Time Series Analytics Notes

Stationarity, Integration, and Co-integration

Stationarity, integration, and co-integration are three interrelated concepts that are important in the ARIMA approach to time series. In this set of notes, I will explain the meaning of these three terms and how they interrelate. Let us begin with stationarity.

Stationarity

A time series Y_t is **stationary** if

1. the mean of Y_t is the same for all time periods *t*
2. the variance of Y_t is the same for all time periods *t*, and
3. the autocorrelation structure of Y_t is the same for all time periods *t*.

From this definition, you can see that the intention of stationarity is to describe a time process that is statistically stable – not exactly the same values, but the same behavior on average. If you look at the time series today, tomorrow, or ten years from now, you will see the same structure. The values of the time series will be at about the same level, they will vary around that level by approximately the same amount, and they will depend upon their past in about the same way and affect their future in about the same way.

You can also see that stationarity is similar to some familiar concepts.

- #1 in the definition looks like the **L** specification.
- #2 looks like **H**.

If you noticed that similarity, then good for you! You are correct. In fact, you can verify #1 and #2 by the same qualitative and quantitative tests of L and/or H that you have seen previously. However, #3 is not the **I** specification. #3 is a more general condition than **I**. The **I** specification is a special case of #3 – the case in which the autocorrelations are all zero because the Y_t's are all independent of each other. Tests for the same autocorrelation structure are harder to come by in the general case. You can try calculating and comparing the empirical autocorrelation function, starting at different times *t*.

Notice two other things about the definition of stationarity: It does not require that the distributions of the Y_t's be normal, nor does it require that those distributions even be the same (as in a Random Sample). However, in applications, the distributions of the Y_t's are usually the same, and they are often normal or approximately so. Thus, stationarity is a more general concept than Random Sample. All Random Samples are stationary, but there are many stationary time series that are not Random Samples.

A bit more explanation is in order about the meaning of #3 in the definition. To say that the autocorrelation structure of Y_t is the same means that the way one *Y* affects another *Y* depends only upon the number of time periods between them – and not upon which time period it is. For

example, Y_1 affects Y_4 in the same way that Y_{17} affects Y_{20}; the correlation between each of these pairs is the same number, that is, the autocorrelation at lag 3.

Examples:

- All Random Samples are stationary.
- The period-to-period changes in a Random Walk are stationary.
- The residuals in a valid regression are stationary.
- The period-to-period changes in a Random Sample are stationary.[1]
- No Random Walk is stationary.
- No time series that has a trend is stationary.
- No heteroscedastic time series is stationary.
- No time series that has seasonality is stationary.

The ARIMA approach to time series modeling is to use differencing to change the original time series into a stationary time series. Differencing is applied intelligently at an appropriate number of lags, at an appropriate number of times (differencing of differences) in order to remove features like trends and seasonality that make the original time series nonstationary. Once the original time series has been made stationary, then the only structure that remains is the autocorrelation structure, which the ARIMA machinery is designed to estimate. Then ARIMA produces a (hopefully) valid model of the autocorrelation structure of the differenced data. This means that the residuals from the ultimate ARIMA model are free of autocorrelation and are, in fact, a zero mean Random Sample. To get estimates and forecasts for the original time series, the ARIMA procedure is reversed, including un-differencing.

The reason it is important to make a time series stationary before estimating the autocorrelation structure is that the presence of nonstationary features in a time series can easily induce spurious autocorrelation. For example, suppose Y_t is N(t, σ^2) and the Y_t's are all independent. That is, Y_t would be a Random Sample, except for the linear trend in means. Then a timeplot of observed Y_t's would go upward to the right. The empirical autocorrelation would be positive. The trend induces autocorrelation. But if the trend were removed, say by first differencing or by subtracting t from each observed Y_t, then the time series would be stationary, and the autocorrelation would be zero.

