IS624 Week 6

Aaron Palumbo

Tuesday, July 14, 2015

Contents

HA - 8.1, 8.2, 8.6, 8.8

```r
library(knitr)
library(fpp)

# Suppress messages and warnings in all chuncks
opts_chunk$set(message=FALSE, warning=FALSE)
```

8.1

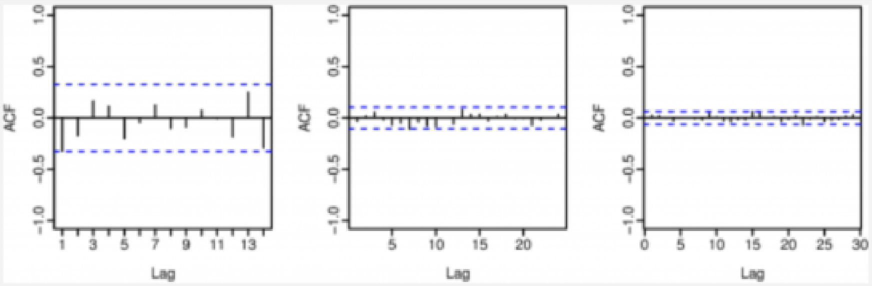


Figure 8.24: Left: ACF for a white noise series of 36 numbers. Middle: ACF for a white noise series of 360 numbers. Right: ACF for a white noise series of 1,000 numbers.

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

8.1 (a)

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

These figures show the correlation between different lags of the series (shown on the x axis). The y axis (the correlation) has the same scale for each plot, but the x axis shows an increasing number of lags as the series gets longer.

If the data are white noise (random) then we expect the correlations to be below the blue line, which indicates a significant lag.

For all the plots, the correlations of the lags shown are all below the significance level so they are all indicitive of white noise.

8.1 (b)

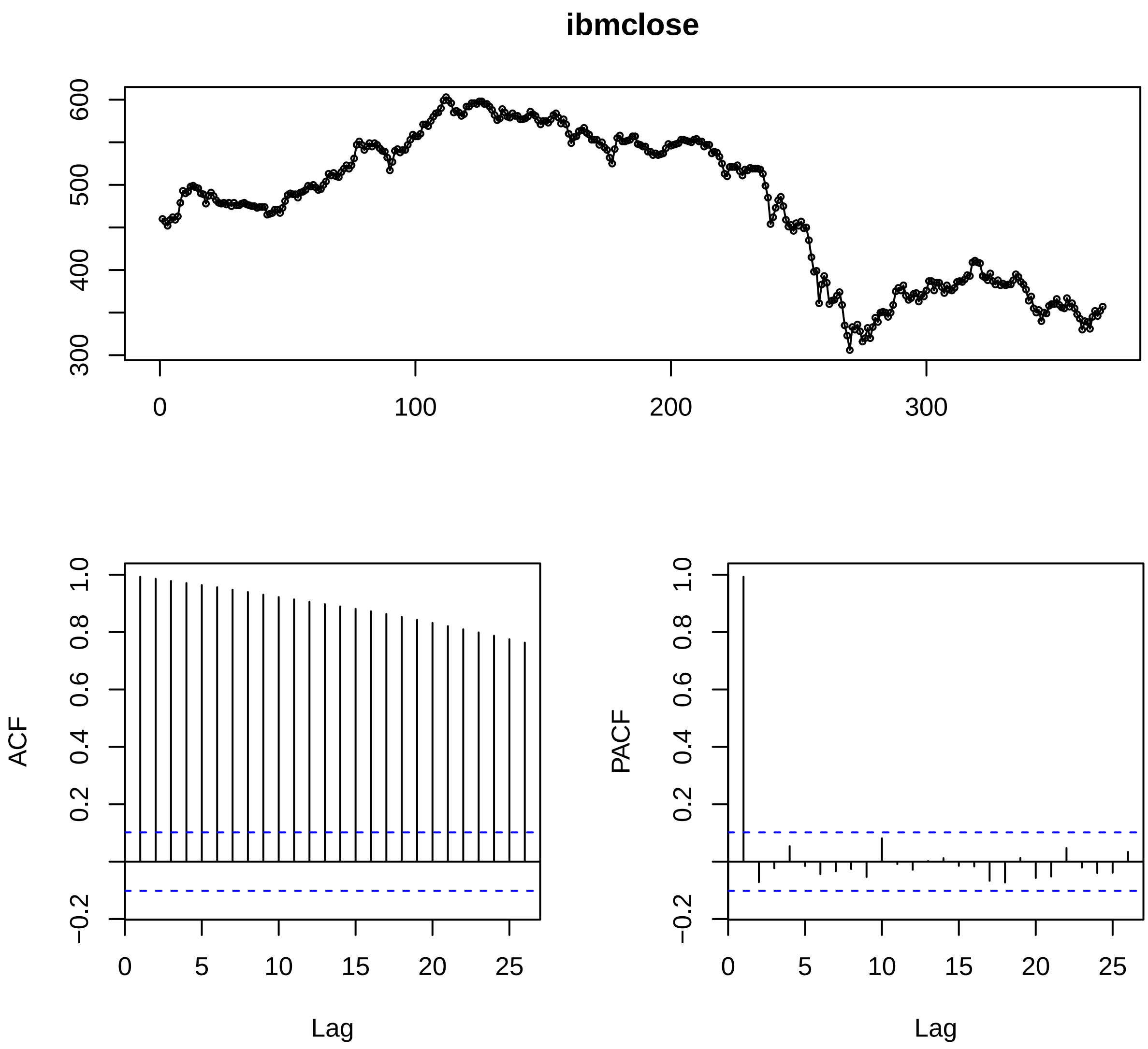
Why are the critical values at different distances from the mean of zero?

