Time Series Introduction

Overview

A time-series is a set of observations on the same object over time. Classically, these observations are assumed to be at regular (i.e. evenly spaced) intervals. (Not a necessary feature, but a useful simplifying assumption.) The primary feature of time series data that distinguishes it from other data is the correlation between measurements.

For example, suppose we record the daily high temperature at UCLA every day for 10 years. Obviously, if it was unusually hot on one day, it is likely to be warm the next as well. Not only that, but temperature varies seasonally, and so if you know that one measurement occurred in, say, Winter, then you know quite a bit about what temperatures are likely to be seen.

The goal in a time series analysis is often prediction. This is attained by choosing and then fitting a good model for the time series. Models have, typically, three components:

- a trend component: is there a general upward/downward trend?
- a seasonal component: are there regular cycles to the data?
- a model for the correlation structure of the data

Earlier in the quarter I said that statistical models consist of two components: one models the deterministic aspect, the other component measures the stochastic aspect. For example, in regression, we say $Y = a + bx + E$. The deterministic part, $a + bx$, tells us that the response Y is linearly dependent on the predictor x. The stochastic part tells us, then, that E is a random error, and further describes the distribution of this error, usually Normal, with mean 0, fixed standard deviation, and independent observations.

The trend and seasonal component of a time-series together describe its deterministic aspects. Once you've explained the trend and the seasonal component, what's left over is stochastic (i.e. random), and as in regression we describe the deviations in terms of a probability distribution, but no longer can we assume the observations are indepedent.

There is only one way for observations to be independent, but there are many ways for them to be dependent. For example, perhaps the correlation between today's temperature and tomorrow's is .8. But the correlation between today's and two days from now will probably be smaller. How much smaller? Eventually, the correlation between today and some day t days in the future will be effectively 0. How rapidily does the correlation fall off? Often, a time series model specifies a functional form which governs the "decay" of correlation.

There are two general approaches to time-series. One is called the time-domain, the other the frequency domain. Analyses in the time-domain are the most straight-forward intuitively, but can be awkward. The general stratgegy is to first fit the trend and season (sometimes called de-trending), and then model the correlation structure on the de-trended, de-seasonalized data. In

many cases, this is more art than science, since the data often do not suggest any natural ways for determining the trend or seasonal variations. Thus, one hopes you are guided by a theory. For example, knowledge of the length of the sunspot cycle can be used to help de-seasonalize data that might depend on sunspot activity. And of course, we know that daily temperature data have an annual cycle which should be removed before fitting the stochastic component.

Frequency domain analyses essentially use fourier transformations (modern analysts might use other similar techniques) to break a time series into its cyclical frequencies. The analysis is a little more direct here, but the intuition is not always as strong.

Time Domain

We'll start out with time-domain analyses. First, some terminology.

Mean function: mu(t) = E(X(t)) (trend)
Variance function: Var(X(t))
Autocovariance: Gamma(t1, t2) = Cov(X(t1), X(t2)) = E(X(t1) - mu(t1))(X(t2) - mu(t2))
Essentially, everything is now allowed to vary with time. Covariance now is a two-dimensional function that describes the covariance between any two points in time.

Note1: Gamma(t1,t1) is Var(X(t1)).

Note2: Intuitively, the meaning of Var(X(t1)) is this; imagine that we had many different instruments, all measuring the same characteristic at the same time. So for example, we have very many thermometers recording temperature from which we shall report the day's high temperature (and we label that day "t1"). Then E(X(t1)) is the mean temperature of those thermometers, and Var(Xt1) measures the variability about that mean.

Stationary Process

Intuitively defined, a process is stationary if the distribution of X(t) is the same for all t; in other words, while the values of X(t) vary with time, the probability structure does not. This has several implications that are worth making explicit:

- The mean does not vary with time (no trend): E(Xt) = constant
- The variance does not vary with time: Var(Xt) = constant
- The autocovariance depends only on the "lag": t2 - t1. Therefore, the autocovariance function really depends on a single variable, call it tau: gamma(tau) = E(X(t+tau) - m)(X(t)- m)

Note that this last bit means that gamma(0) = E(X(t) - m) (X(t) - m) = Var(X(t)).

There are actually shades of "stationary": strictly stationary and weakly stationary, but these are mathematical fine points.

When a process is stationary, we can define an autocorrelation function:

rho(tau) = gamma(tau)/gamma(0) = Correlation between X(t) and X(t + tau).

Note:

- rho(0) = 1
- rho(tau) = rho(-tau)
- -1 <= rho(tau) <= 1

Stationary processes are useful because of this fact:
Assuming that the distribution of X(t) is normal and that a process is stationary, the process is fully described by its autocorrelation function. Hence, identify the autocorrelation, and you have finished your work.

Later, we'll have a little "library" of autocorrelations that will fit a wide class of phenomenon. But most time series observed in nature are NOT stationary. Before we can identify the autocorrelation function, therefore, we must remove the non-stationary parts.

<u>De-Trending</u>

The first step is always to examine a plot of the data and ask yourself three questions:
1) Is there a trend?
2) Is variance changing with respect to the mean?
3) Is there a seasonal effect (cycles)?

If you answer yes to any of these questions, you do not have a stationary cycle and you must first "remove" these features.

How to remove a trend?

First, try a log transformation. While these does not eliminate the trend, it sometimes helps differentiate the trend from the seasonal cycle.

There are various other techniques. But it's useful to note that at this point it is sometimes useful to switch to a frequency domain analysis. Some of the things such an analysis can tell you are helpful in detrending a series. We'll get to that later.

That said, here are some techniques:

1) Fit a curve. For example, if the trend is linear, do a regression in which
X(t) = a + bt. Now base your analysis on the residuals. In additional to fitting linear trend, there is a wide class of curves that one can fit. They do not have to be linear functions.

2) Differencing. Replace each observation with Y(t) = X(t) - X(t-1). In fact, this is a sufficiently useful tool that there is some time-saving notation. Define the difference operator to be delta(Xt)

= X(t) - X(t-1).

And we can carry this notation further: delta^2(Xt) = delta(delta(Xt)) = delta(X(t) - X(t-1))
= delta(X(t)) - delta(X(t-1)) = X(t) - X(t-1) - (X(t-1) - X(t-2)) = X(t) - 2X(t-1) - X(t-2).

If a first order differencing doesn't remove your trend, try a second order.

3) Apply a filter. A filter is a linear combination of some number of past and some (possibly different) number of future observations. A moving average is the classic example:

Y(t) = (1/3)* (X(t-1) + X(t) + X(t+2)).

Choosing a filter is an art, of course. Once you've decided that you're going to use a moving average filter, you still need to decide how many terms to include in the moving average. Include too many and you get a flat line. Include too few, and your time-series will not be affected. Frequency domain analysis can help here.

Removing Cycles

So you've removed the trend. Now you must remove any seasonal cycles. We assume here, sadly, that you know the length of these cycles. The general strategy is to compare each seasonal average with the overall average. Essentially, what we end up doing is applying a filter designed with the characteristics of the cycle in mind.

For monthly data, these smoother is often useful:
S(t) = (1/12) * (.5X(t-6) + X(t-5) +...+ X(t+5) + .5X(t+6)
and then X(t) - S(t) is the "seasonal effect".

Adjustments can be made if the trend is quarterly.

One can also apply difference operators. For example, X(t) - X(t-12) if you have an annual cycle and your data are measured monthy.