**Time Series Analysis**

**Why do we need stationary samples for modelling?**

We can only perfectly model a dataset if all the samples are independent of each other. Because at the end of the day, we are just teaching our model that if we give observation X to it then the result would be Y.

If the observations are not independent, then that would mean that just giving an observation X is not enough. And, we would have to give more observations as well, that we don’t in a normal regression/classification model. Along side this in modelling and prediction, we make the assumption that the samples’ mean and SD would remain constant over time, only then we can extrapolate the time series.

Now, this was about independent observations. Why a time series has to be stationary is because a stationary time series has the same attributes as that of an independent observation. That is all there is.

We can go through the link below to confirm the same: <https://stats.stackexchange.com/questions/19715/why-does-a-time-series-have-to-be-stationary>

One more good example: <https://stats.stackexchange.com/questions/469657/can-someone-explain-the-importance-of-mean-stationarity-in-time-series>

A stationary time series has constant mean and variance that makes it more predictable and reliable. The time series that are not stationary are typically more difficult to predict over time.

**Unit Roots:** A time series is said to be stationary if it:

1. Has a constant mean over time
2. Has constant variance over time
3. Has no seasonal component

If a time series has a unit root then it is said to be not stationary. Having a unit root would mean that the variance of the series increases over time and we don’t want that.

To deal with this, first we have to detect it and then go about dealing with the series to make it stationary.

To detect if the time series has a unit root or not, we run something known as an **Augmented dickey fuller test (this is used when the series cannot be generated using a AR1 model)**. This test gives a p-value for each time series you put through it.

The hypotheses of the ADF test are:

H0: The series has unit root

H1: The series is stationary

So, if the p-value of the test for a series is less than 0.05 then we can reject the null hypothesis and say that the series is stationary.

Just like this test, there is one more test known as Phillips-Perron test that tests for stationarity. The only difference is that it is non parametric and performs worse that Augmented dickey Fuller in finite data.

**Seasonality:** This is the phenomenon of repeating patterns in our time series over the course of time (max: 1 year). Any time series that has seasonality in it is not stationary.

A time series can be made stationary through differencing the time stamps that are separated by the period of seasonality. Eg: if a time series is having yearly seasons then you subtract a value (today’s value) with it’s value at the same time but an year ago

(today - 365)

Seasonality is different from cycles as seasonality happens within a year whereas cycles are much more spread apart and happen over the course of multiple years.

**Differencing:** This process is done when a series is non stationary. In this method, you difference the current value at time ’t’ from the previous value at time ‘t-1’

There is a test known as KPSS test that is done next to tell if differencing is required or not. This test also essentially tells if a series is stationary or not.

The hypotheses of this test are:

H0: The series is stationary

H1: The series is not stationary

This test also gives out the lags, which were used to see if differencing was required or not.

The below link showcases what the different outputs of the ADF and KPSS test can tell us:

<https://www.statsmodels.org/stable/examples/notebooks/generated/stationarity_detrending_adf_kpss.html>

Following testing the series for stationarity, we check for **random walk.** This is done through the Variance-Ratio test or F-Test. This tells if the data will result in a model which is just a random walk or it is predictable.

The hypotheses in this test are:

H0: The series is a random walk

H1: The series is not a random walk

**Plots**

There are two kinds of plots that can be used in time series analysis:

* Auto Correlation plots: These are plots that tell you what is the correlation between the present series and the lagged series. It uses various types of lagged series, it could be lagged at 1 period, 2 periods, 3 periods and so on
* PACF: These are also correlation plots that tell us how correlated the current time series is. Present day’s value can be dependent on yesterday’s and day before yesterday’s values. To find out such facts, we need to use PACF plots. PACF plots are used most often to find out what all lagged time series will be required
  + Partial autocorrelations measure the linear dependence of one variable after removing the effect of other variable(s) that affect both variables. That is, the partial autocorrelation at lag k is the autocorrelation between yₜ and yₜ+yₜ₊ₖ that is not accounted for by lags 1 through k−1. PACF are like PDP wherein we are controlling the effect of the other lag versions of the same variable

This is a link that can be used to read up on these two plots: <https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/>

