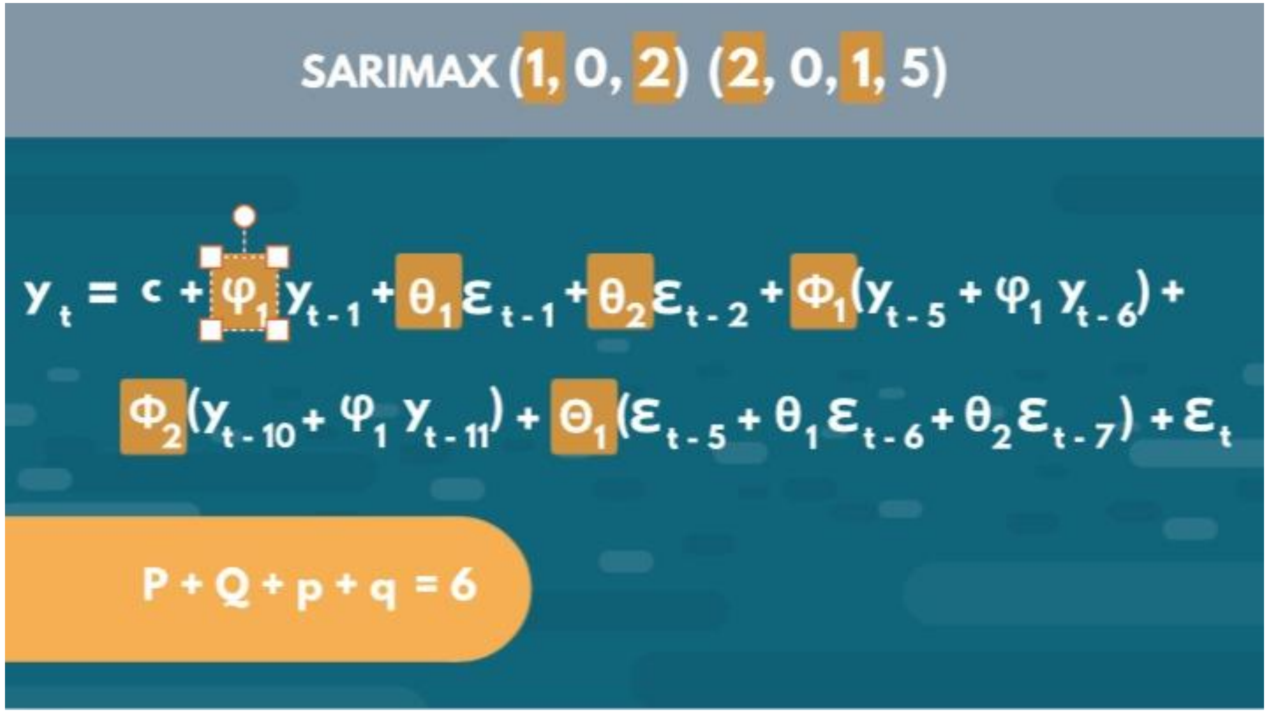
The SARIMAX Model

Full Name:

The **S**easonal **A**uto**r**egressive **I**ntegrated **M**oving **A**verage e**X**ogenous Model

Mathematical Notation:



SARIMAX (1, 0, 2) (2, 0, 1, 5)

$$y_t = c + \varphi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \Phi_1 (y_{t-5} + \varphi_1 y_{t-6}) +$$

$$\Phi_2 (y_{t-10} + \varphi_1 y_{t-11}) + \Theta_1 (\varepsilon_{t-5} + \theta_1 \varepsilon_{t-6} + \theta_2 \varepsilon_{t-7}) + \varepsilon_t$$

$$P + Q + p + q = 6$$

Short Description:

