Important to mention that I use the word observation sometimes instead of time series because it is easier in my head to make the connection.

**What is stationarity, and why is it important for time series data?**

Time series models are much better at forecasting future observations in the short term then a linear regression relationship. Stationarity is a property of a time series. It says that all statistical properties, such as mean, variance, and autocorrelation, remain constant over time and there is no seasonality in the time series.

Achieving stationarity is often a critical first step in time series analysis and modeling, ensuring that the assumptions underlying statistical methods are met and that the insights and forecasts derived from the analysis are valid and reliable.

1. Many statistical models, particularly those used in time series analysis (such as ARIMA, SARIMA, and GARCH), **assume that the data is stationary**. These models rely on the constancy of statistical properties over time.

2. Stationary time series are easier to predict because their properties do not change over time, so models can learn from historical data and apply the knowledge to future data points. Also, you do not have to remove trends, seasonality, or other patterns because they are inexistent.

3. Hypothesis tests, are designed **under the assumption of stationarity**. Non-stationary data can lead to misleading results and incorrect conclusions.

4. Stationarity ensures that the relationships between variables after a regression are genuine.

**Testing for Stationarity**

Augmented Dickey-Fuller (ADF) Test: A statistical test that checks for the presence of a unit root in the time series. The null hypothesis is that the series is non-stationary.

KPSS Test (Kwiatkowski-Phillips-Schmidt-Shin): Another statistical test where the null hypothesis is that the series is stationary. This test complements the ADF test.

Variance and Autocorrelation Function (ACF) Analysis: Analyzing the variance over different time periods and the autocorrelation function can help detect non-stationarity.

You need stationarity of the time series to meaningfully use the Pearson corr, ACF, and PACF. Note that if the time series is non-stationary, then maybe a variable derived from the time series, such as the difference between consecutive points, can be stationary. Or the logs. It’s about being creative in arriving at a stationary time series.

**Correlation**

**The Pearson correlation** tells you the strength of the linear relationship of 2 variables and the direction. However, if the relationship between variables is not linear (e.g., for a 1 unit increase in one, the other increases by 1 sometimes and by 2 the other times), interpreting the Pearson correlation coefficient may not provide meaningful insights into the strength and direction of that relationship. Furthermore, the Pearson correlation takes into account the direct and indirect linear relationships between observations and then spits the number.

**Autocorrelation**

**Autocorrelation (lag=1)** looks at the linear relationship observed between consecutive observations (so the [last observation & the penultimate], [penultimate & the one before penultimate], so on… and the first observation won’t have anything before it that is the order). So you will look at 2 series, and from those 2 series, you pick the [] as seen above. The series have a lag of 1 (show code).

A black screen with white text

Description automatically generated

Suppose now we are dealing with an autocorrelation of lag=2. With autocorrelation, it is important to note that it will account for both the direct impact and the indirect impact. So, if I want the autocorrelation of lag=2 of 2 time series, I will look at X(t-2) & X(t). Now, in between these, there’s the X(t-1) and the autocorrelation of lag=2. We will look at the direct linear relationship between X(t-2) & X(t) but also at the indirect linear relationship that is going from X(t-2) to X(t-1) to X(t).

So, how do we get the direct linear relationship only? Introducing PACF (Partial autocorrelation function).

**Partial autocorrelation**

This is the direct correlation between X(t-2) & X(t), excluding any indirect correlation through X(t-1). It allows us to understand if there is any linear relationship between X(t-2) & X(t). The autocorrelation lags=2 can not help with that because you wouldn’t know where is the linear relationship coming from (like directly or indirectly).

For lag 1, the Autocorrelation Function (ACF) = acf[1] and the Partial Autocorrelation Function (PACF) = pacf[1] will have the same value.

Please note that the PACF's output, except for positions [0] and [1], will not be the Pearson correlation. The ACF’s output is the Pearson correlations for all the lags you specified.

**AR model**

An AR model is a predictive model that spits the value that you need based on previous observations.

The parameter is a constant + the coefficient is constant + the error follows a N(0,std^2)

In an AR1 (auto-regressive) model, the time series looks like this:

y(t) = parameter + coefficient \* y(t-1) + error

In an AR2 (auto-regressive) model, the time series looks like this:

y(t) = parameter + coefficient(1) \* y(t-1) + coefficient(2) \* y(t-2) + error

In both cases:

y(t) = signal + error

y(t) = time series (e.g., 1,2,3,4,5)

y(t-1) = time series lagged by 1 (e.g., NaN,1,2,3,4)

So, if you get data in a time series and you wanna write it as an AR1 model, you need to know the signal (i.e., the relationship) and error. Say u came up with smth now you wanna check the error term, then u need to know that the mean = 0 + the std is constant over time + any correlation between lags of the error must be 0 (i.e., the ACF must be 0 of within the shaded area for all lags, obv it should converge to 0) which basically implies that there shouldn’t be any pattern.

Right, so what about the signal?

For the signal, you can use the PACF to write the relationship. Essentially, if the direct linear relationship between the observation u want and God knows what lagged observation is statistically insignificant (i.e., the value is within the shaded area which means it approximates to 0) then that lagged observation does not go into the predicting relationship i.e., signal.

Here is a picture with the PACF and the relationship. Please observe how the third observation was excluded because the value in the PACF is in the shaded region.

A hand holding a piece of paper with text

Description automatically generated

Here M(t) is the observation you need/look for. Betas are constants. The error appears since this is an AR model. The M(t-3) is not in the signal because there is no stat significant linear relationship between it and M(t).

 **Crack Spread**: Due to the more stable relationship between crude oil and refined product prices, the crack spread can display characteristics of stationarity, assuming stable market conditions.

 **Spread Between Crude Oils**: This spread tends to be non-stationary due to the diverse and dynamic factors influencing the pricing of different crude oils relative to each other.