Code ▾

# What is cointegration of time series data in statistics

Suppose you see two drunks (i.e., two random walks) wandering around. The drunks don't know each other (they're independent), so there's no meaningful relationship between their paths.

But suppose instead you have a drunk walking with her dog. This time there is a connection. What's the nature of this connection? Notice that although each path individually is still an unpredictable random walk, given the location of one of the drunk or dog, we have a pretty good idea of where the other is; that is, the distance between the two is fairly predictable. (For example, if the dog wanders too far away from his owner, she'll tend to move in his direction to avoid losing him, so the two stay close together despite a tendency to wander around on their own.) We describe this relationship by saying that the drunk and her dog form a cointegrating pair.

In more technical terms, if we have two non-stationary time series X and Y that become stationary when differenced (these are called integrated of order one series, or I(1) series; random walks are one example) such that some linear combination of X and Y is stationary (aka, I(0)), then we say that X and Y are cointegrated. In other words, while neither X nor Y alone hovers around a constant value, some combination of them does, so we can think of cointegration as describing a particular kind of long-run equilibrium relationship. (The definition of cointegration can be extended to multiple time series, with higher orders of integration.)

Other examples of cointegrated pairs:

- Income and consumption: as income increases/decreases, so too does consumption.
- Size of police force and amount of criminal activity
- A book and its movie adaptation: while the book and the movie may differ in small details, the overall plot will remain the same.
- Number of patients entering or leaving a hospital

## So why do we care about cointegration?

Someone else can probably give more econometric applications, but in quantitative finance, cointegration forms the basis of the pairs trading strategy: suppose we have two cointegrated stocks X and Y, with the particular (for concreteness) cointegrating relationship X - 2Y = Z, where Z is a stationary series of zero mean. For example, X could be McDonald's, Y could be Burger King, and the cointegration relationship would mean that X tends to be priced twice as high as Y, so that when X is more than twice the price of Y, we expect X to move down or Y to move up in the near future (and analogously, if X is less than twice the price of Y, we expect X to move up or Y to move down). This suggests the following trading strategy: if X - 2Y > d, for some positive threshold d, then we should sell X and buy Y (since we expect X to decrease in price and Y to increase), and similarly, if X - 2Y < -d, then we should buy X and sell Y.

But why do we need the notion of cointegration at all? Why can't we simply use, say, the R-squared between X or Y to see if X and Y have some kind of relationship? The reason is that standard regression analysis fails when dealing with non-stationary variables, leading to spurious regressions that suggest relationships even when there are none.

For example, suppose we regress two independent random walks against each other, and test for a linear relationship. A large percentage of the time, we'll find high R-squared values and low p-values when using standard OLS statistics, even though there's absolutely no relationship between the two random walks. As an illustration, here I simulated 1000 pairs of random walks of length 100, and found p-values less than 0.05 in 77% of the cases:

## So how do you detect cointegration?

There are several different methods, but the simplest is probably the Engle-Granger test, which works roughly as follows: Check that $X_t$ and $Y_t$ are both I(1). Estimate the cointegrating relationship $Y_t = aX_t + e_t$ by ordinary least squares. Check that the cointegrating residuals et are stationary (say, by using a so-called unit root test, e.g., the Dickey-Fuller test).

Also, something else that should perhaps be mentioned is the relationship between cointegration and error-correction mechanisms: suppose we have two cointegrated series $X_t, Y_t$, with autoregressive representations

$$X_t = aX_{t-1} + bY_{t-1} + u_t$$

$$Y_t = cX_{t-1} + dY_{t-1} + v_t$$

By the Granger representation theorem (which is actually a bit more general than this), we then have

$$\Delta X_t = \alpha 1(Y_{t-1} - \beta X_{t-1}) + u_t$$

$$\Delta Y_t = \alpha 2(Y_{t-1} - \beta X_{t-1}) + v_t$$

where $Y_{t-1} - \beta X_{t-1} \sim I(0)$ is the cointegrating relationship. Regarding $Y_{t-1} - \beta X_{t-1}$ as the extent of disequilibrium from the long-run relationship, and the αi as the speed (and direction) at which the time series correct themselves from this disequilibrium, we can see that this formalizes the way cointegrated variables adjust to match their long-run equlibrium.

## So, just to summarize

cointegration is an equilibrium relationship between time series that individually aren't in equilbrium (you can kind of contrast this with (Pearson) correlation, which describes a linear relationship), and it's useful because it allows us to incorporate both short-term dynamics (deviations from equilibrium) and long-run expectations (corrections to equilibrium).