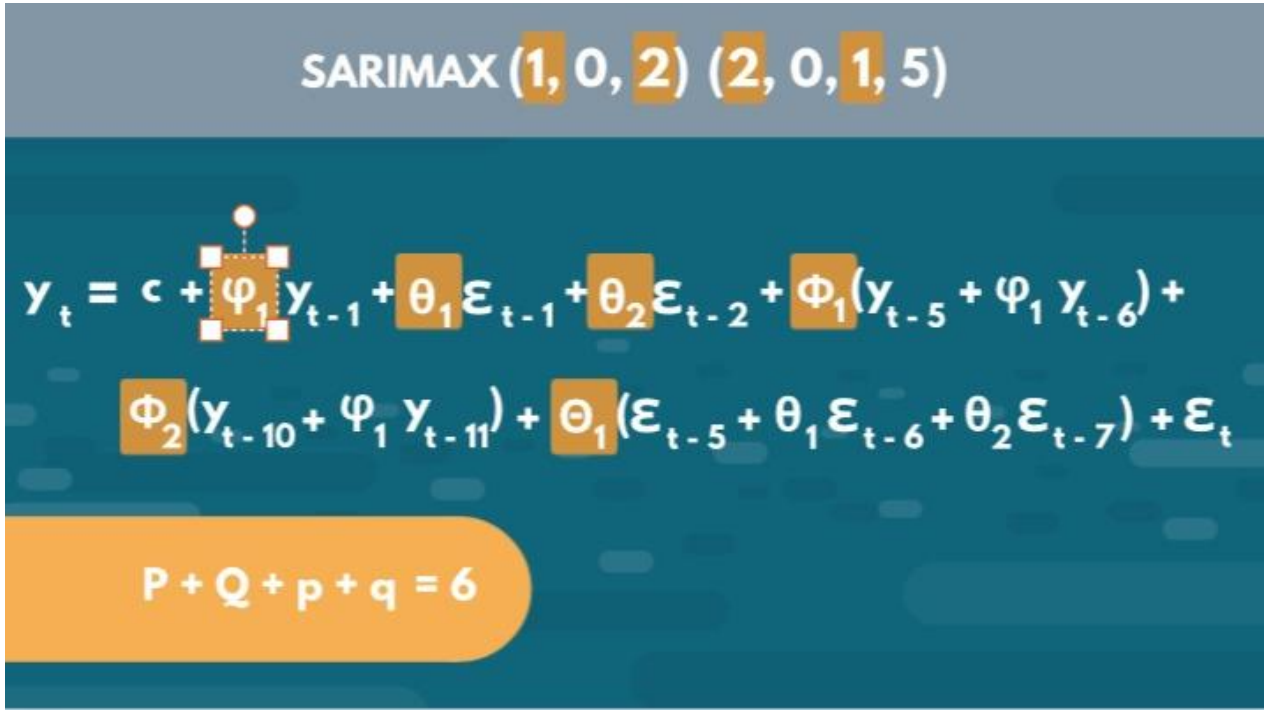
The SARIMAX is the **seasonal** equivalent of the ARIMAX model. Of course, there exist seasonal versions of the other models as well (SARMA, SARIMA, SARMAX, etc.).

Seasonal models help capture patterns which aren't ever-present but appear periodically. For example, the amount of flights leaving an international hub like JFK Airport in NYC are far larger in December compared to October.

That is mainly due to the festive period for many countries in December. Thus, October is far less busy. Therefore, we need a way to account for this expected influx of demand in December and we can do so by checking the values in December of the previous year.

The SARIMAX Model

Equivalents of the SARIMAX:



SARIMAX (1, 0, 2) (2, 0, 1, 5)

$$y_t = c + \varphi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \Phi_1 (y_{t-5} + \varphi_1 y_{t-6}) + \Phi_2 (y_{t-10} + \varphi_1 y_{t-11}) + \Theta_1 (\varepsilon_{t-5} + \theta_1 \varepsilon_{t-6} + \theta_2 \varepsilon_{t-7}) + \varepsilon_t$$

$P + Q + p + q = 6$

The SARIMAX is among the most-complicated models we can have, since it **can** incorporate seasonality, integration and/or exogenous variables.

However, it doesn’t **have** to.

By setting the values of certain orders to 0, or by not providing certain information, the model can be simplified.

For instance, by not including exogenous variables and having no integration, the model automatically becomes equivalent to a SARMA.

The equation on the left is exactly that - a SARIMAX equivalent of a SARMA.

The SARIMAX Model

The Original Equation:

So, a seasonal model has 7 orders split into two parts – seasonal vs nonseasonal: SARIMAX (p, d, q) (P, D, Q, s)

The nonseasonal ones are the ARIMA lags we're already used to: **p**, **d** and **q**. The rest are the seasonal ones – **P**, **D**, **Q** and **s**. The first 3 are obviously the seasonal equivalents of the **p**, **d** and **q**, while **s** is the only new one. It represents the length of the **season**, hence the name – '**s**'.

Now, the seasonal order (P, Q) determines the number of seasons we're going back. For instance, if P = 2 and s = 10, then we're including the values from 1 and 2 seasons ago, which is the same as 10 and 20 periods ago.

Then, if p = 1, we'd be including $X_{t-1}, X_{t-10}, X_{t-11}, X_{t-20}$ and X_{t-21}. That is because for each of the two seasons, we also need to include p-many past values relevant to it. Thus, for each seasons (X_{t-10}, X_{t-20}), we also include 1 additional past value (X_{t-11}, X_{t-21}).

To make it easier, let's see what a SARIMAX (1,0,0) (2,0,0,10) model looks like:

$$X_t = C + \phi_1 X_{t-1} + \phi_{10} X_{t-10} + \phi_{11} X_{t-11} + \phi_{20} X_{t-20} + \phi_{21} X_{t-21} + \epsilon_t$$

The SARIMAX Model

The Modified Equation:

However, the values for ϕ_{11} and ϕ_{21} are restricted. They must be equal to $\phi_1\ \phi_{10}$ and $\phi_1\ \phi_{20}$ respectively.

$$X_t = C + \phi_1 X_{t-1} + \phi_{10} X_{t-10} + \phi_1 \phi_{10} X_{t-11} + \phi_{20} X_{t-20} + \phi_1 \phi_{20} X_{t-21} + \epsilon_t$$

Thus, we can rewrite the equation to get the following:

$$X_t = C + \phi_1 X_{t-1} + \phi_{10}(X_{t-10} + \phi_1 X_{t-11}) + \phi_{20}(X_{t-20} + \phi_2 X_{t-21}) + \epsilon_t$$

For consistency, we like to use distinct notation for the seasonal coefficients as well, so we plug in Φ_1 and Φ_2 for ϕ_{10} and ϕ_{20}.

$$X_t = C + \phi_1 X_{t-1} + \Phi_1(X_{t-10} + \phi_1 X_{t-11}) + \Phi_2(X_{t-20} + \phi_2 X_{t-21}) + \epsilon_t$$

Now, this is the actual model that gets regressed. In other words, Python only complies a constant and 3 coefficients: ϕ_1, Φ_1 and Φ_2. Thus, even though the model uses very many past variables, it only needs to compute (p + P) – many values.

