Forecasting and Time Series Analysis

San Cannon

Rockhurst University

Week 5

Introducing ARIMA models

Last week we saw how exponential smoothing models used trend and seasonality in the data to both forecast and model the data.

Tonight, we'll cover ARIMA models which we use to describe autocorrelations in the data.

Before we do that, we need to discuss a very important topic: **stationarity**

Stationarity

Stationary time series: A time series whose properties do not depend on the time at which the series is observed

Why do we care if a series is stationary or not?

The usual hypothesis tests for regression coefficients do not work when the data are non-stationary.

The p-values from our linear regressions are worthless if the series are not stationary.

Seeking stationarity

Time series with trends, or with seasonality, are not stationary. Home sales, GDP, air passengers are all **non stationary** time series.

But a time series with cyclic behavior (but not trend or seasonality) is stationary.

That is because the cycles are not of fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.

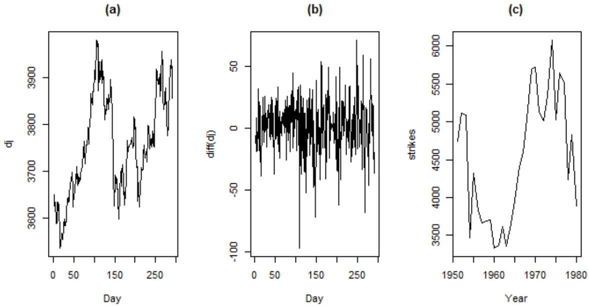
This can be very confusing.

More details

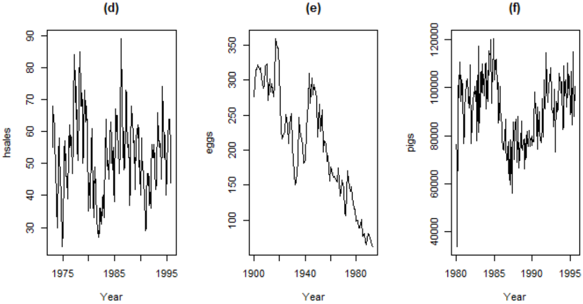
- For seasonal time series, we might know that the series is always higher in the fourth quarter (retail sales). That means that the value is related to the time subscript.
- For a time series with trend, we know that y_t will be less than (or greater than) y_{t+1} so the value is related the time subscript.
- For series that peaks every 8-10 years, we can't really use the time subscript to describe the value. (Which is why it is so hard to pull a "cyclical" component out of time series when doing a decomposition).

In general, a stationary time series will have no obvious observable patterns in the long-term. And by obvious, we mean easily describable in terms of t

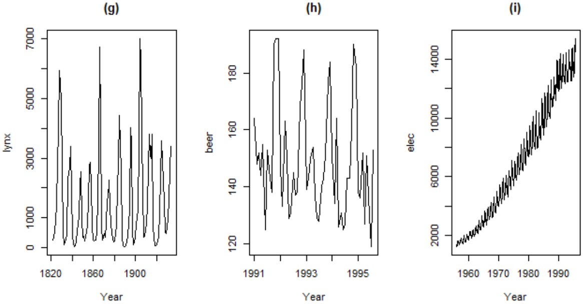
Pictures!


(a)
dj
3600
3700
3800
3900
0
50
150
250
Day
(b)
diff(dj)
-100
-50
0
50
0
50
150
250
Day
(c)
strikes
3500
4000
4500
5000
5500
6000
1950
1960
1970
1980
Year

Pictures!


(d)
hsales
30
40
50
60
70
80
90
1975
1985
1995
Year
(e)
eggs
100
150
200
250
300
350
1900
1940
1980
Year
(f)
pigs
40000
60000
80000
100000
120000
1980
1985
1990
1995
Year

Pictures!


(g)
lynx
0
1000
3000
5000
7000
1820
1860
1900
Year
(h)
beer
120
140
160
180
1991
1993
1995
Year
(i)
elec
2000
6000
10000
14000
1960
1970
1980
1990
Year

How do you know if a series is stationary?

Look at the time plot of the data: the series should be roughly horizontal (although some cyclic behavior is possible) with constant variance.

The ACF plot is also useful for identifying non-stationary time series.

For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly. Also, for non-stationary data, the value of r_1 is often large and positive.

Remember the random walk?

The differenced series is the **change** between consecutive observations in the original series, and can be written as

$$y'_t = y_t - y_{t-1}.$$

The differenced series will have only $T - 1$ values since it is not possible to calculate a difference y'_1 for the first observation.