So differencing plays a key role in ARIMA. The "I" in ARIMA refers to "integration", which means *differencing* (and not the calculus procedure) – but more about that in a minute. In order to see why ARIMA can rely on differencing to remove structure from time series to make the original time series stationary, let us consider a few examples of the efficacy of differencing:

- Suppose Y_t has a linearly increasing mean: 1, 3, 5, 7, 9, …. Then differencing yields the sequence 2, 2, 2, 2, …Thus, differencing adjacent values (called **first-order differencing** in the ARIMA approach) removes a linear trend and leaves a (mean) stationary series.
- Suppose Y_t has a quadratic increasing mean: 1, 4, 9, 16, 25, …. Then differencing yields the sequence 3, 5, 7, 9, … . Differencing again (called **second-order differencing**) yields

[1] But the changes in a RS are not a RS, since they are autocorrelated.

2, 2, 2, 2, … Second-order differencing removes a quadratic trend and leaves a (mean) stationary series.

- Third-order differencing removes a cubic trend – and so forth.
- Suppose Y_t is a Random Walk. Random Walks fail stationarity. But first differencing yields a Random Sample, which is stationary.
- Suppose Y_t is a nonstationary time series. In an attempt to make it stationary, you difference Y_t. But suppose the differences of Y_t turn out to be a Random Walk. Random Walks fail stationarity. But first differencing of the differences will yield a Random Sample, which is stationary. So second differences of Y_t yield a stationary time series.
- Suppose Y_t is quarterly data for which the mean function is periodic (strongly seasonal) with linear trend, for example: 1, 4, 5, 2 | 3, 6, 7, 4 | 5, 8, 9, 6 | 7, 10, 11, 8, … . Then first differencing at lag 4 yields 2, 2, 2, 2 | 2, 2, 2, 2 | 2, 2, 2, 2 … Differencing at an appropriate lag can remove seasonality and possibly a trend and leave a (mean) stationary series.

Differencing is linked to integration, another important concept in ARIMA.

Integration

In the context of time series, integration has nothing to do with calculus. Instead, it refers to differencing. A time series Y_t is **integrated** if it is stationary at some order of differencing. **Zero-order integration** means that Y_t itself is stationary. **First-order integration** means that $Y_t - Y_{t-1}$ is stationary. **Second-order integration** means that $(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2})$ is stationary. A time series can be stationary at multiple orders of integration; we are usually interested in the smallest order for which a time series is stationary. Often, the order of integration means the smallest order of stationarity. There are time series that are not stationary at any order of integration. Examples:

- All Random Samples are zero-order integrated; they are also first-order integrated.
- The residuals in a valid regression are zero-order integrated.
- All Random Walks are first-order integrated.
- Suppose Y_t is independent $N(\mu_t, \sigma^2)$ with the mean μ_t given by the linearly increasing sequence: 1, 3, 5, 7, 9, …. Then differencing yields the sequence of means 2, 2, 2, 2, … The series remains homoscedastic and independent. The Y_t time series is first-order integrated.
- Suppose Y_t is independent $N(\mu_t, \sigma^2)$ with the mean μ_t given by a quadratic increasing sequence: 1, 4, 9, 16, 25, …. Then differencing yields the sequence 3, 5, 7, 9, … . Differencing again yields 2, 2, 2, 2, … After double differencing, the series remains homoscedastic and with stable autocorrelation function.[2] This time series is second-order integrated.

[2] The variance of the doubly-differenced series is $6\sigma^2$ for all *t*; the autocorrelation is -2/3 at lag 1, +1/6 at lag 2, and 0 at all higher lags.

