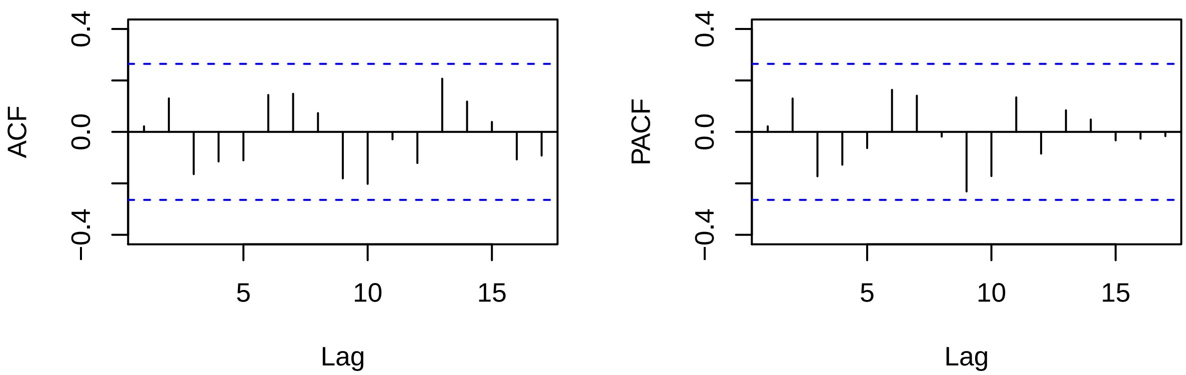
```r
fit5 <- auto.arima(wmurders,seasonal=FALSE)

summary(fit5)
```

```
## Series: wmurders
## ARIMA(1,2,1)
##
## Coefficients:
##           ar1      ma1
##       -0.2434  -0.8261
## s.e.   0.1553   0.1143
##
## sigma^2 estimated as 0.04457:  log likelihood=6.44
## AIC=-6.88   AICc=-6.39   BIC=-0.97
##
## Training set error measures:
##                       ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.01065956 0.2072523 0.1528734 -0.2149476 4.335214 0.9400996
##                     ACF1
## Training set 0.02176343
```

```r
tsdisplay(residuals(fit5))
```

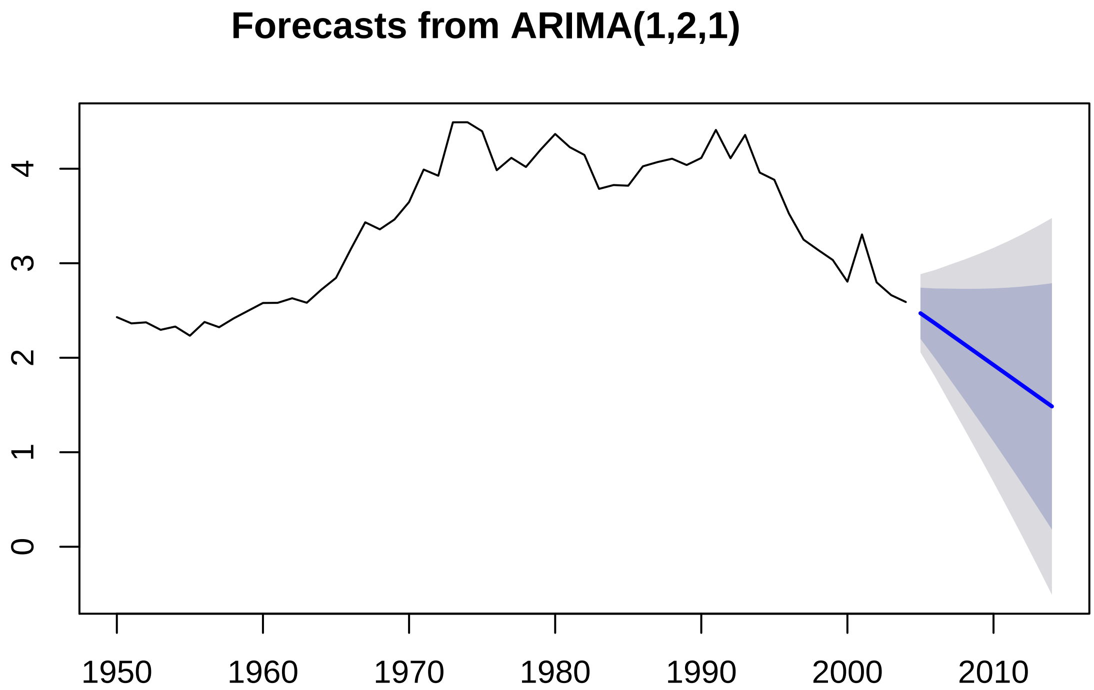





```
Box.test(residuals(fit5), lag=10, fitdf=2, type="L")
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit5)
## X-squared = 12.4191, df = 8, p-value = 0.1335
```

```
fc2 <- forecast(fit5)
plot(fc2)
```