When the differenced series is white noise, the model for the original series can be written as

$$y_t - y_{t-1} = e_t \quad \text{or} \quad y_t = y_{t-1} + e_t \; .$$

So?

A random walk model is very widely used for non-stationary data, particularly financial and economic data.

Random walks typically have:

- long periods of apparent trends up or down
- sudden and unpredictable changes in direction.

The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down. Thus, the random walk model underpins naive forecasts.

Adding drift

A closely related model allows the differences to have a non-zero mean. Then

$$y_t - y_{t-1} = c + e_t \quad \text{or} \quad y_t = c + y_{t-1} + e_t \ .$$

The value of c is the average of the changes between consecutive observations. If c is positive, then the average change is an increase in the value of y_t. Thus y_t will tend to drift upwards. But if c is negative, y_t will tend to drift downwards.

This is the model behind the drift method.

Making stationarity

Text

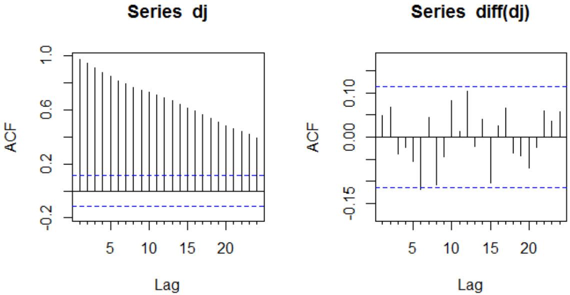
To create a stationary version of a non-stationary time series, you need to transform the data.

Most of the time, you can use **differencing** to achieve this. Subtracting one value from the neighboring value can help stabilize the mean of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality.

Sometimes, logarithms can help stabilize the mean of a series.

Remember our discussion of the Dow Jones Index versus the change in the index?

Pictures!



Series dj
ACF
-0.2
0.2
0.6
1.0
5
10
15
20
Lag
Series diff(dj)
ACF
-0.15
0.00
0.10
5
10
15
20
Lag

What if the differencing doesn't help

Subtracting y_{t-1} from y_t is known as **first differencing**. What if the first differenced series *still* isn't stationary?

Occasionally the differenced data will not appear stationary and it may be necessary to difference the data a second time to obtain a stationary series:

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}. \end{aligned}$$

In this case, y''_t will have $T - 2$ values. Then we would model the *change in the changes* of the original data. In practice, it is almost never necessary to go beyond second-order differences.

What about seasonal data?

For data with strong seasonal patterns, first or second differencing might not be appropriate. We may need to use **seasonal differencing**

A seasonal difference is the difference between an observation and the corresponding observation from the previous year. So

$$y'_t = y_t - y_{t-m} \qquad \text{where } m = \text{number of seasons.}$$

These are also called **lag-m differences** as we subtract the observation after a lag of m periods.

So?

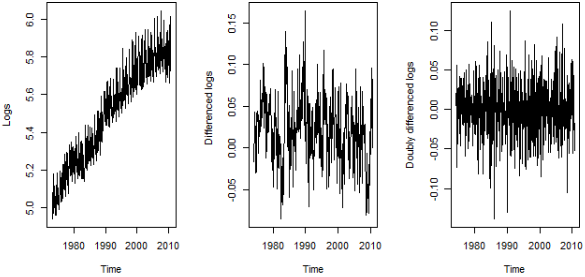
If seasonally differenced data appear to be white noise, then an appropriate model for the original data is

$$y_t = y_{t-m} + e_t.$$

Forecasts from this model are equal to the last observation from the relevant season.
This is what we saw with the seasonal naive forecasts.

Picture!

Monthly net US electricity generation.



Can we combine them?

Yes... but:

If $y'_t = y_t - y_{t-m}$ denotes a seasonally differenced series, then the twice-differenced series is

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-m}) - (y_{t-1} - y_{t-m-1}) \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1} \,. \end{aligned}$$

Does order matter?

When both seasonal and first differences are applied, it makes no difference which is done first —the result will be the same.

But how do you know you need both? If the data have a strong seasonal pattern, it might be that the seasonally differenced series will be stationary and you can stop there.

You're more likely to see seasonality in the first differenced series if you do that first.

You can, but should you?

Transforming data to make a series stationary isn't helpful if you can't interpret the results.

Yes, the lag-19 differenced series is stationary - but what does that mean?

- First differences are the change between one observation and the next. (Interpretation?)
- Second differences are the change in the change. (Interpretation?)
- Seasonal differences are the change between one year to the next. (Interpretation?)