Another link which explains how PACF is used for finding the lags to be used for autoregression: <https://stats.stackexchange.com/questions/281666/how-does-acf-pacf-identify-the-order-of-ma-and-ar-terms>

*Read more on these plots*

**Multivariate Analysis**

1. Causality: This is a process in which we determine if there are any other features’ time series that can be used to predict the value of the target variable. This is done using the Granger Causality F-Test (or t-test, for single features) in which we put in the lagged versions of different features into the test and choose the ones which have a significant p-value

The hypotheses for this test are:

H0: The lagged value of a feature does not affect the target variable, hence no correlation

H1: The lagged value of a feature affects the target variable

This test can be used for feature selection in the models.

1. Cointegration tests: Cointegration tests identify scenarios where two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term. The tests are used to identify the degree of sensitivity of two variables to the same average price over a specified period of time.

The above test can be done to go about feature engineering. This will allow us to combine two features together to make a stationary time series

1. Cross-Correlation: In this process, we take external features one by one and lead/lag differentiate them and see their correlation with their target variable. This tells if this external variable can be used in model building or not

**Panel Data**

A lot of the time, in the dataset we see clustered longitudinal data. Wherein, inside each cluster the behavior of the variables differs.

To model for such datasets, we can do one of the 3 things:

* Ignore the clusters and create a global model for the entire data 🡺 This does not lead to good performance
* Create a model with the cluster ID as part of the dataset or OHE the cluster, This can lead to good performance but the performance starts sucking when the cardinality of the categorical variable becomes very high
* Create a mixed effect model, which strives to learn the idiosyncrasies of each cluster and models for the fixed and random effects. This model gives the best performance, however there are not many implementations of this model

The models that are normally used for mixed effect are the LMER ones, but they have the linearity assumption.

Other than this, the models that are now coming up are MERF that are mixed effects random forest models, which learn the fixed part using RF and the random part using linear model. These models are good as you can deal with the non-linearity as well with RF.

The only thing I want to look at is what kind of EDA is done on top of longitudinal data for regression analysis.

Some questions that need to be answered:

* Do we need to decompose a time series even if we are not making a classical time series model? None of the papers that I read today had this decomposition going on
  + One more reason in panel data, the time series component is ignored because the amount of time related data it has is less
  + Here, in the answer by ttphns, he mentions that in panel data the carry over effect of time series is ignored: <https://stats.stackexchange.com/questions/332411/difference-between-multivariate-time-series-data-and-panel-data>
  + The final answer is that we still need to decompose because we have to make sure that the distribution of training and testing data is the same.

**Time Series Models**

* First type of models are the AR models. In these models, you use the past values of a time series to predict the future value of these time series.

We don’t just put all the past values of the variable into the model because it might lead to overfitting or a needlessly complicated model. To overcome this, we use PACF plots, which tell us what are the most significant variables for predicting.

* One more way of modelling time series is MA models. They are moving average models in which we average out the error terms that were generated in past predictions.
  + How many error terms we should average or find the order of the MA model can be done using ACF. The moment the correlation in ACF graph hits 0 then all the number of lags till that point can be the order for our MA model
* One more type of model that we can go for is the ARMA model, which is the combination of AR and MA model. Here we consider both the past values as well as the moving average of error terms. The order for AR and MA is still found out using PACF and ACF graphs
* ARIMA modes: These models of course combine the AR and MA bits but the ARMA models can only work if the time series is stationary. We can make a time series stationary by differencing manually. However, in ARIMA if you provide the relevant differencing order (this is what I stands for in ARIMA) then the model will do it for you and generate the stationary time series
  + After differencing, you will predict the difference of the time series, you can for sure recover the actual value of the prediction by doing some math. The details are present here: <https://www.youtube.com/watch?v=3UmyHed0iYE&list=PLvcbYUQ5t0UHOLnBzl46_Q6QKtFgfMGc3&index=9>
* Another variant of time series model is SARIMA, which is ARIMA + Seasonal component of the time series. The math is a little confusing but we will dive deep when required
* ARCH models stand for Auto Regressive Conditional Heteroskedasticity, in which we try to reduce the volatility of residuals that we are getting across time