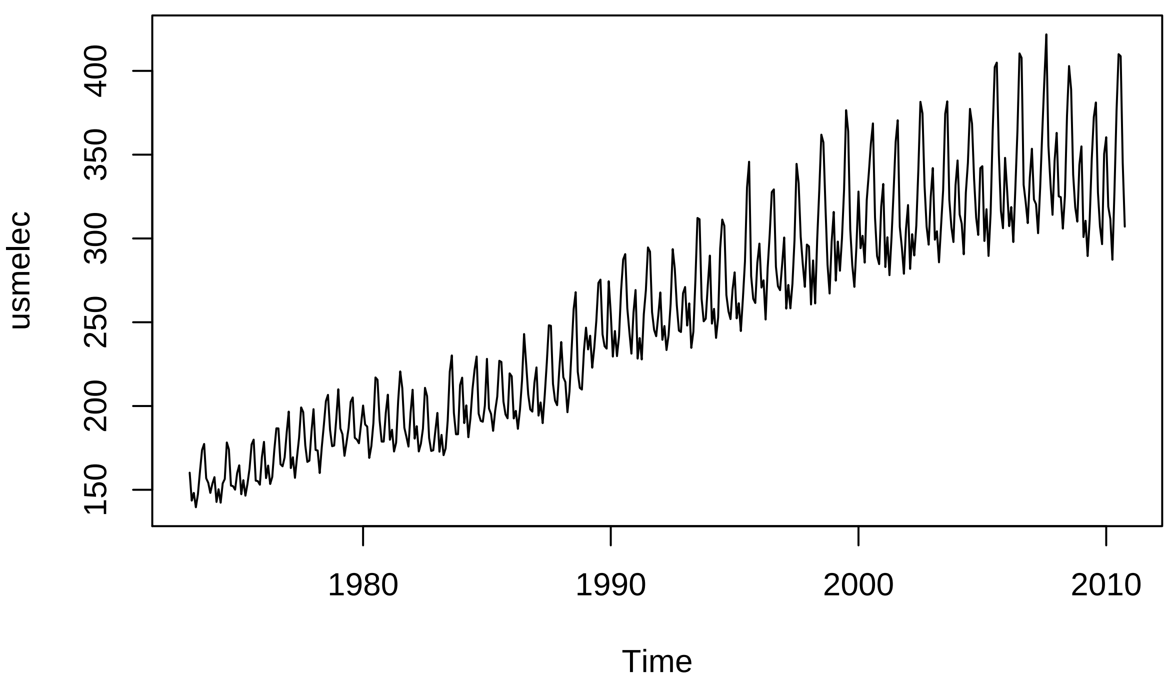


Auto Arima gives Arima(1,2,1). The autoarima model is better because it has a lower pvalue, lower RMSE, and a better forecast.

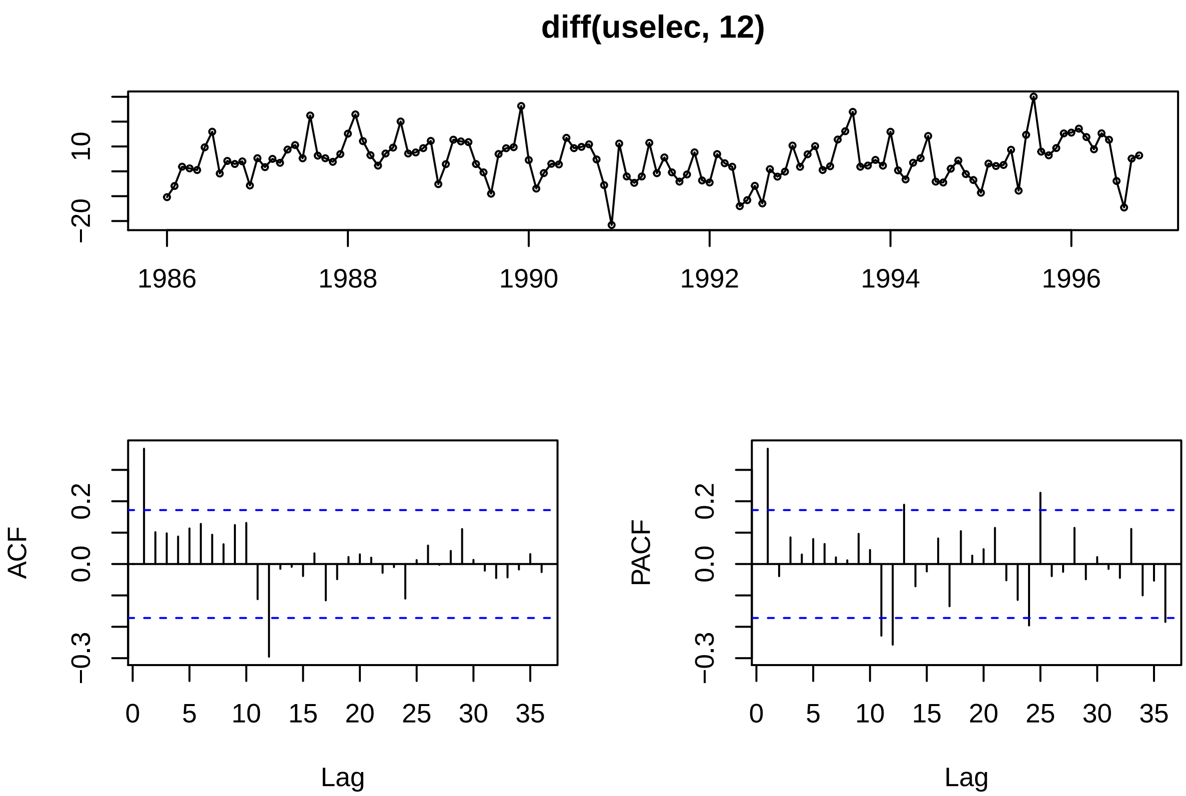
8.8 Consider the total net generation of electricity (in billion kilowatt hours) by the U.S. electric industry (monthly for the period 1985–1996). (Data set usmelec.) In general there are two peaks per year: in mid-summer and mid-winter.