Other lags are unlikely to make much interpretable sense and should be avoided.

Tests for stationarity

Is looking at pictures the only way to know if your transformation worked?

Of course not: we have statistical test!

Unit root tests test the hypothesis of stationarity and are designed for determining whether differencing is required.

What's a unit root?

Consider this data process: $y_t = ay_{t-1} + e_t$
We can write it as: $y_t - ay_{t-1} = e_t$

- If $|a| < 1$, the process is stationary
- If $|a| > 1$, the process is explosive and non stationary
- If $|a| = \cancel{-}1$, the process is a random walk and therefore non-stationary

$|a| = 1$ is the **unit root** and it forms the boundary between stationary and non-stationary.
Therefore: testing for a unit root is the test for stationarity.

ADF

One of the most popular tests is the **Augmented Dickey-Fuller test** where the following regression model is estimated:

$$y'_t = \phi y_{t-1} + \beta_1 y'_{t-1} + \beta_2 y'_{t-2} + \cdots + \beta_k y'_{t-k},$$

where y'_t denotes the first-differenced series, $y'_t = y_t - y_{t-1}$ and k is the number of lags to include in the regression.

If the original series, y_t, needs differencing, then the coefficient $\hat{\phi}$ should be approximately zero. If y_t is already stationary, then $\hat{\phi} < 0$.

How does this work?

The null-hypothesis for an ADF test is that there is a unit root so the data are non-stationary.

Large p-values are mean that we reject the null of non-stationarity, small p-values suggest stationarity.

Using the usual 5% threshold, differencing is required if the p-value is greater than 0.05.

Mathematical aside: backshift notation

The backward shift operator B is a useful notational device when working with time series lags:

$$By_t = y_{t-1}.$$

Sometimes you'll see L for *lag* instead of B for *backshift* In other words, B, when applied to y_t, shifts the data back one period. Two applications of B to y_t shifts the data back two periods:

$$B(By_t) = B^2 y_t = y_{t-2}.$$

For monthly data, if we wish to consider *the same month last year*, the notation is $B^{12} y_t = y_{t-12}$.

Differencing with a B

The backward shift operator is convenient for describing the process of differencing. A first difference can be written as

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t \, .$$

Note that a first difference is represented by $(1 - B)$. Which means second order differences are

$$y''_t = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1 - B)^2 y_t \, .$$

In general, a dth-order difference can be written as

$$(1 - B)^d y_t.$$

Warning: algebra!

Backshift notation is very useful when combining differences as the operator can be treated using ordinary algebraic rules. In particular, terms involving B can be multiplied together.

For example, a seasonal difference followed by a first difference can be written as

$$\begin{aligned}(1-B)(1-B^m)y_t &= (1-B-B^m+B^{m+1})y_t \\ &= y_t - y_{t-1} - y_{t-m} + y_{t-m-1},\end{aligned}$$

the same result we obtained earlier.

For the math heads:

From the unit root discussion: $y_t - ay_{t-1} = e_t$

Using the backshift operator we can say $By_t = y_{t-1}$

Which means we can write: $y_t - aBy_t = e_t$ or $(1 - aB)y_t = e_t$

So the characteristic equation is $1 - ay$ which has a (unique) root at $y = 1/a$

Which is where we get unit root.

If this doesn't make any sense, ignore it.

Autoregressive models

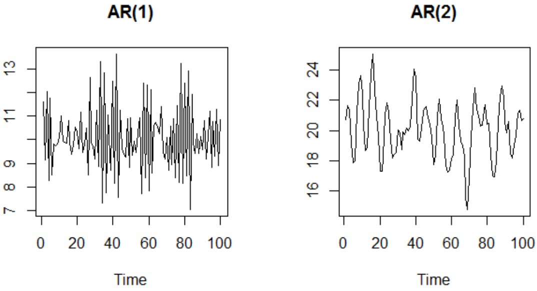
An autoregressive model of order p, denoted AR(p) model, can be written as

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t,$$

where c is a constant and e_t is white noise.

It's multiple regression but with **lagged values** of y_t as predictors.

Pictures!



Left: AR(1) with $y_t = 18 - 0.8y_{t-1} + e_t$.
Right: AR(2) with $y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + e_t$.
e_t is normally distributed white noise with mean zero and variance one.

