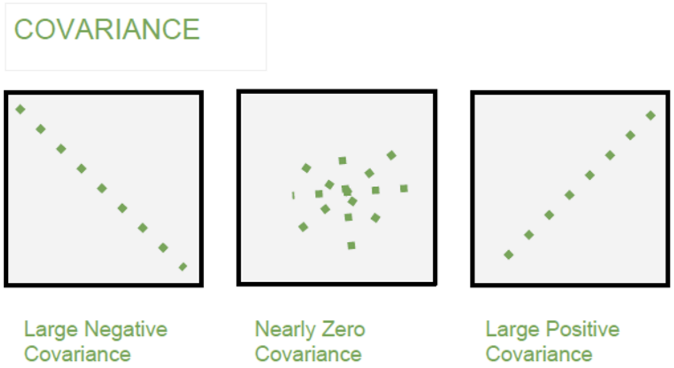
Module Demo.1

COVARIANCE

First, we start off with the concept of covariance. Covariance tells you how two random variables are similar or dissimilar in their behavior, referred as dependence. Covariance captures so-called linear dependence. Graphically here, you can see what covariance looks like.

- The graph on the left represents negative covariance. This tells us that the slope of the relationship between the variable on the x-axis and the variable on the y-axis is negative. That is, decreases in one variable also causes decreases in the other variable.
- The graph on the right shows positive covariance. An increase in one variable results in an increase in the other variable.



Source: https://www.geeksforgeeks.org/mathematics-covariance-and-correlation/

EXAMPLE:

To give an example, Bank of America stock and Chase stock may have a positive covariance with reported earnings. That is, the value of BoA stock may rise when Chase stock rises.

FORMULA:

Here, we can see the formula for covariance.

$Covariance\ (x, y)\ = \frac{1}{n} \Sigma_{\{i=1\}}^{n} \quad (xi\ -\ \underline{x})\ (yi\ -\ \underline{y})$

where

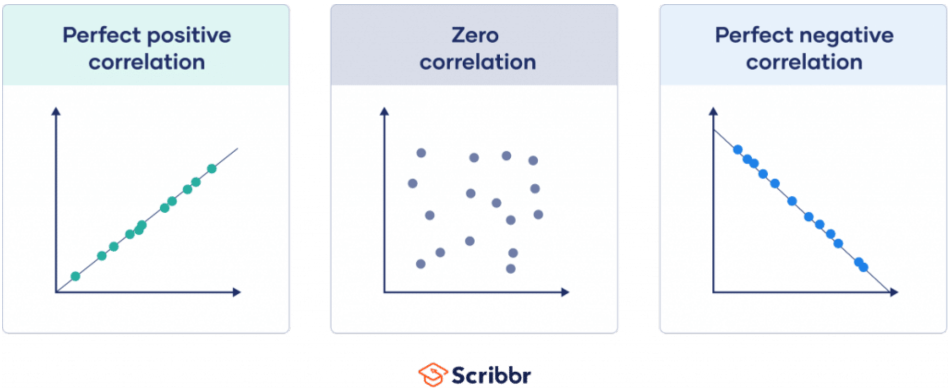
- xi = data value of x
- yi = data value of y
- x̄ = mean of x
- ȳ = mean of y
- n = number of data points

From the formula, you can see that covariance values can range from infinity to negative infinity. The larger the range of X and Y values, the larger the covariance. So it can be hard to interpret covariance. However, covariance is the stepping stone to what we are most familiar with, correlation.

CORRELATION

Correlation is scaled covariance. The values for the correlation measure range from -1 to 1.

- A value of -1 shows perfect negative correlation, that is for a positive increase in one variable is a decrease in the other variable
- A value of 1 shows perfect positive correlation, that is for a positive increase in one variable is also a positive increase in the other variable
- A value of 0 indicates that there is no linear relationship between the two variables.



Source: https://www.scribbr.com/statistics/correlation-coefficient/

EXAMPLE:

For example, there is a positive correlation between the oil price and the stock price of Exxon Mobil because when oil prices rise, oil companies make more money and stock price increases.

FORMULA:

$$Correlation(x,y) \ = \frac{Covariance(x,y)}{\sqrt{Var(x)\,Var(y)}}$$

You can see how correlation is related to covariance here because if you take the covariance and divide by the square root of the product of var(x) var(y), then we get the correlation. Essentially, it is the covariance scaled.

Now that we discussed covariance and correlation, we can talk about autocovariance and autocorrelation, which are concepts specific to time series data.

AUTOCOVARIANCE/AUTOCORRELATION

Time series is a collection of random variables, hence we can define covariance and correlation of any pair of random variables in the time series. Generally, we want to measure their dependence, that is the within time series dependence or so-called autocovariance/autocorrelation. Note that autocorrelation is also referred to as serial correlation. Serial correlation occurs if the time series data in one period in time is correlated with the time series data in other periods. Autocovariance or autocorrelation is a general concept, applying to any time series, regardless of whether stationary or non-stationary. Dependence exists in any time series that is not white noise or independent noise.