a) Examine the 12-month moving average of this series to see what kind of trend is involved.

```
plot(usmelec)
```



```r
tsdisplay(diff(uselec,12))
```



Looks to be a seasonable upward trend of the data.

b) Do the data need transforming? If so, find a suitable transformation.

The data does not need to be transformed.

c) Are the data stationary? If not, find an appropriate differencing which yields stationary data.

The data looks stationary. ACF decreases to zero and there are values that are above the 95% threshold.

d) Identify a couple of ARIMA models that might be useful in describing the time series. Which of your models is the best according to their AIC values?

```r
fit6 <- auto.arima(usmelec)

summary(fit6)
```

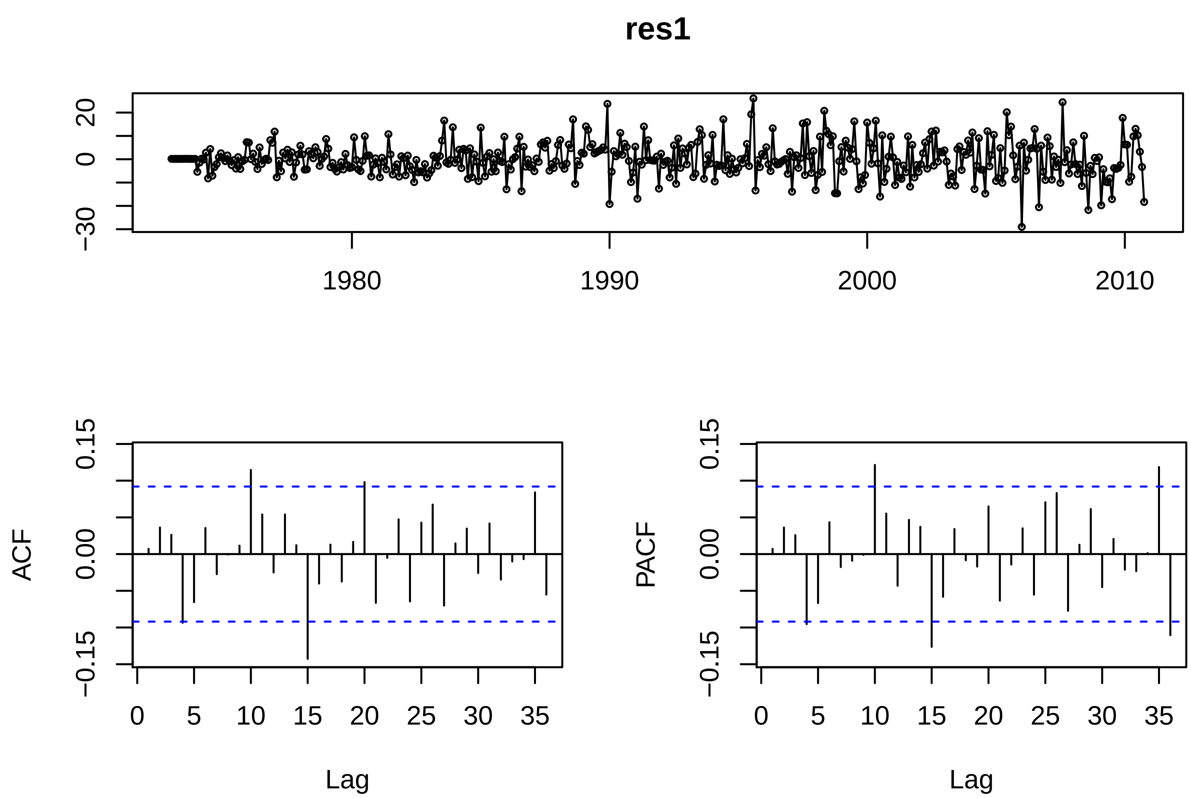
```
## Series: usmelec
## ARIMA(1,0,2)(0,1,1)[12] with drift
##
## Coefficients:
##          ar1      ma1      ma2     sma1   drift
##       0.9586  -0.4153  -0.2735  -0.7041  0.4268
## s.e.  0.0213   0.0529   0.0533   0.0317  0.0686
```

```
##
## sigma^2 estimated as 57.15:  log likelihood=-1478
## AIC=2968   AICc=2968.19   BIC=2992.55
##
## Training set error measures:
##                        ME     RMSE     MAE         MPE     MAPE      MASE
## Training set -0.006862698 7.357142 5.47164 -0.09417419 2.085705 0.6009912
##                     ACF1
## Training set 0.007114792
```