AR(1) models

AR(1) model:

$$y_t = c + \phi_1 y_{t-1} + e_t,$$

You've seen some of these before:

- $\phi_1 = 0$: y_t is equivalent to white noise.
- $\phi_1 = 1$ and $c = 0$: y_t is equivalent to a random walk.
- $\phi_1 = 1$ and $c \neq 0$: y_t is equivalent to a random walk with drift
- $\phi_1 < 0$: y_t tends to oscillate between positive and negative values.

Moving average

A moving average model (**MA(q) model**) uses past forecast errors as predictors:

$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$,
where e_t is white noise.

Do not confuse moving average **smoothing** with moving average **models**

Just in case

The moving average smoothing averages the nearest order periods of each observation. As neighboring observations of a time series are likely to be similar in value, averaging eliminates some of the jitters in the data, leaving a smooth trend-cycle component.[1]

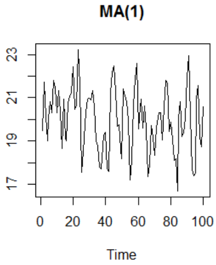
$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j}$

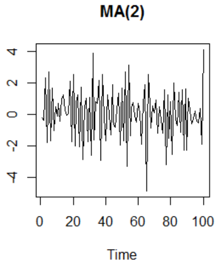
where $k = \frac{m-1}{2}$

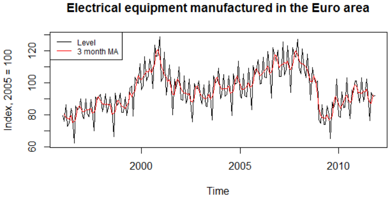
When an even order is specified, the observations averaged will include one more observation from the future than the past (k is rounded up). If center is TRUE, the value from two moving averages (where k is rounded up and down respectively) are averaged, centering the moving average.

[1] Taken from the `ma` function documentation in the package `forecast`

Pictures!


MA(1)


MA(2)


Electrical equipment manufactured in the Euro area

ARMA models

We can combine the AR and MA process and model the data as an ARMA process:
$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t,$

This is an ARMA(p, q) model

- Predictors include both lagged values of y_t and lagged errors (e_t).
- ARMA models can be used for a huge range of stationary time series.
- They model the short-term dynamics.

ARIMA models

An ARMA model applied to **differenced** data is an **ARIMA** model. (The `i` stands for `Integrated`)

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t,$$

where y'_t is the differenced series (it may have been differenced more than once). The predictors on the right hand side include both lagged values of y_t and lagged errors.

We call this an ARIMA(p, d, q) model, where $p =$ order of the autoregressive part; $d =$ degree of first differencing involved; $q =$ order of the moving average part.

In other words

Once we start combining components in this way to form more complicated models, it is much easier to work with the backshift notation.

$$\begin{array}{rl} \overset{\substack{\text{AR}(p)\\ \downarrow}}{(1-\phi_1 B - \cdots - \phi_p B^p)} & \overset{\substack{d \text{ diffs}\\ \downarrow}}{(1-B)^d}\, y_t \\ = & c + \underset{\substack{\uparrow \\ \text{MA}(q)}}{(1+\theta_1 B + \cdots + \theta_q B^q)}\, e_t \end{array}$$

Selecting appropriate values for p, d and q can be difficult. The `auto.arima()` function in R will do it for you automatically.

Many of the models we have already discussed are special cases of the ARIMA:

White noise	⇒	ARIMA(0,0,0)
Random walk	⇒	ARIMA(0,1,0) with no constant
Random walk with drift	⇒	ARIMA(0,1,0) with a constant
Autoregression	⇒	ARIMA(p,0,0)
Moving average	⇒	ARIMA(0,0,q)

Dow Jones again

We know that the DJ is non-stationary. Here's the output from `auto.arima` function in R

```
Series: dj
ARIMA(1,1,1)

Coefficients:
         ar1      ma1
      0.8510  -0.5263
s.e.  0.1383   0.2548

sigma^2 estimated as 0.1474:  log likelihood=-34.69
AIC=75.38   AICc=75.71   BIC=82.41
```

What does this mean? What's with the (1,1,1)?

First differences

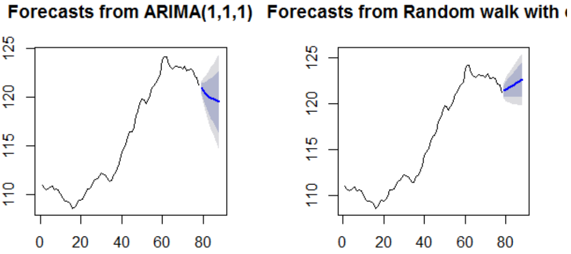
What happens if we take first differences of Dow Jones? Then the output from `auto.arima` function in R is:

```
Series: diff(dj)
ARIMA(1,0,1) with zero mean

Coefficients:
         ar1      ma1
      0.8510  -0.5263
s.e.  0.1383   0.2548

sigma^2 estimated as 0.1472:  log likelihood=-34.69
AIC=75.38   AICc=75.71   BIC=82.41
```

What do you notice?