- Suppose Y_t is a nonstationary time series. In an attempt to make it stationary, you difference Y_t. But suppose the differences of Y_t turn out to be a Random Walk. Random Walks fail stationarity. But first differencing of the Random Walk will yield a Random Sample, which is stationary. So Y_t is second-order integrated.
- Suppose Y_t is independent $N(\mu_t, \sigma^2)$ with the mean μ_t given by a periodic (strongly seasonal) sequence with linear trend, for example: 1, 4, 5, 2 | 3, 6, 7, 4 | 5, 8, 9, 6 | 7, 10, 11, 8, … . Then first differencing at lag 4 yields 2, 2, 2, 2 | 2, 2, 2, 2 | 2, 2, 2, 2 … The differenced series remains homoscedastic and with stable autocorrelation function.[3] The Y_t time series is first-order integrated.
- Suppose that Y_t is independent $N(0, t\sigma^2)$. The mean function is already stationary, but the variance increases proportionally with *t*. This Y_t is not integrated at any order. For example, the first difference $Y_t - Y_{t-1}$ is $N(0, (2t-1)\sigma^2)$, which is heteroscedastic. Continued differencing of differences just increases the heteroscedasticity. For example, $(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = (Y_t - 2Y_{t-1} + Y_{t-2})$ is $N(0, (6t-6)\sigma^2)$. In order to apply the ARIMA machinery, the inherent heteroscedasticity of Y_t should be addressed first, perhaps by some type of transformation. For example, the transformation $Y_t / \sqrt{t}$ is independent $N(0, \sigma^2)$, which is zero-order integrated.

As noted in the preceding section, it is important to stabilize a time series (make it stationary) before applying ARIMA machinery to estimate the autocorrelation structure. If the time series is not stationary, then spurious correlations can be induced by the nonstationary features.

Co-integration

The important role of integration in understanding and forecasting individual time series by the ARIMA method has long been understood. The role of integration in understanding how one time series relates to another has not been understood long. It was long thought that all that was necessary to allow the regression of one time series Y_t on another X_t is that both be integrated at some order.

That this is insufficient can be seen in simple examples. For example, let Y_t and X_t be Random Walks formed from independent sets of changes. To be specific, let C_t and D_t be independent Random Samples of $N(0, \sigma^2)$ random variables. That is, not only are the C_t's independent of each other for all *t*'s and the D_t's are independent of each other for all *t*'s, but also all of the C_t's are independent of all of the D_t's for all *t*'s. Then define $Y_1 = C_1$ and $X_1 = D_1$ and, for *t* > 1, define $Y_t = Y_{t-1} + C_t$ and $X_t = X_{t-1} + D_t$. That is, Y_t is a running total of the C_t's and X_t is a

[3] The variance of the differenced series is $2\sigma^2$ for all *t*; the autocorrelation is -1/2 at lag 4, and 0 otherwise.

running total of the D_t's. Therefore, Y_t and X_t are statistically independent of each other since each is a function only of a time series that is completely independent of the other.

So in theory, the regression of Y_t on X_t should have a slope of zero and an R^2 of zero. However, most actual data – realizations y_t and x_t – will show a strong statistical relationship between the two sets of realizations. The Excel file "*(Simulation) Regression with a Stochastic X.xlsx*" implements precisely this example as a live simulation. About half of the realizations will show positive relationships and about half will show negative relationships. Astonishingly, a large proportion of these realizations have relationships between y_t and x_t that are "statistically significant" according to the usual test criteria! If y_t and x_t are real data, whether the relationship will turn out positive or negative is a matter of random chance. Therefore, no reliance can be placed on the empirically estimated relationship. The relationship is spurious. When y_t and x_t are economically or managerially important variables, this is truly alarming, for important policy decisions may be made based on the mistaken belief that the relationship is real.