The SARIMAX Model

Past Seasons and Past Residuals:

Now, if we decide to include residuals, you need to know that the seasonal orders don’t directly affect one another. To see what we mean, here is what a SARIMA (1,0,2)(2,0,1,10) looks like:

$$X_t = C + \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \phi_{10} X_{t-10} + \phi_{11} X_{t-11} + \phi_{20} X_{t-20} + \phi_{21} X_{t-21} \\ + \theta_{10} \epsilon_{t-10} + \theta_{11} \epsilon_{t-11} + \theta_{12} \epsilon_{t-12} + \epsilon_t$$

We’ve highlighted the new additions in red. We se that simply because we’re adding lags, doesn’t mean we’re expanding the coefficients we’re including. In other words, we’re not including the value for t-12 only because we’re adding the residual for that period.

Additionally, the coefficients θ_{11} and θ_{12} are restricted too and equal $\theta_{10}\theta_1$ and $\theta_{10}\theta_2$ respectively. We can once again plug in and substitute a few things. We don’t plan on going over each step once more, so we eventually reach the following:

$$X_t = C + \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Phi_1 (X_{t-10} + \phi_1 X_{t-11}) + \Phi_2 (X_{t-20} + \phi_1 X_{t-21}) + \Theta_1 (\epsilon_{t-10} + \theta_1 \epsilon_{t-11} + \theta_2 \epsilon_{t-12}) + \epsilon_t$$

The SARIMAX Model

A Quick Look at the Coefficients:

Now that we know what the actual equation of a SARIMAX (1,0,2) (2,0,1,10) looks like, let's make a few remarks:

$$X_t = C + \phi_1 X_{t-1} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Phi_1(X_{t-10} + \phi_1 X_{t-11}) + \Phi_2(X_{t-20} + \phi_1 X_{t-21}) + \Theta_1(\epsilon_{t-10} + \theta_1 \epsilon_{t-11} + \theta_2 \epsilon_{t-12}) + \epsilon_t$$

Even though we are using values from 10 different past values and/or residuals, we're only estimating 6 coefficients (excluding the constant). Therefore, when we fit a model, we only get a coefficient for each order, rather than one for each value we're using.

Additionally, we have to include more past values because if the value from yesterday affects the value today, then the value from 11 days ago affects the one from 10 days ago. This is the entire reason we're not only including X_{t-10}, X_{t-20} and ϵ_{t-10} in the model, but also the values that shape them.

You can think of the values we're including as a time series with a different frequency. Notice how $X_{t-10} + \phi_1 X_{t-11}$ and $X_{t-20} + \phi_1 X_{t-21}$ are essentially the same thing 1 season (10 periods) apart. Then, we just think of seasonal patterns as trends with a different frequency we need to include in order to make good estimations.

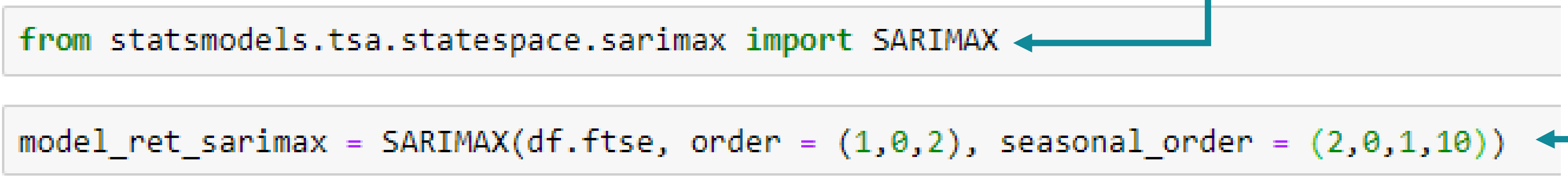
The SARIMAX Model

Implementation of the Model in Python:

The library the *SARIMAX* method comes from

The method we are importing

The seasonal order of the model



```
from statsmodels.tsa.statespace.sarimax import SARIMAX
```

```
model_ret_sarimax = SARIMAX(df.ftse, order = (1,0,2), seasonal_order = (2,0,1,10))
```

The variable storing the model characteristics that we will fit later

The time series we wish to analyse

The non-seasonal order of the model

*For an SARIMAX(p,d,q)(P,D,Q,s) model, simply change the order from (1,0,2) to (p,d,q), and the seasonal order from (2,0,1,10) to (P,D,Q,s).