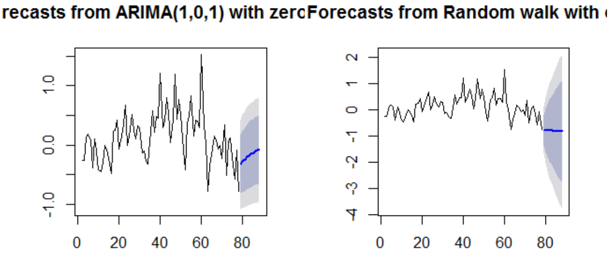
Forecasting

Not only will the information from the ARIMA model explain our time series, we can use it to forecast as well.



Remember what you are forecasting!

Why does this look different?



Understanding ARIMA models

The constant c has an important effect on the long-term forecasts obtained from these models.

- If $c = 0$ and $d = 0$, the long-term forecasts will go to zero.
- If $c = 0$ and $d = 1$, the long-term forecasts will go to a non-zero constant.
- If $c \neq 0$ and $d = 0$, the long-term forecasts will go to the mean of the data.
- If $c \neq 0$ and $d = 1$, the long-term forecasts will follow a straight line.

The value of d also has an effect on the prediction intervals the higher the value of d, the more rapidly the prediction intervals increase in size.

Healthy skepticism

Does anything automatic make you nervous? Maybe it should.

The `auto.arima()` is great for picking out the AR and MA terms but that doesn't absolve you from any critical thought.

You can (and should?) check to see if what comes out of the function makes sense.

You might have a reason to use a different order than the function delivers.

You can get some information from the ACF and PACF plots.

Looking at pictures

It is usually not possible to tell, simply from a time plot, what values of p and q are appropriate for the data. However, it is sometimes possible to use the ACF plot, and the closely related PACF plot, to determine appropriate values for p and q.

Recall that an ACF plot shows the autocorrelations which measure the relationship between y_t and y_{t-k} for different values of k. Now if y_t and y_{t-1} are correlated, then y_{t-1} and y_{t-2} must also be correlated. But then y_t and y_{t-2} might be correlated, simply because they are both connected to y_{t-1}.

Wait? *PACF*?

The **partial autocorrelations** measure the relationship between y_t and y_{t-k} after removing the effects of other time lags – $1, 2, 3, \ldots, k-1$. So the first partial autocorrelation is identical to the first autocorrelation, because there is nothing between them to remove. The partial autocorrelations for lags 2, 3 and greater are calculated as follows:

$$\begin{aligned}\alpha_k &= k\text{th partial autocorrelation coefficient}\\ &= \text{the estimate of } \phi_k \text{ in the autoregression model}\\ y_t &= c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + e_t.\end{aligned}$$

Varying the number of terms on the right hand side of this autoregression model gives α_k for different values of k.

ARIMA in ACF or PACF

The data may follow an ARIMA$(p, d, 0)$ model if the ACF and PACF plots of the differenced data show the following patterns:

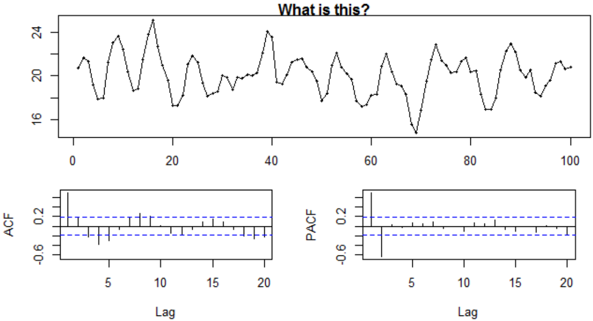
- the ACF is exponentially decaying or sinusoidal[2];
- there is a significant spike at lag p in PACF, but none beyond lag p.

The data may follow an ARIMA$(0, d, q)$ model if the ACF and PACF plots of the differenced data show the following patterns:

- the PACF is exponentially decaying or sinusoidal;
- there is a significant spike at lag q in ACF, but none beyond lag q.