The value of a significant correlation is $\pm 1.96/\sqrt{T}$ where T is the number of data. From this we can see that as the number of data increase the value of a significant correlation decreases.

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.

```r
data(ibmclose)
tsdisplay(ibmclose)
```



When the ACF is slowly decaying, as it is in the graph above, it is a sign that the series may be auto regressive. We then look to the PACF to tell us of what degree. In this case we expect an AR(1) process that will need to be differenced once to be made stationary.

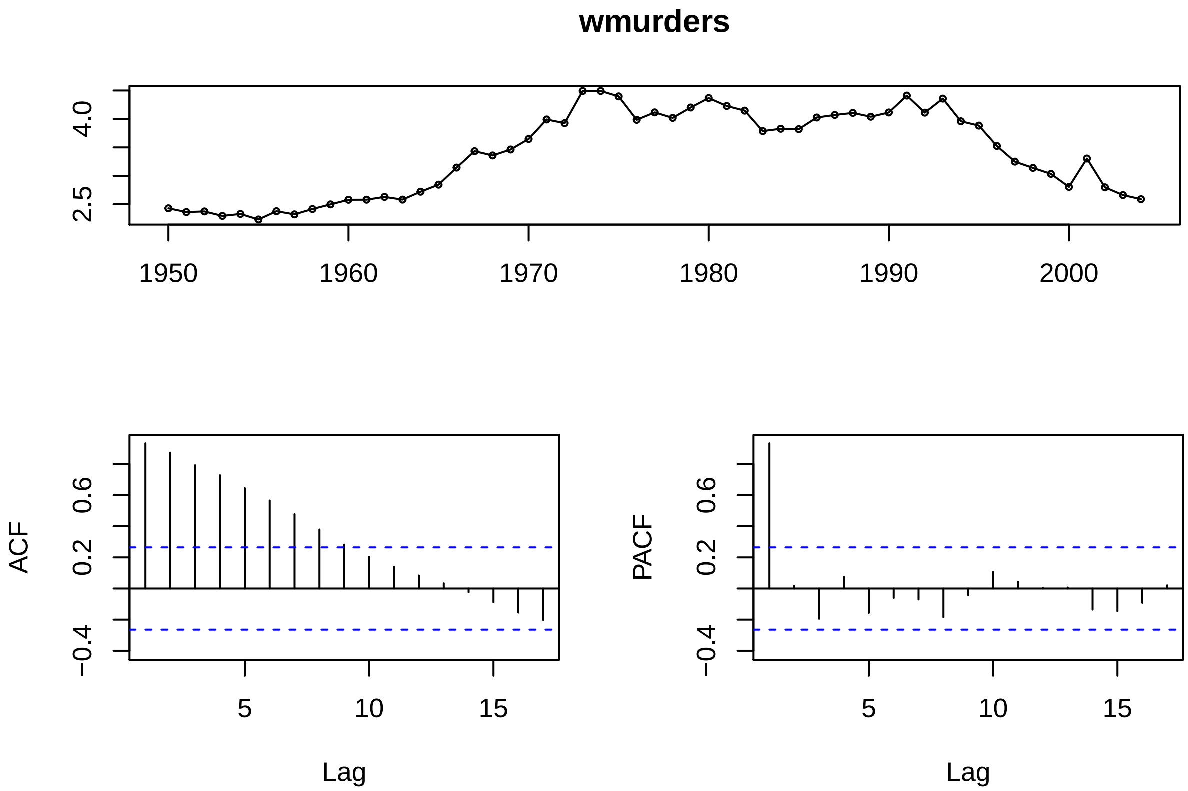
8.6

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

8.6 (a)