Autocorrelation or serial correlation can be calculated by finding the relationship of a time series data at a given time with the time series data at other past time points, which are called lags. That is, the correlation between the present X_t and the previous data point X_{t-1}, and then with X_t and X_{t-2}, etc.

For example, if a stock price has a strong positive autocorrelation, then we would likely assume if the price is up today, it's plausible to be up tomorrow as well.

AUTOCOVARIANCE/AUTOCORRELATION FUNCTION

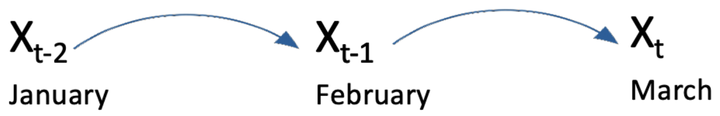
While autocovariance is a measure of dependence of any time series, we define the autocovariance function for **stationary time series** only. For stationary time series, the autocovariance does not change as time shifts and hence it only depends on the difference of the time points, also called lag, and not the time itself. Again, the autocorrelation function is a scaled version of the autocovariance.

To get an intuitive understanding of the autocorrelation function or in short here ACF, let’s look at a specific example.

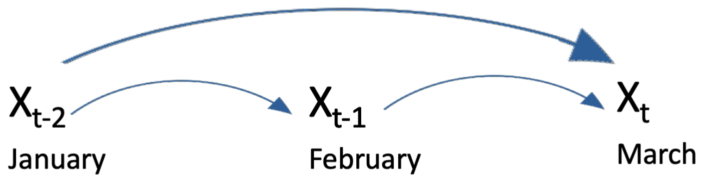
EXAMPLE:
Let’s say we are looking at predicting the average monthly visitors to Yosemite for the month of March, our current month. We’ll just look at a very small window to understand the concept, we have here just January through March.

How many visitors we get this month would depend on how many visitors we got last month, the month prior, etc. As shown in the arrows below,

- The # of visitors in January has an influence on the number of visitors in February.
- The # of visitors in February has an influence on the number of visitors in March.



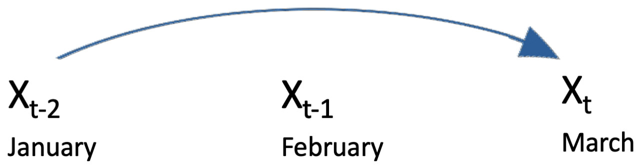
Note that there can also be an influence of the # of visitors in January for the # of visitors in March. That is, there is a direct correlation from two months prior rather than the indirect correlation mentioned above. Why might we have this direct influence from two months previously? For example, there could be a special event for geologists at Yosemite that happens bimonthly in January, March, etc. So because of the special event in January, the # of visitors in January can have a direct effect for # of visitors in March.



When it comes to ACF, we take the effects of all of these correlations, both the indirect and direct.

We also learned about another measure of dependence in a time series called partial autocorrelation or in short PACF. As the name of this measure says, we are capturing partial dependence.

More specifically, we are interested in only the direct effect, which is why PACF is considered conditional. Referring back to the above example, we want to know if directly, the # of visitors in January is a good predictor for the # of visitors in March without considering the indirect effects.



Essentially, we control for the other lags so that we know the exact direct effect of January to March.

If we translate the idea of direct and indirect effects using regression modeling, the ACF is the linear dependence established through a marginal relationship, when regressing the lagged time series, say X_t-h onto X_t using simple linear regression. In contrast, the PACF is the conditional relationship defined through regressing X_t-1, X_t-2, …, X_t-h onto X_t, using a multiple linear regression. The resulting regression coefficient corresponding to X_t-h from this multiple linear regression is the conditional relationship with respect to X_t given X_t-1, X_t-2, …, X_t-(h-1). In short, ACF captures the marginal relationship between X_t and X_t-h while PACF captures the conditional relationship between X_t and X_t-h.

PROVIDE MORE DETAILS USING THE EXAMPLE

With the same Yosemite data example, the ACF is just the simple linear relationship between an observation at time X_t and the observations of previous times, X_{t-h} where h is the lag.

But in order to find the PACF component, we need to start with building a multiple linear regression with the data.

$X_t = \Theta_1 * X_{t-1} + \Theta_2 * X_{t-2}$
OR
March = Θ_1 * February + Θ_2 * January

The # of visitors in our current month of March is equal to some coefficient theta 1 times last month's #of visitors plus some coefficient theta 2 times # of visitors two months ago plus some error term.

Recall from earlier, we are interested in only the direct effect. That is, the direct effect of the # of visitors in January for predicting the # of visitors in March. In this regression, the PACF is the coefficient theta 2 because that is the direct effect of January for our prediction. February effect is captured with the coefficient theta 1.