The simulation in "*(Simulation) Regression with a Stochastic X.xlsx*" provides not only an illustration of the phenomenon but also provides intuition for why the spurious relationship occurs. There, we see that after the simulation has run for a period of time, there are four possible configurations for the final positions of *Y* and *X*:

1. Both *Y* and *X* can end high;
2. Both can end low;
3. *X* can end high and *Y* low;
4. *X* can end low and *Y* high.

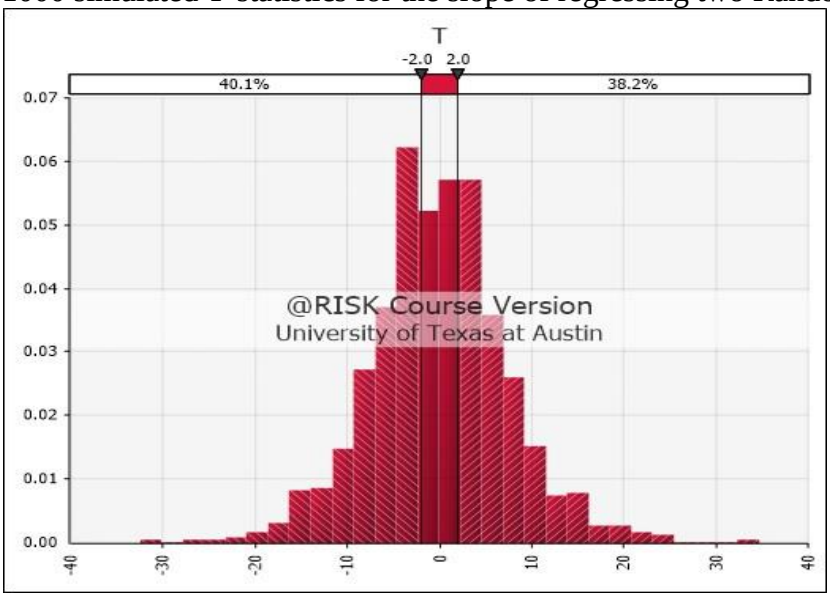
Each configuration happens at random about ¼ of the time.

1. In the first case (both end up), the time path of Y and the time path of X each has more ups than downs – therefore both tend to be moving up together, implying a positive relationship.
2. In the second case (both end down), each time path has more downs than ups – therefore both tend to be moving down together, again implying a positive relationship.
3. In the third case (*X* high, *Y* low), *X* has more ups than downs, and *Y* has more downs than ups – therefore the two tend to be moving opposite each other, implying a negative relationship.
4. In the fourth case (*X* low, *Y* high), *X* has more downs than ups, and *Y* has more ups than downs – therefore the two tend to be moving opposite each other, implying a negative relationship.

Moreover, these imbalances tend to be statistically significant, according to conventional criteria, *far* more often than they should by chance. To confirm this, I iterated the simulation of the regression of Y_t on X_t in the first tab of "*(Simulation) Regression with a Stochastic X.xlsx*" 1000 independent times. I collected the usual T-statistics for testing the significance of the slope of that regression. Figure 1 shows the distribution of those 1000 T-statistics. Since the Y's and X's are independent of each other, the magnitude of the T-statistic should exceed the value 2 only about 5% of the time – namely, 50 times out of the 1000 iterations. However the vertical lines on the histogram at -2 and +2 show that -2 and +2 are each exceeded about 400 times! Even more

remarkable are some of the extreme occurrences. For example, T > +10 occurs 77 times in the simulation. This should not happen at all, for P(T > 10) = 0.00000000000000006.

Figure 1. 1000 simulated T-statistics for the slope of regressing two Random Walks



So – what to do about this? We must be careful when we regress one time series upon another. There must be some *real* connection between the two time series. How do we guarantee that? An answer is provided by the concept of co-integration. Two time series are **co-integrated** if there is a linear combination of the two that is integrated at some order. Ordinarily, the two time series are required to be integrated at order zero.