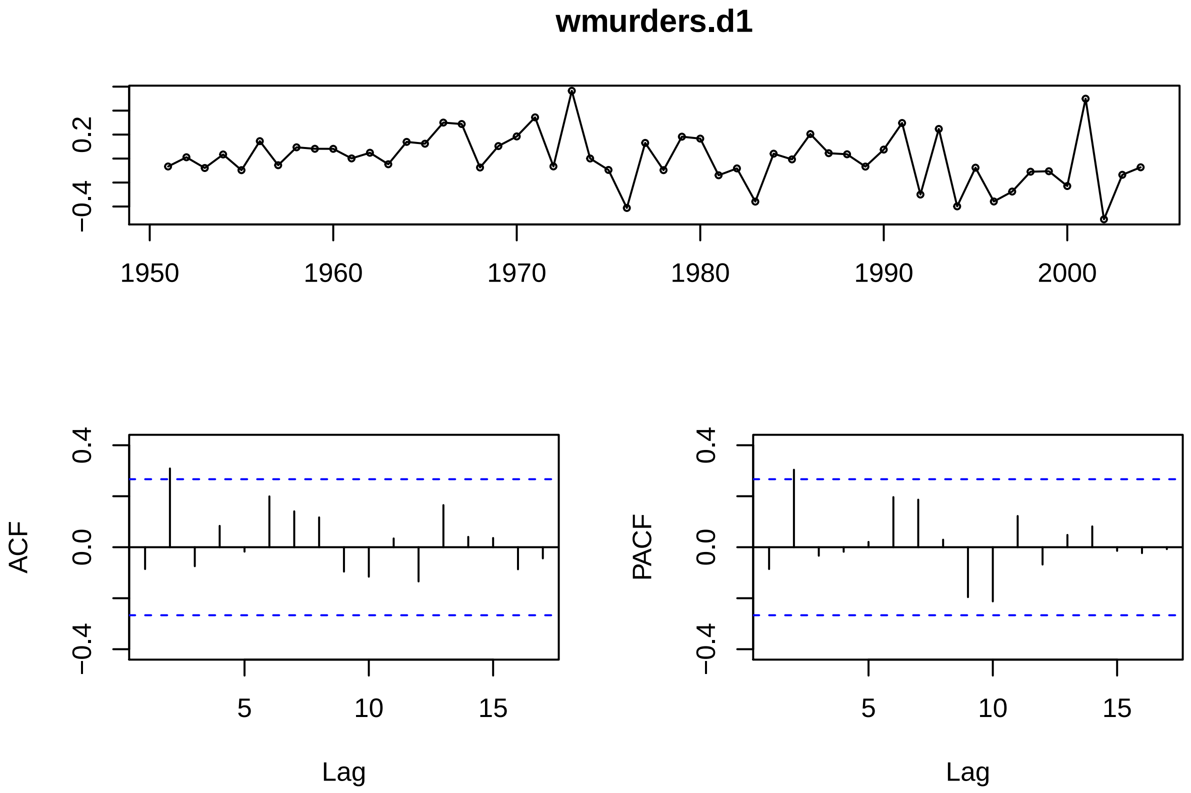
By studying appropriate graphs of the series in R, find an appropriate ARIMA$(\mathbf{p}, \mathbf{d}, \mathbf{q})$ model for these data.

```r
data(wmurders)

tsdisplay(wmurders)
```



This is clearly not stationary. Let's start with taking the first difference.

```r
wmurders.d1 <- diff(wmurders)
tsdisplay(wmurders.d1)
```



This looks much better, but the ACF and PACF still show some significant spikes at a lag of two. Let's try a unit root test:

```
adf.test(wmurders.d1)
```