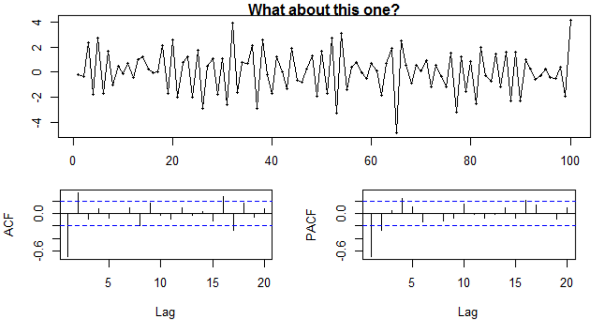
[2] For example Thanks to http://physics.stackexchange.com/questions/36114/sinusoidal-wave-displacement-function

Pop quiz 1



Answer

Pop quiz 2



Answer

MLE

Once the model order has been identified (i.e., the values of p, d and q), we need to estimate the parameters c, $\phi_1, \dots, \phi_p$, $\theta_1, \dots, \theta_q$. When R estimates the ARIMA model, it uses *maximum likelihood estimation* (MLE).

This technique finds the values of the parameters which maximize the probability of obtaining the data that we have observed.

For ARIMA models, MLE is very similar to the *least squares* estimates that would be obtained by minimizing

$$\sum_{t=1}^{T} e_t^2.$$

Step by step

1. Plot the data. Identify any unusual observations.
2. If necessary, transform the data to stabilize the variance.
3. If the data are non-stationary: take first differences of the data until the data are stationary.
4. Examine the ACF/PACF: Is an AR(p) or MA(q) model appropriate?
5. Try your chosen model(s), and use the AICc to search for a better model. (Or try auto.arima() and verify the recommendation.)
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals. If they do not look like white noise, try a modified model.
7. Once the residuals look like white noise, calculate forecasts.

Seasonal ARIMA models

Step two on the checklist was to transform the data to “stabilize the variance” - when is this an issue? One instance is when there is seasonality in the data.

We can apply ARIMA modeling to seasonal data too. We need to include additional seasonal terms in the ARIMA models we have seen so far.

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Non-seasonal part}\\ \text{of the model}\end{array}\right)}}\ \underbrace{(P,D,Q)_m}_{\substack{\uparrow \\ \left(\begin{array}{l}\text{Seasonal part}\\ \text{of the model}\end{array}\right)}}$$

where $m=$ number of periods per season. Following the FPP book, use uppercase notation for the seasonal parts of the model, and lowercase notation for the non-seasonal parts of the model.

Adding the seasonal stuff

The calculations look very similar with lots of scary backshift notation using m to indicate how far back to shift based on the seasonality. And we can still learn things from our ACF/PACF pictures:

The seasonal part of an AR or MA model will be seen in the seasonal lags of the PACF and ACF. For example, an ARIMA(0,0,0)(0,0,1)$_{12}$ model will show:

- a spike at lag 12 in the ACF but no other significant spikes.
- - The PACF will show exponential decay in the seasonal lags; that is, at lags 12, 24, 36, ...

Similarly, an ARIMA(0,0,0)(1,0,0)$_{12}$ model will show:

- exponential decay in the seasonal lags of the ACF
- single significant spike at lag 12 in the PACF.

ARIMA vs ETS

There seems to be an idea that ARIMA models are "better" than exponential smoothing. Let's check the score.

- Linear exponential smoothing models are all special cases of ARIMA models (point for both)
- Non-linear smoothing models have no equivalent ARIMA counterparts (Point for exponential smoothing)
- There are many ARIMA models that have no exponential smoothing versions (Point for ARIMA models)
- Every ETS model is non-stationary but ARIMA models can be stationary (Point to ARIMA)
- WAIT! ETS models can be made stationary either by first differencing or second differencing (if there is seasonality and/or non-damped trend) (Subtract point from ARIMA)

Summary

Topics for tonight:

- Stationarity and differencing
- Unit roots and unit root tests (ADF statistics)
- Backshift notation and ugly algebra
- AR modeling
- MA modeling
- ARMA and (nonseasonal) ARIMA
- ACF and PACF for ARIMA modeling
- Seasonal ARIMA
- ETS vs ARIMA

And now, we code.

Pop quiz answer 1

AR(2) with $y_t = 8 + 1.3y_{t-1} - 0.7y_{t-2} + e_t$.

Back

Pop quiz answer 2

MA(2) with $y_t = e_t - e_{t-1} + 0.8e_{t-2}$

Back