A linear combination $aY_t + bX_t$ is itself a time series, which may – or may not – be integrated at some order. If it is, then there is a real relationship between Y_t and X_t. In order to understand the intuition for why co-integration implies a real relationship, suppose that Y_t and X_t are co-integrated at some order. Then there is a linear combination $aY_t + bX_t$ that is stationary at some level of (possibly) multiple differencing. Think about the mathematical form of those multiple differences: it will also be a linear combination of Y_t and X_t and several of their lags. Because that linear combination is stationary, it has a stable mean and variance. Thus, that linear combination equals a constant (the mean) plus an error that is zero-mean and homoscedastic and (possibly) autocorrelated. Imagine that equation. It has *Y*'s and *X*'s on the left-hand side within the linear combination. On the right-hand side is the mean and an error that is zero-mean and homoscedastic and (possibly) autocorrelated. On the left-hand side, separate the terms of the linear combination. Keep the current value Y_t on the left-hand side and move all of the other terms (lagged *Y*'s and all *X*'s) to the right-hand side. You then have a predictive time series model that expresses Y_t as a function of lagged *Y*'s and *X*'s and zero-mean homoscedastic (possibly) autocorrelated error. This is a real regression relationship between Y_t and lagged *Y*'s and *X*'s that can be estimated.

For example, suppose that Y_t and X_t are co-integrated at zero order. Then there is a linear combination $aY_t + bX_t$ that is stationary, say $aY_t + bX_t = u_t$. Hence, $Y_t = -\frac{b}{a}X_t + \frac{1}{a}u_t = b^*X_t + u_t^*$. This is the form of a regression in which u_t^* looks like the error term. Although u_t^* is not necessarily a Random Sample, it is stationary and so is ready to have its structure recovered by ARIMA.

For a more complex example, suppose that Y_t and X_t are co-integrated at first order. Then there is a linear combination $aY_t + bX_t$ such that its first difference $aY_t + bX_t - (aY_{t-1} + bX_{t-1})$ is stationary, say $aY_t + bX_t - (aY_{t-1} + bX_{t-1}) = u_t$. Hence, $Y_t = Y_{t-1} - \frac{b}{a}X_t + \frac{b}{a}X_{t-1} + \frac{1}{a}u_t =$ $Y_{t-1} + b^*X_t - b^*X_{t-1} + u_t^*$. This is the form of a regression in which u_t^* looks like the error term. Although u_t^* is not necessarily a Random Sample, it is stationary and so is ready to have its structure recovered by ARIMA. Clearly, this process can be continued for still higher orders of co-integration.

The reason that zero-order integration is usually desired for co-integration is that the relationship of interest is usually the relationship between Y_t and X_t, rather than between some order of their differences. You want to rule out a spurious relationship between Y_t and X_t.

The simulation in the second tab of "*(Simulation) Regression with a Stochastic X.xlsx*" provides an example of two zero-order co-integrated time series. In that second tab, Y_t and X_t are nonstationary. X_t is a Random Walk; Y_t is not quite a Random Walk, but is nonstationary. This example is very similar to the non-co-integrated Random Walks in the first tab. The difference is that in the second tab, $Y_t - X_t$ equals a Random Sample, which is stationary. $Y_t - X_t$ is a linear combination, which is therefore zero-order integrated. So Y_t and X_t are co-integrated at order zero in the simulation of the second tab.

SUMMARY

Stationarity expresses what we mean by a time series being statistically stable. A time series is **stationary** if

1. its mean is the same in all time periods
2. its variance is the same in all time periods, and
3. its autocorrelation structure is the same in all time periods.

Stationarity plays a big role in ARIMA modeling. ARIMA focuses upon modeling the autocorrelation structure of a stationary time series. For ARIMA to work properly, the time series first must be made stationary.

Repeated differencing is the primary method for rendering a time series stationary in ARIMA. A time series is **integrated** if it is stationary at some order of differencing. A difference of a difference is a second-order differencing. A difference of a second-order difference is a third-order difference. And so on.

Co-integration was introduced to provide a means to tell when it is justified to regress one time series upon another. Two time series are **co-integrated** if there is a linear combination of them that is integrated at some order of differencing. Regressing one time series upon another can produce spurious relationships if the time series are not co-integrated.