Two best models are (1,0,2) and (0,1,1) according to auto.arima.

e) Estimate the parameters of your best model and do diagnostic testing on the residuals. Do the residuals resemble white noise? If not, try to find another ARIMA model which fits better.

```
res1 <- residuals(fit6)
tsdisplay(res1)
```

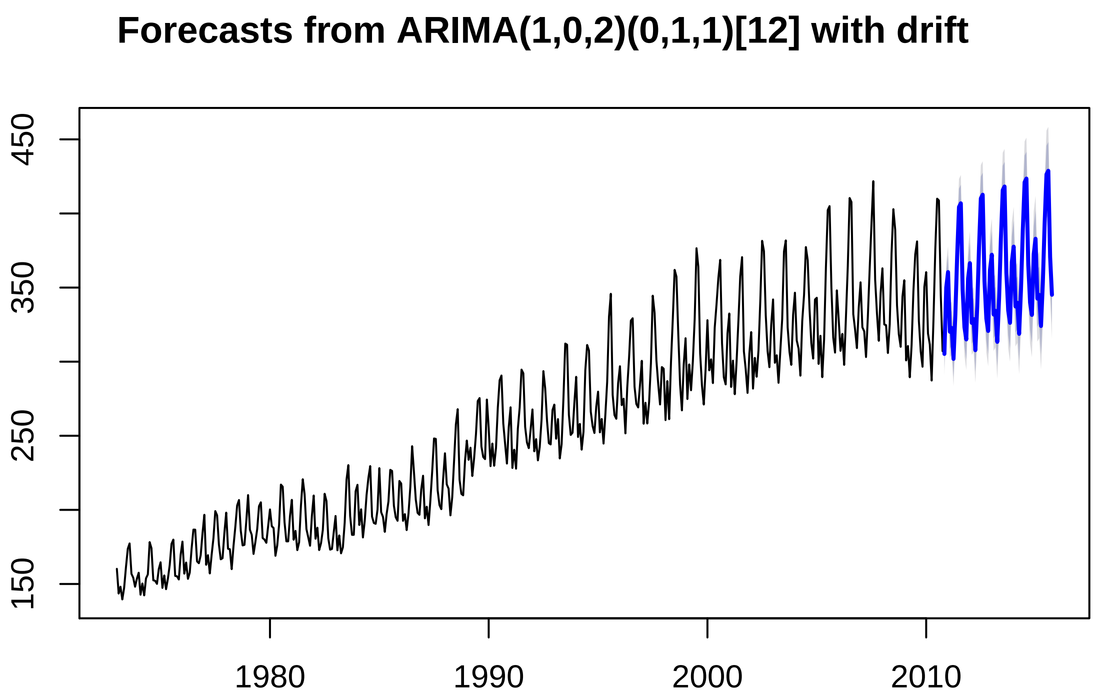



```
Box.test(res1, fitdf=4, lag=60, type="Lj")
```

```
##
##  Box-Ljung test
##
## data:  res1
## X-squared = 98.4912, df = 56, p-value = 0.0003937
```

f) Forecast the next 15 years of generation of electricity by the U.S. electric industry. Get the latest figures from http://data.is/zgRWCO to check on the accuracy of your forecasts.

```r
plot(forecast(fit6, h=60))
```



g) How many years of forecasts do you think are sufficiently accurate to be usable?

Since this data is stationary and seasonal with an upward trend, most forecasts should be relatively accurate. There is not a lot of variance and generally a person can see the way the market is going It seems that the way the market has actually performed http://data.is/zgRWCO is that there is a little bit of leveling off so 15 years is probably the upperbound of what is sufficiently accurate.