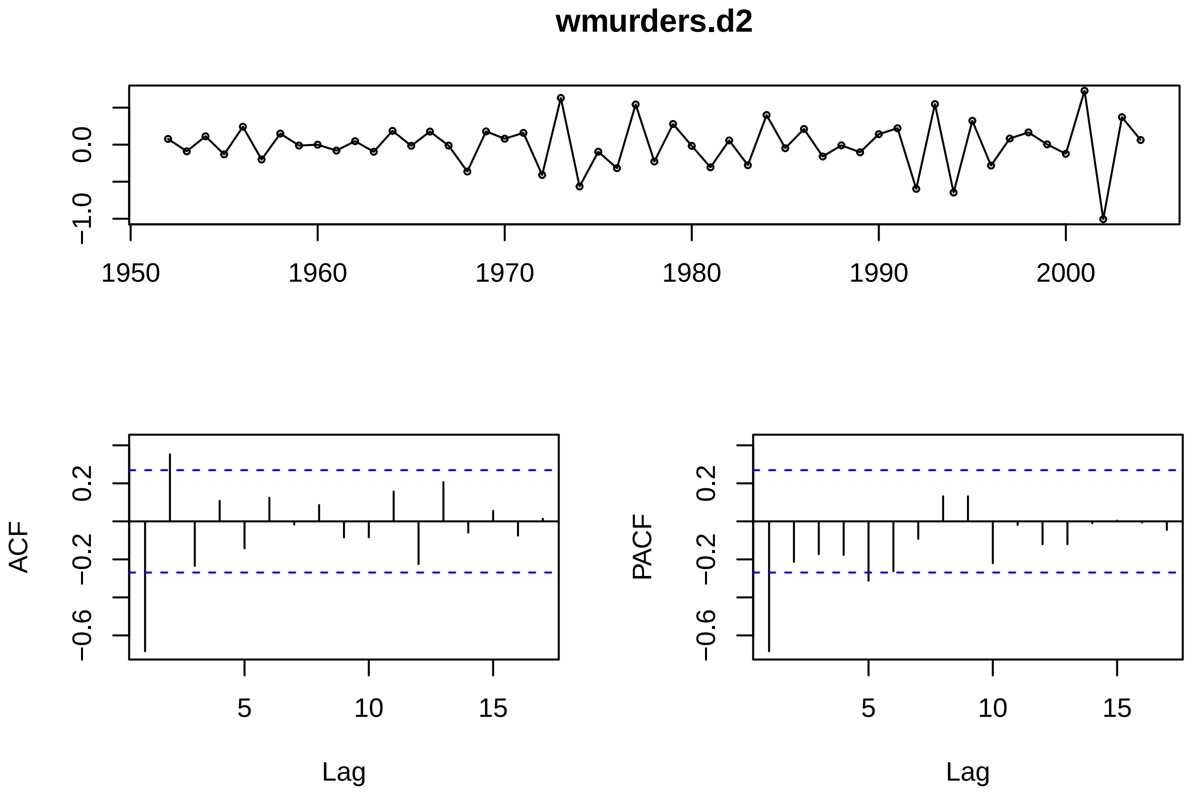
```
##
##  Augmented Dickey-Fuller Test
##
## data:  wmurders.d1
## Dickey-Fuller = -3.7688, Lag order = 3, p-value = 0.02726
## alternative hypothesis: stationary
```

```
kpss.test(wmurders.d1)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  wmurders.d1
## KPSS Level = 0.58729, Truncation lag parameter = 1, p-value =
## 0.02379
```

These tests are telling us different things. An ADF test < 0.5 indicates a stationary series, but a KPSS test < 0.5 indicates a non-stationary series. Take another difference and see what happens:

```r
wmurders.d2 <- diff(diff(wmurders))

tsdisplay(wmurders.d2)
```



```r
# unit root tests
adf.test(wmurders.d2)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  wmurders.d2
## Dickey-Fuller = -5.1646, Lag order = 3, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(wmurders.d2)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  wmurders.d2
## KPSS Level = 0.030483, Truncation lag parameter = 1, p-value = 0.1
```

Now both unit root tests tell us we have a stationary series.

Looking at the ACF and PACF plots, the large spike at 1 tells us we need either p or q to be 1. Let's start with $p = 1$ and test several ARIMA models in that neighborhood:

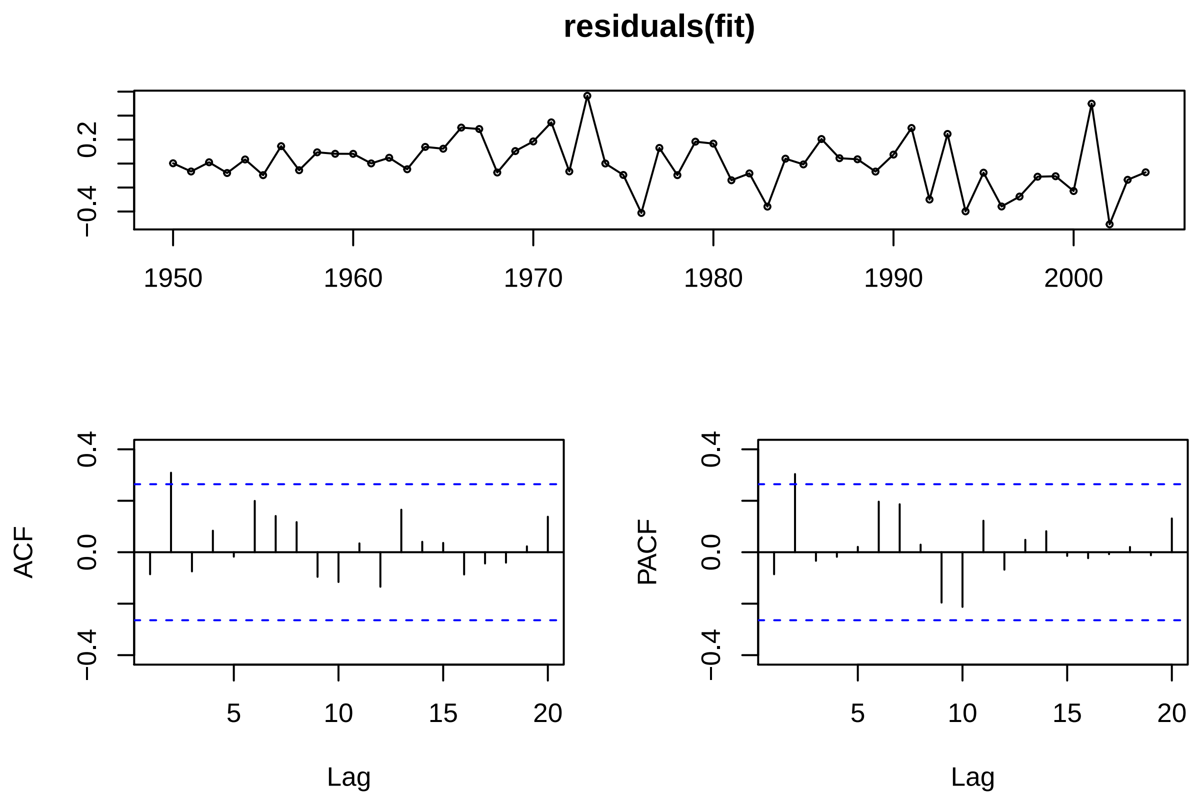
```r
test.arima <- function(t.series, order){
  df <- data.frame(model=paste0("ARIMA(",
                                paste0(order, collapse=","),
                                ")"),
                   AICc=Arima(t.series, order=order)$aicc)
  return(df)
}

df <- test.arima(wmurders.d2, c(1, 2, 0))
df <- rbind(df,
            test.arima(wmurders, c(1, 2, 1)),
            test.arima(wmurders, c(2, 2, 0)),
            test.arima(wmurders, c(1, 1, 1)),
            test.arima(wmurders, c(0, 1, 1)),
            test.arima(wmurders, c(1, 1, 0)),
            test.arima(wmurders, c(0, 1, 0)))
kable(df)
```

model	AICc
ARIMA(1,2,0)	90.2951663
ARIMA(1,2,1)	-6.3899721
ARIMA(2,2,0)	-0.1881352
ARIMA(1,1,1)	-9.8769992
ARIMA(0,1,1)	-9.4650495
ARIMA(1,1,0)	-9.6107488
ARIMA(0,1,0)	-11.3805724

Trying several models, including several with $d = 1$, we find the model with the lowest AICc to be ARIMA(0,1,0)

Let's fit this model and test the residuals:

```r
fit <- Arima(wmurders, order=c(0, 1, 0))
tsdisplay(residuals(fit), lag.max=20)
```



We see in the ACF and PACF the small spike at lag 2. Since it is only one this might be okay. Let's do a portmanteau test:

```
Box.test(residuals(fit), lag=24, fitdf=4, type="Ljung")
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit)
## X-squared = 23.348, df = 20, p-value = 0.272
```

The portmanteau test indicates the residuals are white noise so we conclude that the best model is ARIMA(0, 1, 0).

Note: The auto.arima function returns ARIMA(1, 2, 1) as the best model. We see in out table that the AICc for that model is -6.4 while the AICc for the model we selected is -11.4. As long as the residuals look okay, I don't see why we wouldn't go with the model with the lower AICc.

8.6 (b)

Should you include a constant in the model? Explain.

No. A constant introduces drift into the model, which we do not appear to have in these data.

8.6 (c)

Write this model in terms of the backshift operator.

$$(1 - B)\, y_t = 0$$